Empirics of Relational Contracts

Labour Economics, LSE PhD

John Van Reenen, Spring 2020

Overview

1. **Relational contracts revision**
2. Macchiavello & Morjaria (2015) – Roses in Kenya
3. Blader, Gartenberg & Prat (2016) – Truckers in US
4. Conclusions

Relational contracts

- Many reasons why formal contracts could fail (incompleteness)
- Relational contracts better: supported through threat to renege in repeated game (e.g. “trust”)
- Much theory – Examples: Bull (1987, QJE); Levin (2003, AER) Baker. Gibbons & Murphy (1994, QJE), Malcolmson (2013)
- Case study evidence in Gibbons & Henderson (2013)
- But little systematic quantitative evidence
- Consider 2 recent papers

Overview

1. Relational contracts revision

2. **Macchiavello & Morjaria – Roses in Kenya**

3. Blader, Gartenberg & Prat– Truckers in US

4. Conclusions

Macchiavello & Morjaria (2015, AER)

- Importance of relational contracts with learning about seller type
- Key issue is seller **reputation**
- **Context**: Start with population of rose exports. Major industry for Kenya
  - 80% of all cut flower exports are roses

Data & Design

- Export data is of high quality – can observe price and quantity for 56 local sellers to 71 foreign buyers. 189 "**Direct Relationships**" between 2004-2008
- In addition to direct buyer-seller relationship, seller can also sell on spot market. Netherlands auctions. Useful as gives measurable source of Incentive Compatibility Constraint to construct a lower bound to value of relationship
- Dec 2007 heavily contested Kenyan election. An exogenous supplier shock: as violence erupts
- Focused in Rift Valley & Western Provinces

TABLE 1—DESCRIPTIVE STATISTICS, DIRECT RELATIONSHIPS

Variable	Observations	Mean	SD	Min	Max
Panel A. Relationship characteristics					
Number of transactions	189	60.52	35.70	20.00	140
Number of stems per week (in 1000s)	189	100.24	161.85	0.81	926.49
Average FOB price (Euro cents per stem)	189	11.37	8.37	1.14	58.46
Age (in days)	189	830.27	469.54	24.50	1286
Number of previous transactions	189	309.88	304.51	12.00	1204
Left censored (Yes = 1, No = 0)	189	0.40	0.49	0	1
Panel B. Number of relationships per buyer and seller					
Number of relationships per seller	56	3.38	2.88	1	14
Number of relationships per buyer	71	2.66	2.83	1	14
Panel C. Estimated relationship values (season before the violence)					
Estimated S (/ average weekly revenues)	157	3.84	3.65	0.75	30.72
Estimated U (/ average weekly revenues)	157	2.70	1.77	0.77	10.64
Estimated V (/ average weekly revenues)	157	1.61	3.28	0.00	25.50

Notes: The sample is given by all relationships that had at least 20 transactions in the 20 weeks immediately before the violence. Left censored refers to relationships that were already active before August 2004. Estimated S, U, and V are lower bounds to the value of the relationship as a whole, to the buyer, and to the seller respectively. Details on the computation of S, U, and V are given in Section IIIA.

Source: Authors' calculations from customs records. See online Appendix A for data sources.

Direct Relationships have less variation in prices

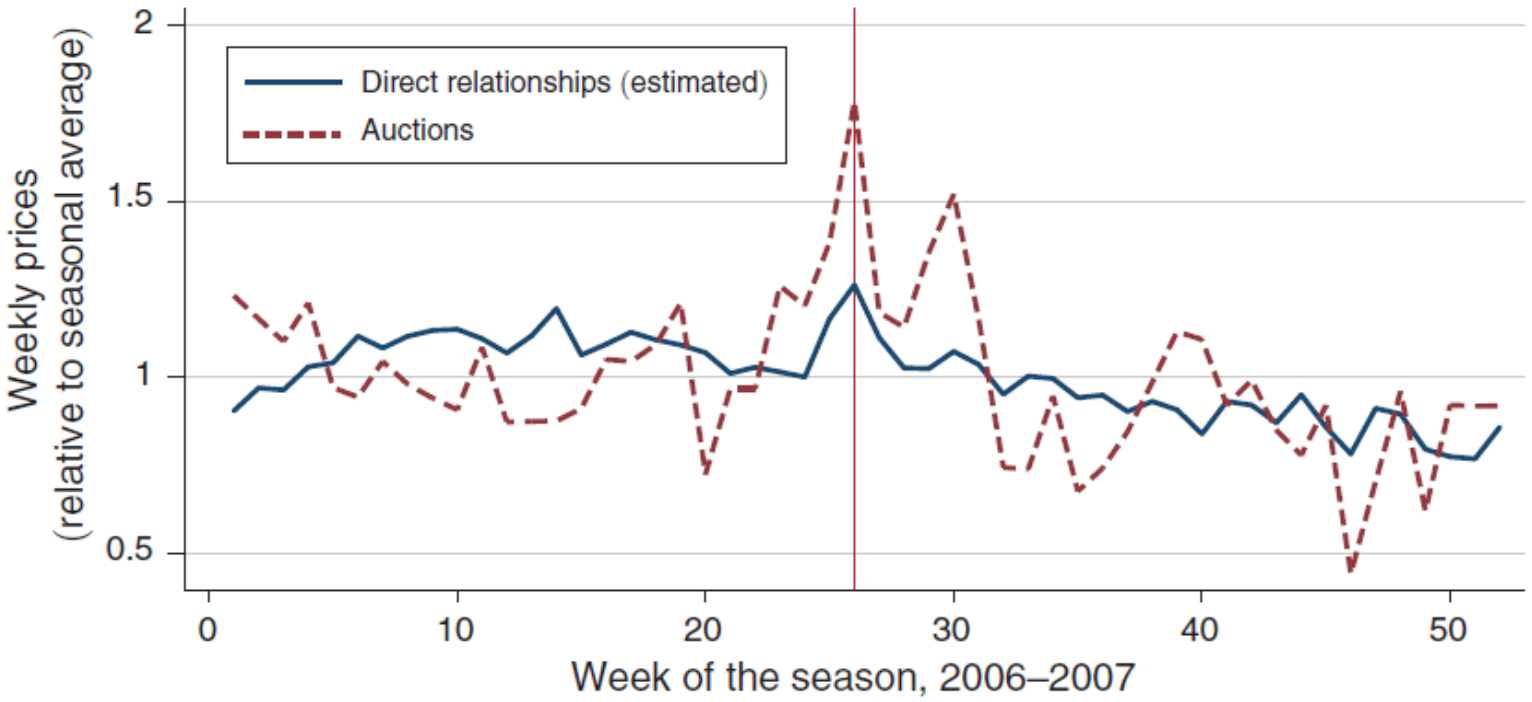


FIGURE 2. FLUCTUATIONS IN PRICES: DIRECT RELATIONSHIPS VERSUS AUCTION

Contracts

- **Contact:** seller agrees to supply quantity (q^r) of roses at a stable price. Buyer agrees to purchase these at this fixed price
- But formal contract hard to enforce in courts
- Incentives to renege
  - **Seller** could sell on spot market when price is high (and simply not deliver what they have promised in contract)
  - **Buyer** could refuse to pay or purchase usual amount when spot price is low (less of problem as these are large MNEs)
- So incomplete contracts need to enforced through relational contracts

Theory

- Looks at several theories, but focus on relational contracts
  - No contract enforcement
  - Extension with foreign buyer's uncertainty over seller quality (uncertainty falls with age of relationship)
  - Extension to examine sellers' reaction to the violence

Predictions

1. If lack of enforcement constrains volume an unexpected increase in spot price should lead to a fall in transaction quantity of roses (elasticity = -1), not an ending of contract

    - Incentive compatibility constraint binds: lack of enforcement constrains rose quantity

2. **Lack of enforcement + learning about sellers' type:** imply relationship value increases with age of the relationship.

    - Consistent with learning over supplier quality

3. **Lack of enforcement + learning about sellers' type:** In violence period suppliers can respond by increasing effort (e.g. replacing lost workers). But this effort is inverse U-shaped in relationship age because:

    - For very young relationships, not worth putting in effort. But as relationship lengthens, effort increases to signal supplier quality
    - But for very long lived relationships don't need to put in effort as supplier types already revealed

Predictions

- For each relationship *i* in season *t*, ω^*_{it} is the week of the season which has highest spot market price $p_{i,t\omega}$ for roses (maximum temptation to renege)

$$\omega^*_{it} = \arg\max_{\omega} \left\{ q^R_{i,t\omega} \times p_{i,t\omega} \right\}.$$

- Lower bound to the value of the relationship *i* in season *t*

$$\hat{S}_{it} = q^R_{i,t\omega^*_{it}} \times p_{i,t\omega^*_{it}}.$$

- This varies across seasons (T=3) and relationships (N=189). Sources of variation in relationship value
  - Week of season where Incentive Constraint binds
  - Price in this week
  - Quantity in this week

When do you think spot price for Kenyan Roses peaks?

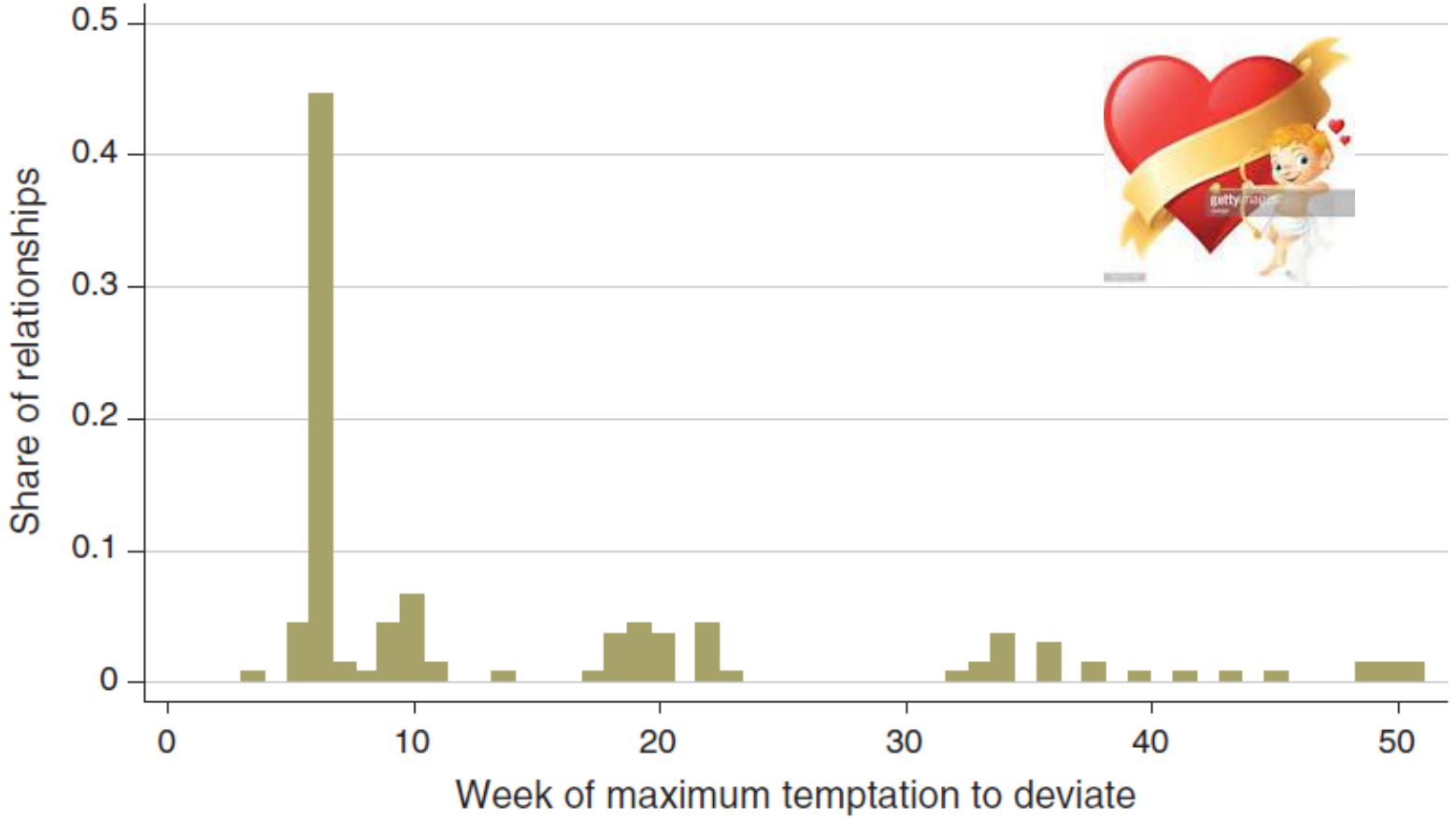
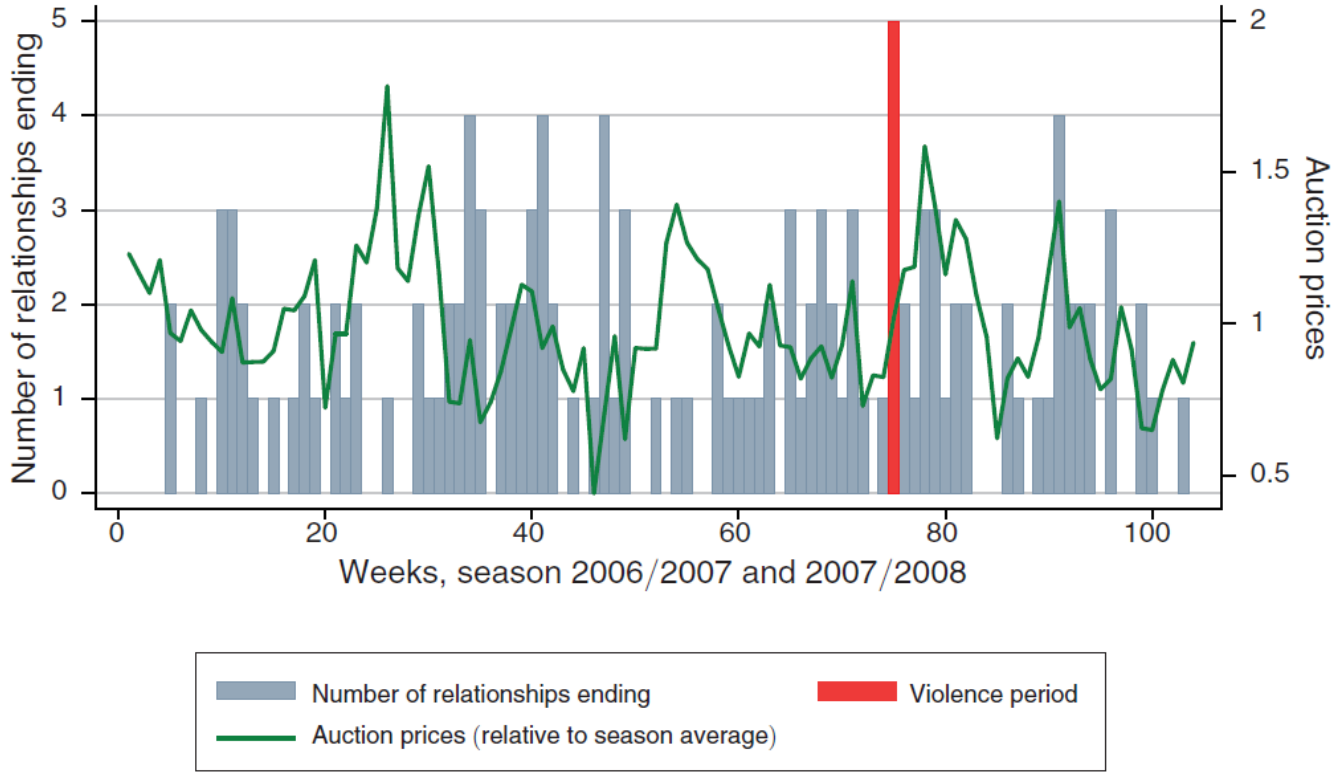


FIGURE 3. DISTRIBUTION OF HIGHEST TEMPTATION WEEKS

- The modal week of maximum temptation is the week **Valentine's Day** falls

Relationships do not disproportionately end when prices are high (suggests constrained optimal contracting)



Test 1: Relationship Volume falls when spot prices unexpectedly spike

TABLE 2—BINDING INCENTIVE COMPATIBILITY CONSTRAINT (TEST 1)

	Trade volume (1)	Relationship value (2)	Price (3)
Price at auction (ln)	−0.936**	0.064	0.313
	(0.371)	(0.371)	(0.193)
Relationship fixed effects	Yes	Yes	Yes
Season fixed effects	Yes	Yes	Yes
Seasonality fixed effects	Yes	Yes	Yes
Adjusted R^2	0.861	0.867	0.606
Observations	430	430	430

Notes: The table reports correlations between prices at the auctions and relationship outcomes at the time of the maximum temptation to deviate. A binding incentive compatibility constraint implies a minus one elasticity of quantity of roses traded in the relationship with respect to auction prices at the time of the largest temptation to deviate (Test 1). All variables are in logs. The outcomes are computed for all seasons before the violence and the sample refers to relationships that were active during the period. The sample excludes relationships that are in the baseline sample but were not active in the season preceding the violence and includes relationships that did not survive until the violence season. Following business practices in the industry, a season starts in mid-August. The three considered seasons are those starting in August 2004, 2005, and 2006. Seasonality fixed effects are dummies for the week of the season in which the maximum aggregate temptation to deviate occurs. Robust standard errors, clustered at the firm level are reported in parenthesis.

- Expectations controlled for by season & seasonality dummies
- Cf. Rotemberg & Saloner (1983): temptation to renege ↑ in booms

Shipments & length of relationship

- *AGE*: How many shipments have been made between the parties in the relationship in the past
- Intuitive but:
  - Mixes up age with selection (include relationship fixed effects on balanced panel to try and deal with this)
  - Mixes up length of time with “size”
- Panel regression (even columns of Table 4) of transaction volume when IC binds (week of highest prices) include these relationship fixed effects

$$\log(\hat{y}_{it}) = \mu_i + \phi_t + \beta \log(AGE_{it}) + \varepsilon_{it},$$

Test 2: More value in longer relationships

TABLE 4—AGE OF THE RELATIONSHIPS AND OUTCOMES (TEST 2)

	Trade volume		Relationship value	
	(1)	(2)	(3)	(4)
Panel A. Age of the relationship (log)				
Age of the relationship	0.675***	0.561***	0.818***	0.617***
	(0.105)	(0.095)	(0.147)	(0.102)
Adjusted R^2	0.779	0.824	0.760	0.854
Panel B. Age of the relationship (level)				
Age of the relationship	0.198***		0.239***	
	(0.052)		(0.071)	
Adjusted R^2	0.741		0.708	
Seller and buyer fixed effects	Yes	—	Yes	—
Relationship fixed effects	No	Yes	No	Yes
Season fixed effects	No	Yes	No	Yes
Observations	156	430	156	430

Notes: The table reports correlations between the relationship outcomes and the age of the relationship. The pure limited enforcement model predicts zero correlation between relationship outcomes and age of the relationship while the learning model predicts positive correlation (Test 2). Age of the relationship is measured as the number of previous shipment in the relationship. In panel A the age is in logs and in panel B it is in levels (hundreds of past transactions). The outcomes are computed for all seasons before the violence and the sample refers to relationships that were active during the period. The sample excludes relationships that are in the baseline sample but were not active in the season preceding the violence and includes relationships that did not survive until the violence season. Following business practices in the industry, a season starts in mid-August. The three considered seasons are those starting in August 2004, 2005, and 2006. Robust standard errors, clustered at the firm level are reported in parenthesis.

Fixed effects control for *some* types of selection

“Reliability”

- % roses shipped in violence areas vs. same period in pre-violence period in the same relationship

$$\hat{\mathcal{R}}_{sb} = \frac{y_{sb}}{y_{sb}^{0}}.$$

- Unsurprisingly, reliability fell (by about 17%) after violence (Tab 6 column (2) regression). Buyer fixed effects.

$$\hat{\mathcal{R}}_{sb} = \alpha_b + \beta\left(\mathrm{I}_s^{C=1}\right) + \gamma\mathbf{Z}_{sb} + \eta\mathbf{X}_s + \varepsilon_{sb},$$

- Firms in violence-affected regions lost an average of 50% of their workers

Interpretation: For few transactions, little effort because relationship unlikely to last. For loads of transactions seller doesn't need to invest in effort to signal "good type" & maintain relationship. Sweet spot in middle

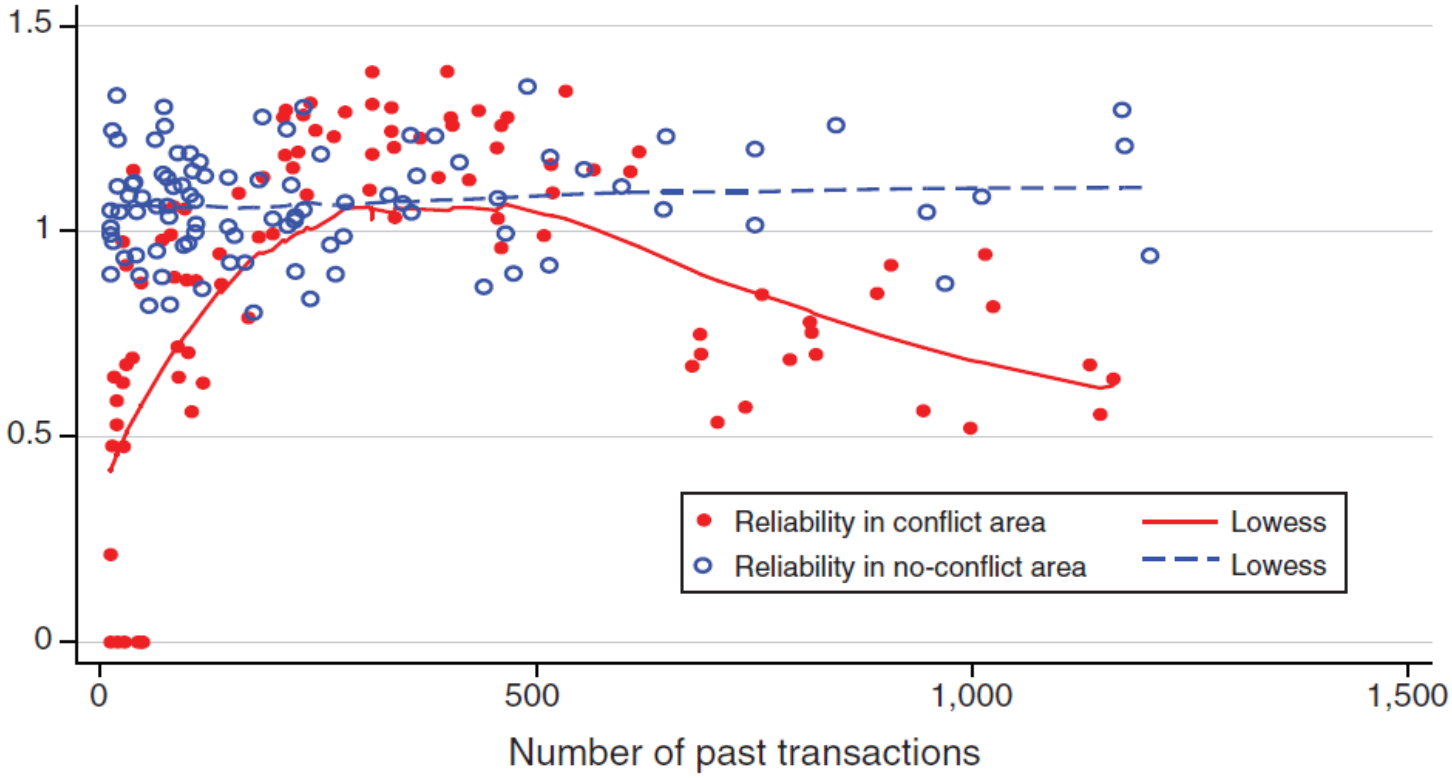


FIGURE 9. RELIABILITY AND CONFLICT IN DIRECT RELATIONSHIPS

Test 3: Indirect Evidence – reliability inverse U shaped in age of relationship in conflict regions

TABLE 6—RELIABILITY AT THE TIME OF THE VIOLENCE (TEST 3)

	Dependent variable: reliability during violence			
Sample	All firms (1)	All firms (2)	Firms in no-conflict regions (3)	Firms in conflict regions (4)
Conflict region	−0.174***	−0.200*	—	—
	(0.031)	(0.115)		
Age of the relationship			−0.010	0.200**
			(0.040)	(0.104)
Age of the relationship (sqrd)			0.0001	−0.020***
			(0.004)	(0.009)
Relationship controls	No	Yes	Yes	Yes
Seller controls	No	Yes	—	—
Buyer fixed effects	No	No	Yes	Yes
Seller fixed effects	No	No	Yes	Yes
Adjusted R^2	0.093	0.027	0.100	0.396
Observations	189	189	189	189

Notes: The table reports differences in estimated reliability between direct relationships of firms located in regions directly affected by the violence against firms located in regions not directly affected. The learning model predicts an inverted-U shape relationship between reliability and age of the relationship (Test 3). Reliability is computed as the ratio between actual shipments volumes during the week of the violence divided by the average volume shipped in the relationship during the control period, i.e., the first 20 weeks of the season. Age of the relationship is measured as the number of previous shipments in the relationship (in hundreds of transactions). Robust standard errors, two-way clustered at the seller and buyer levels, are reported in parenthesis.

Test 3: Direct Evidence (tab 7)

TABLE 7—EFFORT AT THE TIME OF VIOLENCE: DIRECT EVIDENCE

	Reliability during violence		Percent workers lost	
Dependent variable	Relationships		Seller	
Sample	(1)	(2)	(3)	(4)
Conflict region	−0.595***	—	—	—
	(0.112)			
Direct relationship	−0.022	−0.043		
	(0.058)	(0.105)		
Direct relationship × conflict region	0.421***	0.479**		
	(0.117)	(0.194)		
Seller only sells to the auctions			16.783	40.157***
			(13.678)	(12.718)
Seller sells to both marketing channels			12.067	7.978
			(13.748)	(10.315)
Seller fixed effects	No	Yes	—	—
Location fixed effects	No	—	Yes	Yes
Seller controls	No	No	Yes	Yes
Only sellers in conflict region	No	No	Yes	Yes
Adjusted R^2	0.248	0.331	0.039	0.594
Observations	262	262	42	42

Notes: The table reports direct evidence that firms exerted effort to maintain deliveries to direct buyers. Column 1 expands the specification in column 2 of Table 6 including sales to the auctions during the violence. Reliability at the auctions is computed as reliability in direct relationships in Table 6. Firms in the conflict region had significantly lower exports than usual. The column documents a differential effect across the two marketing channels, with significantly lower export reductions to direct relationships. Column 2 includes seller fixed effects. Robust standard errors, clustered at the seller level, are reported in parenthesis in columns 1 and 2. Columns 3 and 4 report results at the seller level focusing on the sample of interviewed sellers located in the conflict region. Correlations between the percentage of workers reported missing during the violence and the marketing channels used by the firm are reported. Workers lost (percent) is the highest percentage reported by the firm throughout the period during and following the eruption of violence. These results also appear in Ksoll, Macchiavello, and Morjaria (2013). Robust standard errors, clustered at the location level, are reported in parenthesis.

Macchiavello & Morjaria: Summary

- Nice design to look at impact of relationship contracts.
- Evidence for imperfect enforceability and learning about supplier type
- 3 tests reasonably intuitive, but
  - No causality (is it really violence or some other region-specific time series change?)
  - Nonlinearities hard to interpret – measurement error, etc. could generate non-linearities; are differences significant?
  - Is it age or selection?
- Assumption that it's learning about type may be wrong – could be more general relationship (e.g. mutual investment)

Overview

1. Relational contracts revision

2. Macchiavello & Morjaria – Roses in Kenya

3. Blader, Gartenberg & Prat (2016) – Truckers in US

4. Conclusions

Blader, Gartenberg & Prat (2019)

- RCT to look at relational contracts
- Illustrates many of the empirical issues in trying to test such theories
- Not focused so much on nuances of theory. These very 2nd order compared to more major empirical challenges of causality

Idea: Collectivist vs individualist cultures

- Some management practices like Relative Performance Evaluation (RPE) work well in “individualist” culture
- But RPE may work poorly/negatively in “collectivist” settings. One reason is that if agent A improves effort then this could come at expense of agent B (as benchmark improves)
  - In Benabou & Tirole (2003) framework, altruistic preferences will deliver this result (like an externality). This is model Blader et al (2003) use
  - But also fear of reprisal by low productivity workers would do the same thing – Bandiera, Barankay & Rasul (2010) find evidence for this
- In either case impact of RPE heterogeneous (like a complementarity)

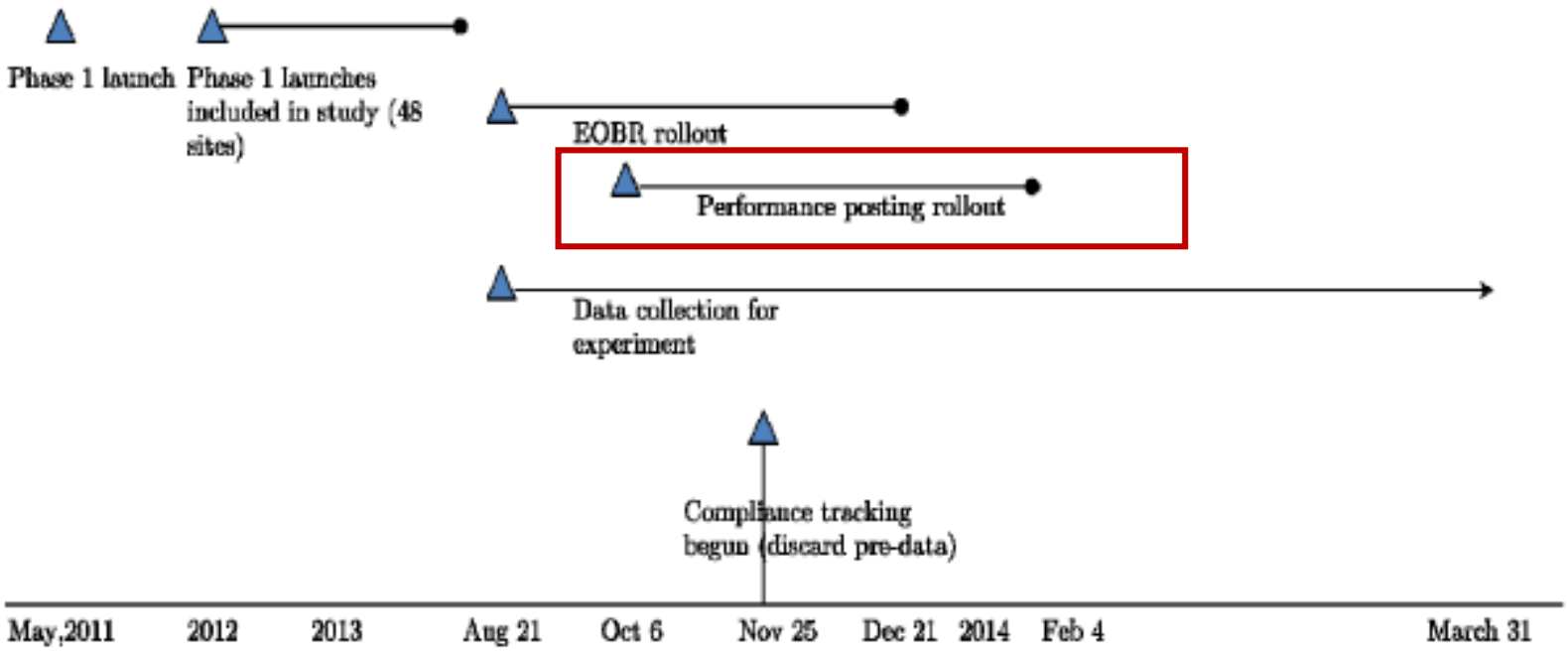
Setting

- Big US Trucking company (“less than truckload” segment)
- RCT across sites within the company: treatment is explicit comparison of the weekly productivity of drivers
  - **Treatment 1:** Public Posting (names)
  - **Treatment 2:** Anonymous Posting (only Driver ID, IDed)
- Technologically feasible because company has Electronic On-Board Recorder (EOBR) which collects this data (from Aug 21 & completed in all sites by Dec 2013)
- Firm has also started roll-out of Toyota “lean manufacturing.” Phase 1 (from May 2011) of this is a cultural intervention to make sites more collectivist based on team work instead of traditional individualist system (“last American cowboys”)
- Employees, not contractors (unlike Baker & Hubbard, 2003). Under 200 miles shipments

Timeline

Lean roll-out: pre-RCT

Figure 1: Timeline of Experiment



Setting

- Researchers randomized the RPE treatments based on whether the site had introduced Phase 1 of Lean at least 3 months ago (since Lean may take a while to have an effect)
- Other WMS elements of lean (Phases 2-5) not introduced yet anywhere
- Had practical problem that firm didn't check compliance at first, so started this in 11/25/13 & discarded earlier data
- ~5,000 drivers in 143 sites - 47 in control; 50 in **named** postings (Treatment 1), 46 in **IDed** postings (Treatment 2, anonymous)
- Look 30 days prior to treatment, 30 days after & discarded 5 days around implementation
- 93,313 driver-days

Figure A2: Phase 1 Evaluation Criteria

Safety	Employee s have a formal avenue to openly voice, share, and regularly address safety concerns at the facility
	Safety concerns are addressed in a timely manner by a cross-functional, integrated team of employees, supervision, and management.
Safety and leadership	What level of leader is involved in the safety journey?
	What organizational levels originated, supported, and have advocated the lean implementation initiative in the facility?
Power distance	Management availability to team members. Do employees feel that management is approachable?
	What percentage of the day do Supervisors spend on the Dock, during normal working hours?
	What percentage of the day do Managers spend on the Dock, during normal working hours?
Employee recognition	Individuals who meet, exceed, or achieve objectives are recognized on a regular basis through an employee recognition program?
	Groups who meet, exceed, or achieve objectives are recognized on a regular basis through a group recognition program?
Management style	Feedback and concerns are encouraged and included before making changes and taking actions.
	Employees, Supervisors, and Managers are encouraged/empowered to try improvement ideas, using innovation and creativity to enrich job responsibilities.
	The organizational level involved in determining and leading facility, function, and CIR Goals.
Teamwork and empowerment	Daily work activities are organized into team functions.
	SME s are utilized as initial point of contact for problem-solving, resolution, and employee directing activities.
	Problem-Solving activities are organized into team functions.
	Employees are empowered, utilized, participate, initiate, and lead problem-solving activities autonomously, without significant management involvement.
Communication	There is an avenue for workers to openly share common concerns, issues, and problems regularly with other employees, supervisors, and management.
	Employee concerns and questions are addressed in a timely manner.
	Are there daily meetings with employees and supervision/management where the daily plans, performance, etc. are shared?

These criteria are taken from a formal assessment tool used by managers to score how successful the business transformation rollout has been at any given site. Sites are assessed formally in a two-day process at least once per year to certify their progression in adopting the new culture and practices.

Performance Measures (Higher value indicates LOWER productivity)

- **Gap Score:** difference between actual & “potential” (optimal EOBR estimate given weather & route characteristics) miles per gallon
- **Shift Score**: “shifting events” due to excess revving, etc.
- **Excess Idle Time:** engine idling wastes fuel
- **Total Fuel Lost**: Aggregate measure of fuel wasted from idling, inefficient shifting, speeding & gearing

- They look at all 4 measures

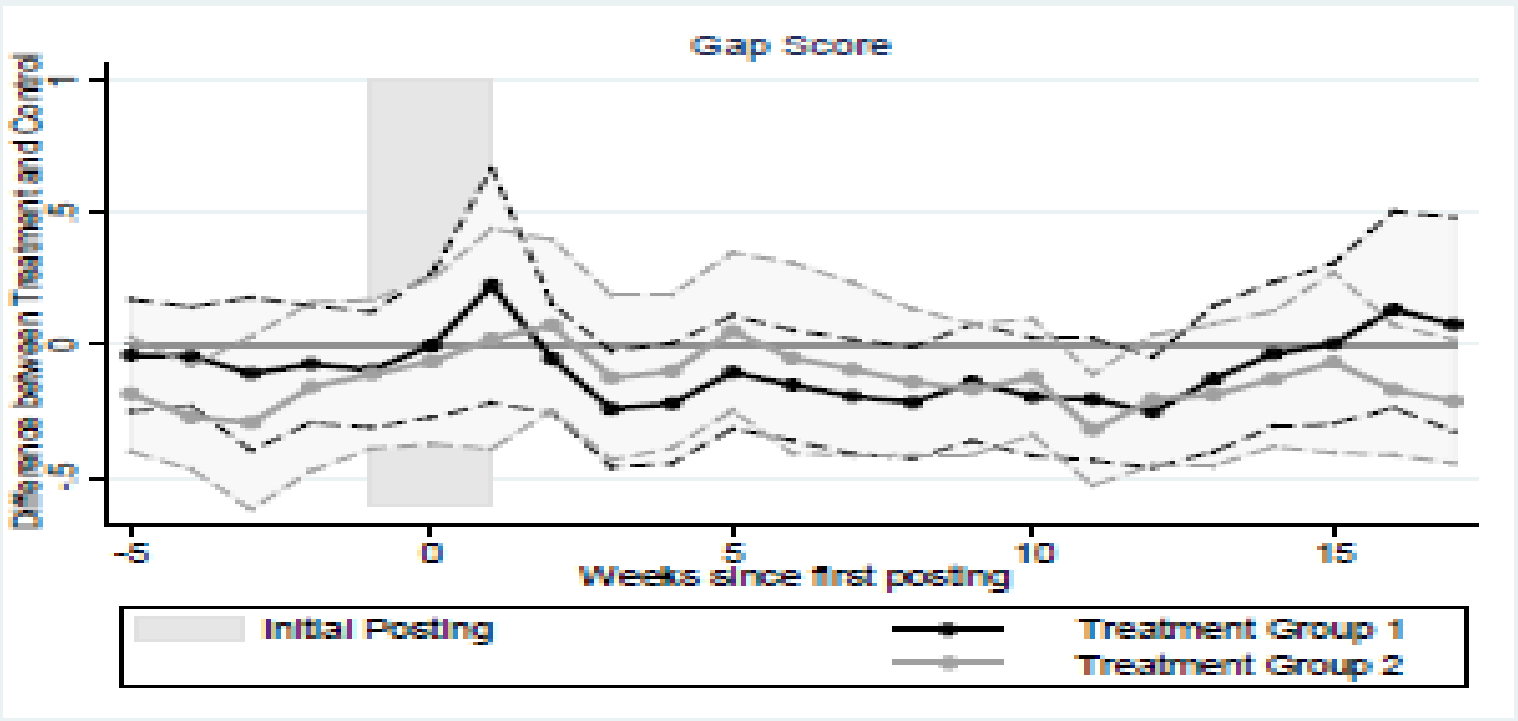
Balance of Experimental Assignment

- Looks reasonably balanced

	Full sample				
	Control Mean	Treatment 1 (names) Mean	Diff p-value	Treatment 2 (IDs) Mean	Diff p-value
Site characteristics					
# sites	47.00	50.00	n/a	45.00	n/a
Phase 1 status	0.30	0.26	0.681	0.47	0.098
Tractors / site	25.00	25.32	0.924	23.73	0.664
Distance / trip	124.08	130.63	0.309	131.24	0.247
Eastern region	0.44	0.44	0.966	0.30	0.149
Central region	0.33	0.22	0.220	0.39	0.607
Western region	0.22	0.34	0.207	0.32	0.313
Pre-treatment driver performance					
Miles per gallon	6.76	6.88	0.247	6.82	0.558
Gap score	2.18	2.14	0.787	1.98	0.310
Shift score	90.77	90.69	0.902	91.79	0.149
Excess idle time	0.12	0.12	0.838	0.14	0.429
Fuel lost	0.34	0.35	0.722	0.31	0.185

Coefficients suggest positive productivity effect (remember lower value better), but not significant

Figure 2: Impact of Rankings on Driver Performance



Difference between treatment and control for all sites (Phase 1 and pre-Phase 1 sites pooled together). See caption to Table 1 for definition of variables. Error bars reflect 90% confidence intervals, clustered by site. See Appendix Figure A4 for other outcome measures.

Look at treatment heterogeneity depending on whether site had Lean Phase 1 (“collectivist spirit”)

- Idea that RPE will be beneficial when there is an individualist culture (“Truckers are America’s last cowboys”), but counter-productive when there is a collectivist culture

For named postings (Treatment 1), performance improves in in sites without Lean (Phase 1), but negative effect in sites with Lean

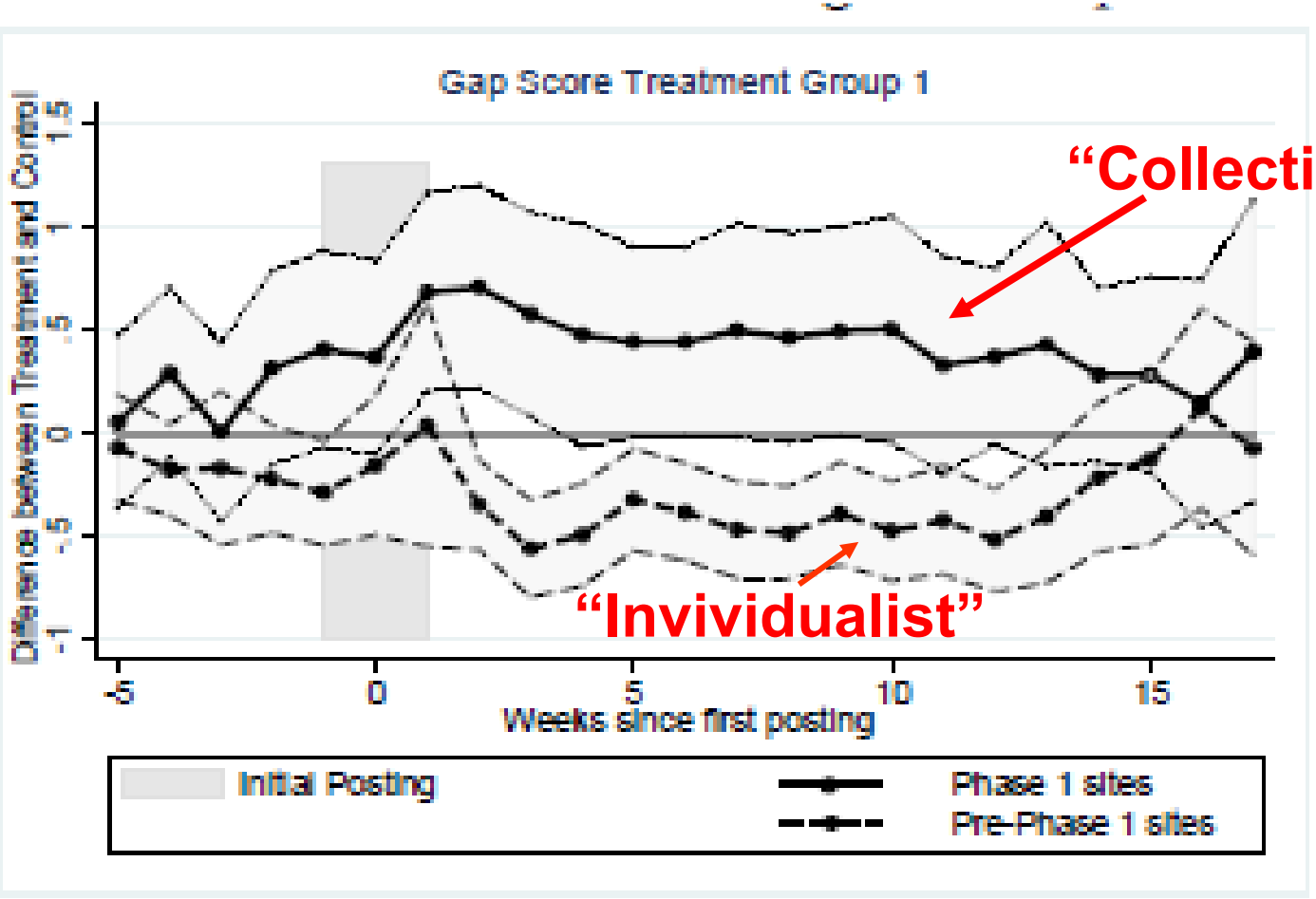


Table 4: Effect of Rankings on Phase 1 and Pre-Phase 1 Sites

Dependent variable:	Log(Gap Score)		Shift Score		Log(Idle Time)		Log(Fuel Lost)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post*Treatment Group 1*Phase 1	0.1351***	0.1364***	1.8567**	1.9631***	0.0397***	0.0354***	0.0607***	0.0549***
	(0.0406)	(0.0375)	(0.7414)	(0.6711)	(0.0125)	(0.0122)	(0.0136)	(0.0121)
Post*Treatment Group 2*Phase 1	0.0207	0.0312	0.5047	0.7213	-0.0096	-0.0098	0.0078	0.0129
	(0.0498)	(0.0451)	(0.7248)	(0.6709)	(0.0157)	(0.0146)	(0.0162)	(0.0158)
Post*Treatment Group 1	-0.0475*	-0.0400*	-0.1718	-0.0960	-0.0149*	-0.0129	-0.0224***	-0.0169**
	(0.0257)	(0.0236)	(0.4198)	(0.3900)	(0.0079)	(0.0078)	(0.0079)	(0.0081)
Post*Treatment Group 2	0.0224	0.0309	0.2686	0.2620	0.0152	0.0144	0.0066	0.0059
	(0.0370)	(0.0331)	(0.4214)	(0.4163)	(0.0128)	(0.0122)	(0.0112)	(0.0119)
Post*Phase 1	-0.0393	-0.0295	-0.9703	-0.9613*	0.0024	0.0016	-0.0122	-0.0096
	(0.0332)	(0.0298)	(0.6016)	(0.5432)	(0.0084)	(0.0082)	(0.0114)	(0.0100)
Treatment Group 1*Phase 1	0.1506*	0.0435	-0.4987	-0.7301	0.0088	-0.0141	0.0147	0.0072
	(0.0814)	(0.0645)	(0.9923)	(0.7370)	(0.0156)	(0.0148)	(0.0228)	(0.0238)
Treatment Group 2*Phase 1	0.0629	0.0897	0.8060	0.4052	0.0029	0.0235*	-0.0004	0.0313
	(0.0871)	(0.0719)	(0.9163)	(0.7771)	(0.0159)	(0.0131)	(0.0246)	(0.0278)

Major problem: Is it really Lean causing the heterogeneous treatment effect

- Recall RCT is over posting treatment, NOT over Lean itself
- Lean Phase 1 not random. Sites which had the early introduction look systematically different from those that did not
- Example: lots more tractor per site (bigger); productivity better in sites where Lean introduced first (Gap, Shift significantly different)

Look at treatment heterogeneity depending on whether site had Lean Phase 1 ("collectivist spirit")

Table 2: Balance Between Phase 1 and Pre-Phase 1 Sites

	Full sample			Matched sample		
	Pre-Phase 1 Mean	Phase 1 Mean	Diff p-value	Pre-Phase 1 Mean	Phase 1 Mean	Diff p-value
Site characteristics						
# sites	95	48	n/a	41	41	n/a
Tractors / site	20.35	33.25	0.000	25.95	27.51	0.581
Distance / trip	128.04	127.53	0.609	128.04	127.53	0.937
Eastern region	0.27	0.39	0.155	0.37	0.38	0.865
Central region	0.41	0.37	0.626	0.44	0.38	0.626
Western region	0.32	0.24	0.357	0.20	0.23	0.701
Control group	0.35	0.29	0.480	0.39	0.27	0.245
Treatment Group 1	0.39	0.27	0.149	0.32	0.24	0.467
Treatment Group 2	0.26	0.44	0.027	0.29	0.49	0.072
Pre-treatment driver performance						
Miles per gallon	6.90	6.72	0.039	6.76	6.71	0.602
Gap score	2.14	2.04	0.537	2.00	2.03	0.838
Shift score	90.35	91.55	0.076	91.62	91.66	0.950
Excess idle time	0.12	0.13	0.781	0.12	0.13	0.815
Fuel lost	0.34	0.33	0.473	0.32	0.33	0.763

See Table 1 caption for variable definitions.

Doesn't look balanced on observables (so implement matching)

Solutions?

1. Use propensity score matching to get common support for Lean and non-Lean sites (82 from 143 sites)
    - But requires conditional independence assumption

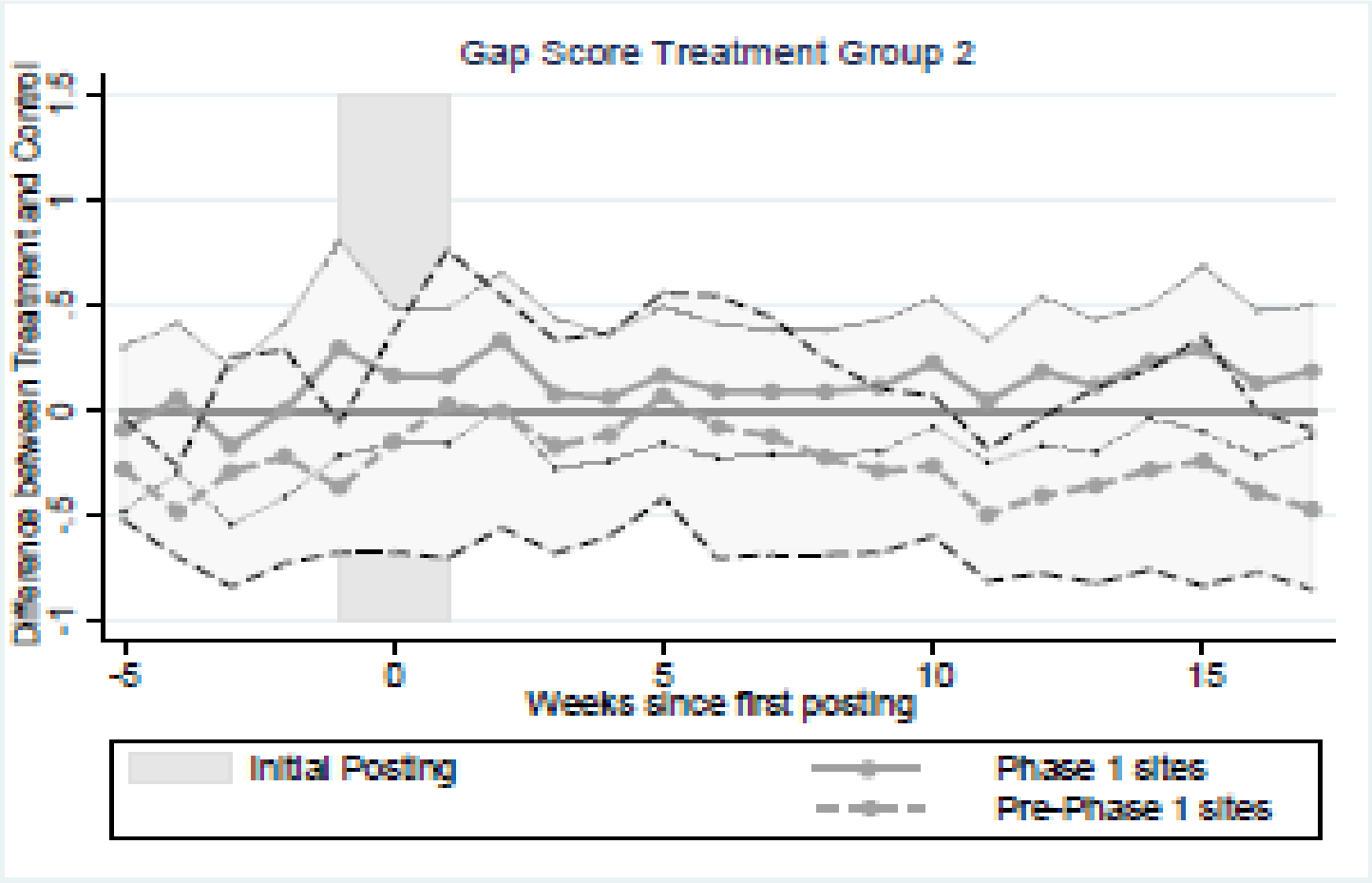
Table 7: Matched Analysis

Dependent variable:	Log(Gap Score)		Shift Score		Log(Idle Time)		Log(Fuel Lost)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post*Treatment Group 1*Phase 1	0.1366***	0.1425***	-0.2841	0.0365	0.0480***	0.0482***	0.0549***	0.0561***
	(0.0504)	(0.0433)	(0.7637)	(0.7349)	(0.0167)	(0.0152)	(0.0156)	(0.0144)
Post*Treatment Group 2*Phase 1	-0.0131	0.0126	-0.5912	-0.3145	-0.0276	-0.0212	0.0006	0.0048
	(0.0606)	(0.0509)	(0.6610)	(0.6081)	(0.0184)	(0.0167)	(0.0183)	(0.0189)
Post*Treatment Group 1	-0.0672*	-0.0563*	0.8533	0.7136	-0.0239**	-0.0239**	-0.0257**	-0.0220*
	(0.0369)	(0.0332)	(0.5163)	(0.5417)	(0.0109)	(0.0108)	(0.0116)	(0.0112)
Post*Treatment Group 2	0.0332	0.0341	0.5088	0.4096	0.0286*	0.0240	0.0069	0.0081
	(0.0491)	(0.0430)	(0.4416)	(0.4429)	(0.0157)	(0.0151)	(0.0152)	(0.0156)

Solutions?

1. Use propensity score matching to get common support for Lean and non-Lean sites (82 from 143 sites)
2. Comparison of Treatment 1 vs. Treatment 2
   - But concern over randomization
   - T2 looks like T1 but noisier
   - Is RPE really anonymous? Informal comparison

For anonymous postings (Treatment 2, IDed), no significant interaction effects



Solutions?

1. Use propensity score matching to get common support for Lean and non-Lean sites (82 from 143 sites)
2. Comparison of Treatment 1 vs. Treatment 2
3. Try to bound the bias by using Altonji et al (2005) approach to say that bias on unobservables can't be bigger than bias on observables
   - Not a strong test
4. Use a separate Employee Engagement Survey
   - Suggests higher collectivism in Lean/Phase 1 sites

Table 11: Effect of Ranking and Engagement on Driver Performance, Driver Fixed Effects

Depedent variable:	Log(Gap Score)			
Category:	Collective Index		Instrumental Index	
	(1)	(2)	(3)	(4)
Post*Treatment Group 1*[Category]	0.0988***	0.1033**	0.0548	0.0398
	(0.0348)	(0.0461)	(0.0359)	(0.0448)
Post*Treatment Group 2*[Category]	0.0227	0.0162	-0.0333	-0.0931
	(0.0399)	(0.0543)	(0.0591)	(0.0957)

Solutions?

1. Propensity score matching
2. Comparison of Treatment 1 vs. Treatment 2
3. Bounding the bias
4. Use a separate Employee Engagement Survey
   - But….
   - Only available for a small sub-sample of sites (43 of the 143)
   - Information gathered after the experiment, so could be outcome
   - Unclear if Phase 1 actually shifted any culture. Indeed, usually thought that culture is very hard to change

Conclusions on Blader et al (2019)

- In my view, probably best attempt so far to get at some causal evidence over relational contracts
  - RCT in clean-ish single firm setting
  - Does show that RPE has roughly zero effect
- But would be better to use a long-standing measure of culture (like employee engagement) PRIOR to the RCT
  - Unclear than having the quasi-experiment of Lean is helpful at all to tackle the problem
  - Unclear if real evidence for relational contract causing heterogeneity
- Would be even better to do a cultural intervention to see if it actually worked in an RCT

Conclusions on empirics of relational contracts

- Despite a lot of theoretical interest little econometric work. Only starting to emerge (Gibbons et al, 2015 on BSI)
- Fruitful area if cultural interventions possible. But by definition likely to be hard to shift in short-run. Heterogeneity of treatment wrt to existing differences
- Relates to issue of collusion in IO: often hard to identify (Chassang & Ortner, 2015)
  - E.g. how to measure discount rates?
- Fruitful area for new work

Back Up

Volume variation during week of maximum temptation to deviate

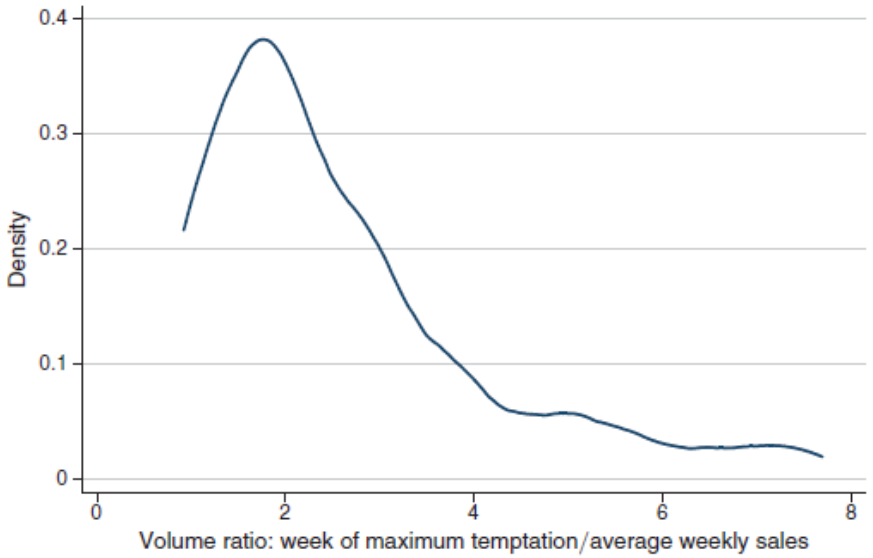


FIGURE 5. STRETCH IN VOLUMES AT THE MAXIMUM TEMPTATION TO DEVIATE

Notes: The figure shows the distribution of the ratio between the volume of roses traded in the week in which the incentive constraint is more likely to bind and the average weekly sales in the relationship. The sample is given by all the relationships active in the season before the violence.

Solutions?

- Treatment 1 reduces variance
  - Doesn't seem to work for Treatment 2

Table 10: Effect of Ranking and Engagement on Driver Performance

Dependent variable:	Log(Gap Score)			
Category:	Collective Index		Instrumental Index	
	(1)	(2)	(3)	(4)
Post*Treatment Group 1*[Category]	0.1101**	0.1146**	0.0669	0.0751
	(0.0425)	(0.0543)	(0.0435)	(0.0572)
Post*Treatment Group 2*[Category]	0.0300	0.0046	-0.0327	-0.1055
	(0.0620)	(0.0811)	(0.0762)	(0.1226)
Post*Treatment Group 1	-0.3991***	-0.4352**	-0.2732*	-0.3214*
	(0.1279)	(0.1667)	(0.1365)	(0.1745)
Post*Treatment Group 2	-0.0506	-0.0556	0.1582	0.3164