Testing Project Analysis Report

Automated Report by Climmob

07 October, 2019

You are reading a report generated by ClimMob. This is a software package to analyze data generated by citizen science or crowdsourcing.

Introduction

In agriculture, the local environmental conditions determine to a large degree which technological solutions are the most suitable. In dry soils, for example, drought-resistant crop varieties will outperform other varieties, but in wet soils these same varieties may do worse than most. Not only drought, but an entire range of problems including excessive heat, floods, new pests and diseases tend to intensify under climate change. This multitude of limiting factors requires multiple technological solutions, tested in diverse environments.

Citizen science is based on the cooperation of citizen scientist or observers (paid or unpaid). Researchers assign microtasks (observations, experiments. . .) that, once completed and gathered, contribute with a great amount of information to science. One of the advantages of citizen science is that agricultural researchers can get access to many environments by crowdsourcing their experiments. As farmers contribute with their time, skills and knowledge to the investigation, researchers are able to do more tests than in a traditional setup. Also citizen scientists acquire new knowledge, abilities and information useful for future challenges of their work.

ClimMob

The primary goal of ClimMob is to help farmers adapt to variable and changing climates. ClimMob was created as part of Bioversity International's research in the CGIAR Research Programme on Climate Change, Agriculture, and Food Security (CCAFS). It serves to prepare and analyze citizen science experiments in which a large number of farmers observe and compare different technological options under a wide range environmental conditions (van Etten 2011).

ClimMob software assigns a limited number of items (typically 3 crop varieties or agricultural practices) to each farmer, who will compare their performance. Each farmer gets a different combination of items drawn from a much larger set of items. Comparisons of this kind are thought to be a very reliable way to obtain data from human observers (Martin 2004). Once the results of the microtasks have been collected, ClimMob builds an image of the whole set of assigned objects, combining all observations. ClimMob not only reconstructs the overall ordering of items, but also takes into account differences and similarities between observers and the conditions under which they observe. It assigns similar observers to groups that each corresponds to a different preference profile. Groups are created on the basis of whichever variables which have been collected, that are found to be significantly linked to the preferences.

ClimMob uses a recently published statistical method to analyze ranking data (Strobl et al. 2011). It automatically generates analytical reports, as well as individualized information sheets for each participant.

How to cite

If you publish any results generated with ClimMob, you should cite a number of articles as the package builds on various contributions. Van Etten (2011) introduced the crowdsourcing philosophy behind ClimMob. It is important to mention that ClimMob is implemented in R, a free, open-source analysis software (R Development Core Team 2012). Methodologically, if you report on the tree results, you should mentioned that ClimMob applies the Bradley-Terry tree method published by Strobl et al. (2011). To cite ClimMob itself, mention Van Etten & Calderer (2015).

Section 1: Headline Results

Overall there were 150 farmers participating in this study. Each farmer assessed three different varieties and ranked them in order of their overall preference. In addition they also provided rankings for 3 additional traits:

Short name	Question
difficult	What dance is the most difficult?
style	What dance looks the best?
fun	What dance is the most fun?

Table 1 provides a list of the varieties assessed within this trial, with the frequency and percentage of farmers who assessed each variety.

Table 2: Table 1: Frequency of varieties Assessed

Variety	Frequency	% of Respondents	Man (n=102)	Woman (n=46)
Ballroom	64	43%	40	23
Hip Hop	65	43%	44	21
Morris Dancing	64	43%	43	20
Rave	64	43%	40	23
Rumba	64	43%	46	17
Salsa	65	43%	48	16
Tap	64	43%	45	18

Overall Differences in Rankings

Overall there were no statistically significant differences found in the rankings of varieties in the overall ranking (p=0.19737)

A summary of the p-values testing the hypothesis that there exist differences in the rankings within each of the Plackett-Luce models fitted for each of the assessed traits, and the list of varieties which were significantly highest and lowest ranked overall, are summarised in Table 1.1.

Table 3: Table 1.1: Summary of Differences Found In varieties by Trait

Ranking	p.value	Best Ranked	Worst Ranked
Overall	0.19737	Rumba, Rave, Morris Dancing, Ballroom, Salsa, Tap, Hip Hop	Hip Hop, Tap, Salsa, Ballroom, Morr
difficult	0.72434	No significant differences	No significant differences
style	0.92946	No significant differences	No significant differences
fun	0.31183	No significant differences	No significant differences

See Section 3 for further details.

Effect of covariates

None of the variables tested were found to have a statistically significant relationship to the overall ranking at the 5 % significance level.

A summary of the univariate signifance levels for all covariates that were able to be tested is shown in Table 1.2.1.

Table 4: Table 1.2.1: Summary of univariate p-values for first split in Plackett-Luce tree model for the overall ranking

Variable	p
age	0.29448
gender	0.99079

See Section 4 for further details.

Relationships between traits

Table 1.3 shows, for each trait in the study, the frequency with which the rankings matched with the overall ranking.

The trait(s) which had the strongest relationship with the overall ranking was difficult. Overall the rankings for difficult matched the rankings for the overall ranking 29.7% of the time.

The trait(s) which had the weakest relationship with the overall ranking was fun. Overall the rankings for fun matched the rankings for the overall ranking only 19.4% of the time.

Table 5: Table 1.3: Relationship between individual trait assessment and overall assesment

trait	Complete Ranking Agreement
fun	19%
style	20%
difficult	30%

See Section 5 for further details.

Section 2: Data Summary & Exploratory Analysis

Short name	Question
age	How old are you?
gender	What is the gender ?

Tables 2.2.1 to 2.2.2 summarise the covariate data collected from the survey, with corresponding bar charts or histograms of the responses.

age

Table 7: Table 2.2.1.1 Missing Values in age

age	Freq	Percent
Complete	149	99.3%

age	Freq	Percent
Missing	1	0.7%

Table 8: Table 2.2.1.2 Summary statistics for age

Variable	gender	Mean	Median	Minimum	Maximum	Standard Deviation
age	Man	41.58	34	2	266	35.88
age	Woman	43.17	36	12	122	22.64
age	Overall	42.15	35	2	266	32.19

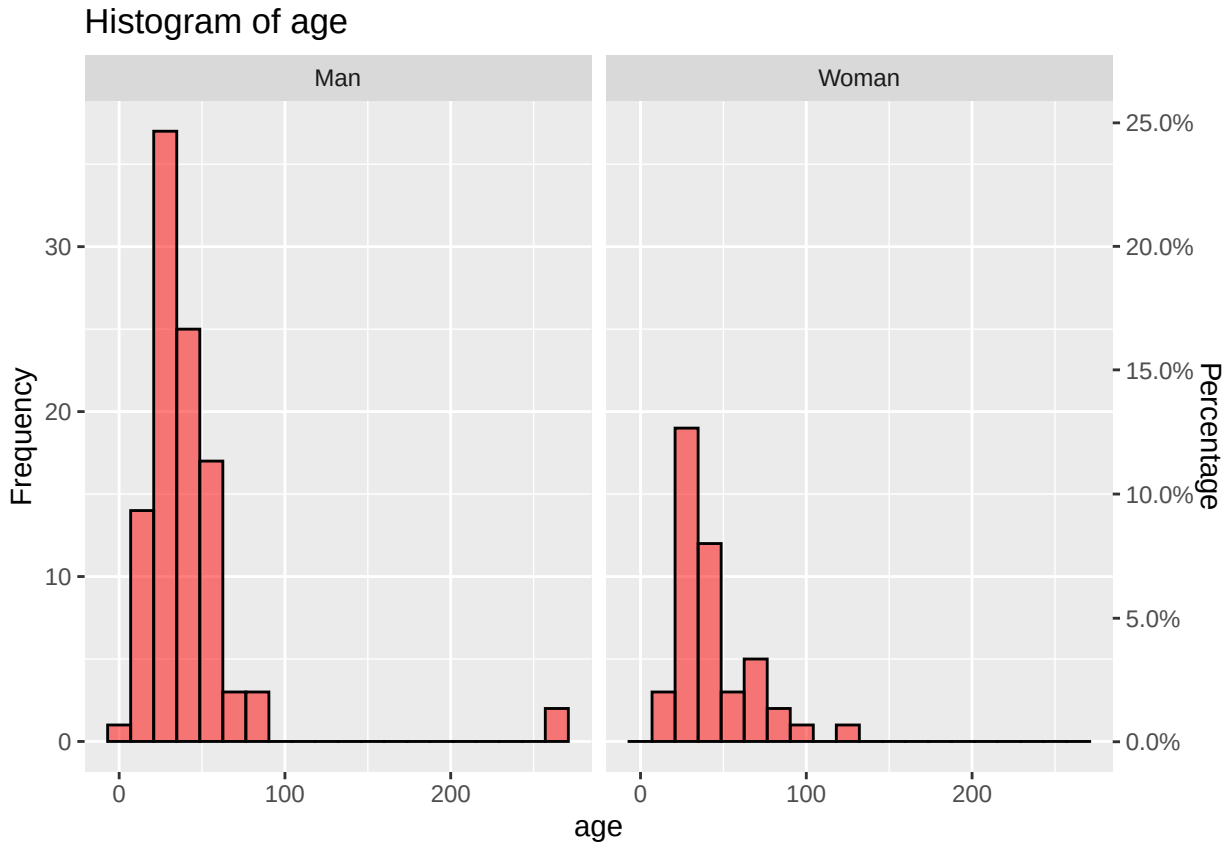


Figure 2.2.1

gender

Table 9: Table 2.2.2.1 Missing Values in gender

gender	Freq	Percent
Complete	148	98.7%
Missing	2	1.3%

Table 10: Table 2.2.2.2 Summary statistics for gender

gender	Frequency	% of Respondents
Man	102	69%
Woman	46	31%

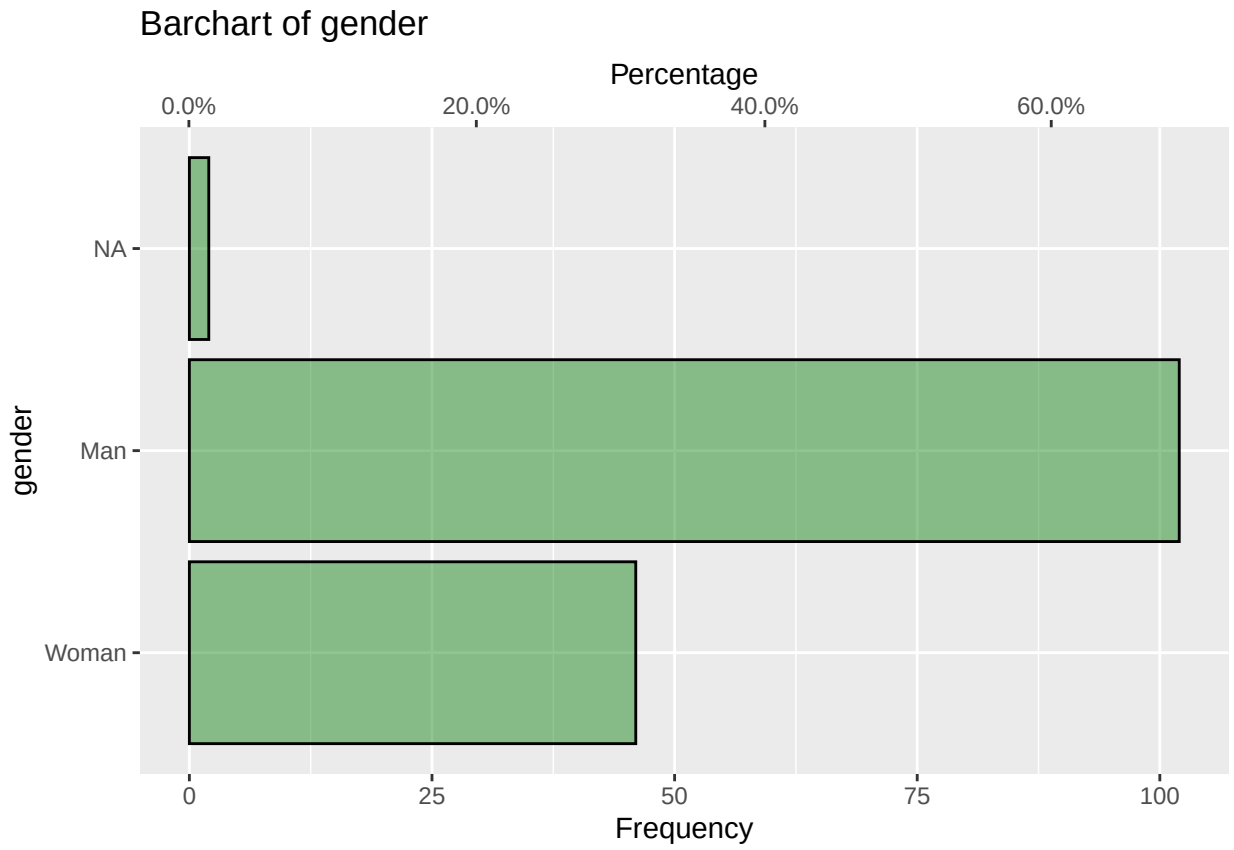


Figure 2.2.2

Location of farmers

Figure 2.3

No coordinates recorded in data

Assessment of varieties

Exploratory analysis within the following section summarises results from the data directly. Given the structure of a climmob trail, where each farmer only assesses 3 of the possible varieties these results may be skewed if certain varieties were randomly assigned to face worse varieties than others. This is particularly a potential issue within a smaller trial, as due to the randomisation process the potential for an unbalanced assignment decreases as the sample size increases. Results from other sections, and in the overall summary use Plackett-Luce models, to adjust for any imbalance [add in references].

Overall

Overall performance of each of the varieties is summarised in Table 2.4.

Table 11: Table 2.4 - Summary of Overall Performance

Variety	N	% Top Ranked	% Bottom Ranked	% Contests Won	Net Favourability Score
Rumba	62	45.2%	21.0%	62.1%	24.2
Rave	64	40.6%	31.2%	54.7%	9.4
Morris Dancing	63	28.6%	28.6%	50.0%	0.0
Ballroom	64	28.1%	32.8%	47.7%	-4.7
Salsa	63	31.7%	36.5%	47.6%	-4.8
Tap	62	33.9%	41.9%	46.0%	-8.1
Hip Hop	63	25.4%	41.3%	42.1%	-15.9

This shows the % of farmers who assessed the variety as their most preferred of the 3 varieties they were provided, the % of farmers who included the variety as their least preferred, the % of 'head to head contests' for which the variety won and the net favourability score. A score of +100 indicates the variety won all 'contests' it was involved in, a score of 0 indicates an equal number of wins and losses, a score of -100 indicates the variety lost all contests.

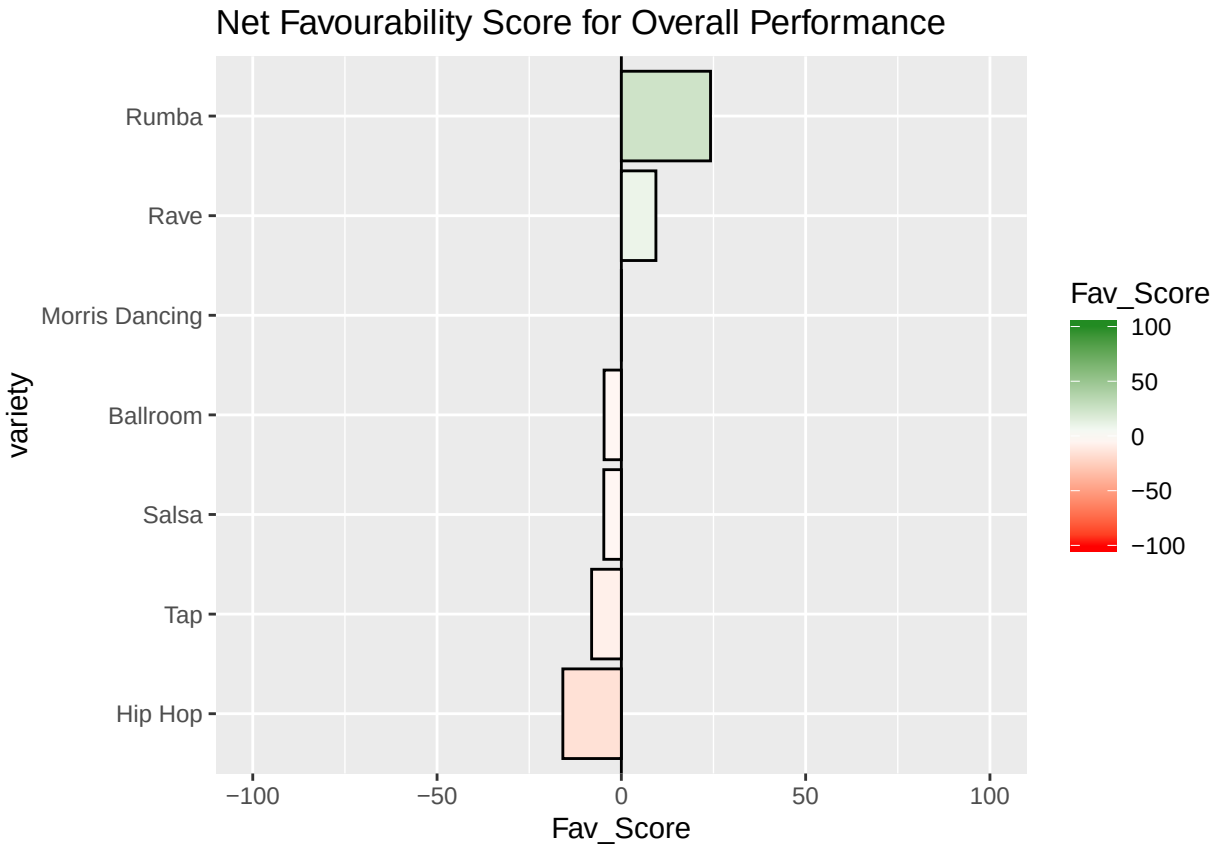


Figure 2.4

The variety Rumba was the most preferred variety overall being ranked highest by 45.2% of the 62 farmers who assessed this variety.

Other varieties with strong positive rankings were

Other Traits

Net favourability scores are shown below for the other traits assessed in this study.

What dance is the most difficult?

Table 12: Table 2.5.1: Favourability Statistics for difficult

variety	N	% Top Ranked	% Bottom Ranked	% Contests Won	Net Favourability Score
Rumba	61	45.9%	32.8%	56.6%	13.1
Morris Dancing	63	39.7%	28.6%	55.6%	11.1
Ballroom	63	31.7%	30.2%	50.8%	1.6
Tap	61	32.8%	32.8%	50.0%	0.0
Salsa	61	34.4%	36.1%	49.2%	-1.6
Rave	63	23.8%	31.7%	46.0%	-7.9
Hip Hop	63	25.4%	41.3%	42.1%	-15.9

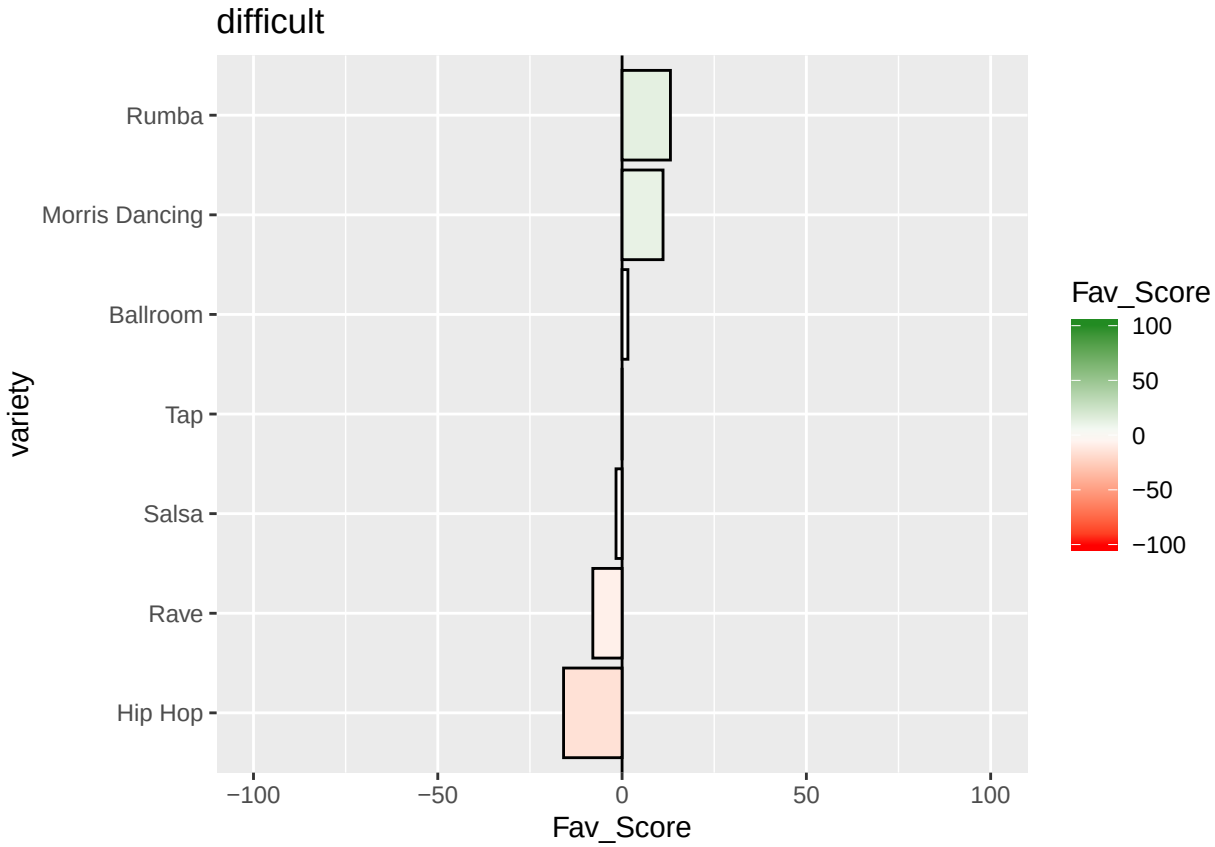


Figure 2.5.1

What dance looks the best?

Table 13: Table 2.5.2: Favourability Statistics for style

variety	N	% Top Ranked	% Bottom Ranked	% Contests Won	Net Favourability Score
Tap	60	33.3%	26.7%	53.3%	6.7
Rave	61	39.3%	34.4%	52.5%	4.9
Ballroom	61	31.1%	26.2%	52.5%	4.9
Morris Dancing	61	31.1%	34.4%	48.4%	-3.3

variety	N	% Top Ranked	% Bottom Ranked	% Contests Won	Net Favourability Score
Rumba	59	37.3%	40.7%	48.3%	-3.4
Hip Hop	62	29.0%	33.9%	47.6%	-4.8
Salsa	59	32.2%	37.3%	47.5%	-5.1

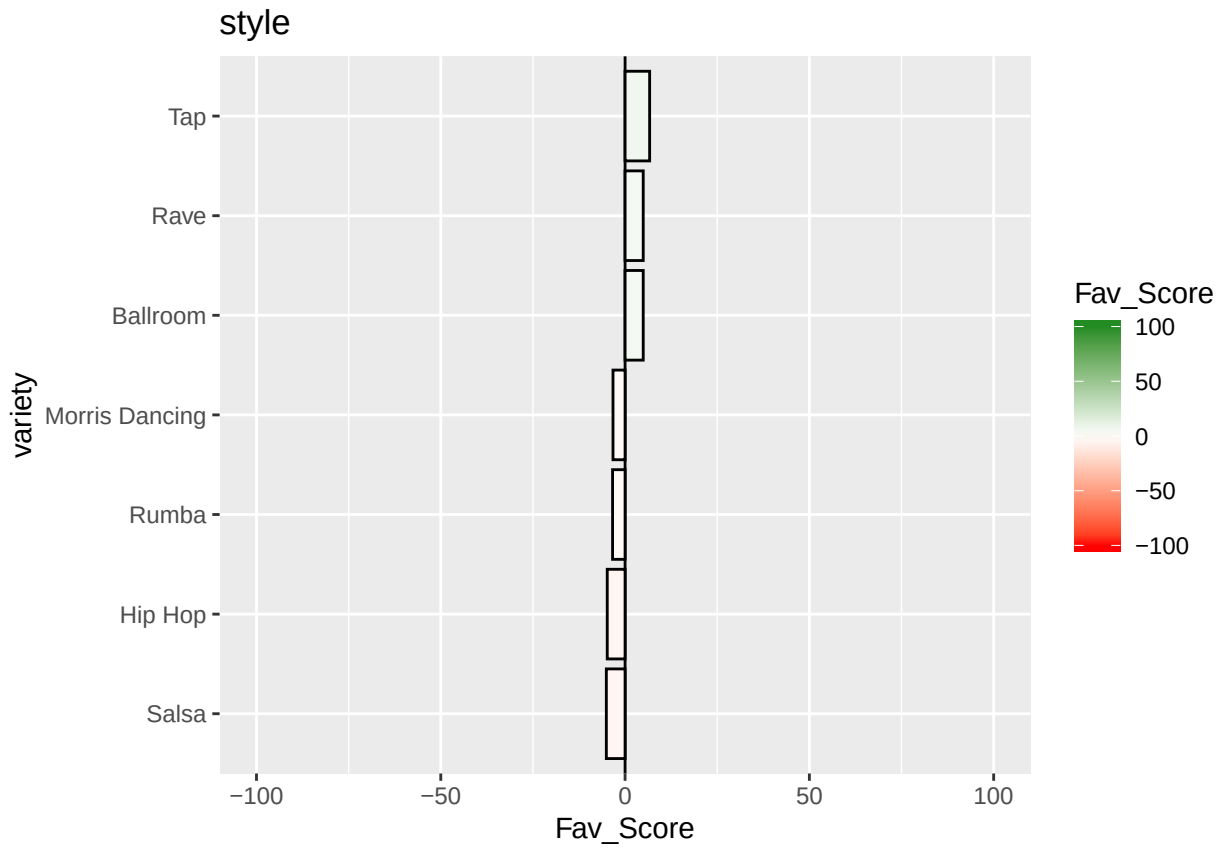


Figure 2.5.2

What dance is the most fun?

Table 14: Table 2.5.3: Favourability Statistics for fun

variety	N	% Top Ranked	% Bottom Ranked	% Contests Won	Net Favourability Score
Rumba	59	39.0%	25.4%	56.8%	13.6
Tap	58	41.4%	29.3%	56.0%	12.1
Hip Hop	61	34.4%	26.2%	54.1%	8.2
Morris Dancing	60	40.0%	35.0%	52.5%	5.0
Ballroom	61	32.8%	41.0%	45.9%	-8.2
Salsa	60	26.7%	35.0%	45.8%	-8.3
Rave	61	19.7%	41.0%	39.3%	-21.3

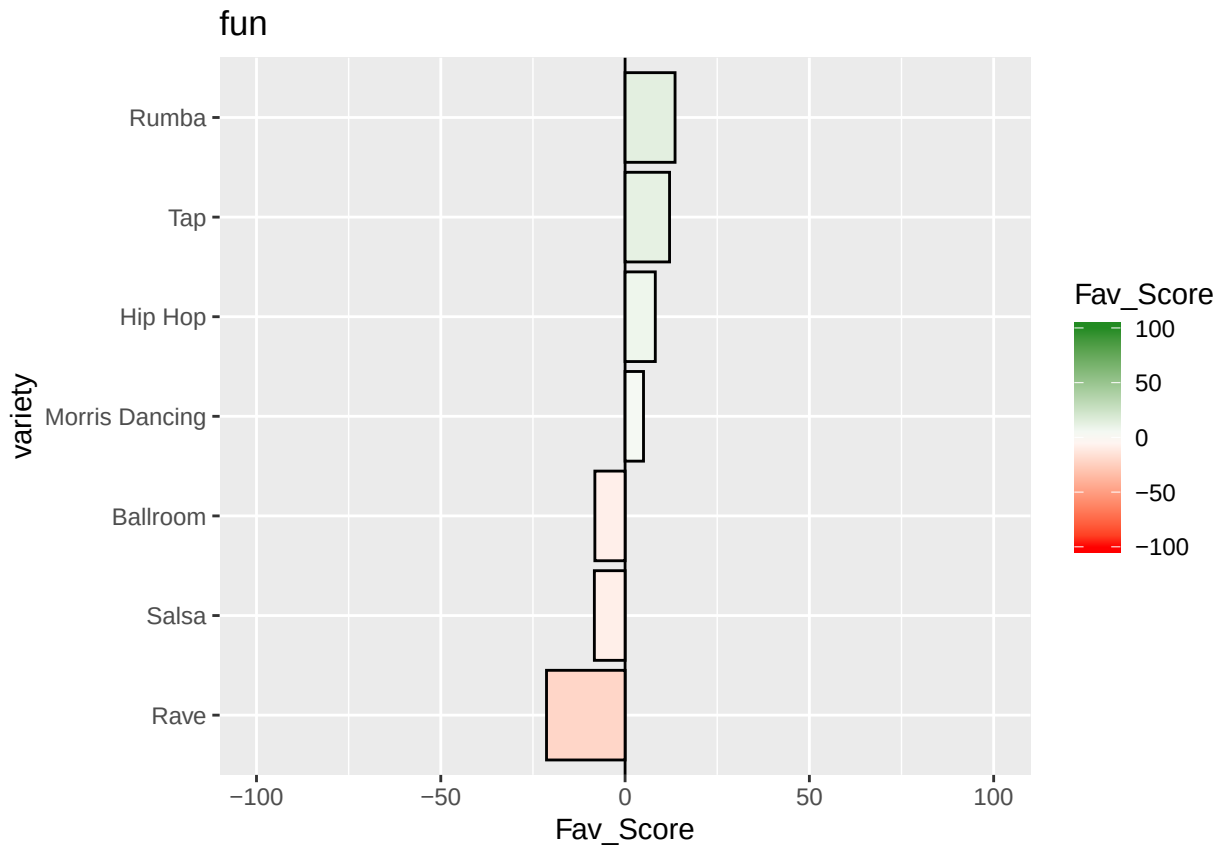


Figure 2.5.3

Pairwise Contests

Appendix B contains the full data for each pairwise comparison of the varieties, which is summarised in the plots below.

Overall

Figure 2.6 shows the outcomes of all pairwise contests between the varieties included in the trial. Each panel shows the performance of one variety against all the other varieties, and shows the % of the times in which the panelled variety was ranked above the varieties shown as bars. The most preferred variety is shown in the top left panel and the least preferred is shown in the bottom right panel

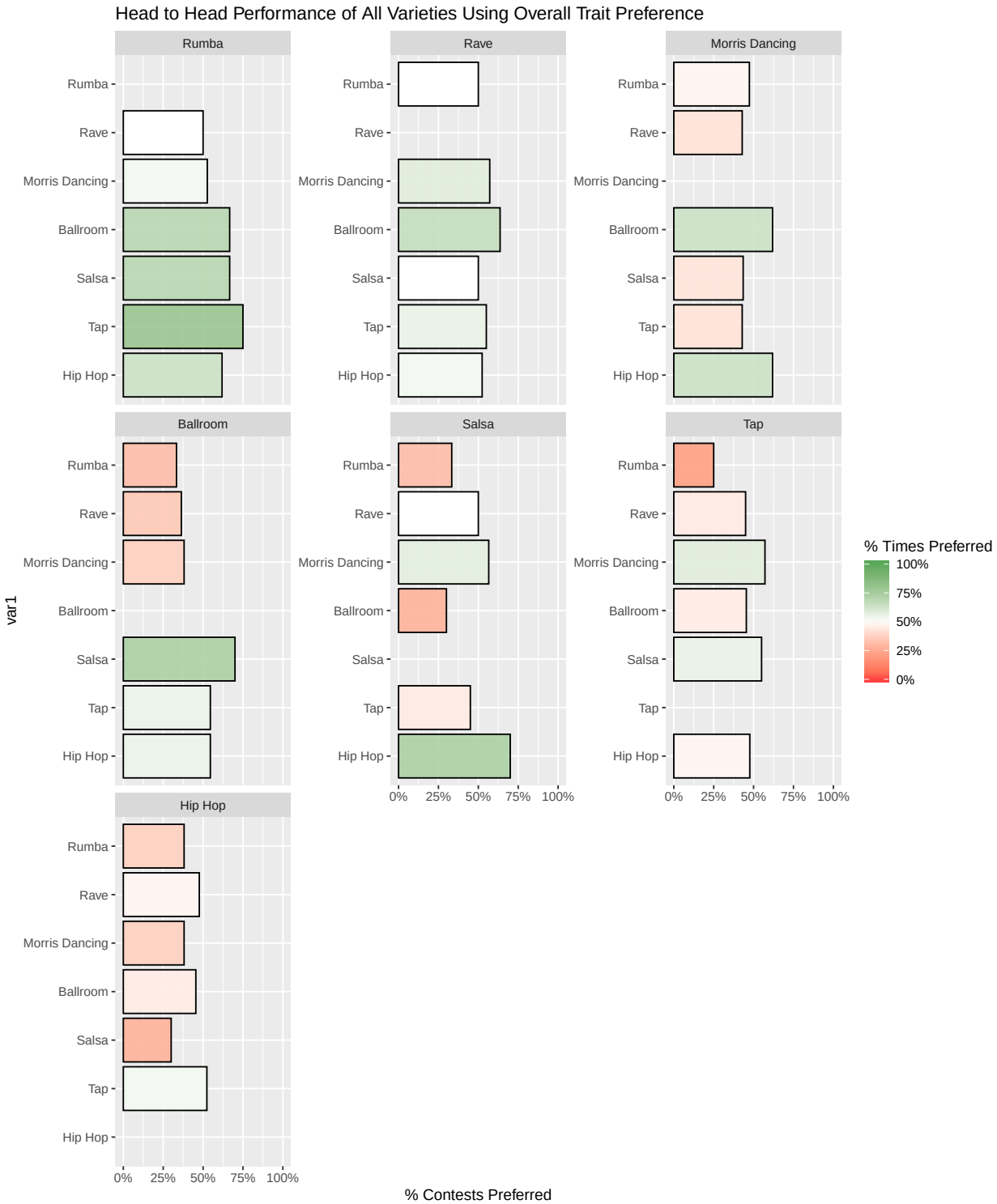


Figure 2.6

Other Traits

Results from the pairwise contests of the other individual traits assessed are shown below.

What dance is the most difficult?

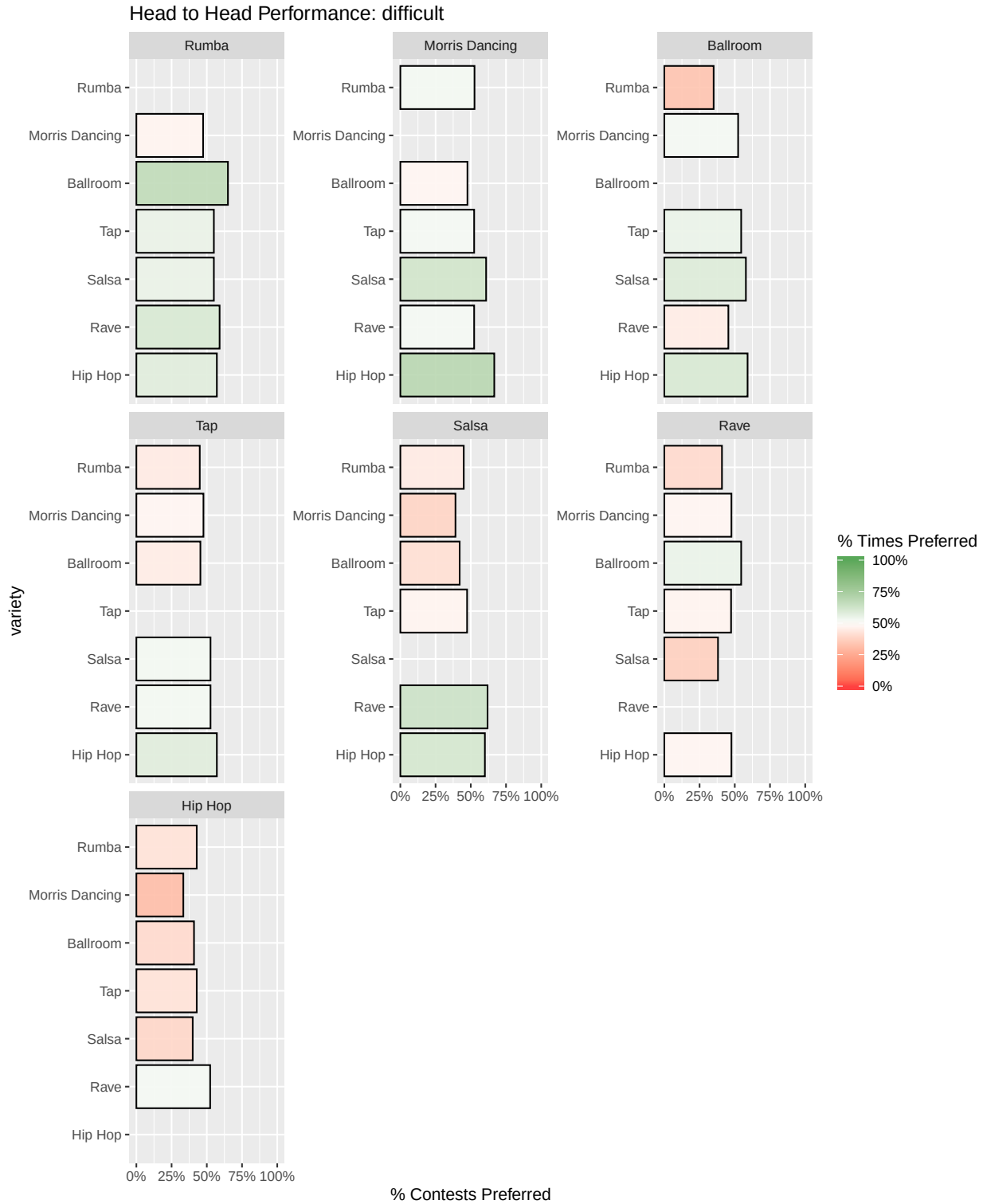


Figure 2.8.1

What dance looks the best?

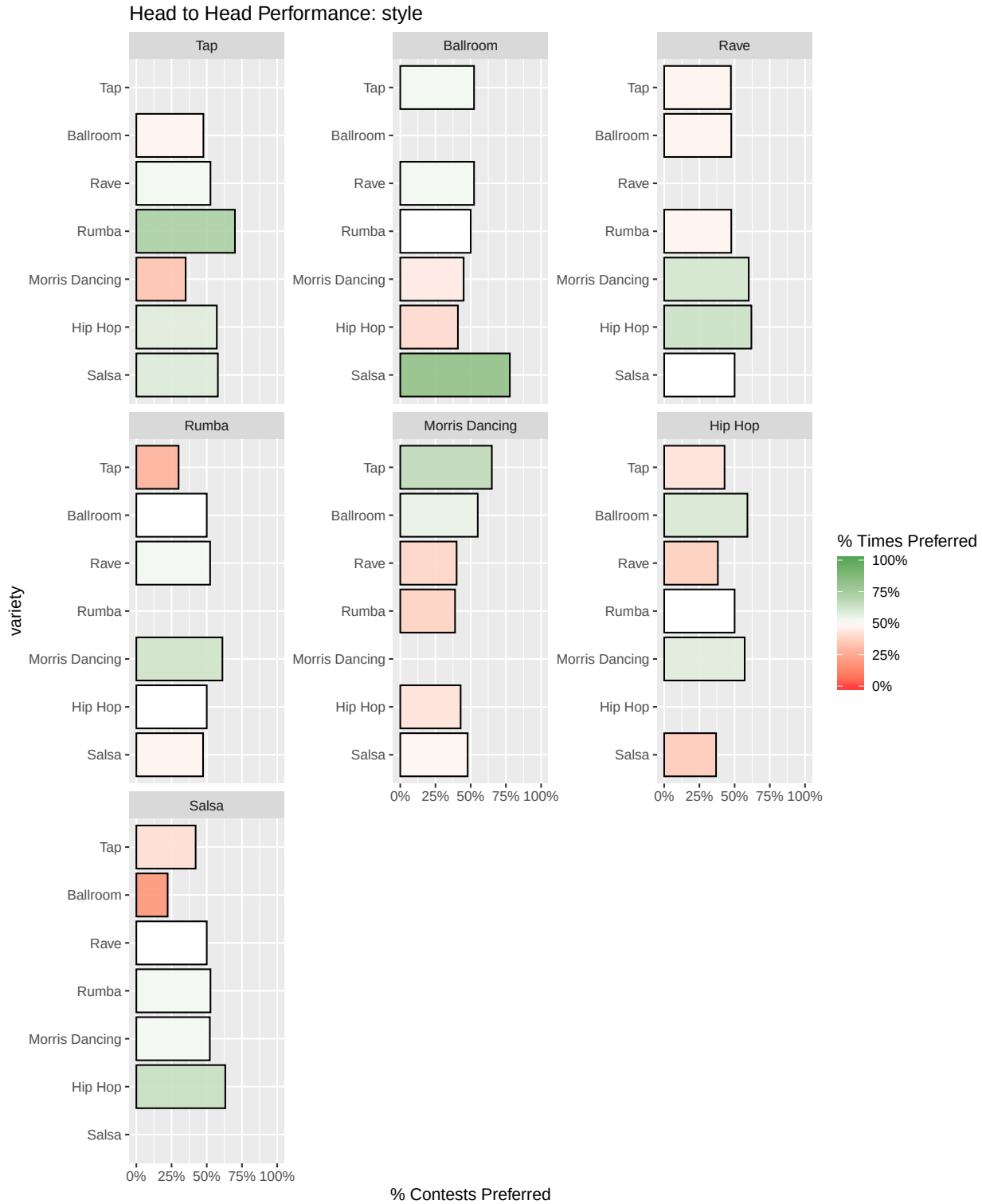


Figure 2.8.2

What dance is the most fun?

Figure 2.8.3

Relationship between traits

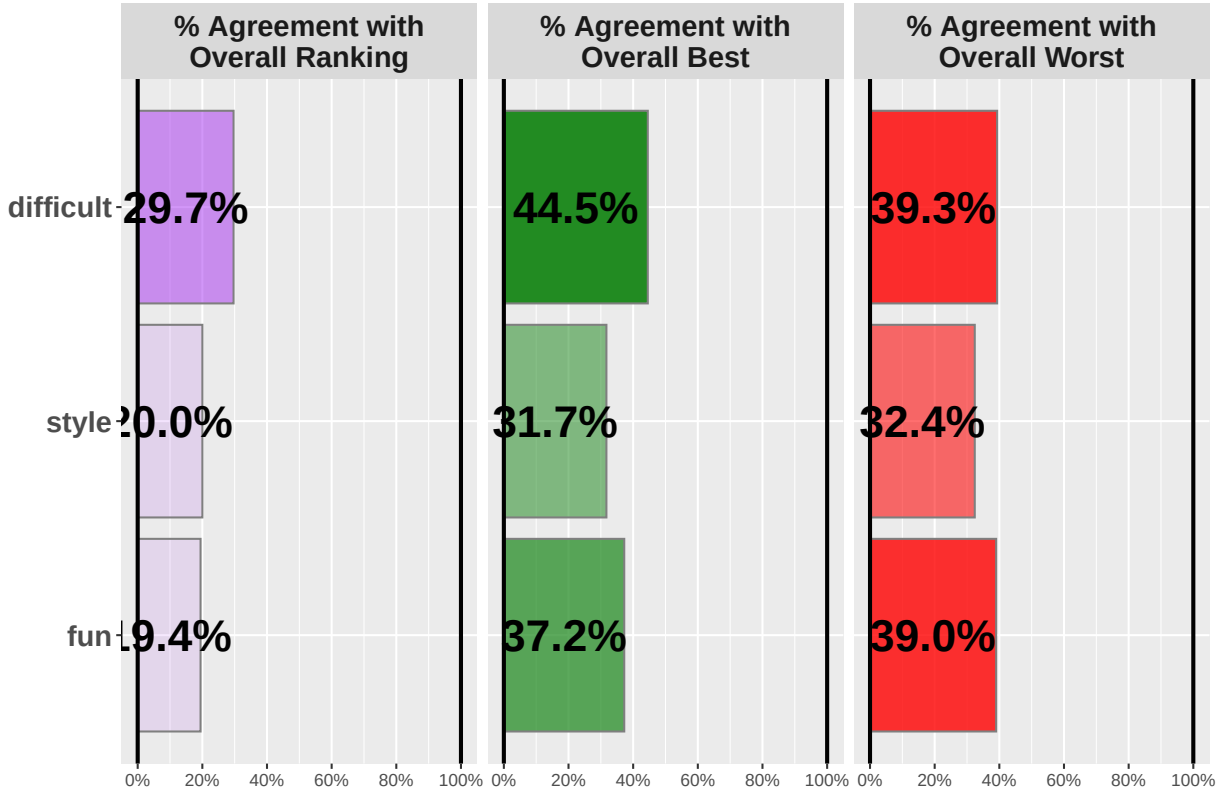
Table 2.9 shows the relationship between the individual trait rankings and the overall rankings. Complete agreement represents the percentage of respondents for which the ranking of the three varieties in

respect to the trait is an exact match to the overall ranking. Given that there are 6 possible ways of ranking the three varieties that each farmer assesses, it would be expected that, if there was no relationship between the ranking of the trait and the overall ranking, that the complete rankings would agree 17% of the time by chance. Best/worst agreement represents the percentage for which the best/worst variety for the trait matched the overall best/worst. Given that there are 3 possible varieties that each farmer assesses, it would be expected that, if there was no relationship between the ranking of the trait and the overall ranking, that the best/worst rankings would still agree 33% of the time by chance.

Table 15: Table 2.9: Relationship between individual trait assessment and overall assessment

trait	Complete Ranking Agreement	Agreement with Overall Best	Agreement with Overall Worst
fun	19%	37%	39%
style	20%	32%	32%
difficult	30%	45%	39%

Figure 2.9



Overall the trait which had the strongest relationship with the overall ranking was difficult, with identical rankings being given as the overall ranking 30% of the time.

Section 3: Plackett-Luce Models of Ranking Differences

Overall Ranking

Table 3.1 shows the results from the likelihood ratio test from the Plackett-Luce model for overall rankings of the different varieties. The hypothesis being tested is that there is no difference in the assessments of any of the different varieties.

Table 16: Table 3.1: Likelihood ratio test results from overall model

model	logLikelihood	DF	Statistic	Pr(>Chisq)
NULL	-263.3886	441	NA	NA
mod_overall	-259.0887	435	8.599806	0.19737

p=0.19737. This suggests that there is not enough evidence to conclude that there are differences between varieties

Figure 3.1 shows the estimates of the model coefficients with 84% confidence intervals. The purpose of this graph is to be able to best distinguish between the relative strength of each of the varieties assessed. As such the coefficient estimates themselves are not directly interpretable, but it can be concluded that a higher value for the coefficient indicates that a variety has been more preferred. The 84% confidence width is chosen so that non-overlapping confidence intervals could be interpreted as indicating significant differences at the 5% significance level. This may not match exactly with the mean separation groupings, as these groupings also take into account multiple testing through the Benjamini and Hochberg adjustment. [add references here]. Mean separation analysis was also conducted to indicate, using letters, which varieties are significantly more preferred than others: when varieties have at least one letter in common, there is not enough evidence from the experiment to be confident about their relative order of preference at the 5% significance level.

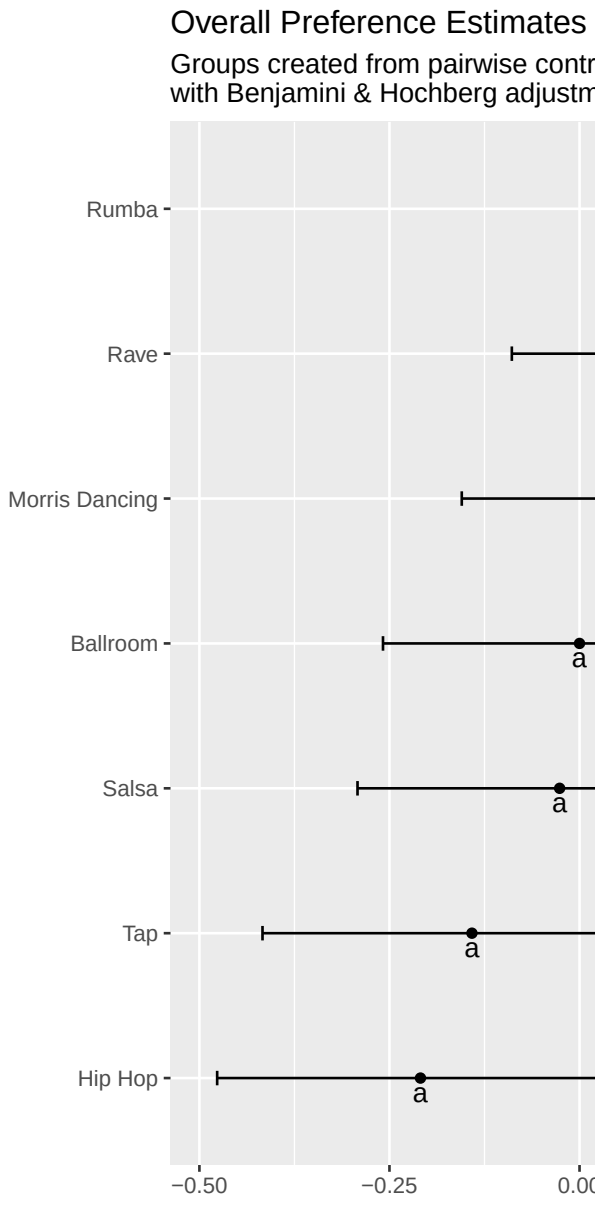


Figure 3.1 - Overall Ranking: Model Coefficients and Mean Separation

The same information as Figure 3.1 is shown in Table 3.2 below

Table 17: Table 3.2 - Model Coefficients and Mean Separation of varieties at 5% Level with Benjamini & Hochberg adjustment

	Estimate	quasiSE	Group
Rumba	0.48	0.19	a
Rave	0.18	0.19	a
Morris Dancing	0.10	0.18	a
Ballroom	0.00	0.18	a
Salsa	-0.03	0.19	a
Tap	-0.14	0.20	a
Hip Hop	-0.21	0.19	a

Table 3.3 and Figure 3.2 use the coefficients from the model to estimate the probability of each variety being considered to be the top ranked variety in a direct comparison between all of the possible varieties

Table 18: Table 3.3: Percentage probability of being the highest ranked overall

	Variety	% Probability
Rumba	Rumba	21.3%
Rave	Rave	15.8%
Morris Dancing	Morris Dancing	14.6%
Ballroom	Ballroom	13.2%
Salsa	Salsa	12.9%
Tap	Tap	11.5%
Hip Hop	Hip Hop	10.7%

Figure 3.2 - Overall Ranking: Probability of Being The Highest Ranked Overall

What dance is the most difficult?

Table 19: Table 3.4.1

model	logLikelihood	DF	Statistic	Pr(>Chisq)
NULL	-259.8051	435	NA	NA
mod_difficult	-257.9817	429	3.646831	0.72434

p=0.72434 , this suggests that there is not enough evidence to conclude that there are differences between

varieties

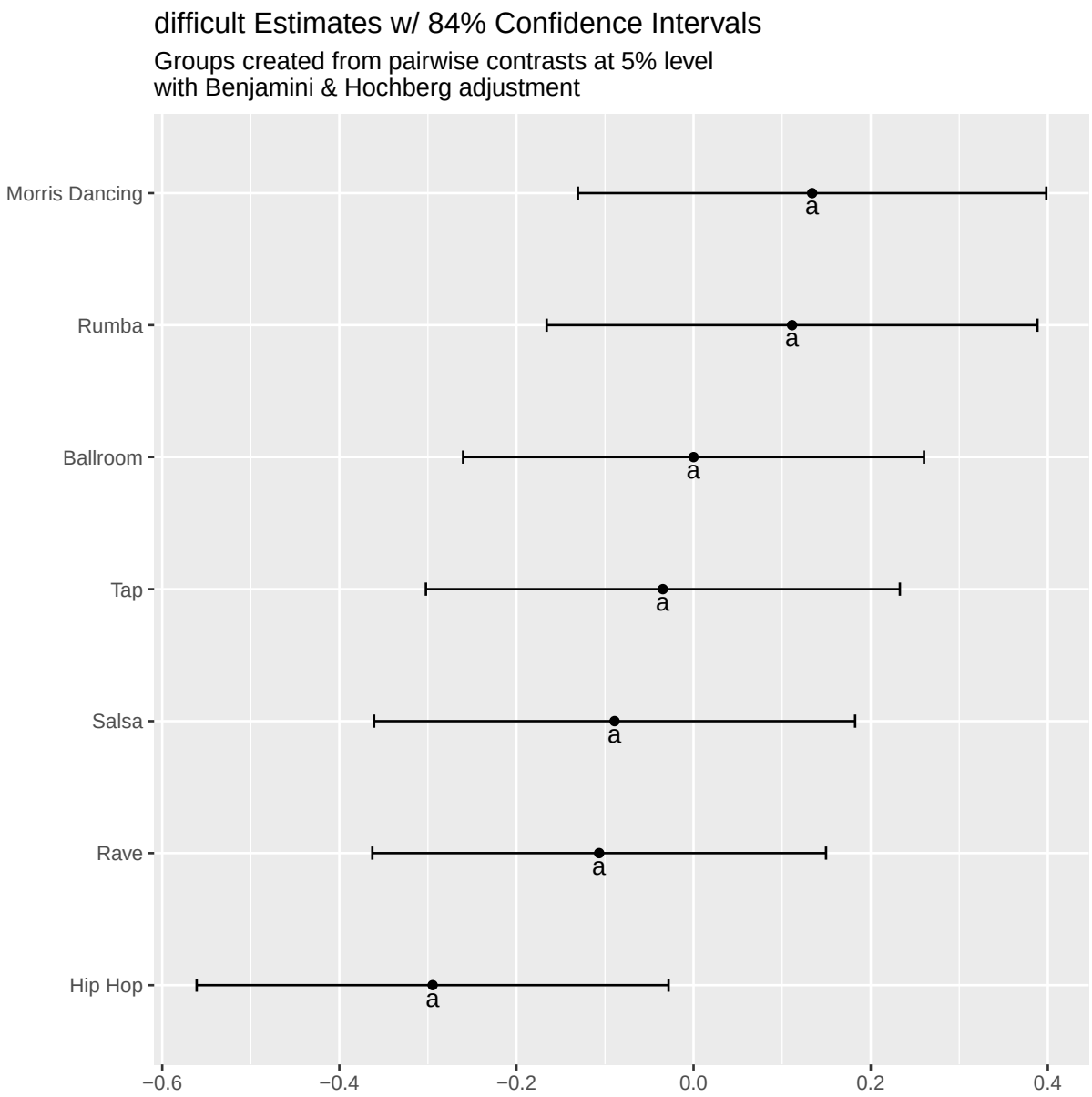


Table 20: Table 3.5. 1 difficult model parameter estimates

	Estimate	quasiSE	Group
Morris Dancing	0.13	0.19	a
Rumba	0.11	0.20	a
Ballroom	0.00	0.19	a
Tap	-0.03	0.19	a
Salsa	-0.09	0.19	a
Rave	-0.11	0.18	a
Hip Hop	-0.29	0.19	a

Probability that variety is the highest ranked difficult

Morris Dancing: 16.85%
Rumba: 16.47%
Ballroom: 14.74%
Tap: 14.24%
Salsa: 13.48%
Rave: 13.25%
Hip Hop: 10.98%

What dance looks the best?

Table 21: Table 3.4.2

model	logLikelihood	DF	Statistic	Pr(>Chisq)
NULL	-252.6381	423	NA	NA
mod_style	-251.6926	417	1.890871	0.92946

p=0.92946 , this suggests that there is not enough evidence to conclude that there are differences between varieties

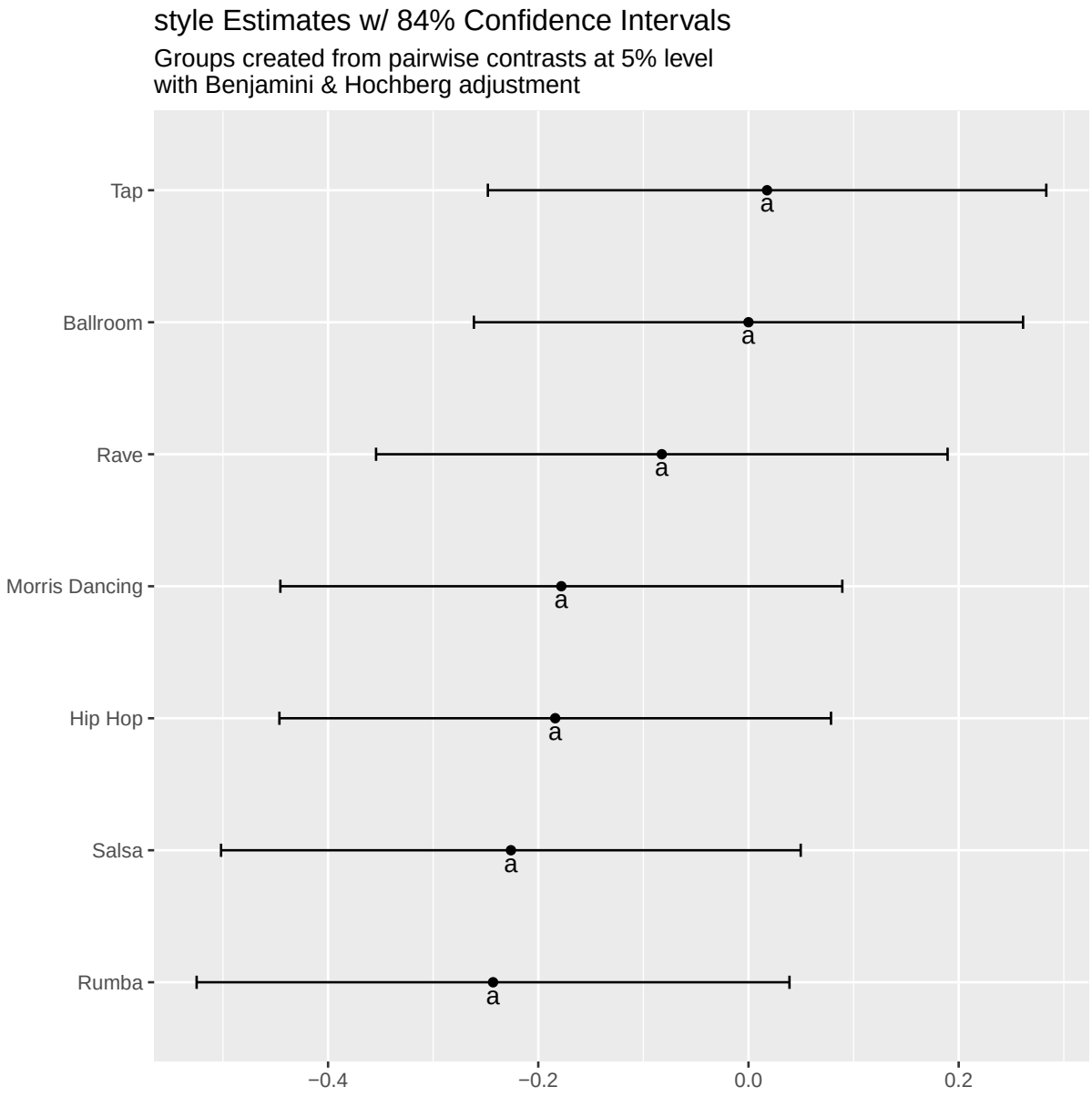
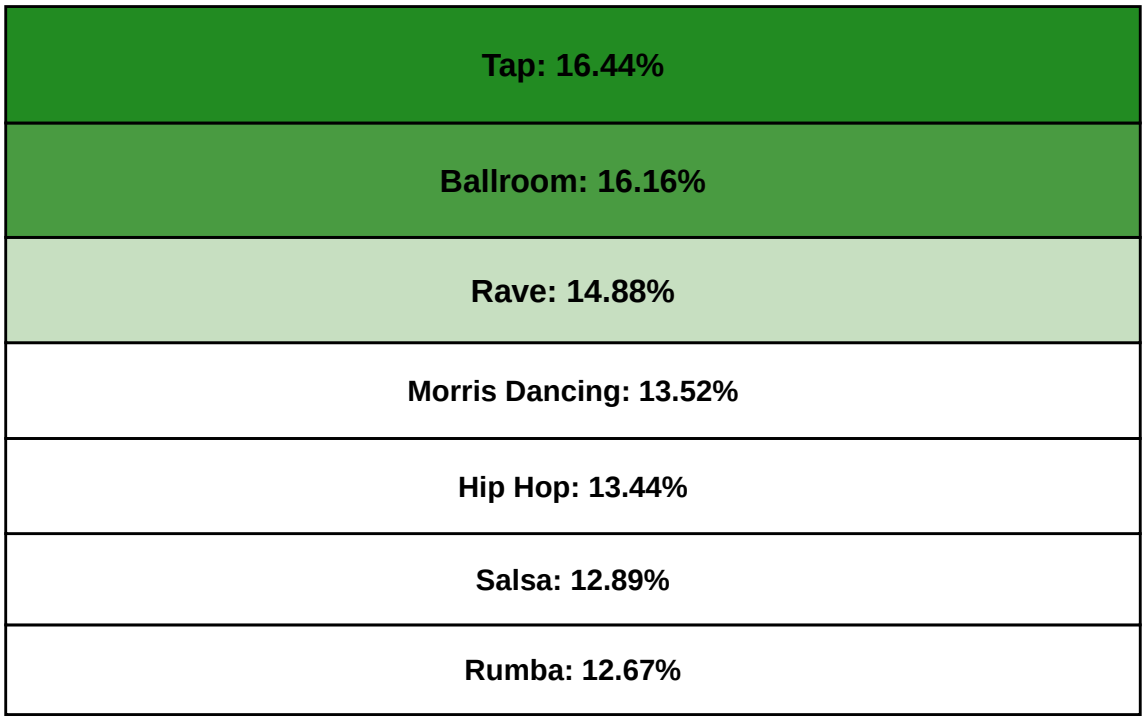


Table 22: Table 3.5. 2 style model parameter estimates

	Estimate	quasiSE	Group
Tap	0.02	0.19	a
Ballroom	0.00	0.19	a
Rave	-0.08	0.19	a
Morris Dancing	-0.18	0.19	a
Hip Hop	-0.18	0.19	a
Salsa	-0.23	0.20	a
Rumba	-0.24	0.20	a

Probability that variety is the highest ranked style



What dance is the most fun?

Table 23: Table 3.4.3

model	logLikelihood	DF	Statistic	Pr(>Chisq)
NULL	-250.8463	420	NA	NA
mod_fun	-247.2971	414	7.098535	0.31183

p=0.31183 , this suggests that there is not enough evidence to conclude that there are differences between varieties

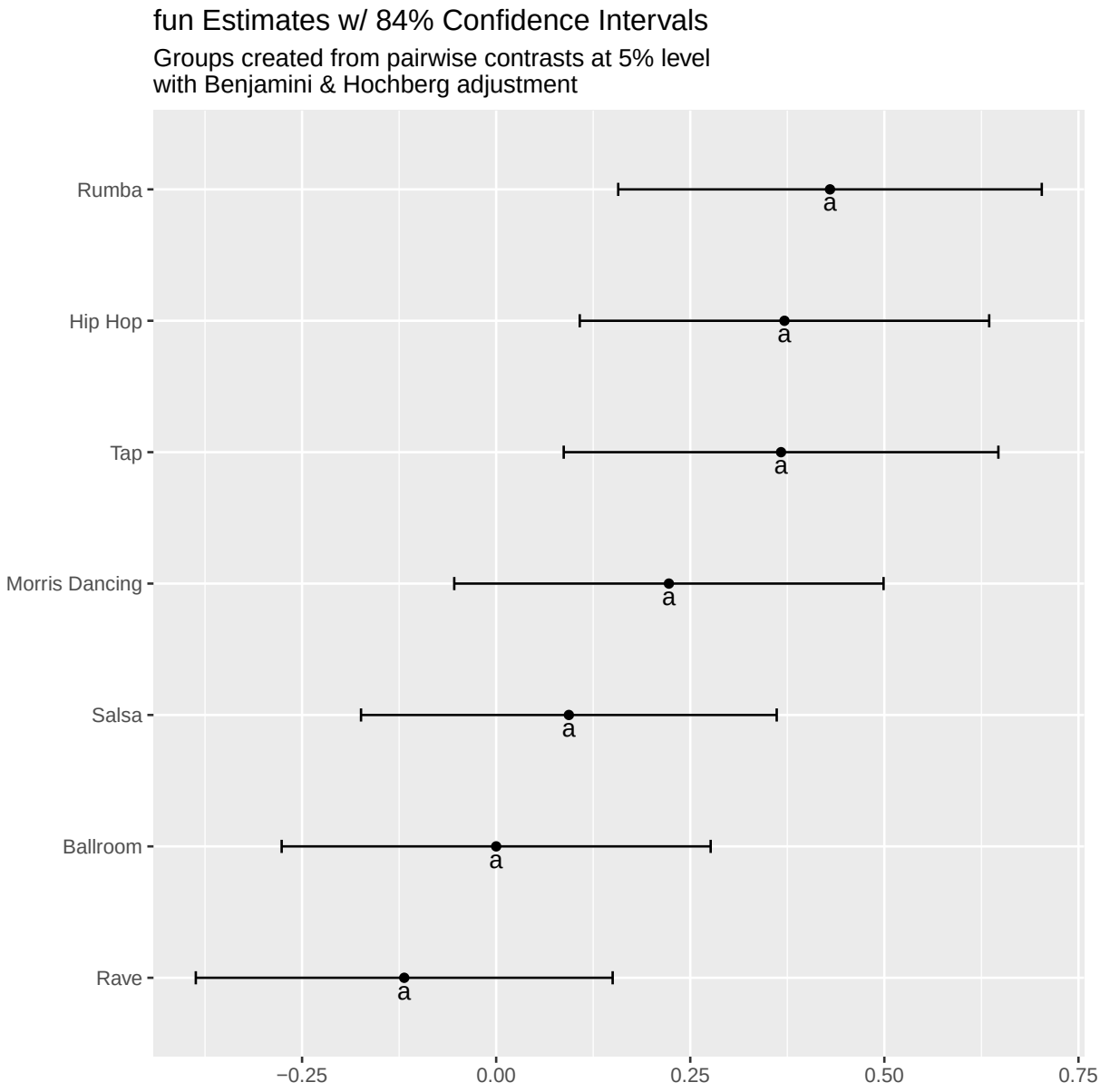
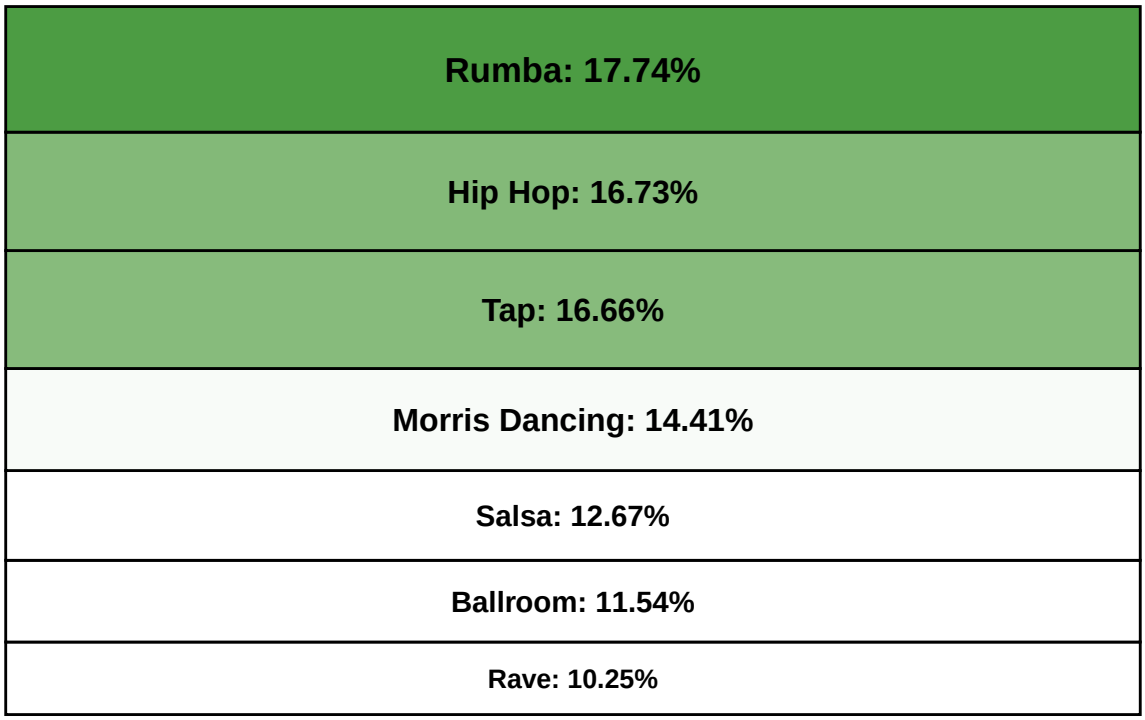


Table 24: Table 3.5. 3 fun model parameter estimates

	Estimate	quasiSE	Group
Rumba	0.43	0.19	a
Hip Hop	0.37	0.19	a
Tap	0.37	0.20	a
Morris Dancing	0.22	0.20	a
Salsa	0.09	0.19	a
Ballroom	0.00	0.20	a
Rave	-0.12	0.19	a



Section 4: Plackett Luce Models With Covariates

Overall Ranking

A classification tree approach was used to determine which of the covariates, if any had significant relationships with the rankings. This approach identifies sub-groups in the data for which the rankings of the different varieties are significantly different to each other. Table 4.1 shows the p-values for each of the covariates tested, one-by-one, showing whether or not the covariate could be used to define sub-groups with significantly different rankings.

Table 25: Table 4.1: Univariate p-values for first split in Plackett-Luce tree model for the overall ranking

Variable	p
age	0.29448
gender	0.99079

Figure 4.1 shows the partitioning of the rankings based on the most significantly different sub-groups which could be identified from the data using a 5% significance level. At the top of the tree is the full dataset, then working down through the different levels of the tree shows the combinations of variables which define each subgroup. The model parameters are shown for the final subgroups ("terminal nodes") in the plots at the bottoms of the tree.

Figure 4.1 - Overall Ranking Plackett-Luce Tree Considering All Possible Covariates

No figure produced: no significant covariates identified

The highest and lowest performing variety within each sub-group is identified within Table 4.2.

The model coefficient estimates, along with 84% confidence intervals are provided in Figure 4.2. This will help identification of which varieties were better suited to particular sub-groups identified by the analysis.

Figure 4.2 - Coefficient Estimates Within Each Identified Terminal Node Subgroup

No figure produced: no significant covariates identified

Table 4.3 outlines the p-values for each covariate at each of the nodes in the tree, outlining whether an additional significant split could be determined from within the existing sub-group at that node.

Table 26: Table 4.3: p-values for effect of each covariate at each node

parameter	Node	statistic	p
age	1	14.022196	0.29448
gender	1	2.163712	0.99079

Univariate analysis

age

- –
- –

Covariate could not be included within statistical model

gender

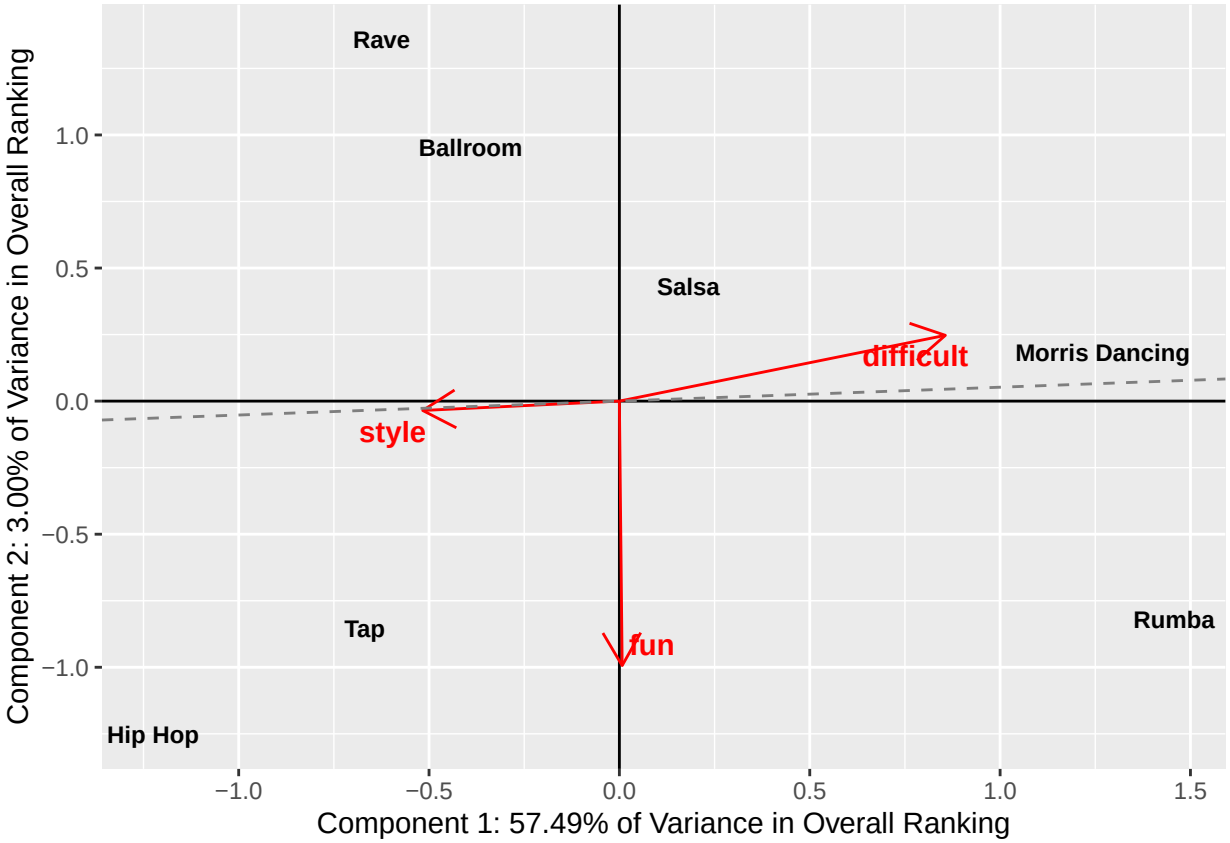
- –
- –

Covariate could not be included within statistical model

Section 5: Relationship between Traits

Partial least squares regression was used to determine relationship between the specific traits and the overall rankings.

Figure 5.1 - Partial Least Squares Biplot of Relationship Between Traits and Overall Performance



The first two components recombining the specific traits are able to explain 60.5% of the variability in the overall ranking. The dashed line represents the overall ranking, with an increase in performance as the x and y axes increase.

Varieties positioned close to the dashed line will be performing equally across all traits; varieties positioned further away from the dashed line, on either side, will have varying performance in different traits. Better performance in traits will correspond with arrows pointing in the direction away from the dashed line and worse performance in traits directed on the opposite side.

Traits with arrows pointing in similar directions have a similar relationship to the overall ranking. Traits with arrows more closely following the dashed line will have a stronger relationship to the overall ranking.

Table 5.1 - Partial Least Squares Regression Model Of Traits Against Overall Ranking

```
## Response Overall (3 comps):
##            Estimate Std. Error Df t value Pr(>|t|)
## difficult  0.150974   0.098069  6  1.5395   0.1746
## style     -0.098222   0.086556  6 -1.1348   0.2998
## fun       -0.041788   0.114473  6 -0.3651   0.7276
```

Table 5.1 summarises the regression model between the rankings of each trait of an individual variety and the overall ranking. The most positive value will be associated with the trait which can explain the largest proportion of the variability in the overall ranking. Due to the strong correlation in rankings for different traits it would be expected for some of these coefficients to be negative. An individual negative value does not neccesarily indicate a negative relationship overall between that trait and the overall ranking; rather that holding all other rankings constant and increase ranking for that trait would not neccesarily be expected to result in an increase in the overall ranking.

Appendix A: Summary of Fixed Parameters Used

TBD

Appendix B: All Pairwise Treatment Contests

Table 27: Pairwise: % of Contests Where Variety 1 Was Preferred to Variety 2

Variety 1	Variety 2	Number of Contests	Overall	difficult	style	fun
Ballroom	Hip Hop	22	54.5%	59.1%	40.9%	40.9%
Ballroom	Morris Dancing	21	38.1%	52.4%	45.0%	40.0%
Ballroom	Rave	22	36.4%	45.5%	52.4%	57.1%
Ballroom	Rumba	21	33.3%	35.0%	50.0%	40.0%
Ballroom	Salsa	20	70.0%	57.9%	77.8%	72.2%
Ballroom	Tap	22	54.5%	54.5%	52.4%	28.6%
Hip Hop	Ballroom	22	45.5%	40.9%	59.1%	59.1%
Hip Hop	Morris Dancing	21	38.1%	33.3%	57.1%	60.0%
Hip Hop	Rave	21	47.6%	52.4%	38.1%	55.0%
Hip Hop	Rumba	21	38.1%	42.9%	50.0%	40.0%
Hip Hop	Salsa	20	30.0%	40.0%	36.8%	55.0%
Hip Hop	Tap	21	52.4%	42.9%	42.9%	55.0%
Morris Dancing	Ballroom	21	61.9%	47.6%	55.0%	60.0%
Morris Dancing	Hip Hop	21	61.9%	66.7%	42.9%	40.0%
Morris Dancing	Rave	21	42.9%	52.4%	40.0%	66.7%
Morris Dancing	Rumba	19	47.4%	52.6%	38.9%	50.0%
Morris Dancing	Salsa	23	43.5%	60.9%	47.8%	43.5%
Morris Dancing	Tap	21	42.9%	52.4%	65.0%	55.6%
Rave	Ballroom	22	63.6%	54.5%	47.6%	42.9%
Rave	Hip Hop	21	52.4%	47.6%	61.9%	45.0%
Rave	Morris Dancing	21	57.1%	47.6%	60.0%	33.3%
Rave	Rumba	22	50.0%	40.9%	47.6%	38.1%
Rave	Salsa	22	50.0%	38.1%	50.0%	40.0%
Rave	Tap	20	55.0%	47.4%	47.4%	36.8%
Rumba	Ballroom	21	66.7%	65.0%	50.0%	60.0%
Rumba	Hip Hop	21	61.9%	57.1%	50.0%	60.0%
Rumba	Morris Dancing	19	52.6%	47.4%	61.1%	50.0%
Rumba	Rave	22	50.0%	59.1%	52.4%	61.9%
Rumba	Salsa	21	66.7%	55.0%	47.4%	65.0%
Rumba	Tap	20	75.0%	55.0%	30.0%	42.1%
Salsa	Ballroom	20	30.0%	42.1%	22.2%	27.8%
Salsa	Hip Hop	20	70.0%	60.0%	63.2%	45.0%
Salsa	Morris Dancing	23	56.5%	39.1%	52.2%	56.5%
Salsa	Rave	22	50.0%	61.9%	50.0%	60.0%
Salsa	Rumba	21	33.3%	45.0%	52.6%	35.0%
Salsa	Tap	20	45.0%	47.4%	42.1%	47.4%
Tap	Ballroom	22	45.5%	45.5%	47.6%	71.4%
Tap	Hip Hop	21	47.6%	57.1%	57.1%	45.0%
Tap	Morris Dancing	21	57.1%	47.6%	35.0%	44.4%
Tap	Rave	20	45.0%	52.6%	52.6%	63.2%
Tap	Rumba	20	25.0%	45.0%	70.0%	57.9%

Variety 1	Variety 2	Number of Contests	Overall	difficult	style	fun
Tap	Salsa	20	55.0%	52.6%	57.9%	52.6%

References

TBD