Peer Evaluation and Team Performance: An Experiment on Complex Problem Solving

John Morgan[a], Susanne Neckermann[b] and Dana Sisak[c]

[a]University of California, Berkeley

[b]University of Chicago & ZEW

[c]Erasmus University Rotterdam
& Tinbergen Institute

UIB, Palma de Mallorca, March 3rd, 2017

Motivation

- Organizations succeed when they are capable of solving complex, non-routine problems.
- Often, these tasks are done by teams of individuals, usually after the individuals alone have had a chance to think through the issues and possibilities.
- The interplay of incentives and performance on complex choices is not well understood, neither theoretically nor empirically.
- In particular, an objective measure of performance is often not available, and thus less-studied incentives relying on subjective evaluation are needed.
- We study incentives for individual and group performance in a novel complex and non-routine task: guesstimations

Preview of Results I: Unraveling Group Decisions

Sequential design: First subjects work individually, then decide on final answer in group.

Results:

- Compared to the median quality guess, the group significantly improved.
- Compared to the best individual guess, the group performs significantly worse.
- Groups outperform "mechanical" ways of aggregating individual answers.
- Groups especially valuable when individual answers do not "straddle" truth.
- Individual characteristics not predictive of group outcomes in our setting.

Treatments: add group and individual incentives

- Group piece rate by closeness to truth
- Payoff relevant peer evaluation
    - Each individual votes for most valuable group member.
    - The vote is made on the basis of perceived performance (no performance feedback).
    - The winner received biggest share of group surplus.
    - **Pro:** May help mitigate the free rider problem by introducing individual incentives.
    - **Con:** May encourage showing off, sabotage and other performance reducing behaviors.

Result: Treatments affected group atmosphere but performance was not significantly affected.

Literature

We consider a complex and non-routine task

- Groups vs. individuals (Blinder and Morgan 2005, 2008; Laughlin *et al.* 2006; Charness *et al.* 2015; Sniezek 1989, Thompson and Wilson 2015)

- Optimal group composition (Hoogendoorn and van Praag 2012; Barrick *et al.* 1998; Williams and O'Reilly 1998; Bell 2007; Hamilton *et al.* 2003)

- Group incentive schemes (Charness and Grieco, 2014; Ramm *et al.*, 2013)

- Individual incentives for groups (Bracha and Fershtman, 2013; Honeywell-Johnson and Dickinson, 1999)

Our contributions: (1) novel, complex task (2) study of peer evaluation (3) understanding a complex group production process

The experiment was

- conducted at the Erasmus University Rotterdam
- in May and June 2014
- with a total of 231 students
- for three treatments (93, 78 and 60)

Total duration: approx. 1.5 - 2 h

1. Briefing in plenum & individual task
2. Group task times three (groups of three, separate rooms)
3. Elicitation of social preferences, personality and demographics as well as questionnaire in plenum
4. Payment

First Stage

Duration: approx. 30 min

- Briefing in plenum, example of guesstimation and distribution of fact sheet/calculators.
- Subjects solve a guesstimation problem individually (ability proxy, 7 minutes, incentivized)

Ezafus

Group Stage with Treatment

Duration: approx. 1 h

- Groups of 3 enter into group room.
- At individual tables subjects spend 5 minutes coming up with a first answer which they note on their individual answer sheet (non-incentivized).
- Subjects have an additional 10 minutes to agree on a group answer at the group table which is submitted on the group answer sheet (incentivized).
- Subjects are re-matched twice for in total 3 guesstimations.

Ezafung

Guesstimations

Used in assessment centers and resemble tasks for example in consulting jobs.

- **Individual "Ability"** How many dogs are there in the United States of America? (A: 73.4 million)
- **Group 1 "Toothpaste"** How many liters of toothpaste are used in the United Kingdom every year? (A: 46.3 million liters)
- **Group 2 "Weddings"** How many weddings were there in Germany in June 2006? (A: 49 500)
- **Group 3 "Cycling"** What is the total distance cycled in Amsterdam per day? (A: 2 million km)

Advantage: guesstimations have definite, known, answers. Possible to grade performance in an objective fashion.

Example of Guesstimation answer sheet

Grading and Group Reward Scheme

- The maximum reward is 10 Euro for the individual guesstimation and 35 Euro per group for the group guesstimations.
- We implement a piece rate by closeness to right answer.
- The piece rate group reward is then split amongst group members according to treatment rules.

Group score	Construction
0	Guesstimation is more than $+/-$ 80% of the true answer
0.2	Guesstimation is within $+/-$ 80% of the true answer
0.4	Guesstimation is within $+/-$ 60% of the true answer
0.6	Guesstimation is within $+/-$ 40% of the true answer
0.8	Guesstimation is within $+/-$ 20% of the true answer
1	Guesstimation is within $+/-$ 10% of the true answer

Individual Incentives and Treatments

- **FLAT:** No incentive, just a flat rate per question.
- **EQUAL:** Group piece rate by closeness to "truth". Exogenous group sharing rule. Total payment is randomly allocated in shares of 50%, 30% and 20%.
- **MVP:** Endogenous sharing through peer evaluation. Subject voted best by both team members receives 50%, subject who receives one vote 30% and subject with no vote 20%. Ties are broken randomly. Note that subjects are not informed about their performance at the time of the evaluation, and thus evaluation is subjective.

Duration: approx. 25 minutes

- Questionnaire about group atmosphere
- Social preferences: Social value orientation (Murphy, Ackermann and Handgraaf, 2011)
- Personality: Big 5 index (BFI-10)
- Basic demographics

Erafins

Hypotheses

Hypothesis 1: Group effort improves the quality of guesses. In all treatments, the quality of individual guesses should be lower than the quality of the group guess.

Hypothesis 2: Incentives increase group performance. FLAT $<$ EQUAL, FLAT $<$ MVP.

Hypothesis 3: Individual incentives increase individual performance. When measuring individual performance prior to each group choice, FLAT $<$ EQUAL $<$ MVP.

Ezafing

Overview Analysis

- Preliminaries
- Part I: Unraveling Group Decisions
  - Groups vs. Individuals / Aggregates
  - Straddle vs. Non-Straddle
- Part II: Treatment Effects
  - Group Atmosphere
  - Performance

Summary Statistics Guesses by Question

	Cycling (in 10,000)	Toothpaste (in 1,000,000)	Weddings (in 1,000)
# Observations	275	267	270
Mean	**639.88**	**1,154.15**	**1,175.21**
Maximum	39,647.06	242,027.00	77,620.33
Minimum	3.75	0.00	0.05
Standard deviation	2,661.11	14,868.36	6,256.38
1st Percentile	6.00	0.00	1.10
5th Percentile	37.64	0.40	4.00
10th Percentile	54.00	6.37	12.75
25th Percentile	109.55	30.13	33.75
50th Percentile	**191.00**	**64.19**	**71.74**
75th Percentile	420.00	130.00	300.00
90th Percentile	966.00	348.99	1,257.48
95th Percentile	1,814.86	640.00	2,828.00
99th Percentile	9,175.52	5,344.09	29,840.88
True Answer	**200.00**	**46.30**	**49.50**

Note: Includes all group and individual guesses (77 group guesses and the rest individual guesses). Since not all individuals always made an individual guess, the number of observations are lower than 308 per question.

Summary Statistics and Balance Table I

	FLAT	EQUAL	MVP
Observations	60	93	78
Demographics			
Female	0.367	0.337	0.436
	(0.482)	(0.473)	(0.496)
Age	21.333	21.391	21.064
	(2.370)	(2.643)	(2.457)
Dutch	0.700	0.685	0.782
	(0.458)	(0.465)	(0.413)
Economics Student	0.767	0.685	0.731
	(0.423)	(0.465)	(0.444)
Econometrics Student	0.117	0.065	0.064
	(0.321)	(0.247)	(0.245)
Bachelor 1	0.283	0.239	0.295
	(0.451)	(0.427)	(0.456)
Bachelor 2	0.150	0.228	0.231
	(0.357)	(0.420)	(0.421)
Bachelor 3	0.333	0.304	0.218
	(0.471)	(0.460)	(0.413)
Master	0.233	0.228	0.256
	(0.423)	(0.420)	(0.437)
Previous Experience Task	0.117	0.097	0.128
	(0.321)	(0.296)	(0.334)
Number of Quantitative Classes	5.167	3.651*	3.944
	(6.282)	(2.800)	(2.761)
Average Grade	7.142	7.182	7.096
	(0.725)	(0.762)	(0.780)

Summary Statistics and Balance Table II

	FLAT	EQUAL	MVP
Social Value Orientation			
Individual/Competitive	0.458	0.391	0.434
	(0.498)	(0.488)	(0.496)
Big 5 Inventory			
Extraversion	6.983	6.793	7.192
	(1.396)	(1.757)	(1.721)
Agreeableness	7.200	7.293	7.333
	(1.527)	(1.441)	(1.345)
Conscientiousness	7.169	7.478	7.295
	(1.544)	(1.593)	(1.691)
Neuroticism	4.700	5.118	4.872
	(2.011)	(2.141)	(2.134)
Openness to Experience	6.417	6.882[*]	6.923[*]
	(1.544)	(1.621)	(1.673)
Ability (Dog Question)			
Ability	0.347	0.324	0.318
	(0.329)	(0.299)	(0.323)

Erafus

Overview Analysis

- Preliminaries
- **Part I: Unraveling Group Decisions**
  - Groups vs. Individuals / Aggregates
  - Straddle vs. Non-Straddle
- Part II: Treatment Effects
  - Group Atmosphere
  - Performance

Performance Measures

- Percentage error:

$$\text{P.E.} = \frac{|\text{Guess} - \text{Truth}|}{\text{Truth}}$$

  Note: Smaller numbers mean better performance.
- (Hypothetical) payoffs

Ezafıng

Groups vs. Individuals

Group performs...	Best individual	Second best individual	Random individual
... better than	30% (62)	70% (147)	65% (135.67)
... same as	6% (13)	5% (10)	4% (8.67)
... worse than	64% (134)	25% (52)	31% (64.67)

Note: Sample includes all groups where at least two people made an individual guess (209 observations). The probablities of the group guess being different than the individual guesses are $p < 0.000$, $p < 0.000$ and $p < 0.000$.

Groups vs. Individuals: Percentage Error

	Cycling		Toothpaste		Weddings		Pooled	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Group								
Percentage error	1.005	1.560	1.326	2.440	2.657	6.013	1.661	3.903
Individual								
Best p.e.	0.532**	1.004	0.673**	0.890	1.411**	4.669	0.871***	2.824
Second best p.e	1.502**	2.101	2.094**	3.041	8.813***	14.956	4.140***	9.473
Random indiv. p.e	1.484**	1.486	2.357***	2.179	8.336***	10.143	4.059***	6.766
# Observations	72		67		70		209	

Note: Only groups included of which at leas two individuals made a guess. Having made a guess is ranked better than no guess. Percentage error is the 90% winsorised absolute percentage error of the guess compared to the true answer. Significance stars are added to the individual percentage error, indicating the p-value of a paired t-test for the two sided hypothesis: *group percentage error $\neq$ individual percentage error*. Significance levels are denoted as follows : * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Groups vs. Individuals: Payoffs

	Cycling		Toothpaste		Weddings		Pooled	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Group								
Payoff	4.00	3.167	3.73	3.301	3.60	3.661	3.78	3.370
Individual								
Best Payoff	5.14***	3.251	4.69**	3.403	5.00***	3.583	4.95***	3.403
Second Best Payoff	2.64***	2.703	1.25***	1.995	1.49***	2.376	1.81***	2.450
Random Payoff	3.09**	2.281	2.30***	1.877	2.55***	2.123	2.65***	2.121
# Observations	72		67		70		209	

Note: Only groups where at least two individuals made a guess are included. Having made a guess is ranked better than no guess. Significance stars are added to the individual payoffs, indicating the p-value of a paired t-test for the two sided hypothesis: *group payoff ≠ individual payoff*. Significance levels are denoted as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Result 1: Groups in most cases do not improve on the best individual guess but do outperform the second best individual guess as well as a randomly chosen one. Thus group effort is valuable.

Ezafino

Groups vs. Statistical Aggregates: Percentage Error

	Cycling		Toothpaste		Weddings		Pooled	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Group								
Percentage error	1.005	1.560	1.326	2.440	2.657	6.013	1.661	3.903
Measure of Central Tendency								
Arithmetic Mean	1.607**	2.243	2.539***	3.355	12.142***	18.353	5.434***	11.829
Geometric Mean	0.920	1.328	1.666	2.708	4.479***	7.310	2.351***	4.800
Location. Median	0.808	1.215	1.517	2.647	4.887**	8.185	2.401**	5.307
# Observations	72		67		70		209	

Note: Only includes groups of which at least 2 individuals made a guess. Significance stars are added to mean and median errors, indicating the p-value of a paired t-test with the hypothesis of equal means compared with group error. Significance levels are denoted as follows : * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Groups vs. Statistical Aggregates: Payoffs

	Cycling		Toothpaste		Weddings		Pooled	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Group								
Payoff	4.00	3.167	3.73	3.301	3.60	3.661	3.78	3.370
Measure of								
Central Tendency								
Arithmetic Mean Payoff	3.36	3.453	2.36***	3.132	2.03***	3.292	2.59***	3.333
Geometric Mean Payoff	4.14	3.345	2.87*	3.397	2.43**	3.553	3.16**	3.494
Locational Median Payoff	4.22	3.383	2.90**	3.196	2.77*	3.410	3.31*	3.383
# Observations	72		67		70		209	

Note: Only groups where at least two individuals made a guess are included. Arithmetic Mean Payoff is calculated as the payoff corresponding to the arithmetic mean of the individual guesses. Geometric Mean Payoff is calculated as the payoff corresponding to the geometric mean of the individual guesses. Locational Median Payoff is calculated as the payoff corresponding to the second highest guess if there were three guesses or the geometric mean of the guesses if there were two. Significance stars are added to the mean payoffs, indicating the p-value of a paired t-test for the two sided hypothesis: *group payoff $\neq$ mean payoff*. Significance levels are denoted as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Individual characteristics may lead to a disproportionate influence on the group guess:

- Best guess vs. median sized guess
- "Leader guess" vs. most individual steps
- Most extrovert vs. most agreeable vs. most "able"

Estimation of influence weight of characteristic k:

$$\log[G] = \frac{a}{N}(\sum_{n=1}^{N} \log[G_n]) + \sum_{k=1}^{K} c_k \log[G^k] + \epsilon$$

Expect "weights" to sum to one if model fit is good.

Group Decision Making and Individual Characteristics

	(1)	(2)	(3)
Geo. mean ind. guesses	0.629***	0.904***	1.003***
	(0.240)	(0.161)	(0.178)
Best ind. guess	0.244		
	(0.171)		
Median sized ind. guess	0.127		
	(0.339)		
"Leader" guess		-0.062	
		(0.105)	
Max ind. steps guess		0.157	
		(0.111)	
Max extrovert guess			0.010
			(0.103)
Max agreeable guess			-0.007
			(0.125)
Max ability guess			-0.006
			(0.021)
Observations	209	173	185
Clusters	15	15	15
Sum Coefs	1.000	1.000	1.000
R^2	0.811	0.857	0.827

Note: The table reports OLS estimates of the log group guess on the log geometric mean of the individual guesses. Sample consists of groups with at least two individual guesses. The coefficients have been constrained so that they sum to one. Robust standard errors clustered by session are reported in parentheses. Asterisks indicate significance at the 10/5/1 percent level.

Result 2: Groups perform (somewhat) better than statistical aggregates of individual guesses. Individual characteristics do not help predict the weight of an individual guess in the determination of the group guess.

- Measures of central tendency perform well if individual guesses bracket, or "straddle" the truth.
- In these cases group members only need to "compromise".
- When all group members under- or overestimated they need to move beyond the convex combination of their individual guesses.
- If groups are able to do this, this will not be well captured by a measure of central tendency.
- In these situations the group phase is most valuable.

Group vs. Statistical Aggregates: Straddle vs. Non-Straddle

	Arithmetic Mean	Geometric Mean	Locational Median
Groups outperform...			
Straddle	51%	41%	48%
Non-Straddle	74%	77%	64%

Note: Sample includes all groups where at least two people made an individual guess (209 observations).

Relation of Individual Guesses Relative to "Truth"

Distribution of the Individual Guesses in a Group relative to the Truth
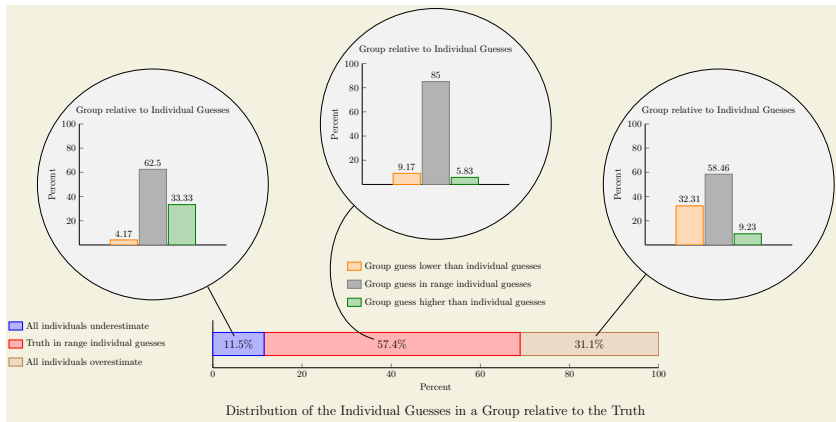
Determinants of the Group Guess (Straddle)

	(1)	(2)	(3)
Geo. mean indiv. guesses	0.812***	0.769***	1.073***
	(0.252)	(0.137)	(0.186)
Best indiv. guess	0.064		
	(0.200)		
Median sized indiv. guess	0.124		
	(0.308)		
Leader guess		0.068	
		(0.075)	
Max indiv. steps guess		0.163*	
		(0.093)	
Max extrovert guess			-0.026
			(0.104)
Max agreeable guess			-0.035
			(0.133)
Max ability guess			-0.012
			(0.033)
Observations	120	98	105
Clusters	15	15	15
Sum Coefs	1.000	1.000	1.000
R^2	0.752	0.827	0.778

Note: The table reports OLS estimates of the log group guess on the log geometric mean of the individual guesses. Sample consists of groups with at least two individual guesses which straddle the truth. The coefficients have been constrained so that they sum to one. Robust standard errors clustered by session are reported in parentheses. Asterisks indicate significance at the 10/5/1 percent level.

Determinants of the Group Guess (Non-Straddle)

	(1)	(2)	(3)
Geo. mean indiv. guesses	0.361	1.217***	0.517
	(0.232)	(0.247)	(0.415)
Best indiv. guess	0.725***		
	(0.122)		
Median sized indiv. guess	-0.086		
	(0.260)		
Leader guess		-0.391*	
		(0.200)	
Max indiv. steps guess		0.174	
		(0.240)	
Max extrovert guess			0.291
			(0.286)
Max agreeable guess			0.202
			(0.250)
Max ability guess			-0.009
			(0.024)
Observations	89	75	80
Clusters	15	15	15
Sum Coefs	1.000	1.000	1.000
R^2	0.918	0.901	0.879

Note: The table reports OLS estimates of the log group guess on the log geometric mean of the individual guesses. Sample consists of groups with at least two individual guesses which do not straddle the truth. The coefficients have been constrained so that they sum to one. Robust standard errors clustered by session are reported in parentheses. Asterisks indicate significance at the 10/5/1 percent level.

Result 3: Group effort is most valuable when individual guesses do not straddle the truth.

Ezafino

- Groups are worse than the best individual of the group on average.
- Groups outperform a random individual as well as statistical aggregates of individual guesses.
- In general individual characteristics do not help predict the weight of an individual guess in the determination of the group guess.
- Groups are especially valuable when individual guesses do not "straddle" the truth.

Erafung

Overview Analysis

- Preliminaries
- Part I: Unraveling Group Decisions
  - Groups vs. Individuals / Aggregates
  - Straddle vs. Non-Straddle
- **Part II: Treatment Effects**
  - Group Atmosphere
  - Performance

Perception of Atmosphere by Treatment
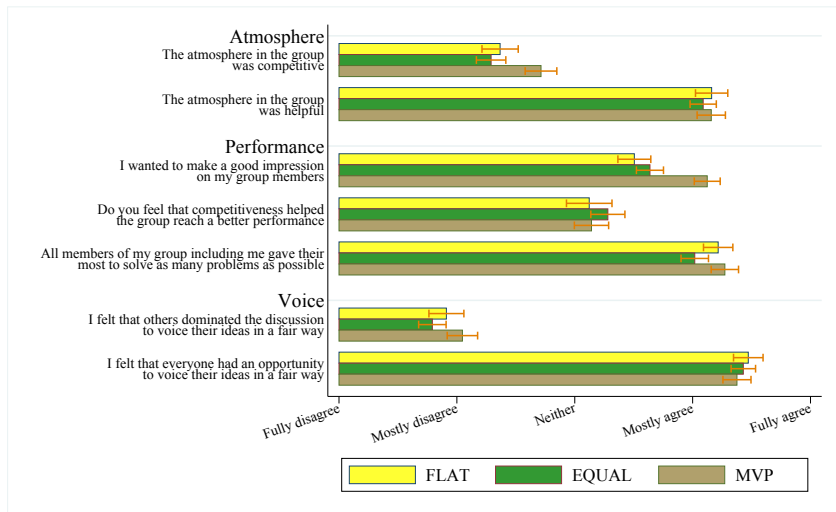
Effects of Treatments on Group Performance

	Percentage Error		Payoff	
	(1)	(2)	(3)	(4)
Group Incentives	0.495	0.535	0.232	0.323
	(0.574)	(0.565)	(0.685)	(0.720)
Group Incentives x Individual Incentives	-0.002	-0.126	0.070	0.007
	(0.830)	(0.522)	(0.401)	(0.486)
Additional Covariates		Yes		Yes
Observations	231	231	231	231
Clusters	15	15	15	15

Note: The table reports OLS estimates of the impact of incentives and performance rewards on various performance measures. The percentage error columns use windsorized errors. Robust standard errors clustered by session are reported in parentheses. Additional covariates consist of the mean group ability score, a measure for group SVO, a measure of group gender, the average age of group member, group same nationality, percentage first and second year bachelor students, number of people with experience with the task, and average grade. Asterisks indicate significance at the 10/5/1 percent level.

Effects of Treatments on Individual Performance

- We find no support for Hypothesis 2: (group) incentives did not increase group performance significantly.
- To test Hypothesis 3, we need to consider individual performance.
- Comparing EQUAL to MVP might have made individuals work harder, but at the same time lead to more grandstanding in the group phase, canceling out the positive effect at the individual phase.

Effects of Treatments on Individual Performance

	Percentage Error		Payoff	
	(1)	(2)	(3)	(4)
Group Incentives	1.129	1.086	0.505	0.525
	(0.842)	(0.840)	(0.484)	(0.501)
Group Incentives x Individual Incentives	-1.286**	-1.339**	-0.461	-0.464
	(0.554)	(0.472)	(0.300)	(0.315)
Additional covariates		Yes		Yes
Observations	581	578	581	578
Clusters	15	15	15	15

Note: The table reports OLS estimates of the impact of incentives and performance rewards on various performance measures. The dependent variable of columns (1) and (2) is the windsorized percentage error while the dependent variable of columns (3) and (4) is the group payoff. Robust standard errors clustered by session are reported in parentheses. Additional covariates consist of the mean group ability score, a measure for group SVO, a measure of group gender, the average age of group member, group same nationality, percentage first and second year bachelor students, number of people with experience with the task, and average grade. Asterisks indicate significance at the 10/5/1 percent level.

Result 4:

- Subjects did not perceive a difference between FLAT and EQUAL. They did perceive EQUAL and MVP differently.
- Neither group nor individual performance are significantly affected by group or individual incentives.

Ezafus

MVP Treatment: Is Individual Performance Rewarded?

- Why did incentives not affect performance?
- MVP treatment: look at individual voting behavior.
- If good individual performance is not rewarded, incentives are *de facto* absent.
- Best individual was only voted winner in 20.5% of cases.
- In 15.4% of cases, a tie was the group outcome.

$\Rightarrow$ Individual performance is not sufficiently rewarded.

MVP Treatment: Voter Characteristics

	(1)	(2)
Vote Strategically	-0.229	-0.136
	(0.251)	(0.259)
Reward Quality	-0.300	-0.259
	(0.284)	(0.287)
Vote Strategically * Reward Quality	0.164	0.043
	(0.602)	(0.627)
Gender		0.052
		(0.175)
Best		0.230
		(0.197)
Leader		0.303*
		(0.182)
Missing Guess		-0.342
		(0.217)
_cons	0.186*	0.052
	(0.095)	(0.167)
N	234	234

Note: This table presents probit estimates of the impact of voter characteristics on whether an individual voted for their best team member (the individual who had the lowest percent error) in the MVP treatment. Asterisks indicate significance at the 10/5/1 % levels. Standard errors are reported in parenthesis. Groups where neither individual submitted an individual guess were excluded.

MVP Treatment: Candidate Characteristics

	(1)	(2)
Worst Ind not Missing Guess	-0.226	-0.229
	(0.168)	(0.173)
Best Teammate Older		0.060
		(0.235)
Best Teammate more Extroverted		0.166
		(0.250)
Best Teammate Female		0.559
		(0.371)
Best Teammate Leader		0.222
		(0.277)
Worst Teammate Female		-0.750**
		(0.362)
Worst Teammate Leader		-0.709**
		(0.279)
_cons	0.253*	0.359**
	(0.130)	(0.153)
N	234	234

Note: This table presents probit estimates of the impact of candidate characteristics on whether an individual voted for their best group member (the individual who had the lowest percentage error in their individual guess) in the MVP treatment. Asterisks indicate significance at the 10/5/1 % levels. Standard errors are reported in parenthesis. Groups where neither individual submitted an individual guess were excluded

- Individual performance is not sufficiently rewarded (at least as much as we can observe it) and thus individual incentives may not be effective.
- We do not find any systematic voter or candidate characteristics that predict merit based voting decision.

Ezafing

Discussion of Results

- Our results validate firms use of groups to arrive at decisions collectively.
- They also validate the decision not to compensate on the basis of solution quality.
- But why don't incentives matter?
  - Incentives were too small to make a difference?
  - Individuals cannot regulate "creative" effort in the way that they might do for routine, non-complex tasks?
    - Insight into complex, unique problems largely outside the conscious control of the individual.
    - Incentives, good or bad, have no effect as conscious effort is only loosely tied to performance.
    - While spark of insight may be outside the control of an individual, working in groups creates tinder that makes such sparks more likely.

Thank you!