Online Appendix

Salience Bias in Crowdsourcing Contests

Table A1a. *Learning Phase Model – with reminder dummy*

	(1)
$Time_{ijc}$	0.00006239
	(0.009369)
log ($\#Feedback_{ijc}$)	0.05765***
	(0.003809)
$Reminder_c \times$ log ($\#Feedback_{ijc}$)	-0.04739**
	(0.01945)
Contestant Fixed Effect	Yes
Contest Fixed Effect	Yes
R^2	0.254
Observations	695622

***p<0.01, **p<0.05, *p<0.1.

Table A1b. *Submission Phase Model – with reminder dummy*

	(1)
$Time_{ijc}$	1.3138***
	(0.04698)
$\widetilde{RankInTeam}_{ijc}^{Public}$	14.786***
	(0.2157)
$\widetilde{RankInTeam}_{ijc}^{Private}$	0.6831***
	(0.09093)
$Reminder_c \times \widetilde{RankInTeam}_{ijc}^{Public}$	-4.3793***
	(0.7518)
$Reminder_c \times \widetilde{RankInTeam}_{ijc}^{Private}$	0.8611*
	(0.4885)
Pseudo-R^2	0.659
Observations	670838

***p<0.01, **p<0.05, *p<0.1.

Table A2. *Submission Phase Model – Including the Moderating Effect by* $log(\#TotalFeedback_{ic})$

$Time_{ijc}$	1.1502***
	(0.04567)
$RankIn\widetilde{Team}_{ijc}^{Public}$	7.9206***
	(0.6390)
$RankIn\widetilde{Team}_{ijc}^{Private}$	-0.6628**
	(0.3338)
$\log(\#TotalFeedback_{ic}) \times RankIn\widetilde{Team}_{ijc}^{Public}$	2.2656***
	(0.2406)
$\log(\#TotalFeedback_{ic}) \times RankIn\widetilde{Team}_{ijc}^{Private}$	0.4766***
	(0.1152)
Pseudo-R^2	0.661
Observations	670838

***p<0.01, **p<0.05, *p<0.1.

Table A3a. *Learning Phase Model - Sub-sample Analysis*

	(1) Exclude contests that mention dataset difference	(2) Exclude contests with smaller training set	(3) Exclude contests with over 50% of test set for deriving public score	(4) All Criteria Combined
$Time_{ijc}$	-0.01599	-0.01538	0.003219	-0.01127
	(0.01147)	(0.01239)	(0.01004)	(0.01555)
log ($\#Feedback_{ijc}$)	0.06260***	0.08595***	0.06021***	0.08822***
	(0.005139)	(0.005269)	(0.004032)	(0.007038)
Contestant Fixed Effect	Yes	Yes	Yes	Yes
Contest Fixed Effect	Yes	Yes	Yes	Yes
R^2	0.212	0.233	0.264	0.223
Observations	416683	414921	635021	240069

***p<0.01, **p<0.05, *p<0.1.

Table A3b. *Submission Phase Model - Sub-sample Analysis*

	(1) Exclude contests that mention dataset difference	(2) Exclude contests with smaller training set	(3) Exclude contests with over 50% of test set for deriving public score	(4) All Criteria Combined
$Time_{ijc}$	1.2697***	1.3723***	1.3469***	1.3191***
	(0.05809)	(0.05658)	(0.04811)	(0.06852)
$\widetilde{RankInTeam}_{ijc}^{Public}$	14.917***	13.193***	14.315***	13.293***
	(0.2618)	(0.2181)	(0.2092)	(0.2682)
$\widetilde{RankInTeam}_{ijc}^{Private}$	0.6300***	0.5404***	0.7019***	0.4037***
	(0.1154)	(0.1003)	(0.08933)	(0.1278)
Pseudo-R^2	0.666	0.630	0.652	0.630
Observations	401613	399326	612452	230833

***p<0.01, **p<0.05, *p<0.1.

Table A4a. *Learning Phase Model – Contestants who participated in cross-validation discussion*

	(1) All contests	(2) The contests after cross-validation discussions occur
$Time_{ijc}$	0.08750***	-0.0009789
	(0.03218)	(0.03554)
log ($\#Feedback_{ijc}$)	0.05176***	0.05130***
	(0.01119)	(0.01516)
Contestant Fixed Effect	Yes	Yes
Contest Fixed Effect	Yes	Yes
R^2	0.454	0.499
Observations	85105	33139

***p<0.01, **p<0.05, *p<0.1.

Table A4b. *Submission Phase Model – Contestants who participated in cross-validation discussion*

	(1) All contest	(2) The contests after cross-validation discussions occur
$Time_{ijc}$	2.2194***	2.2144***
	(0.1720)	(0.2592)
$\widetilde{RankInTeam}_{ijc}^{Public}$	12.021***	11.226***
	(0.5977)	(0.8178)
$\widetilde{RankInTeam}_{ijc}^{Private}$	1.3831***	0.7898*
	(0.2685)	(0.4282)
Pseudo-R^2	0.560	0.544
Observations	84514	32852

***p<0.01, **p<0.05, *p<0.1.

Table A5. *Learning Phase Model – With IV*

	(1)	(2)	(3)
$Time_{ijc}$	-0.005343	-0.006058	-0.006120
	(0.01134)	(0.01131)	(0.01131)
$\log(\#Feedback_{ijc})$	0.05857***	0.05767***	0.05866***
	(0.004780)	(0.004825)	(0.004844)
$Winner_{ic} \times \log(\#Feedback_{ijc})$		0.03747	0.06249*
		(0.02705)	(0.03195)
$\log(\#Team_c) \times \log(\#Feedback_{ijc})$			-0.01979***
			(0.003664)
$Winner_{ic} \times \log(\#Team_c) \times \log(\#Feedback_{ijc})$			0.07283***
			(0.02497)
Contestant Fixed Effect	Yes	Yes	Yes
Contest Fixed Effect	Yes	Yes	Yes
R^2	0.254	0.254	0.255
Observations	695622	695622	695622

***p<0.01, **p<0.05, *p<0.1.

Table A6a. *Learning Phase Model – Subsample with only Winners*

	(1)	(2)
$Time_{ijc}$	-0.06225	-0.03301
	(0.08124)	(0.07628)
log $(\#Feedback_{ijc})$	0.1010***	0.1195***
	(0.03646)	(0.04364)
$\log(\#Team_c) \times \log(\#Feedback_{ijc})$		0.05558**
		(0.02755)
Contestant Fixed Effect	Yes	Yes
Contest Fixed Effect	Yes	Yes
R^2	0.198	0.201
Observations	19767	19767

***p<0.01, **p<0.05, *p<0.1.

Table A6b. *Submission Phase Model – Subsample with only Winners*

	(1)	(2)
$Time_{ijc}$	5.6377***	5.4365***
	(1.0235)	(1.0351)
$\widetilde{RankInTeam}_{ijc}^{Public}$	10.151***	11.863***
	(1.1987)	(2.1617)
$\widetilde{RankInTeam}_{ijc}^{Private}$	3.5338***	5.5889***
	(0.7127)	(1.1379)
$\log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Public}$		1.1017
		(1.3155)
$\log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Private}$		1.7828**
		(0.8094)
Pseudo-R^2	0.548	0.553
Observations	19749	19749

***p<0.01, **p<0.05, *p<0.1.

Table A7. *Learning Phase Model – Dynamic Number of Teams*

	(1)
$Time_{ijc}$	-0.0006463
	(0.009365)
$\log(\#Feedback_{ijc})$	0.05459***
	(0.003861)
$Winner_{ic} \times \log(\#Feedback_{ijc})$	0.05382*
	(0.02895)
$\log(\#team_{ijc}) \times \log(\#Feedback_{ijc})$	-0.003228
	(0.002043)
$Winner_{ic} \times \log(\#team_{ijc}) \times \log(\#Feedback_{ijc})$	0.03853***
	(0.01484)
Contestant Fixed Effect	Yes
Contest Fixed Effect	Yes
R^2	0.254
Observations	695622

***p<0.01, **p<0.05, *p<0.1.

Table A8a. *Learning Phase Model – Including Contests with only Kaggle Points as Reward*

	(1)	(2)	(3)
$Time_{ijc}$	-0.0009699	-0.001068	0.0003313
	(0.009334)	(0.009321)	(0.009317)
$\log(\#Feedback_{ijc})$	0.05606***	0.05472***	0.05507***
	(0.003782)	(0.003795)	(0.003794)
$Winner_{ic} \times \log(\#Feedback_{ijc})$		0.04146	0.06698**
		(0.02681)	(0.03177)
$\log(\#Team_c) \times \log(\#Feedback_{ijc})$			-0.01960***
			(0.003206)
$Winner_{ic} \times \log(\#Team_c) \times \log(\#Feedback_{ijc})$			0.07730***
			(0.02453)
Contestant Fixed Effect	Yes	Yes	Yes
Contest Fixed Effect	Yes	Yes	Yes
R^2	0.256	0.256	0.257
Observations	699236	699236	699236

***p<0.01, **p<0.05, *p<0.1.

Table A8b. *Submission Phase Model – Including Contests with only Kaggle Points as Reward*

	(1)	(2)	(3)
$Time_{ijc}$	1.3100***	1.3097***	1.3129***
	(0.04639)	(0.04640)	(0.04705)
$\widetilde{RankInTeam}_{ijc}^{Public}$	14.242***	14.302***	15.352***
	(0.2084)	(0.2085)	(0.2183)
$\widetilde{RankInTeam}_{ijc}^{Private}$	0.8463***	0.7879***	0.6025***
	(0.09188)	(0.09171)	(0.09214)
$Winner_{ic} \times \widetilde{RankInTeam}_{ijc}^{Public}$		-3.1256**	-2.6151
		(1.2709)	(2.3141)
$Winner_{ic} \times \widetilde{RankInTeam}_{ijc}^{Private}$		3.4585***	6.0268***
		(0.8022)	(1.1317)
$\log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Public}$			2.6996***
			(0.1537)
$\log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Private}$			-0.1673**
			(0.08466)
$Winner_{ic} \times \log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Public}$			-1.6608
			(1.3381)
$Winner_{ic} \times \log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Private}$			2.1604***
			(0.7176)
Pseudo-R^2	0.655	0.655	0.660
Observations	673901	673901	673901

***p<0.01, **p<0.05, *p<0.1.

Table A9. *Submission Phase Model – Alternative Specification*

	(1)	(2)	(3)
$Time_{ijc}$	1.3102***	1.3102***	1.3112***
	(0.04683)	(0.04685)	(0.04745)
$\widetilde{RankInTeam}_{ijc}^{Public}$	15.253***	15.258***	16.050***
	(0.2012)	(0.2022)	(0.2100)
$\widetilde{RankDifference}_{ijc}$ †	-0.7217***	-0.6571***	-0.5345***
	(0.08929)	(0.08946)	(0.09277)
$Winner_{ic} \times \widetilde{RankInTeam}_{ijc}^{Public}$		0.1675	3.3693
		(1.4401)	(2.5623)
$Winner_{ic} \times \widetilde{RankDifference}_{ijc}$		-3.7230***	-6.0402***
		(0.8209)	(1.1453)
$\log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Public}$			2.4894***
			(0.1600)
$\log(\#Team_c) \times \widetilde{RankDifference}_{ijc}$			0.07634
			(0.08523)
$Winner_{ic} \times \log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Public}$			0.6337
			(1.3212)
$Winner_{ic} \times \log(\#Team_c) \times \widetilde{RankDifference}_{ijc}$			-1.9316**
			(0.7602)
Pseudo-R^2	0.658	0.658	0.663
Observations	670838	670838	670838

† $RankDifference_{ijc} = \widetilde{RankInTeam}_{ijc}^{Public} - \widetilde{RankInTeam}_{ijc}^{Private}$

***p<0.01, **p<0.05, *p<0.1.

Table A10a. *Learning Phase Model – Including Reward Size and Experience*

	(1)
$Time_{ijc}$	-0.0004221
	(0.009278)
$\log(\#Feedback_{ijc})$	0.05575***
	(0.003724)
$Winner_{ic} \times \log(\#Feedback_{ijc})$	0.06914**
	(0.03191)
$\log(Reward_{ic}) \times \log(\#Feedback_{ijc})$	0.007425***
	(0.002406)
$\log(Experience_{ic}) \times \log(\#Feedback_{ijc})$	-0.003316
	(0.002944)
$\log(\#Team_c) \times \log(\#Feedback_{ijc})$	-0.02503***
	(0.003627)
$Winner_{ic} \times \log(\#Team_c) \times \log(\#Feedback_{ijc})$	0.07874***
	(0.02464)
Contestant Fixed Effect	Yes
Contest Fixed Effect	Yes
R^2	0.255
Observations	695622

***p<0.01, **p<0.05, *p<0.1.

Table A10b *Submission Phase Model – Including Reward Size and Experience*

	(1)		
$Time_{ijc}$	1.3214***	$Winner_{ic} \times \widetilde{RankInTeam}_{ijc}^{Public}$	-2.2577
	(0.04834)		(2.3213)
$\widetilde{RankInTeam}_{ijc}^{Public}$	16.230***	$Winner_{ic} \times \widetilde{RankInTeam}_{ijc}^{Private}$	5.8949***
	(0.2717)		(1.1395)
$\widetilde{RankInTeam}_{ijc}^{Private}$	0.1655	$\log(\#Team_c)$	2.4056***
	(0.1233)	$\times \widetilde{RankInTeam}_{ijc}^{Public}$	(0.2006)
$\log(Rewardsize_c)$	0.2531*	$\log(\#Team_c)$	-0.3271***
$\times \widetilde{RankInTeam}_{ijc}^{Public}$	(0.1367)	$\times \widetilde{RankInTeam}_{ijc}^{Private}$	(0.1105)
$\log(Rewardsize_c)$	0.2548***	$Winner_{ic} \times \log(\#Team_c)$	-1.1334
$\times \widetilde{RankInTeam}_{ijc}^{Private}$	(0.07085)	$\times \widetilde{RankInTeam}_{ijc}^{Public}$	(1.3451)
$\log(Experience_{ic})$	-0.7264***	$Winner_{ic} \times \log(\#Team_c)$	1.9243**
$\times \widetilde{RankInTeam}_{ijc}^{Public}$	(0.2054)	$\times \widetilde{RankInTeam}_{ijc}^{Private}$	(0.7759)
$\log(Experience_{ic})$	0.3980***		
$\times \widetilde{RankInTeam}_{ijc}^{Private}$	(0.08464)		
Pseudo-R^2	0.663		
Observations	670838		

***p<0.01, **p<0.05, *p<0.1.

Table A11a. *Learning Phase Model – Two-way Clustered Standard Error*

	(1)	(2)	(3)
$Time_{ijc}$	-0.0007494	-0.0008435	0.0006821
	(0.01424)	(0.01422)	(0.01399)
$\log(\#Feedback_{ijc})$	0.05633***	0.05499***	0.05529***
	(0.01639)	(0.01636)	(0.01549)
$Winner_{ic} \times \log(\#Feedback_{ijc})$		0.04155	0.06762**
		(0.02672)	(0.02787)
$\log(\#Team_c) \times \log(\#Feedback_{ijc})$			-0.02062*
			(0.01131)
$Winner_{ic} \times \log(\#Team_c) \times \log(\#Feedback_{ijc})$			0.07869***
			(0.02484)
Contestant Fixed Effect	Yes	Yes	Yes
Contest Fixed Effect	Yes	Yes	Yes
R^2	0.254	0.254	0.255
Observations	695622	695622	695622

***p<0.01, **p<0.05, *p<0.1.

Table A11b. *Submission Phase Model – Two-way Clustered Standard Error*

	(1)	(2)	(3)
$Time_{ijc}$	1.3102***	1.3102***	1.3112***
	(0.1138)	(0.1129)	(0.1128)
$\widetilde{RankInTeam}_{ijc}^{Public}$	14.531***	14.600***	15.516***
	(0.8622)	(0.8756)	(0.1.0895)
$\widetilde{RankInTeam}_{ijc}^{Private}$	0.7217***	0.6571***	0.5345**
	(0.2175)	(0.2205)	(0.2261)
$Winner_{ic} \times \widetilde{RankInTeam}_{ijc}^{Public}$		-3.5554*	-2.6709
		(1.8176)	(3.0129)
$Winner_{ic} \times \widetilde{RankInTeam}_{ijc}^{Private}$		3.7230***	6.0402***
		(0.8688)	(0.8630)
$\log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Public}$			2.5658***
			(0.7429)
$\log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Private}$			-0.07634
			(0.1815)
$Winner_{ic} \times \log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Public}$			-1.2979
			(1.7627)
$Winner_{ic} \times \log(\#Team_c) \times \widetilde{RankInTeam}_{ijc}^{Private}$			1.9316***
			(0.6812)
Pseudo-R^2	0.658	0.658	0.663
Observations	670838	670838	670838

***p<0.01, **p<0.05, *p<0.1.

Table A12a. *Occupation of the Survey Respondents*

Occupation	**Count**
Undergraduate	12
Graduate Student	24
Data Scientist	31
Programmer	18
Professor/ Teacher	5
Self-employed	4
Other	7

Table A12b. *Educational Level of the Survey Respondents*

Educational Level	**Data Analytics Related**	**Non Data Analytics Related**
High School Diploma	3	1
Associate Degree	0	0
Bachelor Degree	13	9
Master Degree	31	19
Doctorate Degree	15	8
Other	2	0
Total	64	37

Table A12c. *Analytics and Kaggle Experience of Survey Respondents*

	Minimum	Maximum	Mean	Standard Deviation
Years of experience in data analytics	0.1	15	3.37	3.45
Years on Kaggle	0.5	7	1.98	1.26
Number of contests participated in	1	73	7.98	12.76
Number of contests won	0	3	0.17	0.51

Appendix A1. Survey of Kaggle Contestants

Principal Investigator: Jan Stallaert

Co-Investigators: Brian Lee, Sulin Ba, Xinxin Li

Title of Study: Design of Crowdsourcing Contests

You are invited to participate in this survey regarding your experience on Kaggle. We are a group of researchers at the University of Connecticut who study how the design of crowdsourcing contests impacts the outcome. The results of this study can inform crowdsourcing contest design improvement. We would like to get your feedback based on your experience with Kaggle contests. What the study is about: Your participation in this study involves the completion of a survey. The objective of this survey is to understand how the feedback system on Kaggle, more specifically the public score displayed for each solution, impacts the final solutions submitted by the contestant. Compensation: If you complete the whole survey, you may opt to enter a pool from which each participant has a 1 in 25 chance to receive a $25 Visa gift card from us. The winners of the gift cards will be notified via email.

Risks and benefits: This survey does not involve any risk to you. However, your participation may help us improve the design of crowdsourcing platforms like Kaggle.

Anyone over 18 years old who has a Kaggle account created before 05/23/2017 and has participated in at least one contest on Kaggle is qualified to participate in this study. Filling out the survey will take approximately 5 minutes. Your participation will be voluntary and anonymous unless you choose to leave your Kaggle ID and email address to enter into the reward pool. Anyone who fills out the survey more than once will be automatically withdrawn from the reward pool. The survey will expire by 05/30/2017. Please fill out the survey by 05/30/2017.

If you have questions about this study, you may contact Brian Lee at brian.lee@business.uconn.edu. If you have any questions about your rights as a research participant, you may contact the University of Connecticut Institutional Review Board (IRB): http://research.uconn.edu/irb. The IRB seeks to create a collaborative relationship with the research community to assure that research with human subjects is conducted in accordance with legal requirements and ethical principles of Respect for Persons, Beneficence and Justice.

Which one of the following best describes your occupation?

- ❍ Undergraduate Student
- ❍ Graduate Student
- ❍ Data Scientist
- ❍ Programmer
- ❍ Professor/Teacher
- ❍ Self-employed
- ❍ Other (Please specified): _____

What is the highest degree that you have achieved/are pursuing?

- ❍ High School Diploma
- ❍ Associate Degree
- ❍ Bachelor Degree
- ❍ Master Degree
- ❍ Doctorate Degree
- ❍ Other (Please specified) _____

Do you have an educational background related to data analytics/predictive modeling/machine learning?

- ❍ Yes
- ❍ No

How many years of experience do you have in data analytics/predictive modeling/machine learning? Please choose 15 if you have 15 or more years of experience.

_____ Number of Years

How many years have you been a member on Kaggle?

_____ Number of Years

How many contests have you participated in on Kaggle before? (Please put a number)

How many contests have you won on Kaggle before? (Please put a number)

Are you familiar with the public scores provided by the contest holder during a Kaggle contest?

- ❍ Yes
- ❍ No
- ❍ I am not sure what public scores mean

Are you familiar with the private scores provided by the contest holder when a Kaggle contest is over?

- ❍ Yes
- ❍ No
- ❍ I am not sure what private scores mean

From your understanding, which score does Kaggle use to determine the prize winners of a Kaggle contest?

- ❍ Public score
- ❍ Private score
- ❍ I am not sure

Consider the following statement: “A solution with a high public score may not have a high private score.” From your understanding, is this statement:

- ❍ True
- ❍ False
- ❍ I am not sure what public scores or privates scores mean

Here is Kaggle’s explanation of what a public score and private score mean:

"Kaggle competitions are decided by your model's performance on a test data set. Kaggle has the answers for this data set, but withholds them to compare with your predictions. Your Public score is what you receive back upon each submission (that score is calculated using a statistical evaluation metric, which is always described on the Evaluation page). BUT: Your Public Score is being determined from only a fraction of the test data set -- usually between 25-33%. This is the Public Leaderboard, and it shows some relative performance during the competition.

When the competition ends, Kaggle takes your selected submissions and score your predictions against the REMAINING FRACTION of the test set, or the private portion. The score computed by this remaining fraction of the test set is known as Private Score. You never know about your private score until the contest is over. " Is your understanding of public scores consistent with the above explanation?

- ❍ Yes
- ❍ No (Please briefly explain) _____

Is your understanding of private scores consistent with the above explanation?

- ❍ Yes
- ❍ No (Please briefly explain) _____

In a Kaggle contest that provides public scores, do you use cross-validation (e.g. K-fold cross-validation, leave-p-out cross-validation, repeated random sub-sampling validation, forward chaining or similar methods) to evaluate your solutions?

- ❍ Always
- ❍ Most of the time
- ❍ Sometimes
- ❍ Never

When evaluating solutions, how reliable do you think cross-validation is compared to the feedback given by the public score?

- ❍ Considerably more reliable
- ❍ Somewhat more reliable
- ❍ Equally reliable
- ❍ Somewhat less reliable
- ❍ Considerably less reliable
- ❍ I do not do cross-validation for Kaggle contests.

If a contest does not mention how the test set is generated, do you assume that the test set and the training set are drawn from the same population?

- ❍ Definitely yes
- ❍ Probably yes
- ❍ Probably not
- ❍ Definitely not
- ❍ I don't know

If you participate in a Kaggle contest, which ranking is more important to you: the public leaderboard ranking during the contest or the final ranking after the contest is over?

- ❍ Public leaderboard ranking during the contest is more important
- ❍ The two rankings are equally important
- ❍ Final ranking after the contest is over is more important
- ❍ Neither ranking matters to me

If you have participated in Kaggle contests in the past, in retrospect, which solution is more likely to have a high private score, the solution selected based on the public scores or the solution selected based on cross-validation?

- ❍ The solutions selected based on public scores
- ❍ The solutions selected based on cross-validation
- ❍ I don't know
- ❍ I have not participated in any Kaggle contest.

If you have participated in Kaggle contests in the past, in retrospect, which methods would you have emphasized more to improve your performance in those Kaggle contests? (Check all that apply)

- ❍ Emphasize more on cross-validations (e.g. K-fold cross-validation, leave-p-out cross-validation, repeated random sub-sampling validation, forward chaining or similar methods)
- ❍ Emphasize more on public scores
- ❍ Others (Please specify) _____
- ❍ I have not participated in any Kaggle contest.