Justifying Social Choice Mechanism for Improving Participant Satisfaction: Supplementary Material

Sharadhi Alape Suryanarayana, David Sarne, Sarit Kraus

1 Generating Feature-based Explanations

We demonstrate the application of Algorithm 1 in generating explanations using Instance 1 (Figure 1) of the six instances generated. The winning candidate is Cariot. In order to generate explanations

#	6 Voters	4 Voters	4 Voters	7 Voters	4 Voters	4 Voters
Rank 1	Branflakes	Branflakes	Cariot	Cariot	Shugi	Shugi
Rank 2	Cariot	Shugi	Branflakes	Shugi	Branflakes	Cariot
Rank 3	Shugi	Cariot	Shugi	Branflakes	Cariot	Branflakes

Figure 1: Instance 1 of the 6 instances generated. The winning candidate is Cariot.

for the winning candidate, Cariot, we need to express the concepts of this problem in theoretical terms as required by the algorithm, Algorithm 1, for Picking the top-n feature-based explanations. The set of solutions, S in the algorithm, is the set of candidates $\{Branflakes, Cariot, Shugi\}$ and the solution to be explained S_i is the winning candidate $Cariot$. The set of features J used to devise the explanations are given in Section 3 of the paper. The table summarizing the values of the different features is given in Figure 2. The first four features are positively correlated with superiority (the higher the better) and the last two are negatively correlated with superiority (the lower the better). All the

Candidate / Feature	Branflakes (BF)	Cariot (CR)	Shugi (SH)
First Place Votes/Plurality (P)	10	11	8
Borda (B)	57	61	56
Head-to-Head (H)	BF vs CR = 14 BF vs SH = 14	CR vs BF = 15 CR vs SH = 17	SH vs BF = 15 SH vs CR = 12
Bucklin Rule (BR)	18	21	19
Greatest Pairwise Opposition/Minimax Condorcet (M)	15	14	17
Last Place Votes (L)	11	8	10

Figure 2: Values of the different features for all of the candidates in Instance 1.

features, except head-to-head comparisons can be readily computed since they are not dependent on

any combination/opponent. The calculation of the values for feature head-to-head comparisons is dependent on the pair of winning candidate and opponent. Given a winning candidate S_i and an alternate candidate S_t, we compute the values of d_i^j and d_t^j as follows

- d_i^j = Number of votes in favor of candidate S_i in a pairwise comparison with S_t.
- d_t^j = Number of votes in favor of candidate S_t in a pairwise comparison with S_i.

Selection of Features We demonstrate the functioning of this part of the algorithm with the help of two instances, Instance 1 (Figure 1) and Instance 3a (Figure 6) for the feature *First Place Votes*. The same logic **(Lines 2-11 in Algorithm 1)** is used for all of the other features.

The values for the feature *First Place Votes* for the 3 candidates in Instance 1, where the winning candidate, S_i is *Cariot* (Figure 1) are as follows:-

- Branflakes $(d_t^j) = 10$
- Cariot $(d_i^j) = 11$
- Shugi $(d_t^j) = 8$

The values for the feature are retrieved and upon the comparison of d_i^j with both of the values of d_t^j, d_i^j is not inferior to either of the two d_t^j's and hence the Feature *First Place Votes* is not excluded (is included) from the list of possible features for explanations. From Figure 2, it can be observed that the winning candidate, Cariot, has the best score with respect to all of the features and hence, all of the features are possible contenders for devising explanations.

The values for the feature *First Place Votes* for the 3 candidates in Instance 3a where the winning candidate, S_i is *Shugi* (Figure 6) are as follows:-

- Branflakes $(d_t^j) = 5$
- Cariot $(d_t^j) = 13$
- Shugi $(d_i^j) = 11$

The values for the feature are retrieved and upon the comparison of d_i^j with the value of d_t^j for *Cariot*, d_i^j is inferior and hence the feature *First Place Votes* is excluded from the list of possible features for explanations.

Calculation of Score We demonstrate this phase of the algorithm with the help of Instance 1 (Figure 1) where the winning candidate, S_i, is Cariot. Since the feature, j = *First Place Votes* is included in the list of features for explanations for Instance 1 (Figure 1), the score is calculated as follows.

$$score_i^j = (11 - 10) + (11 - 8) = 4 \tag{1}$$

This score is normalized by dividing it by the maximum possible value of the feature which is the number of total voters i.e. 29. Therefore, $score_i^j = 4/29$.

The same procedure is followed for all of the features that are included in the list of possible explanations. The calculation for all of the scores is given in Figure 3. **This logic is expressed in lines 12-16 of Algorithm 1.** These scores are sorted in the descending order **(Line 17 in Algorithm 1)** and the top 3 features are returned **(Line 18 in Algorithm 1)**.

The top 3 features to be considered for explanations are highlighted in bold. Thus, the Feature-based explanations are:

1. Cariot has won 2 rounds of head-to-head comparison. 15 voters prefer Cariot over Branflakes while 14 voters prefer Branflakes over Cariot. 17 voters prefer Cariot over Shugi while 12 voters prefer Shugi over Cariot.
2. Cariot is ranked the last choice by 8 voters compared to Branflakes by 11 voters and Shugi by 10 voters.
3. Cariot is ranked either the first or the second choice by 21 voters compared to Branflakes by 18 voters and Shugi by 19 voters.

Feature	Calculation Of Score	Score
First Place Votes/Plurality (P)	$\frac{(11-10)+(11-8)}{29}$	$\frac{4}{29}$
Borda Score (B)	$\frac{(61-57)+(61-56)}{87}$	$\frac{3}{29}$
Head-to-Head Score (H)	$\frac{(15-14)+(17-12)}{29}$	$\mathbf{\frac{6}{29}}$
Bucklin Rule (BR)	$\frac{(21-18)+(21-19)}{29}$	$\mathbf{\frac{5}{29}}$
Greatest Pairwise Opposition/Minimax Condorcet (M)	$\frac{(17-14)+(15-14)}{29}$	$\frac{4}{29}$
Last Place Votes (L)	$\frac{(11-8)+(10-8)}{29}$	$\mathbf{\frac{5}{29}}$

Figure 3: Score Calculation for all of the features for Instance 1.

2 List of Feature-based and Crowdsourced Explanations

Six instances of voting tables and winning candidates were generated. Instance 1 (Figure 1) and the feature-based explanations for instance 1 are given in Section 1. The list of crowdsourced explanations for Instance 1 are:-

1. Cariot is voted number one by the largest group of voters.
2. Across all 6 columns with the number 1 spot, Cariot received the most votes with 11 votes.
3. Cariot has more first and second place votes than the other two options, with 21 votes.

The list of feature-based explanations for Instance 2 (Figure 4) are as follows:

#	6 Voters	2 Voters	8 Voters	5 Voters	4 Voters	4 Voters
Rank 1	Branflakes	Branflakes	Cariot	Cariot	Shugi	Shugi
Rank 2	Cariot	Shugi	Branflakes	Shugi	Branflakes	Cariot
Rank 3	Shugi	Cariot	Shugi	Branflakes	Cariot	Branflakes

Figure 4: Instance 2 of the 6 instances generated. The winning candidate is Cariot.

1. Cariot has won 2 rounds of head-to-head comparison. 17 voters prefer Cariot over Branflakes while 12 voters prefer Branflakes over Cariot. 19 voters prefer Cariot over Shugi while 10 voters prefer Shugi over Cariot.
2. The highest pairwise score against Cariot is 12 compared to Branflakes with 17 and Shugi with 19.
3. Cariot is ranked the last choice by 6 voters compared to Branflakes by 9 voters and Shugi by 14 voters.

4. Cariot is ranked either the first or the second choice by 23 voters compared to Branflakes by 20 voters and Shugi by 15 voters.

There are four possible explanations due to a tie in the score for the features *Last Place Votes* and *Bucklin Score*. The tie is broken randomly and either of the two features is considered for explanation.

The list of crowdsourced explanations for Instance 2 (Figure 4) are as follows:

1. A total of 13 people have ranked Cariot as the best (better than both Branflakes and Shugi).
2. 10 people ranked Cariot as at least better than Shugi or Branflakes (second best).
3. If you took a weighted average of all of the votes, Cariot would rank the best with a 1.76 average ranking.

The list of feature-based explanations for Instance 3 (Figure 5) are as follows:

#	1 Voters	4 Voters	7 Voters	6 Voters	7 Voters	4 Voters
Rank 1	Branflakes	Branflakes	Cariot	Cariot	Shugi	Shugi
Rank 2	Cariot	Shugi	Branflakes	Shugi	Branflakes	Cariot
Rank 3	Shugi	Cariot	Shugi	Branflakes	Cariot	Branflakes

Figure 5: Instance 3 of the 6 instances generated. The winning candidate is Cariot.

1. Cariot is ranked the first choice by 13 voters compared to Branflakes by 5 voters and Shugi by 11 voters.

The list of crowdsourced explanations for Instance 3 (Figure 5) are as follows:

1. Cariot is favored number 1 by 13 voters, which is greater than 11 for Shugi and 5 for Branflakes.
2. The biggest group (tied, 7) and second biggest group (6) of voters vote for Cariot as their first choice.
3. Cariot beat the other cereal Shugi by two votes so it won.

The list of feature-based explanations for Instance 3a (Figure 6) are as follows:

#	1 Voters	4 Voters	7 Voters	6 Voters	7 Voters	4 Voters
Rank 1	Branflakes	Branflakes	Cariot	Cariot	Shugi	Shugi
Rank 2	Cariot	Shugi	Branflakes	Shugi	Branflakes	Cariot
Rank 3	Shugi	Cariot	Shugi	Branflakes	Cariot	Branflakes

Figure 6: Instance 3a of the 6 instances generated. The winning candidate is Shugi.

1. Shugi has won 2 rounds of head-to-head comparison. 17 voters prefer Shugi over Branflakes while 12 voters prefer Branflakes over Shugi. 15 voters prefer Shugi over Cariot while 14 voters prefer Cariot over Shugi.
2. Shugi is ranked either the first or the second choice by 21 voters compared to Branflakes by 19 voters and Cariot by 18 voters.

3. Shugi is ranked the last choice by 8 voters compared to Branflakes by 10 voters and Cariot by 11 voters.

The list of crowdsourced explanations for Instance 3a (Figure 6) are as follows:

1. Shugi is selected as either first or second choice by 21 voters. The tally for Cariot being in either first or second place is 18, and the tally for Branflakes in first or second place is 19.
2. Shugi received more points in the ranked voting system.
3. 11 people have rated Shugi their top cereal.

The list of feature-based explanations for Instance 4 (Figure 7) are as follows:

#	5 Voters	2 Voters	6 Voters	4 Voters	6 Voters	6 Voters
Rank 1	Branflakes	Branflakes	Cariot	Cariot	Shugi	Shugi
Rank 2	Cariot	Shugi	Branflakes	Shugi	Branflakes	Cariot
Rank 3	Shugi	Cariot	Shugi	Branflakes	Cariot	Branflakes

Figure 7: Instance 4 of the 6 instances generated. The winning candidate is Cariot.

1. Cariot is ranked the last choice by 8 voters compared to Branflakes by 10 voters and Shugi by 11 voters.
2. Cariot is ranked either the first or the second choice by 21 voters compared to Branflakes by 19 voters and Shugi by 18 voters.
3. Cariot has won 2 rounds of head-to-head comparison. 16 voters prefer Cariot over Branflakes while 13 voters prefer Branflakes over Cariot. 15 voters prefer Cariot over Shugi while 14 voters prefer Shugi over Cariot.

The list of crowdsourced explanations for Instance 4 (Figure 7) are as follows:

1. Cariot is ranked either first or second place by the most people (21 first or second place votes for Cariot, compared to 19 for Branflakes and 18 candidates for Shugi).
2. While Shugi has more first-place votes than Cariot, more people have ranked Cariot above Shugi than ranked Shugi above Cariot (15 people ranked Cariot above Shugi, but 14 people ranked Shugi above Cariot).
3. Cariot has the most points when added up.

The list of feature-based explanations for Instance 4a (Figure 8) are as follows:

#	5 Voters	2 Voters	6 Voters	4 Voters	6 Voters	6 Voters
Rank 1	Branflakes	Branflakes	Cariot	Cariot	Shugi	Shugi
Rank 2	Cariot	Shugi	Branflakes	Shugi	Branflakes	Cariot
Rank 3	Shugi	Cariot	Shugi	Branflakes	Cariot	Branflakes

Figure 8: Instance 4a of the 6 instances generated. The winning candidate is Shugi.

1. Shugi is ranked the first choice by 12 voters compared to Branflakes by 7 voters and Cariot by 10 voters.

The list of crowdsourced explanations for Instance 4a (Figure 8) are as follows:

1. Shugi is the cereal selected by the largest number of voters (12 voters) as their first choice.
2. 18 of 29 voters prefer Shugi as the first or second place cereal brand.
3. 16 voters prefer Shugi over Branflakes and 14 voters prefer Shugi over Cariot.

3 Description of the Statistical Tests

3.1 ANOVA with Aligned Rank Transform

The answers given to the questions on "Satisfaction" and "Acceptance" were measured on a Likert Scale of 1-5 with 5 being the most desired value and 1 being the least desired value. We use Analysis of Variance (ANOVA) with Aligned Rank Transforms (ART) since such a test is more appropriate in a setting where Likert Scale data is used [3]. The analysis is carried out using the ARTool R package. We have two factors, Explanations (No explanations, Crowdsourced explanations and Feature-based explanations) and Preference Index (Second preference and Third preference), in our ANOVA which results in a (3×2) design with respect to the factors. **Figure 2 in the paper is intended to give a visual representation of the same overall scenario**.

Tables 1 and 3 provide the results of the 3×2 ANOVA to measure the significance of the factors, Explanations and Preference Index, and the interaction effect of Preference Index and Explanation on Satisfaction and Acceptance respectively. Tables 2 and 4 provide the estimation of pairwise contrasts across the three different levels of the factor *Explanations* and the value of the effect size for each pair of treatments. The values of partial eta-squared in Tables 1 and 3 as well as the value of Cohen's-d in the Tables 2 and 4 provide the value of the effect size for the F test and the pairwise comparison of the explanations respectively. The effect sizes are also calculated using the ARTool package [2].

Factor	Df	Df.res	F value	Pr($> F$)	Partial eta-sq
PreferenceIndex	1	459	44.1704	$8.5546e-11$	0.0877841
Explanation	2	459	4.5561	0.010982	0.0194660
PreferenceIndex:Explanation	2	459	1.2947	0.274991	0.0056096

Table 1: Results of the 3×2 ANOVA for the comparison of Satisfaction across different levels of Explanations and Preference Index.

Contrast	Estimate	SE	df	t.ratio	p.value	Effect Size(d)
Feature-Crowd	-1.74	15.2	459	-0.114	0.9928	-0.0131
Feature-None	38.40	15.1	459	2.551	0.0297	0.2890
Crowd-None	40.14	15.1	459	2.661	0.0219	0.3021

Table 2: Estimation of Pairwise Contrasts for all levels of Explanation and effect size for the 3×2 ANOVA for the comparison of Satisfaction across different levels of Explanations and Preference Index.

Factor	Df	Df.res	F value	Pr($> F$)	Partial eta-sq
PreferenceIndex	1	459	27.8200	$2.0555e-07$	0.057146
Explanation	2	459	5.9870	0.0027115	0.025424
PreferenceIndex:Explanation	2	459	3.9967	0.0190197	0.017117

Table 3: Results of the 3×2 ANOVA for the comparison of Acceptance across different levels of Explanations and Preference Index.

3.1.1 Pairwise ANOVA for Satisfaction

We further conduct 2×2 ANOVA with Aligned Rank Transform for pairwise comparisons of the explanations along with the effect of Preference Index for Satisfaction.

Contrast	Estimate	SE	df	t.ratio	p.value	Effect Size(d)
Feature-Crowd	27.5	15.0	459	1.832	0.1603	0.209
Feature-None	51.5	14.9	459	3.459	0.0017	0.392
Crowd-None	24.0	14.9	459	1.608	0.2432	0.183

Table 4: Estimation of Pairwise Contrasts for all levels of Explanation and effect size for the 3×2 ANOVA for the comparison of Acceptance across different levels of Explanation and Preference Index.

Pairwise Comparison of No Explanations to Feature-based Explanations Table 5 denotes the results of the 2×2 ANOVA to compare the effect of Explanations and Preference Index. Table 6 denotes the estimation of pairwise contrasts for the two types of explanations, None and Feature-based along with the estimation of Cohen's d. The value of partial eta-squared in Table 5 and Cohen's d in Table 6 denote the values of effect size for the F test and pairwise estimation of contrasts respectively.

Factor	Df	Df.res	F value	$\Pr(>F)$	Partial eta-sq
PreferenceIndex	1	308	18.6940	$2.0748e-05$	0.0572218
Explanation	1	308	7.2127	0.0076314	0.0228820
PreferenceIndex:Explanation	1	308	1.9377	0.1649268	0.0062518

Table 5: Results of the 2×2 ANOVA for the pairwise comparison of Satisfaction across Explanations (Feature-based Explanations and No Explanations) and Preference Index.

Contrast	Estimate	SE	df	t.ratio	p.value	Effect Size(d)
Feature-None	27.1	10.1	308	2.686	0.0076	0.304

Table 6: Estimation of pairwise contrasts for the two levels of Explanation for Satisfaction, None and Feature-based.

Pairwise Comparison of No Explanations to Crowdsourced Explanations Table 7 denotes the results of the 2×2 ANOVA to compare the effect of Explanations and Preference Index. Table 8 denotes the estimation of pairwise contrasts for the two types of explanations, None and Feature-based along with the estimation of Cohen's d. The value of partial eta-squared in Table 7 and Cohen's d in Table 8 denote the values of effect size for the F test and estimation of pairwise contrasts respectively.

Factor	Df	Df.res	F value	$\Pr(>F)$	Partial eta-sq
PreferenceIndex	1	307	16.6306	$5.7919e-05$	0.0513875
Explanation	1	307	6.5196	0.011153	0.0207948
PreferenceIndex:Explanation	1	307	1.8635	0.173219	0.0060335

Table 7: Results of the 2×2 ANOVA for the pairwise comparison of Satisfaction across Explanations (Crowdsourced Explanations and No Explanations), and Preference Index.

Contrast	Estimate	SE	df	t.ratio	p.value	Effect Size(d)
Crowd-None	25.8	10.1	307	2.553	0.0112	0.29

Table 8: Estimation of pairwise contrasts for the two levels of Explanations for Satisfaction, None and Crowdsourced.

Pairwise Comparison of Crowdsourced and Feature-based Explanations Table 9 denotes the results of the 2×2 ANOVA to compare the effect of Explanations and Preference Index. Table 10 denotes the estimation of pairwise contrasts for the two types of explanations, Crowdsourced and Feature-based along with the estimation of Cohen's d. The value of partial eta-squared in Table 9

and Cohen's d in Table 10 denote the values of effect size for the F test and estimation of pairwise contrasts respectively.

Factor	Df	Df.res	F value	Pr($> F$)	Partial eta-sq
PreferenceIndex	1	303	8.571939	0.0036725	0.02751191
Explanation	1	303	0.047317	0.82794663	0.00015614
PreferenceIndex:Explanation	1	303	0.067198	0.7956374	0.00022173

Table 9: Results of the 2×2 ANOVA for the pairwise comparison of Satisfaction across Crowdsourced Explanations and Feature-based Explanations, and Preference Index.

Contrast	Estimate	SE	df	t.ratio	p.value	Effect Size(d)
Feature-Crowd	-2.18	10	303	-0.218	0.8279	-0.0248

Table 10: Estimation of pairwise contrasts for the two levels of Explanations for Satisfaction, Crowdsourced and Feature-based.

3.1.2 Pairwise ANOVA for Acceptance

We also conduct 2×2 ANOVA with Aligned Rank Transform for pairwise comparisons of the Explanations along with the effect of Preference Index for Acceptance.

Pairwise Comparison of No Explanations to Feature-based Explanations Table 11 denotes the results of the 2×2 ANOVA to compare the effect of Explanations and Preference Index. Table 12 denotes the estimation of pairwise contrasts for the two types of explanations, None and Feature-based along with the estimation of Cohen's d. The value of partial eta-squared in Table 11 and Cohen's d in Table 12 denote the values of effect size for the F test and estimation of pairwise contrasts respectively.

Factor	Df	Df.res	F value	Pr($> F$)	Partial eta-sq
PreferenceIndex	1	308	11.7504	0.00069104	0.036749
Explanation	1	308	7.0117	0.00851484	0.022259
PreferenceIndex:Explanation	1	308	3.3098	0.06983972	0.010632

Table 11: Results of the 2×2 ANOVA for the pairwise comparison of Acceptance across Feature-based Explanations and No Explanations and Preference Index.

Contrast	Estimate	SE	df	t.ratio	p.value	Effect Size(d)
Feature-None	26.5	10	308	2.648	0.085	0.3

Table 12: Estimation of pairwise contrasts for the two levels of Explanations for Acceptance, Feature-based and None.

Pairwise Comparison of No Explanations to Crowdsourced Explanations Table 13 denotes the results of the 2×2 ANOVA to compare the effect of Explanations and Preference Index. Table 14 denotes the estimation of pairwise contrasts for the two types of explanations, None and Crowdsourced along with the estimation of Cohen's d. The value of partial eta-squared in Table 13 and Cohen's d in Table 14 denote the values of effect size for the F test and estimation of pairwise contrasts respectively.

Pairwise Comparison of Feature-based Explanations to Crowdsourced Explanations Table 15 denotes the results of the 2×2 ANOVA to compare the effect of Explanations and Preference Index. Table 16 denotes the estimation of pairwise contrasts for the two types of explanations, Feature-based and Crowdsourced along with the estimation of Cohen's d. The value of partial eta-squared in Table 15 and Cohen's d in Table 16 denote the values of effect size for the F test and estimation of pairwise contrasts respectively.

Factor	Df	Df.res	F value	Pr($> F$)	Partial eta-sq
PreferenceIndex	1	307	7.5248	0.0064432	0.023924
Explanation	1	307	6.0556	0.014412	0.019344
PreferenceIndex:Explanation	1	307	3.9220	0.0485506	0.012614

Table 13: Results of the 2×2 ANOVA for the pairwise comparison of Acceptance across Crowdsourced Explanations and No Explanations, and Preference Index.

Contrast	Estimate	SE	df	t.ratio	p.value	Effect Size(d)
Crowd-None	24.7	10	307	2.461	0.0144	0.279

Table 14: Estimation of pairwise contrasts for the two levels of Explanations for Acceptance, Crowdsourced and None.

Factor	Df	Df.res	F value	Pr($> F$)	Partial eta-sq
PreferenceIndex	1	303	$1.9538e+01$	$1.3757e-05$	$6.0575e-02$
Explanation	1	303	$7.5120e+00$	0.0064929	$2.4192e-02$
PreferenceIndex:Explanation	1	303	$7.8333e-04$	0.9776902	$2.5852e-06$

Table 15: Results of the 2×2 ANOVA for the pairwise comparison of Acceptance across Crowdsourced Explanations and Feature-based Explanations, and Preference Index.

Contrast	Estimate	SE	df	t.ratio	p.value	Effect Size(d)
Crowd-None	26.8	9.78	303	2.741	0.0065	0.313

Table 16: Estimation of pairwise contrasts for the two levels of Explanations for Acceptance, Crowdsourced and Feature-based.

Inference The aim of the study is to device effective explanations that increase the satisfaction and acceptance of the participants from the winning candidate while considering their utility from the winning candidate (Preference Index). As mentioned in the paper, the concept of utility is central to social choice mechanisms. The results of the 3×2 ANOVA (Tables 1 and 3) indeed reveal that there is a statistically significant difference between the individual levels of both of these factors. The values of the partial eta-squared in Tables 1 and 3 and Cohen's d in Tables 2 and 4 show that the explanations (Feature-based and Crowdsourced) have a non-negligible impact on Satisfaction and Acceptance [1]. The pairwise ANOVA for the comparison of the effect of explanations (Feature-based and Crowdsourced) to no explanations also reveal a statistically significant difference in satisfaction and acceptance between explanations (Feature-based and Crowdsourced) and no explanations (Tables 5, 7, 11 and 13). Since the difference is non-negligible (Tables 6, 8, 12 and 14) This provides us with the motivation for conducting post-hoc tests to uncover significant patterns when providing the participant with explanations. Section 3.2 provides with the detailed analysis of the same. In fact, we do observe that when the winning candidate is the participant's third preference, there is a positive impact on Satisfaction and Acceptance i.e., when the participant has the least utility from the winning candidate, providing her with explanations increases her Satisfaction and Acceptance as compared to not providing her with explanations.

A comparison of the Feature-based explanations to Crowdsourced explanations is essential to ascertain the practicality of the former. Feature-based explanations can be considered an alternative to crowdsourced explanations only if they are at least as effective as the latter with respect to the parameters Satisfaction and Acceptance. The pairwise ANOVA on the Satisfaction (Table 9) show that there is no such significant difference while the pairwise ANOVA on acceptance show the opposite (Tables 15 and 16). Section 3.3 provides a detailed breakdown of the comparison of Crowdsourced and Feature-based explanations. We do observe that the significant difference observed in acceptance does not translate in the post-hoc tests i.e., there is no statistically significant difference between Feature-based and Crowdsourced Explanations when comparing them per Preference Index and per Instance. Hence, the Feature-based explanations perform on par with the Crowdsourced Explanations.

3.2 Statistical Tests for the Impact of Explanations

The aim of providing explanations of any form is to increase the Satisfaction from and Acceptance of the winning candidate. From Section 3.1, we observe that there is a statistically significant difference in satisfaction due to the two levels of Preference Index and hence the data is split on the basis of Preference Index. A one-tailed Mann-Whitney-Wilcoxon test is conducted with the hypothesis that "Providing explanations increases satisfaction in comparison to not providing explanations". We test the two pairs Feature-based Explanations & No Explanations and, Crowdsourced Explanations & No Explanations when the winning candidate is the Second Preference and the Third Preference of the participant. Tables 17 and 18 show the p-values obtained from the one-tailed Mann-Whitney-Wilcoxon test for the two pairs of explanations when the winning candidate is the participant's second preference and third preference respectively. **A visual representation of the values of satisfaction and acceptance with respect to Preference Index is provided in Figures 5 and 6 respectively in the paper**.

	Treatments Compared	
Parameter	None vs Feature	None vs Crowd
Satisfaction	0.157	0.168
Acceptance	0.264	0.288

Table 17: p-values for the one-tailed Mann-Whitney-Wilcoxon Test when the winning candidate is the participant's second preference.

	Treatments Compared	
Parameter	N vs M	N vs H
Satisfaction	0.01	0.005
Acceptance	0.029	0.0042

Table 18: p-values for the one-tailed Mann-Whitney-Wilcoxon Test when the winning candidate is the participant's third preference.

We also conduct an instance-wise tests for increase in Satisfaction and Acceptance for interesting insights. Tables 19 and 20 provide us with the results of the same. **The graphs for instance-wise values of Satisfaction and Acceptance are given in Figures 3 and 4 in the paper.**

	Treatments Compared					
Parameter	#1	#2	#3	#3a	#4	#4a
Satisfaction	0.3832	0.2074	0.4879	0.0574	0.0155	0.6079
Acceptance	0.4898	0.2004	0.3604	0.0734	0.0349	0.3787

Table 19: p-values for the Mann-Whitney-Wilcoxon Test comparing the feature-based explanations to no explanations for all of the 6 instances .

Inference We can observe that when the winning candidate is the participant's third preference, explanations, feature-based or crowdsourced, increase the participant's satisfaction and acceptance (Table 18). From the instance-wise comparison, we can see that in instances 3a (Figure 6) and 4 (Figure 7), explanations do lead to an increase in the satisfaction and acceptance (Tables 19 and 20). We report the same in the paper.

An additional measure for the impact of explanations is that the participant needs to be convinced that no other candidate is a more justified winner than the current winning candidate. This is measured by the response to Question 2a on "Alternate Winner". The proportion of participants that feel that an alternate winner is justified is calculated across the three treatments of explanations. A one-tailed Z-test for proportions is conducted to compare if explanations result in decreasing the

	Treatments Compared					
Parameter	#1	#2	#3	#3a	#4	#4a
Satisfaction	0.1512	0.263	0.525	0.015	0.1054	0.7522
Acceptance	0.1678	0.3833	0.3568	0.0793	0.1272	0.2405

Table 20: p-values for the Mann-Whitney-Wilcoxon Test comparing the crowdsourced explanations to no explanations for all of the 6 instances .

proportion of participants that feel so. The p-value while comparing no explanations to feature-based explanations is 0.006. The p-value for the same comparison between no explanations and crowdsourced explanations is 0.2994. This reveals that the feature-based explanations perform better in comparison to the crowdsourced explanations with respect to convincing the participant. **Figure 7 in the paper denotes the values of the aforementioned proportions and the superiority of feature-based explanations with respect to the same is clearly visible.**

3.3 Comparison of Feature-based and Crowdsourced Explanations

The goal behind developing feature-based explanations is to replace the costly alternative of crowdsourced explanations. This is possible only when the performance of feature-based explanations is on par with that of crowdsourced explanations i.e. the former explanations need to result in similar levels of satisfaction and acceptance as the latter explanations. In Section 3.1, we observe that the factor Preference Index has a statistically significant difference between its two levels (Second Preference and Third Preference). We also observe that for the parameter Acceptance, there is a statistically significant difference between Feature-based Explanations and Crowdsourced Explanations. To check for significant patterns, we conduct a two-tailed Mann-Whitney-Wilcoxon Test comparing these two explanations on the basis of the aforementioned parameters while splitting them on the basis of the Preference Index. Table 21 denote the p-values for the comparison of the two different explanations based on the parameters Satisfaction and Acceptance, across Preference Index, Second and Third. **Figures 5 and 6 in the paper enable a visualization of this.** It is important to note that the comparison of no explanations to explanations (Feature-based and crowdsourced) was one-tailed because we want explanations to increase the value of satisfaction and acceptance. The comparison of Feature-based explanations and crowdsourced explanations on the other hand is two-tailed because we want the feature-based explanations to perform as well as the crowdsourced ones.

	Parameter	
Preference Index	Satisfaction	Acceptance
Second	0.975	0.96
Third	0.9521	0.9165

Table 21: p-values for the two-tailed Mann-Whitney-Wilcoxon Test to compare Feature-based and Crowdsourced Explanations across Preference Index.

An instance-wise, two-tailed Mann-Whitney-Wilcoxon Test is conducted to check for statistically significant differences between the two types of explanations. Table 22 provides us with the p-values for the same and we can observe that there is indeed no statistically significant difference. **A visual representation of the same is given by Figures 3 and 4 in the paper.** The results obtained by tests based on Preference Index and the 6 different instances establish that the feature-based explanations perform on par with the crowdsourced explanations which is duly reported in the paper.

Inference From tables 21 and 22 it can be observed that there is no statistically significant difference in the performances of the two types of explanations. It is interesting to note that while the results of the pairwise ANOVA in Table 15 shows a statistically significant difference between the Feature-based and Crowdsourced explanations, comparison based on individual levels of Preference Index do not show the same. We attribute this to the inherent difference between the questions that both these

	Treatments Compared					
Parameter	#1	#2	#3	#3a	#4	#4a
Satisfaction	0.5083	0.904	0.9494	0.4724	0.3785	0.8774
Acceptance	0.4211	0.5462	0.9588	0.8646	0.5237	0.648

Table 22: p-values for the Mann-Whitney-Wilcoxon Test comparing the crowdsourced explanations to Feature-based explanations for all of the 6 instances .

tests aim to answer. Using ANOVA, the influence on the dependent/response variable (Acceptance in our context) due to the Predictor variable/Factor (Explanation and Preference Index in our context) is calculated. The hypothesis is that individual levels of the factors do not affect the values of the response variable. The primary reason behind the usage of ANOVA is to identify the comparisons that could offer interesting insights. Several estimates are calculated to test the hypotheses which also results in the loss of degrees of freedom. Post-hoc tests, on the other hand are tailor-made to cater to specific populations. In most cases, they test if the distribution of the two specific populations is significantly different. In other words, post-hoc tests need the *protection* of ANOVA to be conducted in the first place and the presence of no significant difference is not an anomaly. Based on these results, we arrive at the conclusion that the Feature-based explanations perform as well as the Crowdsourced explanations and the former can hence be used to replace the latter which is rather expensive. We also conduct a Z-test for proportions to compare the proportion of candidates preferring an *Alternate Winner* to the winning candidate. The two-tailed Z-test comparing the proportions for feature-based explanations and crowdsourced explanations results in the p-value of 0.0621. However, conducting a one-tailed test based on the hypothesis that the feature-based explanations lead to a reduced proportion of candidates feeling that an *Alternate Winner* is justified compared to the crowdsourced explanations results in a p-value of 0.031, establishing the superiority of feature-based explanations. **Figure 7 in the paper offers a visual representation of the same.**

References

[1] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Tech. rep. 1988.

[2] Matthew Kay. *Effect Sizes with ART*. 2021. URL: https://cran.r-project.org/web/packages/ARTool/vignettes/art-effect-size.html.

[3] Jacob O Wobbrock et al. "The aligned rank transform for nonparametric factorial analyses using only anova procedures". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2011, pp. 143–146.