Fallacies of Agreement: A Critical Review of Consensus Assessment Methods for Gesture Elicitation

Supplementary Material

Theophanis Tsandilas

February 2018

A Results of Additional Simulation Experiments

This supplementary material presents results from seven additional Monte Carlo experiments.

A.1 Effect of Bias on Agreement

Experiment A.1. We repeat Experiment 3.2 by taking the linear combination of a Zipf-Mandelbrot and a discrete half-normal distribution with equal weights. Results are as follows:

Distribution:	$.5z_1 + .5n_1$	$.5z_2 + .5n_2$	$.5z_3 + .5n_3$
AR (mean):	.190	.095	.047
Fleiss' p_e (mean):	.191	.095	.049
Fleiss' κ_F (mean):	$-.001$	$-.001$	$-.001$
Krippendorff's α (mean):	$-.000$	$-.000$	$-.000$

Notice that observed agreement rates are slightly lower than ones for the distributions z_i or n_i alone (see results from Experiment 3.2). Again, Fleiss' κ_F and Krippendorff's α eliminate the effect of bias on agreement. One can run the same experiment by choosing random weights at each iteration. Results will not change.

A.2 Evaluation of the V_{rd} Statistic

Experiment A.2. We repeat Experiment 6.1 for sign preference distributions produced by the linear combination of a Zipf-Mandelbrot and a discrete half-normal distribution with equal weights. Type I error rates are as follows:

AR	$\alpha = .05$	$\alpha = .01$
.10	.19	.09
.20	.33	.17
.30	.39	.28
.40	.46	.33
.50	.51	.39
.60	.55	.47
.70	.61	.47
.80	.66	.50
.90	.68	.53

Type I error rates are consistently in between error rates produced by Zipf-Mandelbrot and discrete half-normal distributions individually (see Table VI). Choosing random weights at each iteration produces very similar results.

Experiment A.3. We repeat Experiment 6.1 for Zipf-Mandelbrot only distributions by varying the number of participants. Type I error rates are as follows:

	$\alpha = .05$			$\alpha = .01$		
AR	$(n = 10)$	$(n = 20)$	$(n = 30)$	$(n = 10)$	$(n = 20)$	$(n = 30)$
.10	.14	.31	.40	.05	.17	.24
.20	.25	.42	.53	.13	.31	.41
.30	.37	.51	.60	.21	.40	.49
.40	.42	.56	.65	.27	.45	.53
.50	.42	.56	.66	.34	.48	.58
.60	.47	.61	.65	.40	.51	.61
.70	.50	.62	.70	.42	.52	.60
.80	.55	.67	.76	.52	.50	.58
.90	.49	.67	.76	.50	.53	.68

We observe that the Type I error rate of the V_{rd} statistic increases as we add more participants.

Experiment A.4. We attempt to repair the V_{rd} statistic by applying Cochran's Q test on independent only agreement pairs, such as the consecutive agreement pairs shown as blue edges in Figure 7. The following table presents estimations of Type I error rates for this alternative application of Cochran's Q test (*alt. Cochran's Q*). We run a Monte Carlo experiment for $n = 20$ participants, Zipf-Mandelbrot distributions, and 1600 iterations. Type I error rates are contrasted with the ones of the Jackknife technique:

	$\alpha = .05$		$\alpha = .01$	
AR	alt. Cohran's Q	Jackknife	alt. Cohran's Q	Jackknife
.10	.064	.011	.003	.000
.20	.079	.019	.012	.001
.30	.092	.039	.018	.004
.40	.11	.051	.025	.009
.50	.11	.048	.027	.012
.60	.12	.065	.039	.019
.70	.14	.059	.053	.018
.80	.20	.053	.039	.016
.90	.17	.019	.009	.003

We observe that this approach results in considerably lower Type I error rates. However, error rates are still unsatisfying. A possible explanation is that the approach does not fully capture the way independent observations are generated. In particular, participants make decisions on their proposals rather than on whether they agree or disagree with each other. We do not provide any further discussion here, but readers are invited to conduct their own simulation experiments.

A.3 Evaluation of the V_b Statistic

Experiment A.5. We repeat Experiment 6.3 for sign preference distributions produced by the linear combination of a Zipf-Mandelbrot and a discrete half-normal distribution with equal weights. Type I error rates are as follows:

AR	$\alpha = .05$	$\alpha = .01$
.10	.06	.02
.20	.27	.11
.30	.41	.24
.40	.50	.37
.50	.60	.41
.60	.60	.49
.70	.62	.50
.80	.68	.44
.90	.66	.26

As for the V_{rd} statistic, Type I error rates are again in between error rates produced by Zipf-Mandelbrot and discrete half-normal distributions individually (see Table VIII). Choosing random weights at each iteration produces very similar results.

Experiment A.6. According to the experimental procedure for Experiment 6.3, sampling is performed at two steps. First, we repeatedly generate smaller populations of 100 participants from a given source population. From these 100 participants, we then randomly draw two equal samples of 20 participants. This two-step sampling approach is closer to the one described by Vatavu and Wobbrock [2016]. Yet, one could argue that it might affect the Type I error of the statistical test. In order to clarify this issue, we repeat the experiment, but now, we directly draw two samples of 20 participants from the original population. Type I error rates are as follows:

	$(\alpha = .05)$		$(\alpha = .01)$	
AR	Zipf-Mandelbrot	Half-Normal	Zipf-Mandelbrot	Half-Normal
.10	.15	.03	.05	.008
.20	.39	.14	.22	.04
.30	.52	.27	.37	.12
.40	.59	.35	.44	.20
.50	.60	.43	.52	.27
.60	.63	.48	.52	.40
.70	.64	.55	.52	.46
.80	.72	.61	.44	.45
.90	.69	.69	.27	.27

Results are almost identical with the ones reported in Table VIII. We conclude that our two-step sampling approach does not have any effect on the Type I error of the test.

A.4 Jackknifing vs. Bootstrapping

Experiment A.7. The article uses the bootstrap methods to derive the confidence interval of the agreement difference of independent participants groups. We mentioned, however, that the technique can be also used to derive the confidence interval of the agreement different of related samples, i.e., of the same participants but for different groups of referents (see Section 6.4). Evaluating the bootstrap method is computationally expensive. However, we can run experiments to compare the jackknife technique on agreement rates (AR), which require lighter computations. Below, we present the Type I error rates for an experiment with the same parameters as Experiment A.3 (within-participants hypothesis testing by assuming Zipf-Mandelbrot distributions):

		$\alpha = .05$			$\alpha = .01$		
	AR	$(n = 10)$	$(n = 20)$	$(n = 30)$	$(n = 10)$	$(n = 20)$	$(n = 30)$
Jackknife	.10	.006	.011	.018	.001	.000	.001
	.20	.014	.019	.024	.002	.001	.002
	.30	.024	.039	.048	.006	.004	.005
	.40	.042	.051	.043	.010	.009	.009
	.50	.061	.048	.053	.019	.012	.013
	.60	.078	.065	.061	.034	.019	.016
	.70	.078	.059	.069	.026	.018	.018
	.80	.059	.053	.073	.011	.016	.014
	.90	.014	.019	.023	.002	.003	.004
Bootstrap	.10	.016	.037	.043	.003	.006	.008
	.20	.041	.053	.048	.008	.017	.016
	.30	.060	.072	.047	.015	.015	.013
	.40	.062	.076	.066	.013	.021	.016
	.50	.062	.073	.061	.018	.023	.019
	.60	.052	.081	.059	.014	.024	.022
	.70	.028	.061	.062	.008	.026	.024
	.80	.011	.059	.081	.003	.019	.017
	.90	.003	.017	.028	.001	.000	.010

We observe that under different parameters, the precision of each technique can be slightly better or worse, but overall, error rates are comparable.