Quality Disclosure and Gaming: Do Employee Incentives Matter?

Mara Lederman (University of Toronto)

Silke Forbes (UC, San Diego)

Trevor Tombe (Wilfred Laurier University)

FTC Microeconomics Conference

November 3, 2011

Motivation

- Disclosure programs provide systematic information about product quality
  - E.g.: hospitals (report cards), schools (test scores), restaurants (hygiene scores)

- Empirical analysis has found these programs improve product quality but also that firms attempt the **"game"** the programs
  - Improve reported dimensions potentially at the expense of other dimensions
  - If reported measure(s) imperfectly correlated with what consumers care about, gaming may lead to inefficient allocation of resources and distort information
  - Possible since consumers may be heterogeneous in what they care about and program design faces a tradeoff between information quantity vs. usability

- Potential for gaming will depend not only the design of the program but also on characteristics of the product and the incentives in place at the firm
  - What dimensions of quality are measured?
  - How and by whom can those dimensions be manipulated?
  - Do those in a position to manipulate have incentives to do so?

What We Do in This Paper

- **Investigate the relationship between gaming and the incentives provided to the employees most likely to carry out the gaming**
  - Disclosure environment held constant but cross- and within-firm variation in extent of explicit incentives based on firm's performance in disclosure program
- Consider a specific empirical context – government rankings of airline on-time performance
  - But issues relevant in other settings in which disclosure programs do or could exist
- Department of Transportation (DOT) counts a flight as being "late" if it arrives 15 or more minutes later than scheduled; otherwise it's "on-time"
- Based on this, DOT creates monthly rankings of airlines which are often picked up in the media

Four Useful Features of this Setting

1. **Program design gives airlines clear incentive to game**
   - Reduce delays on flights expected to land just over 15 minutes late

2. **But, airlines cannot predict in advance which flights will land 13 vs. 15 vs. 17 minutes late. Thus, gaming must take place in real-time**
   - Makes consideration of employee incentives important

3. **Five airlines have implemented firm-wide employee bonus programs based explicitly on the airline's rank in the government program**
   - All face free-rider problem, but differ in ease of achieving target

4. **Great data and clean identification strategy**
   - Observe millions of flights and observe every stage of each flight
   - Can estimate every flight's expected delay and look for evidence of gaming on specifically those flights that are expected to be right around 15 minutes late

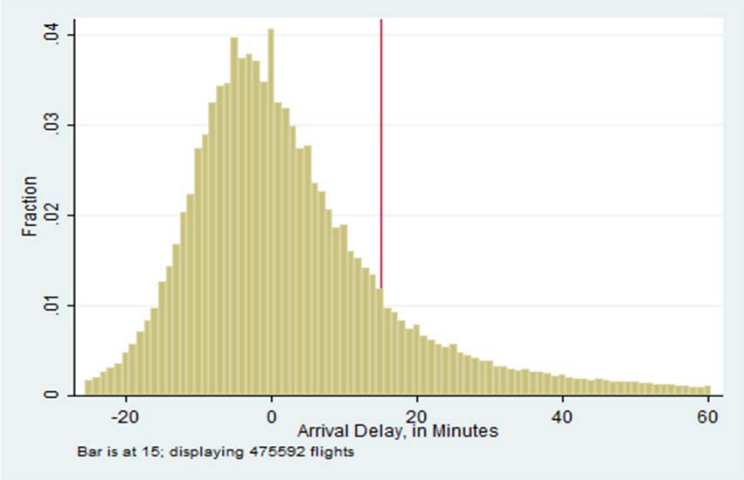
Preview of Findings

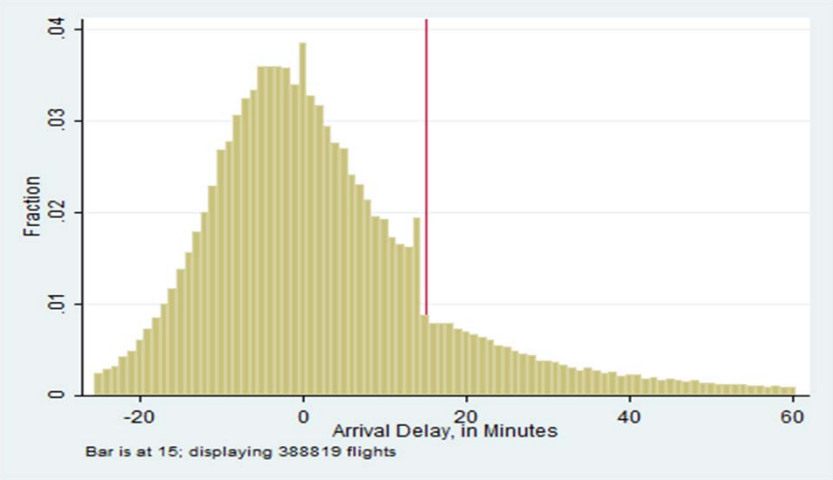
1. **No** evidence of gaming by airlines **without** employee bonus programs in place

2. **No** evidence of gaming by airlines with employee bonus programs that are based on targets that **could not realistically be achieved**

3. **Strong** evidence of gaming by airlines with employee bonus programs based on targets that are **could be – and were - achieved**

Arrival Delays for Continental Airlines, Before and After Bonus Program:





Disclosure of Airline On-Time Performance

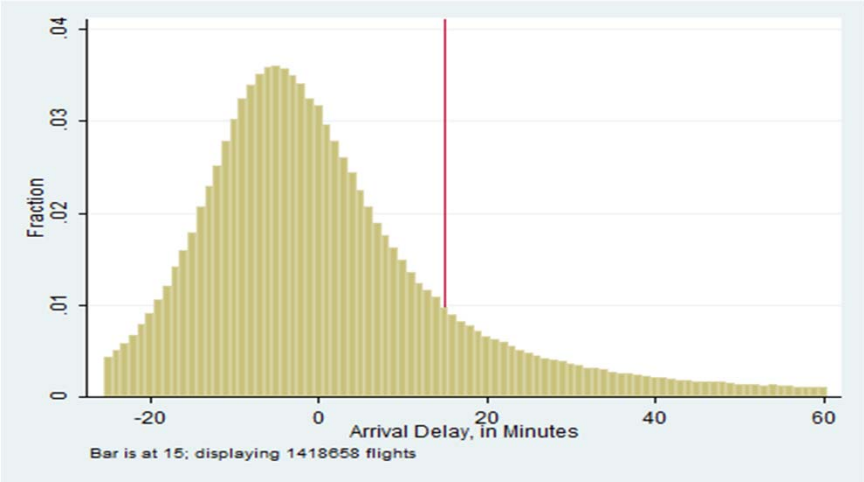
- 1987: airlines accounting for >1/% of domestic passenger revenues must report flights' **scheduled and actual departure and arrival times** to DOT
  - Over time, more airlines have met reporting requirements (10 in 1995, peaked at 20, now 16)
  - 1995: expanded to include additional variables - taxi-out, airborne and taxi-in times

- Flight is considered "late" if arrives 15 or more minutes behind schedule
  - DOT creates monthly rankings based on % of flights "on time" using this metric
  - Media frequently report the DOT's ranking (example)
  - Evidence that demand responds to on-time performance (Forbes, 2008)

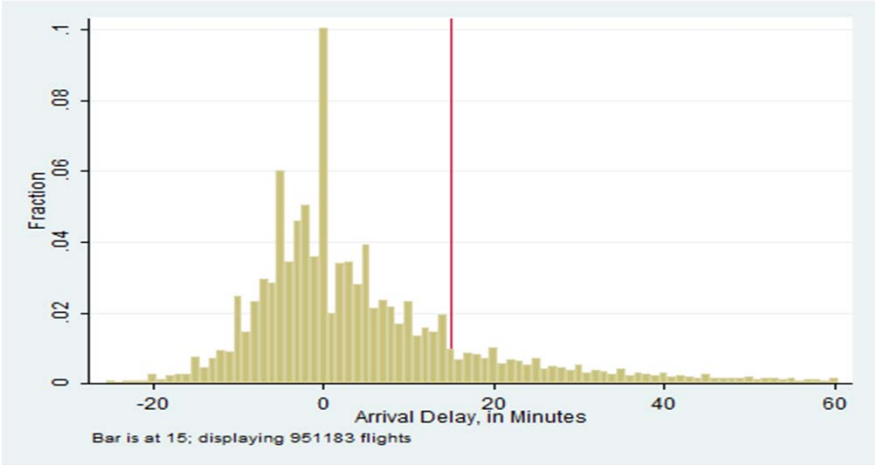
- During our sample period, airlines could report on-time data in 3 ways:
  1. Manually – i.e.: an employee records the arrival time
  2. Automatically – if aircraft has a technology called ACARS
  3. Combination of manual and automatic if some if its planes have ACARS
  - For combo reporters, don't know which planes are manual vs. auto but have developed approach to try to distinguish

Histograms of Arrival Delays, by Reporting Status (1998)

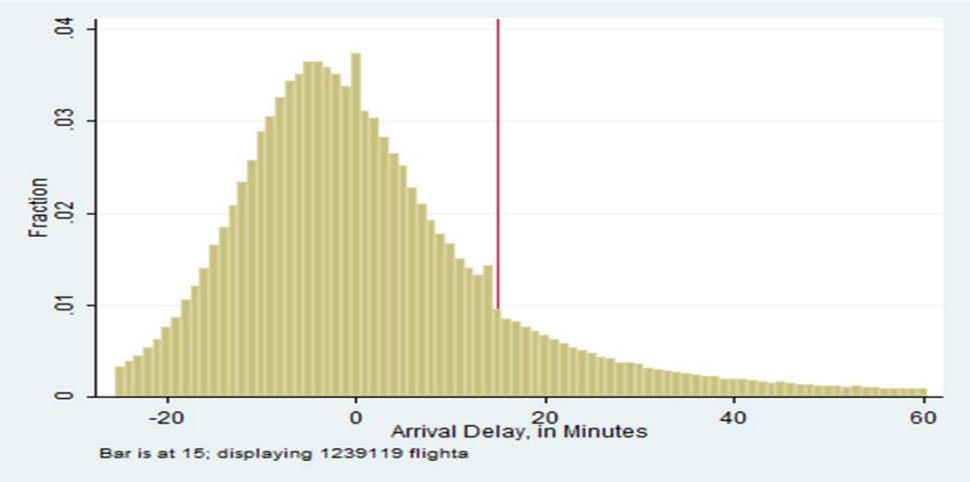
Automatic (AA, NW, UA, US)



Manual (WN, HP, AS)



Combination (CO, DL, TW)



Employee Bonus Programs

Pay between $65 and $100 to **each employee** in months in which the **airline** is near or at the top of the DOT ranking

Airline	Payment Structure	# Airlines Ranked	Airline's Average Rank in Year Before Bonus
Continental (1995)	$65 per employee if airline **ranks among top 5**. Since 1996: $65 for rank 2 and 3; $100 for rank 1.	10	7.1
TWA (1996)	$65 per employee if airline **ranks among top 5** in on-time, baggage and complaints. $100 if it also ranked 1st in one of the categories. In 1999: $100 if on-time performance exceeds fixed threshold of 80%. In 2000: Seasonal targets: 85% summer, 80% winter.	10	8.1
American (2003)	$100 per employee if **airline ranks 1st**. $50 if **airline ranks 2nd.** Since 2009: Bonus based on internal metric that excludes delays that are not under the employees' control.	17	3.1
US Airways (2005)	$75 per employee **if airline ranks 1st**.	19	9.8
United (Jan 2009)	$100 per employee if **airline ranks 1st**. $65 if **airline ranks 2nd.**	20	14.7

Empirical Approach

- Objective is to estimate whether **airlines systematically reduce delays on flights they expect to arrive slightly above the threshold to be considered on-time.** Requires 3 things:
  1. A way to identify which flights the airline expects to be close to the threshold
  2. A way to measure whether the airline reduces delays on those particular flights
  3. A way to measure the counterfactual delay those flights would have had absent incentive to game

<u>How we do each of these:</u>

1. Construct a measure of each flight's **predicted delay** at touchdown at arrival airport
   - Based on delays incurred so far and estimate for what happens next

2. Estimate whether subsequent delays (=taxi-in times) are systematically reduced for flights predicted to be close to threshold
   - Note that likelihood that flight is close to the threshold not known by airline in advance and – for a given flight - will vary from day to day

3. Flights just outside threshold (e.g.: predicted to be 13 or 18 minutes late) provide one counterfactual for what delay would have been absent incentive to game
   - If costs of delay are convex, flights with very long expected delays provide another possible counterfactual

Calculation of Predicted Delay

- We construct measure of each flight's predicted delay when its wheels touch down:

Predicted Delay = (Wheels down Time + Predicted Taxi-in Time) – Sched Arrival Time

Predicted Arrival Time

- Predicted taxi-in time is median taxi-in time for that particular flight in the quarter
- EX: Flight #236 by DL between BOS-ATL in March 1997; Sched arrival at 4:30 pm
  - If wheels down is 4:36 pm and median taxi-in time for that flight in Q1 of 1997 is 4 minutes, then predicted arrival time is 4:40 pm and predicted delay is 10 minutes
  - Results robust to other ways of predicting taxi-in time
- Then construct dummy variables for different levels of predicted delay
  - <10 min, 10-11 min, 11-12 min,… 15-16 min, 16-17 min, … >25 min (16 "bins")
- Construct bins separately for airlines without bonus program and for each airline with a bonus program (pre and post if possible)
  - Mutually exclusive, not additive

Taxi-Time Regressions

- **Estimate flight-level regressions that relate a flight's taxi-in time (in logs) to its predicted delay at wheels-down, captured by the predicted delay bins**

- Regressions include carrier-arrival airport-day FEs
  - Comparing taxi-in times for a carrier's flights arriving at a given airport on a given day that land with different predicted delays
  - Variation in whether flight is threshold flight driven by factors influencing delays at departure and in the air

- Controls: arrival hour of day, arrive/depart from carrier's hub, distance

- Cluster standard errors at arrival airport-date

- Look for evidence of a non-monotonicity right around 15 minutes
  - Test: Bin 15 vs. Bin12; Bin 15 vs. Bin 18; Bin 15 vs. Bin25+

- Three separate samples to investigate different programs; flights on every 5th day

Taxi-In Time as a Function of *Predicted* Delay, 1995-1998 (Table 3A)

Non -bonus Carriers			
Predicted Delay			
[10,11) min	-0.0218***	[18,19) min	-0.0392***
	(0.00199)		(0.00283)
[11,12) min	-0.0201***	[19,20) min	-0.0405***
	(0.00204)		(0.00291)
[12,13) min	-0.0235***	[20,21) min	-0.0467***
	(0.00212)		(0.00293)
[13,14) min	-0.0324***	[21,22) min	-0.0363***
	(0.00230)		(0.00306)
[14,15) min	**-0.0310*** **	[22,23) min	-0.0411***
	(0.00241)		(0.00316)
[15,16) min	**-0.0346*** **	[23,24) min	-0.0436***
	(0.00244)		(0.00331)
[16,17) min	**-0.0390*** **	[24,25) min	-0.0425***
	(0.00254)		(0.00338)
[17,18) min	-0.0413***	>25 min	-0.0489***
	(0.00265)		(0.00145)

No evidence of gaming by carriers WITHOUT bonus programs in place.

Coefficient tells the ~% change in taxi-in time for flights with the given level of predicted delay relative to flights predicted to be <10 minutes late

Taxi-In Time as a Function of *Predicted* Delay, 1995-1998 (Table 3A)

	Non-bonus Carriers	CO post-Bonus
Predicted Delay		
[11,12) min	-0.0201***	-0.0562***
	(0.00204)	(0.00566)
[12,13) min	-0.0235***	-0.0563***
	(0.00212)	(0.00587)
[13,14) min	-0.0324***	-0.0772***
	(0.00230)	(0.00621)
[14,15) min	**-0.0310*** **	**-0.105*** **
	(0.00241)	**(0.00660)**
[15,16) min	**-0.0346*** **	**-0.140*** **
	(0.00244)	**(0.00707)**
[16,17) min	**-0.0390*** **	**-0.144*** **
	(0.00254)	**(0.00781)**
[17,18) min	-0.0413***	-0.132***
	(0.00265)	(0.00935)
[18,19) min	-0.0392***	-0.0874***
	(0.00283)	(0.00929)
[19,20) min	-0.0405***	-0.0857***
	(0.00291)	(0.00880)
>25 min	-0.0489***	-0.0489***
	(0.00145)	(0.00366)

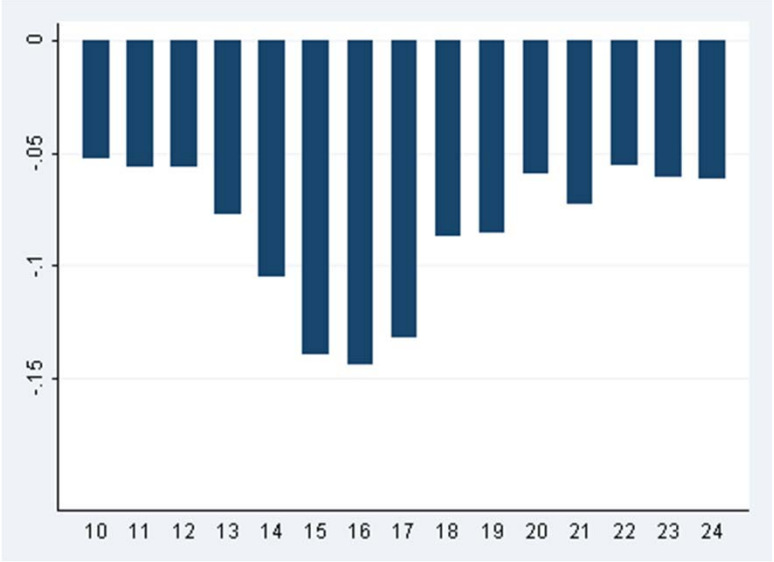
- **Continental's flights predicted to be 15-16 minutes late have taxi-in times that are ~13% shorter than the taxi-in times of its flights predicted to be <10 minutes late.**
- **Its flights predicted to 25 or more minutes late have taxi-in times that are only ~4% shorter**

Taxi-In Time as a Function of *Predicted* Delay, 1995-1998 (Table 3A)

	Non-bonus Carriers	CO post-Bonus	TWA pre-Bonus	TWA post-Bonus
Predicted Delay				
[11,12) min	-0.0201***	-0.0562***	-0.0373**	-0.0530***
	(0.00204)	(0.00566)	(0.0132)	(0.0106)
[12,13) min	-0.0235***	-0.0563***	-0.00858	-0.0757***
	(0.00212)	(0.00587)	(0.0142)	(0.0109)
[13,14) min	-0.0324***	-0.0772***	-0.0502***	-0.115***
	(0.00230)	(0.00621)	(0.0141)	(0.0119)
[14,15) min	**-0.0310*** **	**-0.105*** **	**-0.0726*** **	**-0.116*** **
	(0.00241)	**(0.00660)**	**(0.0158)**	**(0.0133)**
[15,16) min	**-0.0346*** **	**-0.140*** **	**-0.0516** **	**-0.145*** **
	(0.00244)	**(0.00707)**	**(0.0163)**	**(0.0133)**
[16,17) min	**-0.0390*** **	**-0.144*** **	**-0.0160**	**-0.165*** **
	(0.00254)	**(0.00781)**	**(0.0162)**	**(0.0161)**
[17,18) min	-0.0413***	-0.132***	-0.0648***	-0.140***
	(0.00265)	(0.00935)	(0.0178)	(0.0167)
[18,19) min	-0.0392***	-0.0874***	-0.0564**	-0.139***
	(0.00283)	(0.00929)	(0.0175)	(0.0179)
[19,20) min	-0.0405***	-0.0857***	-0.0764***	-0.0835***
	(0.00291)	(0.00880)	(0.0178)	(0.0174)
>25 min	-0.0489***	-0.0489***	-0.0841***	-0.0883***
	(0.00145)	(0.00366)	(0.00978)	(0.00846)

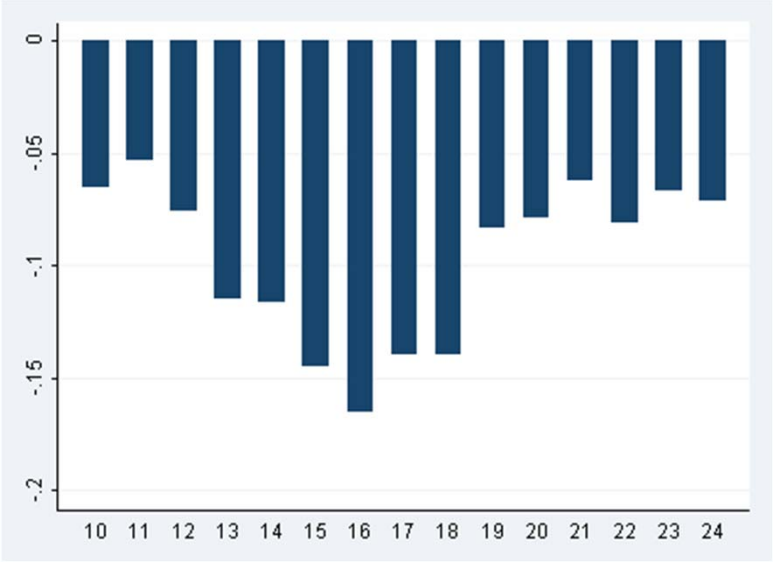
Plots of Regression Coefficients

Continental post-bonus



Predicted Delay

TWA post-bonus



Predicted Delay

Taxi-In Time as a Function of *Predicted* Delay, 2002-2006/2008-2010 (Table 3B)

	AA post-Bonus	US post-Bonus	UA post-Bonus
Predicted Delay			
[11,12) min	-0.0351***	-0.0275**	-0.0343*
	(0.00654)	(0.0104)	(0.0139)
[12,13) min	-0.0486***	-0.0260*	0.000440
	(0.00699)	(0.0116)	(0.0147)
[13,14) min	-0.0467***	-0.0211	-0.0288
	(0.00735)	(0.0118)	(0.0170)
[14,15) min	**-0.0507*** **	**-0.0273* **	**-0.00304**
	(0.00766)	**(0.0115)**	**(0.0169)**
[15,16) min	**-0.0685*** **	**-0.0363** **	**-0.00278**
	(0.00781)	**(0.0124)**	**(0.0170)**
[16,17) min	**-0.0521*** **	**-0.0258* **	**-0.00686**
	(0.00839)	**(0.0130)**	**(0.0183)**
[17,18) min	-0.0586***	-0.0306*	0.00393
	(0.00858)	(0.0138)	(0.0161)
[18,19) min	-0.0465***	-0.0403**	-0.0340
	(0.00843)	(0.0131)	(0.0188)
[19,20) min	-0.0762***	-0.0255	-0.0429*
	(0.00914)	(0.0133)	(0.0184)
>25 min	-0.0579***	-0.0617***	-0.0470***
	(0.00360)	(0.00512)	(0.00567)

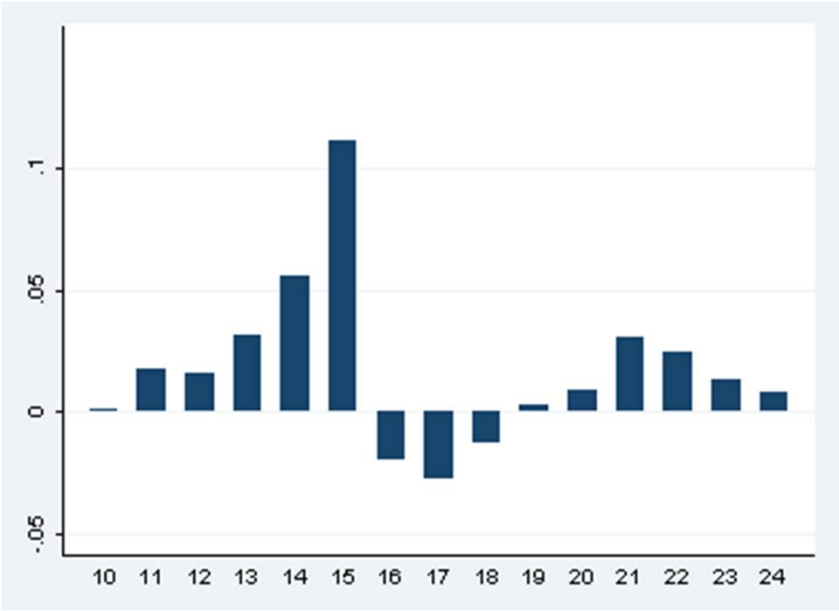
- **NO evidence of gaming by these carriers following the introduction of their bonus programs**

When Gaming Occurs, Does it "Work"?

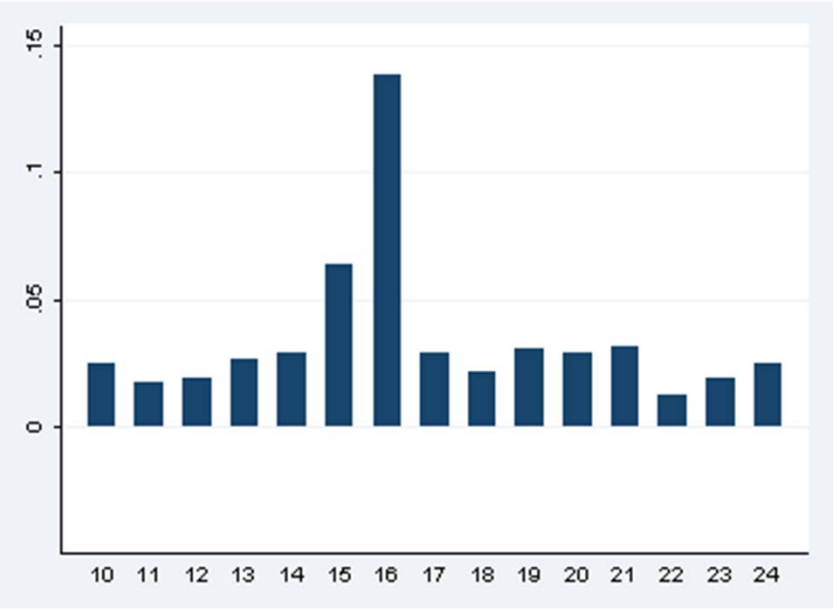
- Run same regression but replace LHS variable with dummy that equals one if flight lands one minute earlier than predicted
- Do same thing for landing two minutes earlier than predicted
- Coefficients measure the change in the probability of being one/two minute(s) earlier than predicted for flights in a given predicted delay bin relative to the probability for flights with predicted delay <10 minutes
- Put differently, these regressions test whether we are systematically worse at predicting delay for specifically those flights in the critical threshold

Probability of Arriving One/Two Minute(s) Earlier than Predicted

Continental – 1 min. earlier



Continental – 2 min. earlier



- **Flights predicted to be 15-16 minutes late are 11 percentage points more likely to arrive 1 minute earlier than predicted - average prob(1 min early) for CO flights is ~20%**

- **Flights predicted to be 16-17 minutes late are 14 percentage points more likely to arrive 2 minutes earlier than predicted – average prob(2 min early) for CO flights is ~10%**

Identifying Manual Planes

- Histograms of manual reporters show tendency to round arrival delays at zero and the 5s. Histograms for CO and TWA in early years show some of this as well
  - Rounding only possible for manual planes

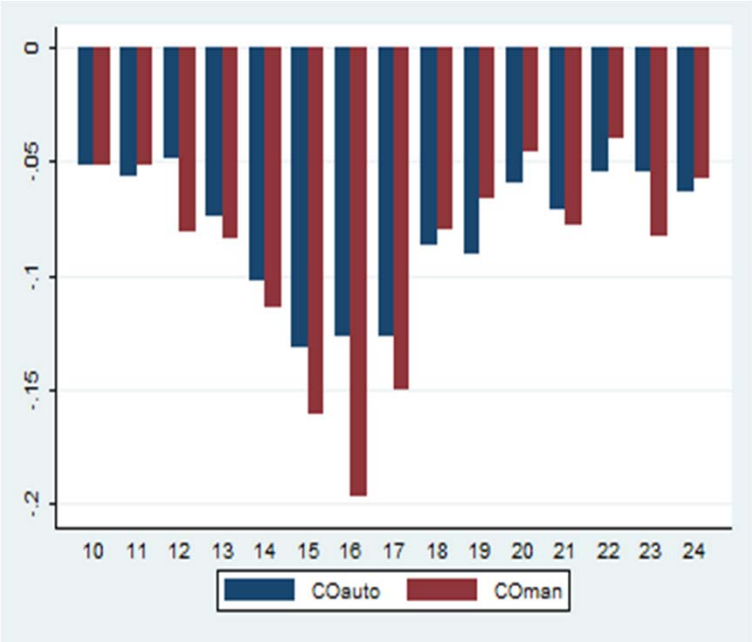
- So, in each year, calculate a variable equal to the likelihood that a **given plane** has an arrival delay of exactly zero minutes
  - Able to do this because starting in 1995, data includes plane's tail#

- Look at differences in the distribution of this variable for carriers who report automatically, manually and combination carriers

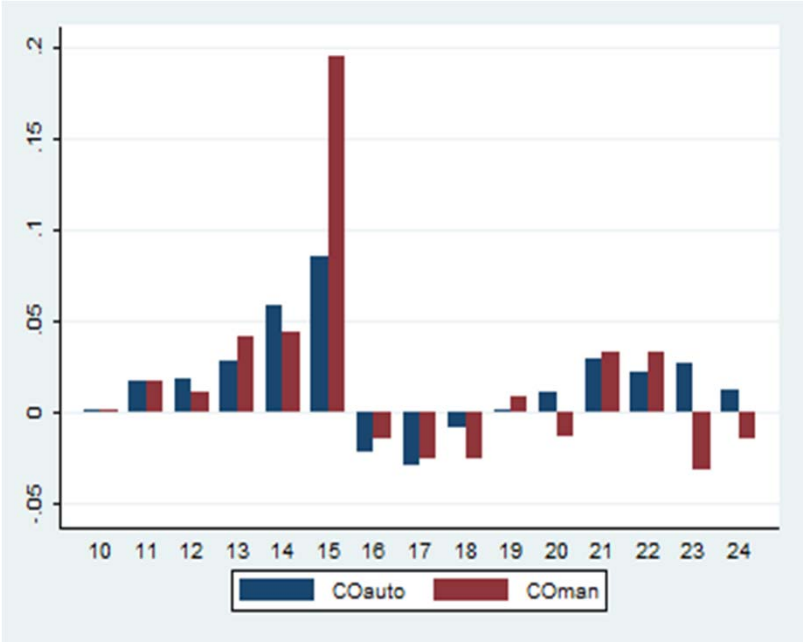
- Define a cutoff above which we assume that a plane is manual: if a plane lands with zero delay more often than is "typical" for an automatic reporter, we classify it as manual
  - We take a conservative approach; rather classify an auto plane as manual than vice versa

Taxi-time Results: Manual vs. Automatic Planes

Continental – Taxi-time results



Continental – 1 min. earlier results



	Predicted delay 15-16 min	**Predicted delay 16-17 min**
Automatic	~13% shorter taxi-in times	12% shorter taxi-in times
Manual	~16% shorter taxi-in times	~19% shorter taxi-in times

Discussion: Early vs. Late Bonus Programs

Why do we observe gaming in response to the two early programs but not in response to the three later program?

Possible Explanations

1. **Misreporting**: At least of some of the gaming by CO and TWA seems to be misreporting. AA, US and UA could not misreport because they were reporting automatically

2. **Much weaker incentives**: CO and TWA programs awarded bonus if airline ranked among top 5 at a time when only 10 airlines were ranked. AA, US and UA only awarded first (in some cases, second) spot at a time when 18 airlines were ranked
   - And, some of those consistently outperformed all others by wide margin – e.g.: Hawaiian Airlines ranked first in almost every month after it qualified
   - Even if gaming can lead to a one or two spot improvement, wasn't likely to move carrier into range where bonus would be awarded

Summary

- Structure of DOT program creates clear incentives for gaming because rank is based on a very blunt and transparent metric – flights arriving <15 minutes late
- But, those flights cannot be identified in advance because difference between 14, 15, and 16 minutes randomly determined once flight is in progress
  - Gaming must occur in real-time by employees who may not have incentives to do so
- Despite clear incentive to game, we find no evidence of gaming by airlines without bonus programs or with programs with unrealistic targets
- But find strong evidence of gaming by the two airlines who introduced programs with targets that could be – and were – met
- Simulations (not shown here) show that small reductions in taxi-in times – if applied to right flights – can meaningfully impact the metrics consumers see
  - Since metric only imperfectly correlated with what consumers care about may lead consumers to make the "wrong" decisions

Concluding Thoughts

- Paper contributes to the growing empirical literature on gaming of disclosure programs

- First to explicitly consider link between gaming and changes in the incentives provided to the employees whose effort is required to carry out the gaming

  - Highlights importance of considering interaction between program design, product characteristics and internal organization and incentives
  - Relevant to the policy discussion on use of disclosure programs (and potentially incentives based on these programs) to improve quality – e.g.: No Child Left Behind
  - Begins to link the *external incentives* provided by the disclosure program (to the firm) with the *internal incentives* provided by the firm (to its employees)

- Also provides evidence that really high-powered incentives do not affect behaviour – precisely because employees do not believe reward can be achieved