Supplement for Estimated Judge Reliabilities for Weighted Bradley-Terry-Luce Are Not Reliable

Andrew F. Dreher
The University of Texas at Austin
Austin, Texas, USA
afdreher@utexas.edu

Etienne Vogua
The University of Texas at Austin
Austin, Texas, USA
vouga@cs.utexas.edu

Donald S. Fussell
The University of Texas at Austin
Austin, Texas, USA
fussell@cs.utexas.edu

The following are additional data for the paper *Estimated Judge Reliabilities for Weighted Bradley-Terry-Luce Are Not Reliable* to provide transparency.

1 BROKEN SYMMETRY

In this section, we show results for all 20, $n^2 - n$, inversions for the example described in section 4.3 where every judge has an inversion from the $s^\star$ order at a single location based on their index, but one judge has two inversions such that they now agree with another judge in the set. As previously mentioned, since this scenario creates a pattern where the majority always agrees, the standard Bradley-Terry-Luce method of eq. (1) finds the $s^\star$ order, shown in table 1.

We show results both with, $\lambda = 1$, and without, $\lambda = 0$, Mease's penalty. We caution that the results here are merely examples of what could happen; different solutions based on different starting positions are possible. The starting conditions, parameters, and tolerances are neither optimized nor cherry-picked.

Tables 1 and 2 show the results for the BTL method. Note that when $\lambda = 1$, $s^\star$ is no longer reliably recovered.

Tables 3 and 4 show the results for Crowd-BT. An important observation here is that the least reliable judge is often in $\mathcal{D}$: 85% of the time when $\lambda = 0$ and 100% of the time when $\lambda = 1$.

Tables 5 and 6 show the results for O'Donovan *et al.*'s method. With the $\lambda = 0$, the least reliable judge is a member of $\mathcal{D}$ in 40% of the trials, but this increases to 100% when $\lambda = 1$. Since the least reliable judge is always in $\mathcal{D}$ for $\lambda = 1$, the judge who shares that inversion often is part of $\mathcal{D}$ too. Interestingly, there is always some setting for which the least reliable judge can be elevated into $\mathcal{D}$.

2 FONTS

In section 4.3, we discussed the results from repeating O'Donovan *et al.*'s result along with a variation where the attributes were treated as independent scales. Here, we provide the full tables of Pearson correlation coefficients for both Crowd-BT, table 7, and O'Donovan *et al.*'s method, table 8, along with histograms fig. 1 and fig. 2.

3 SIMULATION DATA

Using the procedure we explained in section 6 and appendix A.1, we generated fig. 5. Tables 9 and 10 show $\omega = \{0.02, 0.1\}$, respectively, for 8 judges per pair, which closely approximate several of the Fonts attributes. In the tables, we provide the mean value of τ, μ_τ, as well as the standard deviation, σ_τ. In each, we highlight the 800 pair line, since this represents data that are the same size as the Fonts dataset.

In tables 11 and 12, we provide a statistical comparison between the two experimental methods, Chen *et al.*'s Crowd-BT and

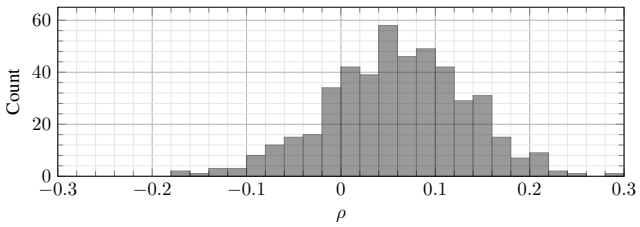

Figure 1: Distribution of ρ values for independently estimated scales using Crowd-BT. Tabular data is in table 7.

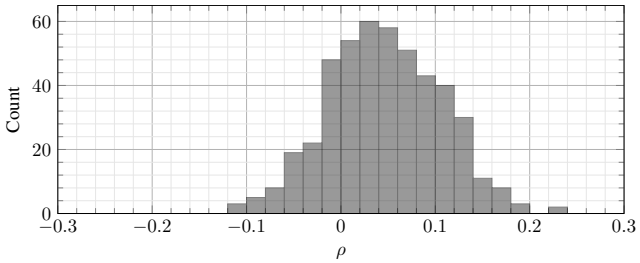

Figure 2: Distribution of ρ values for independently estimated scales using O'Donovan *et al.*'s method. Tabular data is in table 8.

O'Donovan *et al.*'s method, against the unweighted BTL order on a matched pair basis for the simulated data with 8 judges per pair, $\omega = \{0.02, 0.1)$, and $\lambda = 0$. For each, we use a Bayesian Paired Samples T-Test, computed using JASP 0.16 [1]. We use Student-T where $H_a : \tau_{\text{experimental}} \neq \tau_{\text{BTL}}$, and the prior is Cauchy with scale 0.707. Values of $\log(\text{BF}_{10})$ are traditionally interpreted using the scale provided by Jeffreys [2].

We note that many values show strong evidence of being different from the BTL result; however, most are very slight underperformances. When $\omega = 0.02$, both methods underperform except when the pair count is 200. For $\omega = 0.1$, both methods perform much better than BTL when the pair count is 200 and still perform slightly better for a little while longer. However, when we reach the size of Fonts (pair count of 800), both methods underperform BTL very slightly. This trend of underperformance continues throughout the rest of the table but is often quite small and not statistically interesting. Regardless, of the two methods, Crowd-BT tends to behave more similar to BTL than O'Donovan *et al.*'s method.

These results suggest that, at least for reliable judges, one would only want to use the weighted methods for very low number of pairs (a very low data density). Future work remains as to whether some

Judge	"error"	Stimuli Positions						$\hat{s} = s^\star$
		ς_1	ς_2	ς_3	ς_4	ς_5	ς_6	
u_1	$s_2 \succ s_3$	-2.368	-0.982	-0.576	0.810	2.197	3.583	Yes
	$s_3 \succ s_4$	-2.531	-1.145	0.242	0.647	2.033	3.420	Yes
	$s_4 \succ s_5$	-2.694	-1.308	0.078	1.464	1.870	3.256	Yes
	$s_5 \succ s_6$	-2.858	-1.472	-0.086	1.301	2.687	3.092	Yes
u_2	$s_1 \succ s_2$	-2.204	-1.799	-0.412	0.974	2.360	3.746	Yes
	$s_3 \succ s_4$	-2.531	-1.145	0.242	0.647	2.033	3.420	Yes
	$s_4 \succ s_5$	-2.694	-1.308	0.078	1.464	1.870	3.256	Yes
	$s_5 \succ s_6$	-2.858	-1.472	-0.086	1.301	2.687	3.092	Yes
u_3	$s_1 \succ s_2$	-2.204	-1.799	-0.412	0.974	2.360	3.746	Yes
	$s_2 \succ s_3$	-2.368	-0.982	-0.576	0.810	2.197	3.583	Yes
	$s_4 \succ s_5$	-2.694	-1.308	0.078	1.464	1.870	3.256	Yes
	$s_5 \succ s_6$	-2.858	-1.472	-0.086	1.301	2.687	3.092	Yes
u_4	$s_1 \succ s_2$	-2.204	-1.799	-0.412	0.974	2.360	3.746	Yes
	$s_2 \succ s_3$	-2.368	-0.982	-0.576	0.810	2.197	3.583	Yes
	$s_3 \succ s_4$	-2.531	-1.145	0.242	0.647	2.033	3.420	Yes
	$s_5 \succ s_6$	-2.858	-1.472	-0.086	1.301	2.687	3.092	Yes
u_5	$s_1 \succ s_2$	-2.204	-1.799	-0.412	0.974	2.360	3.746	Yes
	$s_2 \succ s_3$	-2.368	-0.982	-0.576	0.810	2.197	3.583	Yes
	$s_3 \succ s_4$	-2.531	-1.145	0.242	0.647	2.033	3.420	Yes
	$s_4 \succ s_5$	-2.694	-1.308	0.078	1.464	1.870	3.256	Yes

Table 1: Full results from adding an additional error to various judges solved using BTL with $\lambda = 0$.

Judge	"error"	Stimuli Positions						$\hat{s} = s^\star$
		ς_1	ς_2	ς_3	ς_4	ς_5	ς_6	
u_1	$s_2 \succ s_3$	-1.142	-0.310	-0.446	0.052	0.572	1.340	No
	$s_3 \succ s_4$	-1.269	-0.479	0.099	-0.099	0.479	1.269	No
	$s_4 \succ s_5$	-1.340	-0.572	-0.052	0.446	0.310	1.142	No
	$s_5 \succ s_6$	-1.381	-0.626	-0.138	0.293	0.847	0.907	Yes
u_2	$s_1 \succ s_2$	-0.907	-0.847	-0.293	0.138	0.626	1.381	Yes
	$s_3 \succ s_4$	-1.269	-0.479	0.099	-0.099	0.479	1.269	No
	$s_4 \succ s_5$	-1.340	-0.572	-0.052	0.446	0.310	1.142	No
	$s_5 \succ s_6$	-1.381	-0.626	-0.138	0.293	0.847	0.907	Yes
u_3	$s_1 \succ s_2$	-0.907	-0.847	-0.293	0.138	0.626	1.381	Yes
	$s_2 \succ s_3$	-1.142	-0.310	-0.446	0.052	0.572	1.340	No
	$s_4 \succ s_5$	-1.340	-0.572	-0.052	0.446	0.310	1.142	No
	$s_5 \succ s_6$	-1.381	-0.626	-0.138	0.293	0.847	0.907	Yes
u_4	$s_1 \succ s_2$	-0.907	-0.847	-0.293	0.138	0.626	1.381	Yes
	$s_2 \succ s_3$	-1.142	-0.310	-0.446	0.052	0.572	1.340	No
	$s_3 \succ s_4$	-1.269	-0.479	0.099	-0.099	0.479	1.269	No
	$s_5 \succ s_6$	-1.381	-0.626	-0.138	0.293	0.847	0.907	Yes
u_5	$s_1 \succ s_2$	-0.907	-0.847	-0.293	0.138	0.626	1.381	Yes
	$s_2 \succ s_3$	-1.142	-0.310	-0.446	0.052	0.572	1.340	No
	$s_3 \succ s_4$	-1.269	-0.479	0.099	-0.099	0.479	1.269	No
	$s_4 \succ s_5$	-1.340	-0.572	-0.052	0.446	0.310	1.142	No

Table 2: Full results from adding an additional error to various judges solved using BTL with $\lambda = 1$.

number of indifferent or adversarial judges changes that decision point.

REFERENCES

[1] JASP Team. 2021. JASP (Version 0.16)[Computer software]. https://jasp-stats.org/
[2] Sir Harold Jeffreys. 1961. *The Theory of Probability.* https://books-google-com.ezproxy.lib.utexas.edu/books/about/The_Theory_of_Probability.html?id=vh9Act9rtzQC

		Judge Weights						Stimuli Positions						
Judge	"error"	μ_1	μ_2	μ_3	μ_4	μ_5	$\lvert\mathcal{D}\rvert$	ς_1	ς_2	ς_3	ς_4	ς_5	ς_6	$\hat{s} = s^\star$
u_1	$s_2 \succ s_3$	**0.000**	*0.500*	**1.000**	**1.000**	**1.000**	4	-23.738	-7.492	8.116	8.116	8.116	8.116	No
	$s_3 \succ s_4$	0.500	**1.000**	***1.000***	**1.000**	**1.000**	4	-14.777	0.819	1.917	3.016	4.115	5.213	Yes
	$s_4 \succ s_5$	**1.000**	0.500	0.500	***1.000***	0.500	2	-15.843	-15.843	0.495	16.866	-1.690	15.923	No
	$s_5 \succ s_6$	0.500	**1.000**	**1.000**	**1.000**	***1.000***	4	-14.697	0.802	1.901	2.999	4.098	5.196	Yes
u_2	$s_1 \succ s_2$	*0.500*	**0.000**	**1.000**	**1.000**	**1.000**	4	-23.559	-7.828	8.148	8.148	8.148	8.148	No
	$s_3 \succ s_4$	0.500	**1.000**	***1.000***	0.500	0.500	2	-13.422	3.247	3.247	-13.722	2.399	18.137	No
	$s_4 \succ s_5$	**1.000**	0.609	**1.000**	***1.000***	0.723	3	-17.209	-16.167	2.859	3.901	4.688	21.783	Yes
	$s_5 \succ s_6$	**1.000**	**0.000**	**1.000**	**1.000**	*0.500*	4	-13.138	-13.138	2.462	2.462	2.462	18.674	No
u_3	$s_1 \succ s_2$	***1.000***	0.500	**1.000**	0.500	0.500	2	-4.718	-23.270	-5.580	-5.580	10.745	27.431	No
	$s_2 \succ s_3$	0.500	***1.000***	**1.000**	0.500	0.500	2	-11.161	6.716	-11.418	-11.418	4.576	21.994	No
	$s_4 \succ s_5$	0.500	0.500	**1.000**	***1.000***	0.500	2	-19.988	-3.174	12.729	12.729	-10.457	8.624	No
	$s_5 \succ s_6$	0.500	0.500	**1.000**	0.500	***1.000***	2	-26.631	-10.464	5.420	5.420	21.958	5.279	No
u_4	$s_1 \succ s_2$	***1.000***	0.500	0.500	**1.000**	0.500	2	-7.858	-24.247	-7.594	7.772	7.772	24.169	No
	$s_2 \succ s_3$	0.500	***1.000***	0.500	**1.000**	0.500	2	-13.343	3.307	-13.125	2.557	2.557	17.887	No
	$s_3 \succ s_4$	0.500	0.500	***1.000***	**1.000**	0.500	2	-18.763	-1.859	13.943	-3.007	-3.007	12.890	No
	$s_5 \succ s_6$	0.500	0.500	0.500	**1.000**	***1.000***	2	-30.911	-12.578	3.523	20.646	20.646	0.288	No
u_5	$s_1 \succ s_2$	*0.500*	**1.000**	**1.000**	**1.000**	**0.000**	4	-16.310	-0.121	-0.121	-0.121	-0.121	16.757	No
	$s_2 \succ s_3$	0.500	***1.000***	0.500	0.500	**1.000**	2	-15.456	1.664	-16.193	-0.373	15.227	15.227	No
	$s_3 \succ s_4$	0.500	0.500	***1.000***	0.500	**1.000**	2	-19.623	-4.329	10.991	-5.733	9.595	9.595	No
	$s_4 \succ s_5$	0.500	0.500	0.500	***1.000***	**1.000**	2	-27.338	-10.022	6.134	22.825	4.702	4.702	No

Table 3: Full results from adding an additional error to various judges solved using Chen *et al.*'s Crowd-BT with $\lambda = 0$. The *italic* entries represent the judges we expect to be in $\mathcal{D}$ because the new "error" agrees with their position; bold entries are judges in $\mathcal{D}$. The cells with the red background highlight the least reliable judge.

		Judge Weights						Stimuli Positions						
Judge	"error"	μ_1	μ_2	μ_3	μ_4	μ_5	$\lvert\mathcal{D}\rvert$	ς_1	ς_2	ς_3	ς_4	ς_5	ς_6	$\hat{s} = s^\star$
u_1	$s_2 \succ s_3$	**1.000**	***1.000***	**0.000**	0.477	0.291	3	0.699	0.203	-1.787	-0.013	0.027	0.552	No
	$s_3 \succ s_4$	**1.000**	**0.000**	***1.000***	0.303	0.410	3	0.216	-0.640	1.326	-0.883	-0.145	0.219	No
	$s_4 \succ s_5$	**1.000**	0.388	0.322	***1.000***	**0.000**	3	0.344	-0.433	0.076	0.809	-1.405	0.463	No
	$s_5 \succ s_6$	**0.000**	**1.000**	**1.000**	**1.000**	*0.146*	4	-0.848	0.551	0.119	-0.259	-0.754	1.270	No
u_2	$s_1 \succ s_2$	***1.000***	**1.000**	**1.000**	**1.000**	**0.000**	5	-0.058	-0.488	-0.302	-0.236	-0.264	1.608	No
	$s_3 \succ s_4$	0.289	**1.000**	***1.000***	**0.000**	0.541	3	-0.131	0.743	0.265	-1.692	0.214	0.324	No
	$s_4 \succ s_5$	0.429	**1.000**	0.242	***1.000***	0.258	2	-0.350	0.387	-0.443	1.158	-1.138	0.390	No
	$s_5 \succ s_6$	**0.000**	**1.000**	**0.000**	**0.000**	***1.000***	5	0.131	0.592	-0.400	0.137	0.733	-1.309	No
u_3	$s_1 \succ s_2$	***1.000***	**0.000**	**1.000**	0.312	0.332	3	0.852	-1.338	0.507	-0.469	-0.023	0.356	No
	$s_2 \succ s_3$	**0.000**	***1.000***	**1.000**	0.377	0.493	3	-0.515	1.642	-0.303	-0.769	-0.075	0.264	No
	$s_4 \succ s_5$	0.493	0.377	**1.000**	***1.000***	**0.000**	3	-0.264	0.075	0.769	0.303	-1.642	0.515	No
	$s_5 \succ s_6$	0.332	0.312	**1.000**	**0.000**	***1.000***	3	-0.356	0.023	0.469	-0.507	1.338	-0.852	No
u_4	$s_1 \succ s_2$	***1.000***	**0.000**	**0.000**	**1.000**	**0.000**	5	1.309	-0.733	-0.137	0.400	-0.592	-0.131	No
	$s_2 \succ s_3$	0.258	***1.000***	0.242	**1.000**	0.429	2	-0.390	1.138	-1.158	0.443	-0.387	0.350	No
	$s_3 \succ s_4$	0.541	**0.000**	***1.000***	**1.000**	0.289	3	-0.324	-0.214	1.692	-0.265	-0.743	0.131	No
	$s_5 \succ s_6$	**0.000**	**1.000**	**1.000**	**1.000**	***1.000***	5	-1.608	0.264	0.236	0.302	0.488	0.058	No
u_5	$s_1 \succ s_2$	*0.146*	**1.000**	**1.000**	**1.000**	**0.000**	4	-1.270	0.754	0.259	-0.119	-0.551	0.848	No
	$s_2 \succ s_3$	**0.000**	***1.000***	0.322	0.388	**1.000**	3	-0.463	1.405	-0.809	-0.076	0.433	-0.344	No
	$s_3 \succ s_4$	0.410	0.303	***1.000***	**0.000**	**1.000**	3	-0.219	0.145	0.883	-1.326	0.640	-0.216	No
	$s_4 \succ s_5$	**0.000**	**1.000**	**1.000**	***1.000***	**1.000**	5	-1.569	0.322	0.325	0.458	-0.039	0.260	No

Table 4: Full results from adding an additional error to various judges solved using Chen *et al.*'s Crowd-BT with $\lambda = 1$. The *italic* entries represent the judges we expect to be in $\mathcal{D}$ because the new "error" agrees with their position; bold entries are judges in $\mathcal{D}$. The cells with the red background highlight the least reliable judge.

		Judge Weights						Stimuli Positions						
Judge	"error"	μ_1	μ_2	μ_3	μ_4	μ_5	$\lvert\mathcal{D}\rvert$	ς_1	ς_2	ς_3	ς_4	ς_5	ς_6	$\hat{s} = s^\star$
u_1	$s_2 \succ s_3$	-0.235	*0.016*	**10.000**	0.238	0.238	1	-100.000	81.526	99.554	99.132	99.566	100.000	No
	$s_3 \succ s_4$	1.039	-0.022	***10.000***	0.023	0.023	1	-100.000	-99.722	95.926	91.673	95.836	100.000	No
	$s_4 \succ s_5$	**10.000**	-0.022	-0.022	*1.009*	-0.022	1	91.638	91.352	95.676	100.000	-100.000	-95.676	No
	$s_5 \succ s_6$	-0.235	0.238	**10.000**	0.238	*0.016*	1	-100.000	81.526	81.960	81.538	81.972	100.000	No
u_2	$s_1 \succ s_2$	*0.016*	-0.235	**10.000**	0.238	0.238	1	-100.000	-81.972	99.554	99.132	99.566	100.000	No
	$s_3 \succ s_4$	0.023	1.039	***10.000***	0.023	-0.022	1	-100.000	-95.836	-95.558	-99.811	-95.648	100.000	No
	$s_4 \succ s_5$	0.032	**10.000**	0.032	***10.000***	-0.032	2	-100.000	-99.290	-99.290	-98.580	-99.292	99.999	No
	$s_5 \succ s_6$	-0.022	**10.000**	-0.022	-0.022	*1.009*	1	87.314	91.638	91.352	95.676	100.000	-100.000	No
u_3	$s_1 \succ s_2$	***10.000***	0.032	**10.000**	0.032	-0.032	2	-99.289	-100.000	-99.290	-99.290	-98.581	100.000	No
	$s_2 \succ s_3$	0.023	***10.000***	1.040	0.023	-0.022	1	-100.000	-95.794	-100.000	-99.722	-95.560	100.000	No
	$s_4 \succ s_5$	0.032	0.032	**10.000**	***10.000***	-0.032	2	-100.000	-99.290	-98.580	-98.580	-99.292	100.000	No
	$s_5 \succ s_6$	-0.022	-0.022	**10.000**	-0.022	*1.009*	1	87.314	91.638	95.962	95.676	100.000	-100.000	No
u_4	$s_1 \succ s_2$	***10.000***	0.023	0.023	1.063	-0.023	1	-95.909	-100.000	-95.926	-91.853	-91.577	100.000	No
	$s_2 \succ s_3$	-0.022	*1.009*	-0.022	**10.000**	-0.022	1	95.676	100.000	-100.000	-95.676	-95.962	-91.638	No
	$s_3 \succ s_4$	0.023	0.023	***10.000***	1.039	-0.022	1	-100.000	-95.836	-91.673	-95.926	-95.648	100.000	No
	$s_5 \succ s_6$	-0.023	0.023	0.023	1.063	***10.000***	1	-100.000	91.577	95.651	99.724	100.000	95.909	No
u_5	$s_1 \succ s_2$	*0.016*	0.238	0.238	**10.000**	-0.235	1	-100.000	-81.972	-81.538	-81.104	-81.526	100.000	No
	$s_2 \succ s_3$	-0.022	*1.009*	-0.022	-0.022	**10.000**	1	95.676	100.000	-100.000	-95.676	-91.352	-91.638	No
	$s_3 \succ s_4$	0.023	0.023	***10.000***	-0.022	1.039	1	-100.000	-95.836	-91.673	-95.926	99.722	100.000	No
	$s_4 \succ s_5$	-0.023	0.023	0.023	***10.000***	1.062	1	-100.000	91.843	95.922	100.000	95.904	96.180	No

Table 5: Full results from adding an additional error to various judges solved using O'Donovan *et al.*'s method with $\lambda = 0$. The *italic* entries represent the judges we expect to be in $\mathcal{D}$ because the new "error" agrees with their position; bold entries are judges in $\mathcal{D}$. The cells with the red background highlight the least reliable judge.

		Judge Weights						Stimuli Positions						
Judge	"error"	μ_1	μ_2	μ_3	μ_4	μ_5	$\lvert\mathcal{D}\rvert$	ς_1	ς_2	ς_3	ς_4	ς_5	ς_6	$\hat{s} = s^\star$
u_1	$s_2 \succ s_3$	**10.000**	***10.000***	-1.277	-1.153	-1.125	2	0.617	0.588	-0.782	-0.441	-0.141	0.150	No
	$s_3 \succ s_4$	**10.000**	-1.157	***10.000***	-1.118	-1.017	2	0.303	0.274	0.635	-0.743	-0.396	-0.086	No
	$s_4 \succ s_5$	**10.000**	-1.058	-1.147	***10.000***	-1.132	2	0.069	0.036	0.348	0.694	-0.748	-0.408	No
	$s_5 \succ s_6$	**-10.000**	2.327	2.325	2.384	*-1.002*	1	-0.587	0.195	0.006	-0.182	-0.380	1.004	No
u_2	$s_1 \succ s_2$	*-1.112*	**10.000**	**10.000**	-1.218	-1.084	2	0.227	0.547	0.520	-0.776	-0.422	-0.113	No
	$s_3 \succ s_4$	-1.025	**10.000**	*-1.177*	**10.000**	-1.128	2	-0.006	0.297	0.267	0.626	-0.773	-0.431	No
	$s_4 \succ s_5$	-1.025	**10.000**	-1.177	***10.000***	-1.128	2	-0.006	0.297	0.267	0.626	-0.773	-0.431	No
	$s_5 \succ s_6$	-1.101	**10.000**	-1.122	-1.197	***10.000***	2	-0.215	0.079	0.044	0.346	0.677	-0.974	No
u_3	$s_1 \succ s_2$	***10.000***	-1.241	**10.000**	-1.112	-1.109	2	1.003	-0.614	-0.276	-0.310	-0.019	0.271	No
	$s_2 \succ s_3$	-1.242	***10.000***	**10.000**	-1.151	-1.099	2	0.464	0.808	-0.523	-0.550	-0.236	0.061	No
	$s_4 \succ s_5$	0.974	1.001	**10.000**	***10.000***	-1.352	2	-0.511	-0.233	0.033	0.032	-0.370	1.143	No
	$s_5 \succ s_6$	-1.109	-1.112	**10.000**	-1.241	***10.000***	2	-0.271	0.019	0.310	0.276	0.614	-1.003	No
u_4	$s_1 \succ s_2$	***10.000***	-1.197	-1.122	**10.000**	-1.101	2	0.974	-0.677	-0.346	-0.044	-0.079	0.215	No
	$s_2 \succ s_3$	-1.128	***10.000***	-1.177	**10.000**	-1.025	2	0.431	0.773	-0.626	-0.267	-0.297	0.006	No
	$s_3 \succ s_4$	1.034	1.057	***10.000***	**10.000**	-1.321	2	-0.455	-0.175	0.096	-0.274	-0.275	1.196	No
	$s_5 \succ s_6$	-1.198	-1.193	-1.241	**10.000**	***10.000***	2	-0.317	-0.031	0.253	0.553	0.521	-1.039	No
u_5	$s_1 \succ s_2$	***10.000***	-1.215	-1.133	-1.116	**10.000**	2	0.943	-0.727	-0.403	-0.108	0.180	0.144	No
	$s_2 \succ s_3$	-1.132	***10.000***	-1.147	-1.058	**10.000**	2	0.408	0.748	-0.694	-0.348	-0.036	-0.069	No
	$s_3 \succ s_4$	-1.017	-1.118	***10.000***	-1.157	**10.000**	2	0.086	0.396	0.743	-0.635	-0.274	-0.303	No
	$s_4 \succ s_5$	-1.125	-1.153	-1.277	***10.000***	**10.000**	2	-0.150	0.141	0.441	0.782	-0.588	-0.617	No

Table 6: Full results from adding an additional error to various judges solved using O'Donovan *et al.*'s method with $\lambda = 1$. The *italic* entries represent the judges we expect to be in $\mathcal{D}$ because the new "error" agrees with their position; bold entries are judges in $\mathcal{D}$. The cells with the red background highlight the least reliable judge.

	artistic	attention-grabbing	attractive	bad	boring	calm	charming	clumsy	complex	delicate	disorderly	dramatic	formal	fresh	friendly	gentle	graceful	happy	legible	modern	playful	pretentious	sharp	sloppy	soft	strong	technical	thin	warm	wide
angular	0.049	-0.024	-0.040	0.031	-0.128	0.023	0.082	0.072	0.065	0.100	0.007	0.105	0.046	-0.082	-0.046	0.115	-0.021	-0.105	-0.039	0.016	-0.053	-0.106	0.146	0.092	0.039	-0.091	0.063	0.088	-0.051	0.152
artistic		0.099	0.076	0.064	0.122	0.061	-0.035	0.066	0.074	0.020	0.193	0.134	0.056	0.065	0.047	0.089	0.119	0.032	0.076	0.069	0.143	0.150	-0.151	0.042	0.088	0.076	0.006	0.090	-0.028	-0.055
attention-grabbing			0.135	-0.004	0.127	-0.006	0.096	0.038	0.019	0.017	0.080	-0.032	-0.042	-0.019	0.061	-0.102	0.047	-0.005	0.128	-0.047	0.109	0.115	-0.088	0.001	0.012	0.178	-0.038	0.193	-0.001	-0.014
attractive				0.050	-0.145	-0.028	0.076	-0.022	-0.019	0.038	0.066	-0.017	-0.050	0.132	0.160	0.020	0.088	0.188	-0.066	0.028	0.035	0.048	-0.014	-0.008	0.053	0.133	0.115	-0.014	0.049	0.069
bad					-0.145	0.093	-0.010	0.085	0.042	-0.119	0.050	0.031	0.183	0.159	0.072	0.078	0.087	-0.003	0.086	-0.020	-0.033	0.082	0.058	0.046	0.002	-0.008	-0.001	0.028	-0.044	0.023
boring						-0.028	0.072	0.095	0.017	0.116	-0.033	-0.026	-0.020	-0.183	0.011	0.057	0.087	0.130	0.010	-0.062	0.027	0.078	-0.093	-0.054	-0.023	-0.001	-0.054	0.061	0.013	-0.083
calm							0.076	0.145	0.052	0.010	0.015	0.183	0.089	0.048	0.017	0.111	0.074	-0.002	-0.078	0.069	0.130	0.139	-0.018	-0.089	0.115	-0.048	-0.112	0.133	-0.133	0.126
charming								-0.022	0.044	0.027	0.203	0.066	-0.077	0.014	0.090	-0.070	0.123	0.079	-0.001	0.091	0.061	0.040	-0.004	0.108	0.101	0.093	-0.012	0.038	0.004	0.008
clumsy									-0.019	-0.095	0.100	0.139	0.063	0.074	0.016	0.048	0.034	0.184	-0.012	0.049	0.036	**0.261**	0.023	0.095	-0.079	0.138	-0.023	0.044	0.025	0.050
complex										0.038	0.069	0.116	0.071	0.046	-0.066	0.025	0.051	0.225	0.089	0.066	0.116	0.191	0.022	0.142	0.149	0.139	0.058	0.136	0.038	0.023
delicate											0.066	-0.008	0.002	-0.129	-0.062	0.177	0.102	0.048	0.083	-0.025	-0.010	0.025	0.139	0.033	0.126	-0.104	-0.188	-0.064	-0.106	-0.011
disorderly												-0.017	0.139	-0.100	0.026	-0.053	0.149	0.070	0.022	-0.023	0.178	0.003	-0.005	0.218	0.036	0.081	0.019	0.095	0.081	0.124
dramatic													-0.050	0.056	0.047	0.190	0.084	0.037	0.024	0.028	0.152	0.014	0.069	0.064	0.128	0.106	0.126	0.043	0.051	0.041
formal														0.132	0.029	-0.006	0.051	0.023	0.064	0.040	0.050	-0.098	0.027	0.005	-0.033	0.107	-0.078	0.119	0.099	0.174
fresh															0.160	0.057	-0.039	0.006	0.029	0.049	-0.015	0.075	0.054	0.116	0.118	0.087	0.020	0.160	0.112	0.061
friendly																0.020	-0.076	-0.039	0.098	-0.167	-0.025	0.046	-0.026	0.075	0.113	0.093	0.024	0.019	0.087	0.085
gentle																	0.088	0.044	-0.028	-0.021	0.027	0.142	0.091	0.061	0.113	-0.046	-0.111	0.038	-0.030	0.038
graceful																		0.188	0.012	-0.026	0.028	-0.019	-0.019	0.143	0.011	-0.026	-0.097	-0.014	0.027	-0.054
happy																			-0.066	0.034	0.120	0.100	0.121	0.035	0.016	-0.094	-0.013	-0.069	0.076	-0.032
legible																				0.028	-0.004	0.003	-0.040	-0.015	-0.040	0.079	-0.092	0.025	0.042	0.145
modern																					0.035	0.092	0.098	0.025	0.159	0.069	0.126	0.026	0.123	-0.003
playful																						0.048	-0.007	-0.066	0.053	0.116	0.064	0.156	0.017	0.106
pretentious																							-0.014	0.047	0.066	0.127	0.003	0.001	-0.021	0.020
sharp																								-0.008	-0.072	-0.033	0.014	0.035	0.027	0.011
sloppy																									0.053	0.120	0.023	0.084	0.011	0.047
soft																										0.133	0.090	0.062	0.064	0.013
strong																											0.115	0.033	0.112	0.041
technical																												-0.014	0.107	0.126
thin																													0.049	0.030
warm																														0.069

Table 7: Full correlation data for the attributes in Fonts using Crowd-BT. The maximum correlation is between *clumsy* and *pretentious*, shown in bold. Histogram of this data is shown in fig. 1

	artistic	attention-grabbing	attractive	bad	boring	calm	charming	clumsy	complex	delicate	disorderly	dramatic	formal	fresh	friendly	gentle	graceful	happy	legible	modern	playful	pretentious	sharp	sloppy	soft	strong	technical	thin	warm	wide
angular	0.152	-0.007	0.021	0.025	0.007	-0.071	0.069	-0.010	-0.011	0.058	0.002	0.090	-0.016	-0.061	-0.009	0.079	-0.005	0.047	0.041	-0.005	0.081	-0.083	0.049	-0.044	0.062	-0.031	-0.017	-0.021	0.012	0.079
artistic		-0.021	-0.049	0.025	0.047	-0.087	0.010	0.130	-0.112	-0.102	0.000	0.027	-0.021	-0.010	-0.042	-0.107	0.038	0.069	0.070	-0.080	0.068	-0.012	0.020	0.036	0.039	0.068	0.064	0.017	0.062	0.051
attention-grabbing			0.086	-0.017	0.036	0.041	0.109	0.092	0.057	0.121	0.158	0.043	0.026	0.012	0.037	-0.022	0.082	0.007	-0.040	-0.028	0.098	0.022	-0.027	0.138	0.142	0.090	-0.022	0.036	-0.053	0.112
attractive				0.027	-0.017	-0.100	0.023	0.051	0.096	0.049	0.119	0.061	0.097	0.096	0.056	-0.036	-0.001	0.024	0.048	0.078	-0.039	0.083	-0.046	-0.027	0.118	0.058	0.116	0.054	0.031	-0.028
bad					-0.017	0.108	0.047	0.095	0.062	0.001	0.089	-0.008	0.026	-0.007	-0.007	0.027	0.017	0.079	0.018	0.012	0.012	-0.092	-0.022	0.092	0.015	-0.033	0.048	-0.001	0.045	0.106
boring						-0.100	-0.034	0.053	0.103	0.067	0.036	0.029	0.042	0.038	-0.066	0.048	0.078	0.048	0.089	-0.040	0.161	-0.016	0.111	-0.004	0.050	0.069	0.019	0.019	0.043	-0.062
calm							0.023	-0.068	-0.000	-0.028	0.115	0.002	0.102	0.090	0.201	-0.005	0.007	0.069	-0.120	-0.006	0.080	0.005	0.005	0.007	0.007	-0.017	-0.084	-0.003	-0.035	0.033
charming								0.051	-0.017	-0.000	0.174	0.074	-0.074	0.075	0.021	-0.033	0.121	0.112	0.012	0.013	0.088	0.027	-0.024	0.006	0.167	0.049	-0.075	-0.060	0.052	0.068
clumsy									0.096	-0.073	0.044	-0.060	-0.020	0.068	-0.023	-0.042	0.072	-0.011	0.068	0.000	0.016	0.008	-0.123	0.039	0.032	0.092	0.042	0.022	-0.021	0.108
complex										0.049	0.070	0.123	-0.060	0.091	-0.043	-0.100	0.101	-0.056	0.010	-0.056	0.114	0.073	0.134	0.060	-0.008	-0.016	0.016	0.010	0.014	**0.207**
delicate											0.119	0.112	-0.046	-0.024	-0.033	-0.041	0.152	0.013	0.101	0.062	0.034	0.029	0.023	0.056	0.058	-0.026	-0.069	0.005	-0.026	-0.005
disorderly												0.061	0.080	-0.022	-0.024	-0.016	0.065	-0.041	0.023	0.049	0.047	-0.019	-0.028	0.146	-0.019	-0.016	-0.035	0.099	-0.006	0.134
dramatic													0.097	-0.006	0.002	0.050	0.037	0.046	0.089	0.106	-0.007	-0.019	0.008	0.067	0.147	0.022	0.013	0.110	0.037	0.014
formal														0.096	-0.135	0.021	-0.050	0.052	-0.006	0.034	-0.042	0.054	0.047	0.018	0.022	0.003	-0.005	0.096	0.017	0.022
fresh															0.056	0.008	-0.093	-0.044	-0.034	0.095	0.005	0.013	0.009	-0.068	0.020	-0.003	-0.006	0.016	-0.000	0.062
friendly																-0.036	0.004	0.027	0.025	-0.007	0.115	0.140	0.021	-0.042	0.008	-0.010	0.071	0.082	0.036	0.094
gentle																	-0.001	0.049	0.100	-0.066	0.019	0.068	-0.038	0.051	0.075	0.004	0.008	0.012	-0.079	0.095
graceful																		0.024	0.130	-0.004	0.107	0.005	-0.026	0.100	0.119	0.091	-0.090	0.061	0.053	-0.009
happy																			0.048	0.126	0.159	0.105	-0.026	0.089	0.093	0.097	0.147	0.034	0.008	0.041
legible																				0.078	0.036	0.013	-0.068	-0.062	0.136	0.017	0.026	0.026	0.088	0.064
modern																					-0.039	0.093	-0.065	-0.029	0.006	-0.024	-0.023	0.091	-0.006	-0.002
playful																						0.083	0.035	0.058	-0.033	0.058	0.018	0.093	0.081	0.066
pretentious																							-0.046	0.115	-0.078	0.029	-0.027	0.076	0.059	-0.017
sharp																								-0.027	-0.076	0.050	-0.005	-0.041	0.027	0.027
sloppy																									0.118	0.102	0.015	0.076	-0.029	-0.048
soft																										0.058	-0.049	0.105	0.021	-0.034
strong																											0.116	0.034	0.015	-0.014
technical																												0.054	-0.034	0.035
thin																													0.031	-0.073
warm																														-0.028

Table 8: Full correlation data for the attributes in Fonts using O'Donovan *et al.*'s method. The maximum correlation is between *complex* and *wide*, shown in bold.. Histogram of this data is shown in fig. 2

	BTL		Crowd-BT		O'Donovan	
Pair Count	μ_τ	σ_τ	μ_τ	σ_τ	μ_τ	σ_τ
200	0.159	6.388e−2	0.402	6.738e−2	0.326	4.386e−2
300	0.618	3.152e−2	0.630	2.386e−2	0.592	3.982e−2
400	0.712	2.189e−2	0.684	3.075e−2	0.699	2.416e−2
500	0.756	1.682e−2	0.735	3.000e−2	0.750	1.817e−2
600	0.786	1.714e−2	0.795	2.644e−2	0.782	1.746e−2
700	0.809	1.234e−2	0.838	1.766e−2	0.805	1.261e−2
800	0.822	1.071e−2	0.870	1.397e−2	0.819	1.147e−2
900	0.838	9.585e−3	0.887	1.640e−2	0.835	9.458e−3
1000	0.845	1.032e−2	0.904	1.553e−2	0.843	1.152e−2
1500	0.876	5.108e−3	0.941	6.072e−3	0.875	5.897e−3
2000	0.897	5.554e−3	0.954	4.039e−3	0.896	5.649e−3
2500	0.907	5.853e−3	0.962	3.220e−3	0.906	5.728e−3
3000	0.915	5.695e−3	0.968	2.432e−3	0.914	6.206e−3
3500	0.922	3.628e−3	0.970	1.555e−3	0.922	3.708e−3
4000	0.929	4.014e−3	0.973	2.040e−3	0.928	4.288e−3
4500	0.933	3.970e−3	0.976	2.003e−3	0.933	3.996e−3
5000	0.937	4.250e−3	0.977	1.628e−3	0.936	4.467e−3
5500	0.939	3.941e−3	0.979	1.286e−3	0.939	3.883e−3
6000	0.942	3.622e−3	0.980	1.447e−3	0.942	3.605e−3
6500	0.945	3.604e−3	0.981	1.268e−3	0.944	3.800e−3
7000	0.947	4.153e−3	0.982	1.226e−3	0.947	3.987e−3
7500	0.950	3.717e−3	0.983	1.432e−3	0.949	3.381e−3
8000	0.951	3.153e−3	0.984	1.229e−3	0.950	3.214e−3
8500	0.952	2.894e−3	0.984	1.124e−3	0.951	2.890e−3
9000	0.955	2.621e−3	0.985	1.201e−3	0.954	2.763e−3
9500	0.956	3.200e−3	0.986	1.308e−3	0.955	3.240e−3
10000	0.957	2.699e−3	0.986	1.156e−3	0.957	2.574e−3
10500	0.958	1.959e−3	0.986	9.190e−4	0.958	2.110e−3
11000	0.958	3.148e−3	0.987	1.362e−3	0.957	2.964e−3
11500	0.960	2.672e−3	0.987	1.126e−3	0.960	2.619e−3
12000	0.961	2.428e−3	0.988	1.164e−3	0.960	2.653e−3
12500	0.961	1.848e−3	0.988	1.200e−3	0.960	1.873e−3
13000	0.962	2.817e−3	0.988	1.141e−3	0.962	2.986e−3
13500	0.964	2.251e−3	0.989	8.040e−4	0.963	2.193e−3
14000	0.964	1.980e−3	0.989	9.820e−4	0.963	1.904e−3
14500	0.964	1.805e−3	0.989	7.530e−4	0.964	1.918e−3
15000	0.966	1.915e−3	0.990	9.960e−4	0.965	1.999e−3
15500	0.966	1.680e−3	0.989	9.980e−4	0.966	1.758e−3
16000	0.967	1.998e−3	0.990	1.126e−3	0.966	2.124e−3
16500	0.967	2.682e−3	0.990	8.210e−4	0.966	2.692e−3
17000	0.967	2.318e−3	0.991	7.880e−4	0.967	2.306e−3
17500	0.968	2.357e−3	0.991	1.009e−3	0.968	2.391e−3
18000	0.969	2.356e−3	0.991	1.000e−3	0.968	2.238e−3
18500	0.969	2.739e−3	0.991	7.680e−4	0.969	2.808e−3
19000	0.970	1.829e−3	0.991	1.104e−3	0.969	1.894e−3
19500	0.971	1.995e−3	0.991	9.350e−4	0.970	1.906e−3
19900	0.971	2.037e−3	0.991	9.620e−4	0.970	2.207e−3

Table 9: Mean τ, μ_τ, and standard deviation, σ_τ, for the simulated data with 8 judges per pair, $\omega = 0.02$, and $\lambda = 0$.

	BTL		Crowd-BT		O'Donovan	
Pair Count	μ_τ	σ_τ	μ_τ	σ_τ	μ_τ	σ_τ
200	0.282	6.349e−2	0.402	6.738e−2	0.606	2.404e−2
300	0.626	2.331e−2	0.630	2.386e−2	0.681	2.446e−2
400	0.681	2.946e−2	0.684	3.075e−2	0.727	2.403e−2
500	0.741	2.870e−2	0.735	3.000e−2	0.765	1.850e−2
600	0.802	2.443e−2	0.795	2.644e−2	0.808	1.618e−2
700	0.843	1.761e−2	0.838	1.766e−2	0.842	1.732e−2
800	0.873	1.406e−2	0.870	1.397e−2	0.870	1.386e−2
900	0.890	1.510e−2	0.887	1.640e−2	0.889	1.358e−2
1000	0.907	1.367e−2	0.904	1.553e−2	0.905	1.343e−2
1500	0.942	6.226e−3	0.941	6.072e−3	0.940	6.366e−3
2000	0.954	3.868e−3	0.954	4.039e−3	0.953	3.996e−3
2500	0.962	3.301e−3	0.962	3.220e−3	0.961	3.522e−3
3000	0.968	2.553e−3	0.968	2.432e−3	0.967	2.297e−3
3500	0.970	1.604e−3	0.970	1.555e−3	0.970	1.908e−3
4000	0.973	2.049e−3	0.973	2.040e−3	0.972	2.349e−3
4500	0.976	2.032e−3	0.976	2.003e−3	0.975	1.936e−3
5000	0.977	1.625e−3	0.977	1.628e−3	0.976	1.365e−3
5500	0.979	1.261e−3	0.979	1.286e−3	0.978	1.568e−3
6000	0.980	1.486e−3	0.980	1.447e−3	0.979	1.630e−3
6500	0.981	1.258e−3	0.981	1.268e−3	0.980	1.302e−3
7000	0.982	1.176e−3	0.982	1.226e−3	0.981	1.400e−3
7500	0.983	1.383e−3	0.983	1.432e−3	0.982	1.226e−3
8000	0.984	1.277e−3	0.984	1.229e−3	0.983	1.286e−3
8500	0.984	1.211e−3	0.984	1.124e−3	0.984	1.156e−3
9000	0.985	1.255e−3	0.985	1.201e−3	0.985	1.101e−3
9500	0.986	1.284e−3	0.986	1.308e−3	0.985	1.225e−3
10000	0.986	1.124e−3	0.986	1.156e−3	0.986	1.075e−3
10500	0.986	9.010e−4	0.986	9.190e−4	0.986	9.450e−4
11000	0.987	1.346e−3	0.987	1.362e−3	0.986	1.245e−3
11500	0.987	1.104e−3	0.987	1.126e−3	0.987	1.168e−3
12000	0.988	1.201e−3	0.988	1.164e−3	0.988	1.171e−3
12500	0.988	1.248e−3	0.988	1.200e−3	0.988	1.175e−3
13000	0.988	1.060e−3	0.988	1.141e−3	0.988	1.291e−3
13500	0.989	8.260e−4	0.989	8.040e−4	0.988	1.037e−3
14000	0.989	9.140e−4	0.989	9.820e−4	0.988	1.022e−3
14500	0.989	7.530e−4	0.989	7.530e−4	0.989	9.090e−4
15000	0.990	9.540e−4	0.990	9.960e−4	0.989	8.090e−4
15500	0.990	9.980e−4	0.989	9.980e−4	0.989	9.160e−4
16000	0.990	1.079e−3	0.990	1.126e−3	0.990	1.005e−3
16500	0.990	7.200e−4	0.990	8.210e−4	0.990	9.230e−4
17000	0.991	8.070e−4	0.991	7.880e−4	0.990	8.530e−4
17500	0.991	9.970e−4	0.991	1.009e−3	0.990	8.880e−4
18000	0.991	1.000e−3	0.991	1.000e−3	0.990	9.200e−4
18500	0.991	8.640e−4	0.991	7.680e−4	0.991	7.240e−4
19000	0.991	9.870e−4	0.991	1.104e−3	0.991	1.257e−3
19500	0.991	9.400e−4	0.991	9.350e−4	0.991	8.430e−4
19900	0.991	9.840e−4	0.991	9.620e−4	0.991	9.850e−4

Table 10: Mean τ, μ_τ, and standard deviation, σ_τ, for the simulated data with 8 judges per pair, $\omega = 0.1$, and $\lambda = 0$.

	Crowd-BT - BTL			O'Donovan - BTL		
Pair Count	μ_τ	σ_τ	$\log(\mathrm{BF}_{10})$	μ_τ	σ_τ	$\log(\mathrm{BF}_{10})$
200	5.664e−2	5.745e−2	**7.823**	1.671e−1	0.068458	**25.437**
300	−3.695e−2	2.123e−2	**17.792**	−2.567e−2	0.022218	**10.178**
400	−2.775e−2	2.007e−2	**13.264**	−1.348e−2	0.014034	**7.464**
500	−1.403e−2	1.426e−2	**7.792**	−6.184e−3	0.005163	**10.761**
600	−7.377e−3	6.312e−3	**10.362**	−4.134e−3	0.003772	**9.353**
700	−5.236e−3	3.542e−3	**14.512**	−3.568e−3	0.003370	**8.834**
800	−3.869e−3	3.730e−3	**8.537**	−3.307e−3	0.003269	**8.179**
900	−3.598e−3	3.638e−3	**7.865**	−3.022e−3	0.003212	**7.194**
1000	−3.447e−3	3.118e−3	**9.487**	−1.933e−3	0.003129	**2.863**
1500	−1.534e−3	2.472e−3	**2.898**	−1.558e−3	0.002241	**3.849**
2000	−1.588e−3	1.280e−3	**11.353**	−1.702e−3	0.001675	**8.239**
2500	−1.146e−3	1.475e−3	**4.937**	−1.102e−3	0.001742	**3.051**
3000	−7.600e−4	1.342e−3	**2.240**	−9.410e−4	0.001692	**2.115**
3500	−4.590e−4	1.227e−3	0.186	−6.260e−4	0.001114	**2.189**
4000	−5.960e−4	1.034e−3	**2.357**	−5.860e−4	0.001362	0.729
4500	−6.570e−4	1.100e−3	**2.604**	−5.660e−4	0.001164	1.318
5000	−7.810e−4	1.252e−3	**2.933**	−5.860e−4	0.001199	1.349
5500	−1.340e−4	8.270e−4	-1.277	−6.630e−4	0.000971	**3.697**
6000	−5.530e−4	1.145e−3	1.279	−5.800e−4	0.001073	**1.926**
6500	−5.090e−4	9.390e−4	**1.953**	−1.092e−3	0.001005	**9.232**
7000	−6.430e−4	1.079e−3	**2.593**	−6.230e−4	0.000973	**3.149**
7500	−2.510e−4	1.030e−3	-0.832	−5.090e−4	0.001258	0.474
8000	−3.920e−4	8.940e−4	0.810	−6.670e−4	0.000964	**3.806**
8500	−4.360e−4	8.620e−4	**1.525**	−5.190e−4	0.001043	1.444
9000	−6.200e−4	7.800e−4	**5.179**	−9.250e−4	0.000905	**8.316**
9500	−1.070e−4	9.220e−4	-1.451	−5.160e−4	0.001106	1.103
10000	−4.190e−4	7.510e−4	**2.129**	−3.520e−4	0.000853	0.548
10500	−3.920e−4	1.051e−3	0.175	−6.100e−4	0.000972	**2.986**
11000	−3.580e−4	8.270e−4	0.759	−4.960e−4	0.000758	**3.325**
11500	−3.580e−4	5.690e−4	**3.016**	−5.460e−4	0.000914	**2.613**
12000	−2.910e−4	8.060e−4	0.072	−6.030e−4	0.001161	**1.689**
12500	−3.690e−4	8.140e−4	0.955	−7.470e−4	0.000795	**7.182**
13000	−2.980e−4	6.910e−4	0.739	−5.700e−4	0.001045	**1.983**
13500	−2.750e−4	7.670e−4	0.043	−5.860e−4	0.000686	**6.002**
14000	−1.570e−4	7.370e−4	-1.017	−6.130e−4	0.000788	**4.959**
14500	−2.950e−4	6.710e−4	0.817	−8.780e−4	0.000850	**8.467**
15000	−1.680e−4	7.190e−4	-0.902	−4.020e−4	0.000755	**1.834**
15500	−1.410e−4	7.090e−4	-1.100	−3.580e−4	0.000760	1.157
16000	−3.520e−4	7.670e−4	1.019	−7.970e−4	0.000813	**7.741**
16500	−5.400e−5	8.290e−4	-1.580	−4.090e−4	0.000949	0.729
17000	−2.000e−5	5.760e−4	-1.621	−3.180e−4	0.000654	1.321
17500	−1.980e−4	5.680e−4	-0.044	−3.220e−4	0.000834	0.294
18000	−3.220e−4	8.510e−4	0.219	−6.730e−4	0.000770	**6.280**
18500	−2.550e−4	7.450e−4	-0.098	−5.590e−4	0.000825	**3.627**
19000	−2.580e−4	6.070e−4	0.674	−6.030e−4	0.000731	**5.591**
19500	−2.080e−4	5.980e−4	-0.054	−4.760e−4	0.000931	**1.590**
19900	−3.950e−4	6.700e−4	**2.523**	−6.300e−4	0.000557	**9.827**

Table 11: Statistical comparison between the two experimental methods, Crowd-BT and O'Donovan *et al.*'s method, and the unweighted BTL order on a matched pair basis for the simulated data with 8 judges per pair, $\omega = 0.02$, and $\lambda = 0$. A positive value of μ_τ indicates that the experimental method is *better*. We show the mean and standard deviation along with a Bayesian Paired Samples T-Test, computed using JASP 0.16 [1]. We use Student-T where $H_a : \tau_{\textbf{experimental}} \neq \tau_{\textbf{BTL}}$ and the prior is Cauchy with scale 0.707. Values of $\log(\mathrm{BF}_{10})$ are traditionally interpreted using the scale provided by Jeffreys [2]; we highlight values above 1.5 in bold, which are generally regarded as having strong evidence.

	Crowd-BT - BTL			O'Donovan - BTL		
Pair Count	μ_τ	σ_τ	$\log(\mathrm{BF}_{10})$	μ_τ	σ_τ	$\log(\mathrm{BF}_{10})$
200	1.202e−1	8.178e−2	**14.405**	3.240e−1	6.826e−2	**42.488**
300	4.429e−3	8.860e−3	**1.467**	5.543e−2	3.383e−2	**16.550**
400	2.593e−3	1.875e−2	-1.374	4.566e−2	3.443e−2	**12.505**
500	−5.853e−3	1.776e−2	-0.202	2.461e−2	2.906e−2	**5.894**
600	−6.697e−3	1.608e−2	0.588	6.533e−3	1.610e−2	0.483
700	−5.079e−3	6.025e−3	**5.839**	−7.140e−4	7.468e−3	-1.511
800	−2.945e−3	3.559e−3	**5.624**	−2.228e−3	4.498e−3	1.416
900	−3.712e−3	3.830e−3	**7.587**	−1.648e−3	4.042e−3	0.504
1000	−2.647e−3	4.127e−3	**3.159**	−2.054e−3	3.699e−3	**2.102**
1500	−6.767e−4	1.070e−3	**3.046**	−1.776e−3	1.707e−3	**8.577**
2000	−3.451e−4	8.680e−4	0.405	−1.236e−3	1.246e−3	**7.909**
2500	−2.044e−4	6.220e−4	-0.211	−9.580e−4	1.136e−3	**5.847**
3000	−9.045e−5	4.900e−4	-1.171	−9.310e−4	1.202e−3	**4.913**
3500	−8.710e−5	4.810e−4	-1.189	−4.760e−4	9.970e−4	1.218
4000	−1.709e−4	4.690e−4	0.100	−9.950e−4	1.081e−3	**6.910**
4500	4.020e−5	3.290e−4	-1.432	−8.210e−4	8.190e−4	**8.054**
5000	−1.005e−4	4.030e−4	-0.797	−7.770e−4	8.490e−4	**6.845**
5500	1.480e−17	4.180e−4	-1.638	−7.840e−4	6.790e−4	**10.171**
6000	−1.206e−4	2.800e−4	0.734	−8.740e−4	7.270e−4	**10.825**
6500	−1.575e−4	3.050e−4	**1.648**	−8.140e−4	7.640e−4	**8.927**
7000	−1.240e−4	3.370e−4	0.132	−7.600e−4	6.590e−4	**10.155**
7500	−2.680e−5	3.390e−4	-1.551	−3.250e−4	7.120e−4	0.993
8000	3.685e−5	2.320e−4	-1.292	−5.260e−4	7.830e−4	**3.544**
8500	−1.273e−4	2.530e−4	**1.503**	−6.060e−4	6.720e−4	**6.665**
9000	1.005e−5	3.090e−4	-1.623	−3.820e−4	6.880e−4	**2.097**
9500	−3.015e−5	2.820e−4	-1.480	−4.590e−4	5.920e−4	**4.910**
10000	−1.005e−4	2.990e−4	-0.144	−5.830e−4	5.590e−4	**8.617**
10500	−6.030e−5	3.150e−4	-1.137	−6.700e−4	5.680e−4	**10.526**
11000	1.110e−17	2.790e−4	-1.638	−4.760e−4	7.090e−4	**3.535**
11500	−9.380e−5	2.990e−4	-0.327	−2.910e−4	6.490e−4	0.921
12000	6.700e−6	2.470e−4	-1.628	−2.040e−4	5.730e−4	0.029
12500	−1.005e−4	2.420e−4	0.579	−4.790e−4	6.710e−4	**4.102**
13000	−1.206e−4	2.660e−4	0.969	−6.400e−4	6.230e−4	**8.391**
13500	−6.365e−5	2.890e−4	-0.976	−4.150e−4	6.270e−4	**3.426**
14000	−4.020e−5	2.850e−4	-1.363	−3.220e−4	5.760e−4	**2.139**
14500	−7.401e−18	2.090e−4	-1.638	−5.560e−4	6.020e−4	**6.950**
15000	9.380e−5	2.790e−4	-0.149	−4.050e−4	3.710e−4	**9.312**
15500	−1.005e−4	2.490e−4	0.463	−4.660e−4	5.730e−4	**5.422**
16000	2.345e−5	2.140e−4	-1.471	−4.190e−4	5.320e−4	**5.079**
16500	−2.680e−5	2.400e−4	-1.466	−4.860e−4	5.080e−4	**7.396**
17000	2.680e−5	2.360e−4	-1.459	−3.350e−4	6.890e−4	1.317
17500	6.700e−5	3.040e−4	-0.976	−4.050e−4	5.060e−4	**5.267**
18000	−4.355e−5	2.340e−4	-1.164	−4.860e−4	4.930e−4	**7.809**
18500	2.345e−5	2.970e−4	-1.551	−2.610e−4	4.370e−4	**2.625**
19000	−3.350e−5	2.980e−4	-1.463	−4.720e−4	5.080e−4	**7.045**
19500	−6.030e−5	2.260e−4	-0.684	−2.480e−4	3.880e−4	**3.135**
19900	1.005e−5	2.660e−4	-1.618	−3.790e−4	5.460e−4	**3.820**

Table 12: Statistical comparison between the two experimental methods, Crowd-BT and O'Donovan *et al.*'s method, and the unweighted BTL order on a matched pair basis for the simulated data with 8 judges per pair, $\omega = 0.1$, and $\lambda = 0$. A positive value of μ_τ indicates that the experimental method is *better*. We show the mean, the standard deviation, and a Bayesian Paired Samples T-Test, computed using JASP 0.16 [1]. We use Student-T where $H_a : \tau_{\mathbf{experimental}} \neq \tau_{\mathbf{BTL}}$ and the prior is Cauchy with scale 0.707. Values of $\log(\mathbf{BF}_{10})$ are traditionally interpreted using the scale provided by Jeffreys [2]; we highlight values above 1.5 in bold, which are generally regarded as having strong evidence.