# RQ2 RQ3 Stats script

03/31/2025

# Contents



# Import the data

```
# Import the longtable with all results
lt <- read.csv(file="total_longtable.csv")
# Separate out results for RQ2 and RQ3
lt_RQ2 <- lt[lt$initial_visualizations == 1,]
lt_RQ3 <- lt[lt$initial_visualizations == 0,]
```

# RQ2: Do differences in taxonomic visualization design elements correspond to

# differences in comprehension, bias perception, and trust?

## Aggregating scores

We aggregate the longtable scores for comprehension, trust, and operationalized bias across all questions per participant to get their totals for every person.

```
lt_RQ2_aggregate <- lt_RQ2 %>%
  group_by(Pid, Vis_Type, qualitative_model_perception, Gender, Age,
           Familiarity) %>%
```

```
  summarise(
    comprehension = sum(comprehension),
    trust = sum(trust),
    likert.only.trust=sum(likert.only.trust),
    bias.operational = sum(bias.operational),
    correct.output = sum(correct.output),
    correct.pushing = sum(correct.pushing),
    correct.power = sum(correct.power),
    .groups = "drop"
  )
```

## Comprehension

We fit a linear model with the operationalized comprehension score as the dependent variable and the visualization type as the predictor. We run an Analysis of Variance (Anova) test on this model to confirm that the visualization type is a significant predictor of comprehension. We run emmeans to find that participants viewing LIME had the highest comprehension scores, while participants viewing Anchors had the lowest scores.

```
long_table_model <- lm(formula = comprehension ~ Vis_Type,
                       data = lt_RQ2_aggregate)

Anova(long_table_model)
```

```
## Anova Table (Type II tests)
##
## Response: comprehension
##           Sum Sq  Df F value    Pr(>F)
## Vis_Type   24694   5   84.68 < 2.2e-16 ***
## Residuals  25546 438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
emmeans(long_table_model, ~ Vis_Type)
```

```
##  Vis_Type  emmean    SE  df lower.CL upper.CL
##  anchors     23.0 0.876 438     21.3     24.7
##  cp          23.9 0.900 438     22.2     25.7
##  eli5        36.3 0.882 438     34.6     38.1
##  forceshap   38.7 0.900 438     36.9     40.4
##  lime        41.2 0.894 438     39.5     43.0
##  shap        39.7 0.876 438     37.9     41.4
##
## Confidence level used: 0.95
```

We also notice that visualizations with explicit magnitude and direction of feature impact had higher average comprehension scores than those where these must be inferred (41.05 vs. 23.45).

```
mean(lt_RQ2_aggregate[(lt_RQ2_aggregate$Vis_Type != "cp" &
                         lt_RQ2_aggregate$Vis_Type
                       != "anchors"),]$comprehension)
```

```
## [1] 38.96622
```

```
mean(lt_RQ2_aggregate[(lt_RQ2_aggregate$Vis_Type == "cp" |
                         lt_RQ2_aggregate$Vis_Type ==
                         "anchors"),]$comprehension)
```

```
## [1] 23.45946
```

We fit a linear model with the aggregate correctness score as the dependent variable, and the aggregate perception score as the predictor, and find that perceived comprehension is a significant predictor of objective comprehension. That is, the participants were more likely to self-report better comprehension if they did indeed better understand the model.

```
long_table_model <- lm(formula = trust ~ qualitative_model_perception,
                        data = lt_RQ2_aggregate)

Anova(long_table_model)
```

```
## Anova Table (Type II tests)
##
## Response: trust
##                              Sum Sq  Df F value    Pr(>F)
## qualitative_model_perception  13669   1  28.668 1.384e-07 ***
## Residuals                    210747 442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
exp(coef(long_table_model))
```

```
##                  (Intercept) qualitative_model_perception
##                 3.932642e+14                 4.581959e+00
```

```
cor(lt_RQ2_aggregate$qualitative_model_perception,
    lt_RQ2_aggregate$comprehension, method = c("pearson"))
```

```
## [1] 0.3333735
```

We isolate each component of the Comprehension metric. We find that visualization is a significant predictor of whether participants correctly indicated model output. People were most likely to correctly indicate output when looking at LIME.

```
long_table_model <- lm(formula = correct.output ~ Vis_Type,
                        data = lt_RQ2_aggregate)

Anova(long_table_model)
```

```
## Anova Table (Type II tests)
##
## Response: correct.output
##           Sum Sq  Df F value   Pr(>F)
## Vis_Type   37.54   5  3.7505 0.002459 **
## Residuals 876.71 438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
emmeans(long_table_model, ~ Vis_Type)
```

```
##  Vis_Type  emmean    SE  df lower.CL upper.CL
##  anchors     6.91 0.162 438     6.59     7.23
##  cp          7.21 0.167 438     6.88     7.54
##  eli5        7.01 0.163 438     6.69     7.33
##  forceshap   7.28 0.167 438     6.95     7.61
##  lime        7.73 0.166 438     7.40     8.05
```

```
##  shap        7.58 0.162 438     7.26     7.90
##
## Confidence level used: 0.95
```

We find that visualization is a significant predictor of whether participants correctly indicated direction of feature impact. People were most likely to correctly indicate impact direction when looking at LIME.

```r
long_table_model <- lm(formula = correct.pushing ~ Vis_Type,
                       data = lt_RQ2_aggregate)

Anova(long_table_model)
```

```
## Anova Table (Type II tests)
##
## Response: correct.pushing
##           Sum Sq  Df F value    Pr(>F)
## Vis_Type   17930   5  100.16 < 2.2e-16 ***
## Residuals  15681 438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
emmeans(long_table_model, ~ Vis_Type)
```

```
##  Vis_Type  emmean    SE  df lower.CL upper.CL
##  anchors     12.0 0.686 438     10.7     13.4
##  cp          14.0 0.705 438     12.6     15.4
##  eli5        25.2 0.691 438     23.8     26.6
##  forceshap   25.7 0.705 438     24.3     27.1
##  lime        27.8 0.700 438     26.4     29.2
##  shap        26.6 0.686 438     25.2     27.9
##
## Confidence level used: 0.95
```

We find that visualization is a significant predictor of whether participants correctly indicated the most impactful feature. People were most likely to correctly indicate the most impactful feature when looking at LIME.

```r
long_table_model <- lm(formula = correct.power ~ Vis_Type,
                       data = lt_RQ2_aggregate)

Anova(long_table_model)
```

```
## Anova Table (Type II tests)
##
## Response: correct.power
##           Sum Sq  Df F value    Pr(>F)
## Vis_Type  535.18   5  48.864 < 2.2e-16 ***
## Residuals 959.44 438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
emmeans(long_table_model, ~ Vis_Type)
```

```
##  Vis_Type  emmean    SE  df lower.CL upper.CL
##  anchors     4.08 0.170 438     3.75     4.41
##  cp          2.69 0.174 438     2.35     3.04
##  eli5        4.13 0.171 438     3.80     4.47
##  forceshap   5.68 0.174 438     5.34     6.02
```

```
##   lime        5.70 0.173 438     5.36     6.04
##   shap        5.53 0.170 438     5.19     5.86
##
## Confidence level used: 0.95
```

### Trust in the underlying model

We fit a linear model with the operationalized trust score as the dependent variable and the visualization type as the predictor. We run an Analysis of Variance (Anova) test on this model to confirm that the visualization type is a significant predictor of trust. We run emmeans to find that participants viewing Ceteris-Paribus (CP) had the highest trust scores, while participants viewing Shap Force Plots had the lowest scores.

```r
long_table_model <- lm(formula = trust ~ Vis_Type, data = lt_RQ2_aggregate)

Anova(long_table_model)
```

```
## Anova Table (Type II tests)
##
## Response: trust
##           Sum Sq  Df F value  Pr(>F)
## Vis_Type    6160   5  2.4724 0.03174 *
## Residuals 218256 438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
emmeans(long_table_model, ~ Vis_Type)
```

```
##  Vis_Type  emmean   SE  df lower.CL upper.CL
##  anchors     53.9 2.56 438     48.8     58.9
##  cp          58.3 2.63 438     53.1     63.4
##  eli5        51.4 2.58 438     46.3     56.4
##  forceshap   45.8 2.63 438     40.6     50.9
##  lime        50.2 2.61 438     45.0     55.3
##  shap        52.0 2.56 438     47.0     57.0
##
## Confidence level used: 0.95
```

We isolate responses to the question `Do people agree with the statement  "Computer models can be trusted to make human decisions" less over time as  they see a biased model?` We find that there is a significant trend where participants are more likely to say “no” to this question as they see more output instances for the biased model.

```r
model <- glmer(computers.can.make.human.decisions ~ order.seen +
               (1 | Pid), family = binomial, data = lt_RQ2)
summary(model)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: computers.can.make.human.decisions ~ order.seen + (1 | Pid)
##    Data: lt_RQ2
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    2163.9    2182.0   -1078.9    2157.9      3105
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.8925 -0.1343 -0.1028  0.1563  2.9705
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Pid    (Intercept) 30.28    5.503
## Number of obs: 3108, groups:  Pid, 439
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.34406    0.35454  -0.970 0.331836
## order.seen  -0.13335    0.03576  -3.729 0.000192 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr)
## order.seen -0.285
```

## Perceived Bias

We fit a linear model with the operationalized bias perception score as the dependent variable and the visualization type as the predictor. We run an Analysis of Variance (Anova) test on this model to confirm that the visualization type is a significant predictor of bias perception. We run emmeans to find that participants viewing Shap Force Plots had the highest bias perception scores, while participants viewing CP had the lowest scores.

```
long_table_model <- lm(formula = bias.operational ~ Vis_Type,
                       data = lt_RQ2_aggregate)

Anova(long_table_model)
```

```
## Anova Table (Type II tests)
##
## Response: bias.operational
##            Sum Sq  Df F value   Pr(>F)
## Vis_Type    992.2   5  3.5291 0.003868 **
## Residuals 24629.8 438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
emmeans(long_table_model, ~ Vis_Type)
```

```
##  Vis_Type  emmean    SE  df lower.CL upper.CL
##  anchors     11.9 0.860 438    10.20     13.6
##  cp          10.5 0.884 438     8.75     12.2
##  eli5        13.9 0.866 438    12.18     15.6
##  forceshap   15.0 0.884 438    13.24     16.7
##  lime        14.2 0.878 438    12.43     15.9
##  shap        12.6 0.860 438    10.95     14.3
##
## Confidence level used: 0.95
```

## Trust vs. Comperehension

We perform a Pearson correlation test and find comprehension and trust to be significantly negatively correlated - i.e., increased comprehension results in decreased trust.

```
long_table_model <- lm(formula = trust ~ comprehension, data = lt_RQ2_aggregate)

Anova(long_table_model)
```

```
## Anova Table (Type II tests)
##
## Response: trust
##               Sum Sq  Df F value    Pr(>F)
## comprehension  17046   1  36.332 3.512e-09 ***
## Residuals     207370 442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
exp(coef(long_table_model))
```

```
##   (Intercept) comprehension
##  1.244139e+31  5.585087e-01
```

```
cor(lt_RQ2_aggregate$trust, lt_RQ2_aggregate$comprehension,
    method = c("pearson"))
```

```
## [1] -0.2756018
```

## Mediation Analysis

We fit a linear model with operationalized trust as the dependent variable, and comprehension and bias perception as the predictors.We find that the only significant predictor of trust score is bias perception

```
long_table_model <- lm(formula = trust ~ comprehension + bias.operational,
                        data = lt_RQ2_aggregate)

chisq.test(lt_RQ2_aggregate$trust,predict(long_table_model))
```

```
## Warning in chisq.test(lt_RQ2_aggregate$trust, predict(long_table_model)):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  lt_RQ2_aggregate$trust and predict(long_table_model)
## X-squared = 22504, df = 21924, p-value = 0.002973
```

```
Anova(long_table_model)
```

```
## Anova Table (Type II tests)
##
## Response: trust
##                  Sum Sq  Df  F value Pr(>F)
## comprehension        91   1   0.4723 0.4923
## bias.operational 122260   1 633.4949 <2e-16 ***
## Residuals         85110 441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To better understand the above, we fit a model with bias perception as the dependent variable, and comprehension as the predictor. This model shows that aggregate comprehension score is a significant predictor of bias perception, and suggests that bias perception mediates the relationship between comprehension and trust.

```r
long_table_model <- lm(formula = bias.operational ~ comprehension,
                        data = lt_RQ2_aggregate)

chisq.test(lt_RQ2_aggregate$trust,predict(long_table_model))
```

```
## Warning in chisq.test(lt_RQ2_aggregate$trust, predict(long_table_model)):
## Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  lt_RQ2_aggregate$trust and predict(long_table_model)
## X-squared = 3287.9, df = 3192, p-value = 0.1158
```

```r
Anova(long_table_model)
```

```
## Anova Table (Type II tests)
##
## Response: bias.operational
##                 Sum Sq  Df F value    Pr(>F)
## comprehension  3577.3   1  71.725 3.695e-16 ***
## Residuals     22044.7 442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We fit a mediation model with trust score as the dependent variable, the comprehension score as the predictor, and the bias perception score as the mediator. We find that comprehension has a direct positive effect on bias perception, and bias perception had a direct negative effect on trust. That is, increased comprehension indirectly decreases trust by impacting perception of bias.

```r
mediation_model <- '
  # Mediator equation: effect of comprehension on the mediator (bias percep.)
  bias.operational ~ a * comprehension

  # Outcome equation: direct effect of comprehension and effect of the
  # mediator (bias percep.) on trust
  trust ~ c_prime * comprehension + b * bias.operational

  # Indirect effect: the mediation path (a * b)
  indirect := a * b

  # Total effect: sum of the direct and indirect effects
  total := c_prime + indirect
'

# Estimate the mediation model
mediation_results <- sem(mediation_model, data = lt_RQ2_aggregate)

# Summarize the results
summary(mediation_results, standardized = TRUE, fit.measures = TRUE)
```

```
## lavaan 0.6-19 ended normally after 1 iteration
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         5
```

```
##
##   Number of observations                           444
##
## Model Test User Model:
##
##   Test statistic                                 0.000
##   Degrees of freedom                                 0
##
## Model Test Baseline Model:
##
##   Test statistic                               497.251
##   Degrees of freedom                                 3
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    1.000
##   Tucker-Lewis Index (TLI)                       1.000
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -3293.732
##   Loglikelihood unrestricted model (H1)      -3293.732
##
##   Akaike (AIC)                                6597.465
##   Bayesian (BIC)                              6617.944
##   Sample-size adjusted Bayesian (SABIC)       6602.076
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.000
##   90 Percent confidence interval - lower         0.000
##   90 Percent confidence interval - upper         0.000
##   P-value H_0: RMSEA <= 0.050                       NA
##   P-value H_0: RMSEA >= 0.080                       NA
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.000
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Regressions:
##                      Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   bias.operational ~
##     cmprhns    (a)      0.267    0.031    8.488    0.000    0.267    0.374
##   trust ~
##     cmprhns (c_pr)      0.046    0.067    0.690    0.490    0.046    0.022
##     bs.prtn    (b)     -2.355    0.093  -25.255    0.000   -2.355   -0.796
##
```

```
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .bias.operatinl   49.650    3.332   14.900    0.000   49.650    0.860
##    .trust           191.689   12.865   14.900    0.000  191.689    0.379
##
## Defined Parameters:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##     indirect         -0.628    0.078   -8.046    0.000   -0.628   -0.297
##     total            -0.582    0.096   -6.041    0.000   -0.582   -0.276
```

## Qualitative Analysis

We find that in 51.15% of responses, participants thought the model would give them a loan.

```r
mean(lt$this.model.will.give.me.a.loan)
```

```
## [1] 0.5214544
```

We find that in 33.86% of responses, participants thought the model would give them a loan.

```r
mean(lt$this.model.shouldnt.give.me.a.loan)
```

```
## [1] 0.338618
```

We find that when participants felt the model would give them a loan, they trusted it significantly more than when they did not feel this way.

```r
wilcox.test(trust ~ this.model.will.give.me.a.loan, data = lt_RQ2)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trust by this.model.will.give.me.a.loan
## W = 503632, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

We find that when participants also trusted the model more if they felt that it would not give them a loan but they thought this was the right decision.

```r
wilcox.test(trust ~ this.model.shouldnt.give.me.a.loan, data = lt_RQ2)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trust by this.model.shouldnt.give.me.a.loan
## W = 841564, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(trust ~ this.model.will.or.shouldnt.give.me.a.loan, data = lt_RQ2)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trust by this.model.will.or.shouldnt.give.me.a.loan
## W = 346748, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
cohen.d(trust ~ this.model.will.or.shouldnt.give.me.a.loan, data = lt_RQ2)
```

```
## Call: cohen.d(x = trust ~ this.model.will.or.shouldnt.give.me.a.loan,
```

```
##     data = lt_RQ2)
## Cohen d statistic of difference between two means
##       lower effect upper
## trust  1.17   1.26  1.34
##
## Multivariate (Mahalanobis) distance between groups
## [1] 1.3
## r equivalent of difference between two means
## trust
##  0.49
```

```
wilcox.test(bias.operational ~ this.model.will.or.shouldnt.give.me.a.loan, data = lt_RQ2)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bias.operational by this.model.will.or.shouldnt.give.me.a.loan
## W = 1465084, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
cohen.d(bias.operational ~ this.model.will.or.shouldnt.give.me.a.loan, data = lt_RQ2)
```

```
## Call: cohen.d(x = bias.operational ~ this.model.will.or.shouldnt.give.me.a.loan,
##     data = lt_RQ2)
## Cohen d statistic of difference between two means
##                  lower effect upper
## bias.operational -1.07  -0.99  -0.9
##
## Multivariate (Mahalanobis) distance between groups
## [1] 0.99
## r equivalent of difference between two means
## bias.operational
##             -0.4
```

## RQ3: Do there exist causal relationships between comprehension, bias

## perception, and trust?

### Comprehension impacts bias perception

We find that adding explicit indicators of model output, feature impact direction and feature impact magnitude has a significant effect on comprehension, trust, and perceived bias. Cohen's D indicates that there is a positive medium effect on comprehension, a small positive effect on bias perception, and a small negative effect on trust. This indicates that including explicit values increases comprehension and bias perception, and therefore decreases trust.

```
explicit_experiment <- lt[(lt$Vis_Type == "cp" | lt$Vis_Type == "cp_explicit"),]

explicit_experiment$Vis_Type <-
  ifelse(explicit_experiment$Vis_Type == "cp", "inferred",
  ifelse(explicit_experiment$Vis_Type == "cp_explicit", "explicit",
  explicit_experiment$Vis_Type))

wilcox.test(trust ~ Vis_Type, data = explicit_experiment)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trust by Vis_Type
## W = 114060, p-value = 3.012e-05
## alternative hypothesis: true location shift is not equal to 0
```

```r
cohen.d(trust ~ Vis_Type, data = explicit_experiment)
```

```
## Call: cohen.d(x = trust ~ Vis_Type, data = explicit_experiment)
## Cohen d statistic of difference between two means
##       lower effect upper
## trust  0.14   0.26  0.39
##
## Multivariate (Mahalanobis) distance between groups
## [1] 0.26
## r equivalent of difference between two means
## trust
##  0.13
```

```r
wilcox.test(bias.operational ~ Vis_Type, data = explicit_experiment)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bias.operational by Vis_Type
## W = 150618, p-value = 0.0003691
## alternative hypothesis: true location shift is not equal to 0
```

```r
cohen.d(bias.operational ~ Vis_Type, data = explicit_experiment)
```

```
## Call: cohen.d(x = bias.operational ~ Vis_Type, data = explicit_experiment)
## Cohen d statistic of difference between two means
##                  lower effect upper
## bias.operational -0.34  -0.22  -0.1
##
## Multivariate (Mahalanobis) distance between groups
## [1] 0.22
## r equivalent of difference between two means
## bias.operational
##           -0.11
```

```r
wilcox.test(comprehension ~ Vis_Type, data = explicit_experiment)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  comprehension by Vis_Type
## W = 182583, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
cohen.d(comprehension ~ Vis_Type, data = explicit_experiment)
```

```
## Call: cohen.d(x = comprehension ~ Vis_Type, data = explicit_experiment)
## Cohen d statistic of difference between two means
##               lower effect upper
## comprehension  -0.8  -0.68 -0.55
##
```

```
## Multivariate (Mahalanobis) distance between groups
## [1] 0.68
## r equivalent of difference between two means
## comprehension
##         -0.32
```

### Lower bias results in higher trust

We see significant differences in comprehension, trust, and perceived bias when comparing responses for the same visualization with a fair and an unfair underlying model. When looking at effect sizes, we can see that the effect size of comprehension is negligible. This indicates that the comprehension differs very little between the fair and the biased model. However, introducing the fair model does result in a small negative effect on bias perception and subsequently a small positive effect on trust. This indicates that with a high level of comprehension, decreasing the incidence of bias will decrease the perception of bias, and lead to an increase in trust.

```r
fair_experiment <- lt[(lt$Vis_Type == "interactive" |
                        lt$Vis_Type == "interactive_fair"),]

fair_experiment$Vis_Type <-
  ifelse(fair_experiment$Vis_Type == "interactive", "unfair",
  ifelse(fair_experiment$Vis_Type == "interactive_fair", "fair",
  explicit_experiment$Vis_Type))

wilcox.test(trust ~ Vis_Type, data = fair_experiment)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trust by Vis_Type
## W = 170486, p-value = 2.439e-11
## alternative hypothesis: true location shift is not equal to 0
```

```r
cohen.d(trust ~ Vis_Type, data = fair_experiment)
```

```
## Call: cohen.d(x = trust ~ Vis_Type, data = fair_experiment)
## Cohen d statistic of difference between two means
##       lower effect upper
## trust -0.56  -0.44 -0.32
##
## Multivariate (Mahalanobis) distance between groups
## [1] 0.44
## r equivalent of difference between two means
## trust
## -0.21
```

```r
wilcox.test(bias.operational ~ Vis_Type, data = fair_experiment)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bias.operational by Vis_Type
## W = 111874, p-value = 3.478e-08
## alternative hypothesis: true location shift is not equal to 0
```

```r
cohen.d(bias.operational ~ Vis_Type, data = fair_experiment)
```

```
## Call: cohen.d(x = bias.operational ~ Vis_Type, data = fair_experiment)
## Cohen d statistic of difference between two means
##                   lower effect upper
## bias.operational    0.2   0.33  0.45
##
## Multivariate (Mahalanobis) distance between groups
## [1] 0.33
## r equivalent of difference between two means
## bias.operational
##             0.16
```

```
wilcox.test(comprehension ~ Vis_Type, data = fair_experiment)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  comprehension by Vis_Type
## W = 125210, p-value = 0.003201
## alternative hypothesis: true location shift is not equal to 0
```

```
cohen.d(comprehension ~ Vis_Type, data = fair_experiment)
```

```
## Call: cohen.d(x = comprehension ~ Vis_Type, data = fair_experiment)
## Cohen d statistic of difference between two means
##                lower effect upper
## comprehension  0.04   0.16  0.28
##
## Multivariate (Mahalanobis) distance between groups
## [1] 0.16
## r equivalent of difference between two means
## comprehension
##          0.08
```

### Artificially lowered bias perception also results in higher trust

We observe significant differences in perceived bias and trust, and less significant differences in comprehension, between a visualization designed to decrease bias perception, and a visualization designed to increase bias perception. Cohen's D shows that the effect size for comprehension is incredibly negligible, while effect sizes for increased bias perception and decreased trust are small. These effects indicate that even in the case where changes in comprehension are small or negligible, a change in perception of bias can impact trust.

```
bias_experiment <- lt[(lt$Vis_Type == "interactive_lower_bias_percep" |
                         lt$Vis_Type == "interactive_higher_bias_percep"),]
bias_experiment$Vis_Type <-
  ifelse(bias_experiment$Vis_Type == "interactive_lower_bias_percep",
         "interactive_lower_bias_percep",
  ifelse(bias_experiment$Vis_Type == "interactive_higher_bias_percep",
         "interactive_upper_bias_percep",
  explicit_experiment$Vis_Type))
wilcox.test(trust ~ Vis_Type, data = bias_experiment)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trust by Vis_Type
## W = 174636, p-value = 0.0001399
```

```
## alternative hypothesis: true location shift is not equal to 0
```

```r
cohen.d(trust ~ Vis_Type, data = bias_experiment)
```

```
## Call: cohen.d(x = trust ~ Vis_Type, data = bias_experiment)
## Cohen d statistic of difference between two means
##       lower effect upper
## trust -0.31   -0.2 -0.08
##
## Multivariate (Mahalanobis) distance between groups
## [1] 0.2
## r equivalent of difference between two means
## trust
##  -0.1
```

```r
wilcox.test(bias.operational ~ Vis_Type, data = bias_experiment)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bias.operational by Vis_Type
## W = 129698, p-value = 1.126e-06
## alternative hypothesis: true location shift is not equal to 0
```

```r
cohen.d(bias.operational ~ Vis_Type, data = bias_experiment)
```

```
## Call: cohen.d(x = bias.operational ~ Vis_Type, data = bias_experiment)
## Cohen d statistic of difference between two means
##                  lower effect upper
## bias.operational  0.16   0.27  0.39
##
## Multivariate (Mahalanobis) distance between groups
## [1] 0.27
## r equivalent of difference between two means
## bias.operational
##             0.14
```

```r
wilcox.test(comprehension ~ Vis_Type, data = bias_experiment)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  comprehension by Vis_Type
## W = 143796, p-value = 0.01294
## alternative hypothesis: true location shift is not equal to 0
```

```r
cohen.d(comprehension ~ Vis_Type, data = bias_experiment)
```

```
## Call: cohen.d(x = comprehension ~ Vis_Type, data = bias_experiment)
## Cohen d statistic of difference between two means
##               lower effect upper
## comprehension -0.05   0.06  0.18
##
## Multivariate (Mahalanobis) distance between groups
## [1] 0.065
## r equivalent of difference between two means
## comprehension
##          0.03
```