

## **Project Description:**

### **Project Title: Gaming of Public Disclosure Programs – Evidence from the U.S. Airline Industry**

#### **I. Introduction**

Quality disclosure programs are intended to provide consumers with information about credence goods. Many such programs present consumers with quality measures that are based on whether or not a product's quality passes a particular threshold. One implication of designing a disclosure program in this way is that it may encourage firms to manipulate quality right around the relevant threshold. For example, for firms with a wide distribution of product quality, this type of program gives them an incentive to improve the quality of products that can, at relatively low cost, pass the threshold but does not give them an incentive to improve the quality of products that are already well above or below the threshold. A second common feature of these programs is that quality is evaluated relevant to a baseline that can be affected by the firm, giving the firm an incentive to lower the baseline quality in order to more easily pass the threshold. These behaviors may not only lead to an inefficient allocation of resources by the firm but they may also distort the information that the program conveys to consumers.

The proposed research will investigate these issues in the context of a government mandated program that requires airlines to disclose the on-time performance of their flights. Since 1987, airlines have been required to report to the Department of Transportation (DOT) the scheduled and actual arrival times for their domestic flights. Although the DOT collects detailed data about the actual minutes of delay incurred on each flight, it counts a flight as being "late" if it arrives 15 minutes or more behind schedule. The DOT issues monthly reports to consumers that rank airlines based on the percentage of their flights that are late as defined by the 15 minute cutoff. These rankings - or excerpts from these rankings - are frequently reported in media outlets, which is likely the primary source of this information for consumers.

In the first part of the study, we focus on the incentive that this program gives airlines to exert effort to reduce delays on specifically those flights that they can move below the 15 minute threshold at low costs. Typically, these will be flights that would otherwise arrive just over 15 minutes late and that can be sped up by a few minutes. On the other hand, the program provides airlines with little incentive to reduce delays on flights whose delays are expected to be well above (or well below) 15 minutes. While these incentives to "game" are inherent in the design of this program, a particularly interesting – and we believe unique – feature of our setting is that, between 1995 and 2009, several airlines implemented employee bonus programs based on the airline's ranking in the government program. Under a typical bonus program, each airline employee would receive a payment of between \$65 and \$100 in each month in which the airline ranked at or near the top of the DOT ranking. Note that these bonuses are based on the airline's overall performance, not an employee's individual performance. To the extent that there is gaming going on, these bonus programs provide a discrete increase in employees' incentives to engage in this type of behavior.

We develop an empirical approach that allows us to estimate whether airlines systematically exert more effort to reduce delays on flights which would otherwise arrive slightly above the 15 minute threshold, relative to the effort they exert on flights that are expected to arrive well above or below the threshold. Much of our empirical analysis focuses on differences in flights' taxi-in times. We focus on taxi-in times because this represents the final stage of a flight and thus the final point at which delays may

be incurred or reduced. By the time a flight has touched down at the arrival airport of its route, an airline should have a fairly precise estimate of the expected delay that the flight will have and can decide whether or not to exert effort to try to reduce that delay below 15 minutes. Furthermore, we as researchers can also predict fairly precisely the typical delay of a flight based on the time that the plane lands on the runway. This allows us to identify flights that are likely candidates for gaming. While we cannot observe precisely what actions airline employees undertake to reduce arrival delays, we expect that taxi-in times can be reduced through several channels. For example, airlines can preferentially allocate resources – such as ground crew and even gates – across flights. Second, it may be possible for ground crew members to reduce taxi-in times by exerting more effort. Finally, pilots can taxi planes faster even though this may lead to safety concerns.

Our empirical analysis uses the very data that is collected by the DOT under the mandatory disclosure program. In this proposal, we present preliminary results that are based on a dataset that includes a random sample of domestic flights operated by the 10 largest carriers between 1994 and 1998. The full study will include data through the latest available period (currently 2010). We exploit the fact that, starting in 1995, the DOT also began collecting information about each flight's wheels-off and wheels-on times (i.e.: the times at which it leaves the runway and touches down on the runway). These additional pieces of information allow us to construct a measure of every flight's *predicted delay* at the time that it touches down at the arrival airport. Our main set of regressions relate a flight's taxi-in time to its predicted delay and look for evidence of a reduced taxi-in times right around the 15 minute threshold. We also construct pairs of flights that land at the exact same time and investigate whether a flight's own taxi-in time depends on whether it lands at the same time as a flight with a predicted delay around the 15 minute threshold. We estimate these relationships separately for airlines with and without employee bonus programs based on the DOT rankings.

Our preliminary results suggest that airlines do indeed try to game these rankings and that this behavior is enhanced when employee bonus programs based on these rankings are in place. For airlines that do *not* have an employee bonus program in place, taxi-in times of flights that are predicted to be between 15 and 16 minutes late are almost 8 percent shorter than taxi-in times for flights that are predicted to be less than 10 minutes late. Moreover, the estimates reveal a discontinuous relationship between taxi-in times and predicted delay right around the 15 minute threshold. While one might have thought that airlines have the greatest incentive to reduce very long delays (because there may be convex costs of delay), we find that taxi-in times for the flights with predicted delays in the critical 15 minute range are significantly shorter than taxi-in times for flights with longer predicted delays.

There are five airlines that introduced employee bonus programs based on the DOT ranking during our sample period. Our preliminary analysis has estimated the relationship between taxi times and predicted delays for two of these airlines, Continental and TWA for the years 1994 to 1998. We find a similar discontinuous relationship between taxi-in times and predicted delay right around the 15 minute threshold, but the magnitudes are much larger than for the airlines without bonus programs. During the time that Continental Airlines had its bonus program in place, its taxi-in times for flights with predicted delays between 15 and 16 minutes were 13 percent shorter than its taxi-in times for flights with predicted delays of less than 10 minutes. Its flights with predicted delays between 16 and 17 minutes had taxi-times that were 15 percent shorter. We see effects of a similar magnitude when we look at TWA.

The project will explore the costs of this gaming behavior on two dimensions – whether it leads to externalities across flights and whether it distorts the information presented to consumers. With the respect to the first of these, our results so far indicate that more effort is allocated to flights that are near the threshold at 15 minutes than to flights that have very long delays. To estimate externalities across flights, we can identify flights that are not candidates for gaming but arrive at an airport at or around the same time as flights that are candidates for gaming. For the former group of flights, we can estimate whether they experience longer delays than comparable flights that do not arrive at or around the same time as flights that are candidates for gaming. With respect to the second potential cost of gaming in this context, we will carry out a series of simulations to estimate the extent to which these distortions in taxi-in times can affect an airline’s rankings. Our results so far indicate that they can. For example, we find that Continental’s gaming behavior improved its rankings by at least one position in more than a third of the months following the introduction of its employee bonus program.

The second part of the research proposal will investigate the schedule response of airlines as they become subject to the disclosure regulation. Under the regulation, flight delays are evaluated relative to the scheduled arrival time, not relative to a “reasonable” or minimum feasible flight time. Therefore, airlines can easily improve their reported quality by lengthening the time that is scheduled for each flight. The disclosure program includes no penalty for such a strategy. However, airlines that lengthen their scheduled flight times face economic costs, such as increased labor costs and reduced capacity utilization, which may limit this kind of behavior. The goal of the research project is to evaluate how airlines balance this tradeoff. We expect that the airlines’ decisions will depend on route characteristics that affect delays, especially airport congestion and the number of connecting passengers. It is worth noting that such behavior is not limited to the context of the airline industry. There are similar concerns in education, for example, where there have been claims that tests which are used to evaluate student progress have become easier in some states after the introduction of No Child Left Behind, making it easier for all schools in the state to pass the program’s required thresholds.

Airlines may also move their flights to less congested times in response to the introduction of the program. Note that this response would likely be desirable for policy-makers since it would alleviate overall airport congestion. Decomposing the possible schedule responses by airlines to the introduction of the program is necessary to inform policy-makers about whether or not these responses were desirable.

**Results from previous NSF support: NONE.**

## **II. Institutional Background**

### *II.A. Disclosure of Airline On-Time Performance*

All airlines that account for at least one percent of U.S. domestic scheduled passenger revenues have been required to submit information on their on-time performance under Title 14, Part 234 of the Code of Federal Regulations since September 1987. The reporting requirements have increased over time. Originally, airlines were only required to submit information on their scheduled and actual departure and arrival times and on flight cancellations and diversions. The original reporting requirement also did not include flights that were delayed or cancelled because of mechanical problems. The reporting rule was amended in January 1995 to cover flights with mechanical problems. The 1995

amendment also required that additional data be reported, including taxi times and airborne times, as well as the aircraft's tail number. Additional amendments to the reporting rule required airlines to include delay causes for their flights beginning in November 2002 and to report tarmac delays for flights that are subsequently cancelled, diverted or returned to their gate beginning in October 2008.

These reporting requirements cover all of an airline's flights that depart from or arrive at one of 29 reportable airports. The airlines have the option of reporting these data for all of their other flights as well and all airlines have chosen to do so. They have an incentive to report the additional data because their on-time performance on the voluntarily reported flights is generally better than it is on the flights that are subject to the reporting requirement (because the 29 reportable airports include the some of the most congested airports in the U.S.) and the voluntarily reported flights are included in the main ranking that the DOT publishes.

Airlines can record delays either manually or through an automatic device inside the aircraft. Many airlines use a combination of manual and automated recording. While the automated devices are presumably reliable in recording the actual arrival times, there is a possibility that airline employees who record flight delays manually report delays of 14 minutes for flights whose actual delays are 15 minutes. This raises the concern that what we interpret as airlines systematically exerting more effort to improve the on-time performance of threshold flights may just reflect employees lying about the arrival times of those flights. While airlines indicate each month whether they record delays manually, automatically or with a combination of the two, for those carriers that use a combination, we do not have information on which planes report automatically and which report manually. However, we can track planes by tail number and thus we can investigate whether any gaming behavior that we identify for the carriers using combination reporting is concentrated in a particular subset of aircraft (which we would presume were those without the automatic device). The two airlines for whose bonus programs we have preliminary results (Continental and TWA) do use combination reporting. Our early analysis of the distribution of gaming behavior across tail numbers does not suggest that the gaming is concentrated on a particular set of aircraft. In later years of the sample, we observe both Continental and TWA switching from combination reporting to fully automatic reporting. We will investigate if the observed gaming behavior persists after the switch to fully automatic reporting. Furthermore, the airlines that introduce bonus programs later in the sample (American, US Airways and United) all use fully automatic reporting. We will investigate if the gaming behavior we observe for Continental and TWA is similar for these other carriers.

## *II.B. Airline Bonus Programs*

In February 1995, Continental Airlines was the first airline to implement a firm-wide employee bonus program which was based on the DOT's ranking. Under the program Continental would pay \$65 to each full-time employee in every month that the airline was among the top five in the DOT's on-time performance ranking. In 1996, the program rules were changed to pay each employee \$65 in every month that the airline ranked second or third and to pay \$100 in months that the airline ranked first. The bonus program was part of a larger turnaround effort called the "Go Forward Plan" which sought to address poor performance and profitability at the airline. The two other parts of the "Go Forward Plan" which were also related to improving on-time performance were changes in the flight schedule that increased aircraft turnaround time (i.e.: the time between flights) and the replacement or rotation of the senior

manager at every airport. Thus, it is important to keep in mind that changes in on-time performance after the introduction of the bonus program may be the result of a combination of all three changes. However, we have no reason to believe that the increased turnaround time would affect flights near the 15 minute threshold differently than flights that are further from the threshold.

In June 1996, TWA implemented an employee bonus program which closely resembled Continental's. TWA would pay \$65 to each employee in every month in which the airline ranked top five in the following three rankings published by the DOT: on-time performance, baggage handling and customer complaints.<sup>1</sup> The airline would pay a total of \$100 to each employee if the airline also ranked first in at least one of those categories. The program was later amended to reward employees if high rankings were sustained for an entire quarter (instead of a single month) and, in 1999, was changed to reward absolute measures of on-time performance (85% or better during the summer months, 80% or better during the winter months) rather than relative rankings. Like Continental's program, TWA's program was introduced after a period of very poor performance. TWA ranked worst in average on-time performance in 1995 and in 1996 and its baggage handling and customer complaints had been ranking among the worst since the beginning of the DOT's disclosure program in 1987.

Three other airlines introduced similarly structured bonus programs in subsequent years. These were American Airlines in April 2003, US Airways in May 2005, and United Airlines in January 2009. The research project will extend the current preliminary analysis to these airlines. The project will also investigate whether the changes in the payment structure of each of these bonus programs had an effect on observed outcomes.

### **III. Data**

*Data to be used for Part 1 of the project:*

Our empirical analysis for this part uses the flight-level data on on-time performance collected by the U.S. Bureau of Transportation Statistics under the DOT's mandatory reporting program. We have already collected these data for all reporting carriers for every year between 1988 and 2008, inclusive. PI requests an external hard drive for back-up storage of these data and the data to be collected for Part 2 of the project. Our preliminary results are based on data for 1994 through 1998. Our primary sample includes domestic flights operated by the 10 airlines that are large enough to report on-time performance for the entire period: American Airlines, Continental Airlines, Delta Air Lines, Northwest Airlines, TWA, United Airlines, US Airways, Southwest Airlines, America West and Alaskan Airlines. Because this dataset is very large, we only include their flights between the 29 airports for which the airlines are required to report their on-time performance. To further reduce the size of the dataset, we take a random sample of flights by restricting to every fifth day of the year. In addition, we drop flights that meet any of the following conditions: depart more than 15 minutes early (since we suspect this may represent a rescheduled flight), arrive more than 90 minutes early, depart on what appears to be the following calendar day, have a taxi-out time of more than 55 minutes, have a taxi-in time of more than 25 minutes or have distance of less than 25 miles. Our final sample for the years 1994-98 includes 5,165,322 flights.

---

<sup>1</sup> The fourth ranked category, oversales, is a function of the airline's reservation system and not directly related to employee effort.

Some of the summary statistics for the main variables in the data are as follows: The average arrival delay in our sample is about seven minutes. 20% of flights in our sample arrive 15 minutes late or more and thus are considered “late” under the program’s definition. The average air time is 100 minutes, the average taxi-out time is about 14 minutes and the average taxi-in time is 5.5 minutes. Note that taxi-out time includes the time between when an aircraft leaves the gate and when it leaves the ground. Similarly, taxi-in time includes the time between when an aircraft touches the ground and arrives at the gate. Delays incurred waiting for a runway or waiting for an arrival gate will therefore be included in taxi-out and taxi-in times, respectively.

*Data to be used for Part 2 of the project:*

The second part of the project will investigate the schedule response of airlines after the introduction of the disclosure rule. The main data requirement is historical schedule data for the time period prior to and after the introduction of the program. These data are available from the Official Airlines Guide (OAG). PI has worked with this data source in previous projects, but has used a different time period. Data for this time period need to be assembled. This task will be allocated to the graduate student research assistant that has been requested. There is an alternative source for schedule data after the introduction of the program which is data collected under the program itself. We will use this data source to cross-check validity of the OAG data.

The second part will also require data on the number of connecting passengers for each route. These data are publicly available from the DOT’s Databank 1B. These data will also be assembled by the graduate student research assistant.

#### **IV. Methods and Preliminary Results for Part 1 of the project**

##### *IV.A. Overview of Empirical Approach*

This and the following sections outline our empirical approach. The graduate student research assistant will perform some of the data analysis outlined here under the PI’s guidance.

We develop an empirical approach that allows us to estimate the extent to which airlines manipulate delays around the 15 minute threshold. That is, we look for evidence of gaming, which we define as an airline systematically exerting more effort to reduce delays on flights which would otherwise arrive slightly above the threshold to be considered on-time. Our empirical approach exploits the fact that the incentive to game changes discontinuously with a flight’s expected delay. Assuming that it is on average less costly for an airline to push a flight that it expects to be 16 minutes late below the 15 minute threshold than to try to do the same for a flight that it expects to be 26 minutes late, the design of the disclosure program creates an incentive to make a greater effort to reduce delays for those flights that are near the threshold. Under convex costs of delays, this will have to be traded off against the incentive to reduce longer delays.

Before describing how we empirically identify gaming, it is useful to consider when it may take place. Delays can be occurred – and reduced - at several different points in a flight’s progression: at the departure gate, while taxiing out, in the air, or while taxiing in. The delay that is recorded upon a plane’s arrival at the gate - and which is used as the basis for classification of a flight as on-time or late -

represents the sum of delays incurred during all phases of a flight. In theory, an airline that is trying to systematically improve the on-time performance of a flight that it expects to arrive just above the threshold could try to reduce delays during any of the phases. However, we expect that airlines that are trying to game will be more likely to try to reduce delays during the later stages of a flight. This is because, as the flight progresses, the airline knows the delay that has been incurred so far and can therefore more precisely predict the total delay the flight will have. For example, when a flight is airborne, the airline knows how delayed the plane was leaving the ground but must predict both how delayed it will be in the air and how delayed it will be while taxiing in. However, once a flight has touched down at the arrival airport, the airline knows how delayed the plane was leaving the ground and while in the air and must only predict how delayed it will be while taxiing in. For any given predicted level of delay, reducing the amount of noise associated with that prediction increases the likelihood that the airline's effort at reducing a flight's delay will actually result in the flight having a shorter delay. Based on this logic, much of our empirical analysis of gaming focuses on measuring an airline's effort to reduce delays during the final phase of the flight – i.e.: when it is taxiing in to its arrival gate.

A second reason why we focus on taxi-in times is that, based on the time that a plane has landed on the runway, we can predict the flight's arrival delay at the gate fairly precisely. Thus, we can identify flights that are likely candidates for gaming. In contrast, we have much less information than the airline about wind and storm conditions in the air that help predict the flight's airborne time and, therefore, we cannot predict as well which flights would be likely candidates for gaming that occurs while the plane is in the air.

#### *IV.B. Taxi Time Analysis*

The first part of our empirical analysis estimates the relationship between taxi-in times and a flight's expected delay when its wheels touch down at the arrival airport. Intuitively, what we are trying to do is construct a measure of the delay that an airline expects a flight to have at a *given* stage in the flight's progression and then investigate whether the airline's behavior *after* this stage is related to the expected delay in a way that is consistent with gaming. Note that it is the richness of the DOT data – specifically, the fact that the program began collecting taxi-in, taxi-out and airborne times in 1995 – that allows us to do this. To construct a measure of each flight's *predicted* delay at the time that its wheels touch down at the arrival airport, we take the flight's wheels-down time and add to it the median taxi-in time for the airline-airport-month. This gives us a predicted arrival time for the flight. The difference between the predicted arrival time and the scheduled arrival time is the flight's predicted delay. Variation in predicted delay across an airline's flights at a given airport on a given day comes from differences in delays incurred prior to the planes landing at the airport.

We construct a series of dummy variables for each level of predicted delay, in one minute increments. Flights with predicted delays of greater than 25 minutes are grouped together in the top category while flights with predicted delays of less than 10 minutes are used as the excluded group. To investigate whether the employee bonus programs enhance the incentives to game that are inherent in the government program, we construct the predicted delay bins for four mutually exclusive sets of flights: (1) flights by carriers that do not have a bonus program in place; (2) flights by Continental after the introduction of its bonus program (which is introduced in the second month for which we have taxi-time

data); (3) flights by TWA before the introduction of its bonus program; and (4) flights by TWA after the introduction of its bonus program.

We estimate a flight level equation that regresses a flight's taxi-in time on these dummy variables, carrier-airport-day fixed effects and a set of control variables. Because we include carrier-airport-day fixed effects, our coefficients are estimated using variation in predicted delays across an airline's flights that arrive at a given airport on a given day. As mentioned above, this variation results from differences in the delays that flights incur prior to arrival which will largely be driven by factors at the flights' respective departure airports. Our primary interest is in testing whether those flights with predicted delay right around the critical threshold have systematically shorter taxi times than flights that are well above or below the threshold and whether this relationship is affected by the introduction of an employee bonus program. The key identifying assumption of the model is that there are no observable factors that are correlated with a flight having a predicted delay in the threshold range and that affect the flight's taxi-in time. Because evidence of gaming would come from a non-monotonic relationship between predicted delay and taxi time, we can rule out most other possible sources of correlation between predicted delay and taxi time since these are not likely to result in the same type of pattern.

The results of this analysis are presented in Table 1. Each column of the table represents the coefficients on the 16 predicted delay bins for one of the four sets of flights described above. The first column represents the coefficients for airlines without bonus programs, the second column represents the coefficients for Continental, the third column represents the coefficients for TWA prior to the introduction of its bonus program and the final column represents the coefficients for TWA after the introduction of its bonus program. The results clearly indicate that flights that are predicted to arrive just above the critical threshold have systematically shorter taxi-in times than flights with shorter or longer predicted delays. We also observe that the results for the carriers that implemented bonus programs based on the DOT rankings show a very similar pattern but the magnitudes are much larger.

We extend the analysis to consider pairs of flights by the same airline that land at the same airport at the precisely the same time. We focus on pairs in which at least one of the flights lands with an expected delay of 25 minutes or more. We construct a variable that equals one if the "late" flight (i.e.: the one that lands with predicted delay of more than 25 minutes) has a shorter taxi-in time than the "early" member of the pair. We relate this variable to the predicted delay of the early member of the pair by regressing it on the same expected delay bins used in the analysis above. Intuitively, what we are doing is estimating whether the probability that a very late flight has a shorter taxi-in time than an earlier flight that arrives at the exact same time depends on whether the earlier flight is close to the critical threshold. The benefit of this empirical exercise (relative to the regression in Table 1) is that if there is some unobservable that is correlated with both the likelihood of a flight having expected delay in the threshold range and that flight's arrival time, this unobservable should equally affect the threshold flight and the flight with which it is paired because that flight lands at the exact same time. We find that the probability of the late flight having the shorter taxi time is lowest precisely when it is paired with a flight in the critical range.

**Table 1**  
**Taxi Time as a Function of Predicted Delay**

Dependent Variable	<i>Log(Taxi In)</i>			
	<b>Coefficient Estimates for:</b>			
	<b>All Other Carriers</b>	<b>CO post-Bonus</b>	<b>TWA pre-Bonus</b>	<b>TWA post-Bonus</b>
<u>Predicted Delay</u>				
[10,11) min	-0.0333*** (0.00159)	-0.0577*** (0.00482)	-0.0538*** (0.0119)	-0.0705*** (0.0104)
[11,12) min	-0.0376*** (0.00167)	-0.0702*** (0.00517)	-0.0332** (0.0126)	-0.0615*** (0.0102)
[12,13) min	-0.0318*** (0.00175)	-0.0643*** (0.00536)	-0.0301* (0.0130)	-0.111*** (0.0108)
[13,14) min	-0.0359*** (0.00178)	-0.0781*** (0.00579)	-0.0246 (0.0151)	-0.108*** (0.0114)
[14,15) min	-0.0511*** (0.00181)	-0.101*** (0.00580)	-0.0457** (0.0141)	-0.135*** (0.0131)
[15,16) min	-0.0795*** (0.00202)	-0.133*** (0.00656)	-0.0469** (0.0160)	-0.146*** (0.0127)
[16,17) min	-0.0583*** (0.00220)	-0.151*** (0.00723)	-0.0730*** (0.0157)	-0.160*** (0.0149)
[17,18) min	-0.0401*** (0.00225)	-0.147*** (0.00825)	-0.0667*** (0.0182)	-0.186*** (0.0159)
[18,19) min	-0.0380*** (0.00233)	-0.123*** (0.00856)	-0.0661*** (0.0163)	-0.106*** (0.0169)
[19,20) min	-0.0395*** (0.00238)	-0.103*** (0.00878)	-0.0717*** (0.0165)	-0.0775*** (0.0155)
[20,21) min	-0.0425*** (0.00241)	-0.0780*** (0.00871)	-0.0382* (0.0175)	-0.0604*** (0.0149)
[21,22) min	-0.0470*** (0.00251)	-0.0619*** (0.00833)	-0.0602** (0.0191)	-0.0757*** (0.0156)
[22,23) min	-0.0414*** (0.00262)	-0.0689*** (0.00862)	-0.0511* (0.0202)	-0.0742*** (0.0154)
[23,24) min	-0.0365*** (0.00271)	-0.0801*** (0.00870)	-0.0460* (0.0188)	-0.0851*** (0.0153)
[24,25) min	-0.0412*** (0.00278)	-0.0569*** (0.00954)	-0.0450 (0.0242)	-0.0783*** (0.0185)
>25 min	-0.0448*** (0.00108)	-0.0534*** (0.00281)	-0.0757*** (0.00916)	-0.0841*** (0.00697)

*Notes:* Standard errors are in parentheses and are clustered at the level of the arrival airport-day. Columns display coefficients from a single regression of taxi time on four sets of predicted delay “bins” that are defined to be mutually exclusive. Specification includes carrier-arrival airport-day fixed effects and arrival hour and hub controls. Coefficients represent the percentage change in taxi time relative to the taxi time for flights with predicted delay of less than 10 minutes. Calculation of predicted delay is described in the text on page 13. The regression contains 4,143,402 observations.

#### *IV.C. Does it Work?*

The results in Table 1 suggest that airlines are trying to improve the on-time performance of specifically those flights that would otherwise arrive just above the threshold for being on-time. Next, we investigate whether they are successful in doing so. We regress a dummy variable that equals one if a flight arrives one minute earlier than predicted on the same expected delay dummies and controls as in Table 1. We also do the same regression for a dummy that equals 1 if the flight arrives two minutes earlier than predicted. Again, we are looking for a discontinuous relationship right around the relevant threshold, and we find clear evidence of this. For carriers without a bonus program, the probability of arriving 1 minute earlier than predicted is 7 percentage points greater for flights predicted to be 15 to 16 minutes delayed than for the comparison category of flights that are predicted to be less than 10 minutes delayed. For Continental and TWA during their bonus programs, the effect is estimated to be 10 percentage points. All these results are statistically significant. We find smaller positive coefficients of 2 percentage points for flights that are predicted to be 11 to 12 minutes and 21 to 22 minutes late, which we believe reflect rounding by manually reporting carriers. All other coefficients, including the one for long delays over 25 minutes, are less than 1 percentage point in magnitude and most of them are statistically insignificant with small standard errors. We find a very similar pattern for the probability of arriving two minutes earlier than predicted, with a large positive coefficient for flights that are predicted to be 16 to 17 minutes late, and much smaller and mostly insignificant coefficients for flights with different predicted delays. Here, too, the effects are larger for carriers that have bonus programs in place.

#### *IV.D. Implications/Costs of Manipulation*

##### *i. Misallocation of Effort Across Flights*

Assuming that there are convex costs of delays, the effort allocated to reducing delays on threshold flights could be more efficiently allocated to flights with long delays. We will investigate whether the resources allocated to threshold flights are diverted from other flights or whether these resources would otherwise be slack. For example, one of the ways in which taxi-in times could be reduced for threshold flights is by speeding up the process by which the plane is directed to the gate by ground crew. Intuitively, we would like to be able to distinguish whether a flight's taxi-in time is reduced because the ground crew that was allocated a different flight is allocated to the threshold flight or because the crew originally allocated to the threshold flights works harder/faster. To explore these issues, we will look at flights by the same airline that arrive at the same airport around the same time (though not necessarily in the exact same minute). We will test whether flights that arrive around the same time as a threshold flight experience an increase in total delays.

##### *ii. Distortion of Information Conveyed to Consumers*

To investigate whether the distortions in taxi-in times that we find in our regression analysis can actually impact airlines' overall on-time performance and DOT rankings, we perform a counterfactual simulation that estimates what arrival delays and rankings would be absent gaming. To do this, we take the following approach. Our data suggest that taxi-in times are distributed approximately log-normal. We calculate the mean and variance of the log taxi-in time for each carrier-airport-month. Then, for each flight in our data, we replace the actual taxi-in time in the data with a random draw from a log-normal distribution with the mean and variance for the appropriate carrier-airport-month. The idea behind this

exercise is to replace a flight's taxi-in time with the taxi-in time it would likely have absent any incentive for the airline to systematically reduce taxi-in times on threshold flights. After doing this exercise for every flight in our data, we can recalculate the fraction of flights that are 15 or more minutes delayed. This leads to counterfactual measures of on-time performance for each airline and these can be used to create counterfactual rankings of airlines. Repeating the simulation a number of times yields standard errors for our simulated on-time performance measures.

In preliminary results, we find that, averaging across months, the difference between actual and simulated on-time performance for Continental is about one full percentage point – that is, the distortion in taxi-in times results in the fraction of flights delayed 15 minutes or more falling by one percentage point. The difference is about 1.3 percentage points for TWA after it introduces its program. These changes in the fraction of delayed flights directly map into changes in rankings. For example, when we simulate Continental's taxi-in time but leave the others carriers' behavior unchanged, we find that the taxi time distortions result in Continental achieving an improvement in rankings of at least one position in 22 of the 36 months following the introduction of their program. When we simulate Continental as well as all other airlines' taxi-in times, we find that the taxi-time distortions result in Continental achieving an improvement in rankings in 8 of the 36 months. Thus, the results of the simulation exercise indicate that while a 40 second reduction in delay may be small in absolute value (and in terms of the disutility to consumers), when applied to flights that are close to the relevant threshold, the impact on reported rankings can be significant.

## **V. Methods for Part 2 of the project**

Part 2 of the project will estimate if airlines responded to the introduction of the disclosure program by changing their flight schedules. We will assemble flight schedules before and after the introduction of the program. We will use the information on flight numbers, departure and arrival airports and departure and arrival times that is contained in the OAG data in order to create a panel data set that links flights over time. Then, we will compare scheduled flight times (defined as the time between scheduled departure and scheduled arrival) for the same flight over time. We will regress the scheduled flight time on a dummy for the flight itself, a dummy for the time period after the introduction and controls. The estimated coefficient on the dummy on the post-period will show if airlines increased their scheduled flight times after the introduction of the program, and by how much. Controls in the basic regression specification will include route and airport characteristics. Most important among these are measures of airport congestion, such as the number of flights scheduled per hour divided by the airport's capacity (see, e.g., Forbes and Lederman (2010)), and the number of direct and connecting passengers typically travelling on the route (available from the DOT's Databank 1B). We will also interact these controls with the dummy for the post-period in order to test how airlines traded off the incentives to lengthen schedules against the cost of doing so. For example, airlines might have increased scheduled flight times most on low-traffic routes because the costs of doing so were relatively low, whereas the benefit of improving on-time performance was equal across all routes, since all flights enter the DOT on-time statistic separately. We will also include month dummies in the regression to control for seasonality.

Preliminary analysis of the OAG data that we already have (these data are for different years than the ones that will be used for this project) shows that the majority of flights can be linked in a panel based on the information on flight numbers, departure and arrival airports and departure and arrival times.

Flights that cannot be linked are ones that are newly added or eliminated. There are also cases in which an airline maintains service on a route but substantially changes the departure/arrival time for a flight. We will not include such flights in our linked panel. This is because we are interested in detecting changes in flight times that are due to the introduction of the program. If flights are moved to different times of day, the airline might choose to change the scheduled flight time for other reasons, such as different levels of congestion at different times of day. Such effects would confound our analysis of the response to the program.

A second part of the analysis will link routes, but not individual flights, over time in order to investigate this very effect that airlines may choose to move their flights to different times of day in response to the introduction of the program. For each flight, we will construct a measure of the arrival airport's congestion during the hour that the flight is scheduled to arrive. In robustness checks, we will narrow this measure down to the half hour or 15 minute block during the flight is scheduled to arrive, and we will also look at the departure airport's congestion during the flight's scheduled departure. We will then regress this measure of congestion on route-level fixed effects, airline fixed effects, a dummy for the post-period, and controls. The estimated coefficient on the dummy for the post-period will reveal if airlines chose to schedule their flights during less congested times in response to the introduction of the program. Note that such a response could be desirable for policy-makers because it would reduce overall airport congestion. In contrast, if airlines simply responded to the program by increasing their scheduled flight times but not by scheduling at less congested times of day, this would likely not be desirable for policy-makers. Nevertheless, both responses would result in better on-time performance being reported under the DOT's program. It is thus important to decompose these possible responses in order to inform policy-makers about whether this program had desirable effects on airlines' behavior.

## **VI. Contributions**

This project will contribute to the economics literature by investigating gaming of public disclosure programs in the context of airline on-time performance. To our knowledge, this will be the first paper to investigate the link between gaming of public disclosure programs and firm-level bonus programs that reward employees for engaging in such gaming behavior.

There are existing (but until now separate) literatures in economics on gaming of quality ratings and on gaming of employee bonus programs. In the former, we can distinguish three groups of papers: (1) Papers that investigate gaming in the form of selection of better risks (e.g., Haney (2000), Deere and Strayer (2001), Dranove, Kessler, McClellan and Satterthwaite (2003), Jacob (2005), Werner and Asch (2005), Cullen and Reback (2006), Figlio and Getzler (2006)). These papers have focused on applications in hospitals and education. (2) Papers that investigate gaming in the form of moving effort from reported to unreported margins (Jacob (2007) in education and Lu (2009) in nursing homes). (3) Neal and Schanzenbach (forthcoming) investigate gaming by focusing on a critical threshold. This paper is closest in spirit to our work. We contribute to this literature by providing a setting in which counterfactuals can be constructed more precisely than in the settings that have previously been investigated. For example, Neal and Schanzenbach use a student's third-grade test score as a proxy for the student being near the achievement threshold in fifth grade, whereas we compare flights by the same airline arriving at the same airport on the same day (or even during the same minute) to one another.

Oyer (1998), Courty and Marschke (2004) and Larkin (2007) have previously studied gaming of employee bonus programs. We contribute to this literature by investigating a novel setting in which gaming is actually desired by the firm because it improves the firm's reported quality rating. Settings that have previously been studied were ones where gaming was not in the firm's best interest because it distorted employees' actions and by doing so led to potentially lower profits for the firm.

Knez and Simester (2001) have previously studied the effects of Continental's bonus program. They find that it resulted in a significant improvement in on-time performance measured by the fraction of flights that depart less than 15 minutes late. Their paper explore why one might observe such an improvement despite the fact that individual employees have an incentive to free-ride on others. The paper does not focus on the gaming of the disclosure program.

Foreman and Shea (1999) have previously studied the introduction of the airline on-time performance disclosure program. Their work focuses on the stock market response to the publication of on-time information. They do not investigate the airlines' schedule response.

## **VII. Summary**

### **Intellectual Merit:**

The main intellectual contributions of the project will be the following: (1) Compared to previous studies, our setting provides a more convincing counterfactual of what the quality would have been absent gaming behavior. Specifically, we will compare the on-time arrival of flights that are close to the threshold to flights that are identical on observables (in a broad definition: flights by the same airline arriving at the same airport, on the same day; in a narrower definition: flights by the same airline, arriving at the same airport in the same minute), except that they are further away from the threshold. Because of the program's design, the incentives for gaming are highly non-linear, whereas confounding factors tend to have a linear impact on on-time arrival. This allows us to rule out many alternate explanations for the observed behavior. (2) We will investigate the extent to which overall quality is lowered by schedule-padding in response to the program's introduction. (3) We will be the first to study the interaction between the gaming incentives inherent in many quality disclosure programs and financial incentives for employees that directly increase their incentives for gaming. (4) Because we can provide a convincing counterfactual, we can also give a convincing estimate of the externalities from gaming and of the extent to which the information provided by the program is distorted by firms that are gaming the program.

### **Broader impact:**

Quality disclosure programs of the nature studied in this project are widely used by the government and by non-government certification agencies. Gaming of these programs is always a concern, and – as our preliminary results suggest – even more so when employees are paid based on information collected under such programs. One prominent example is education, where schools are evaluated on “Adequate Yearly Progress” (AYP) since the No Child Left Behind legislation. The current “Race to the Top” encourages teacher evaluation and potentially bonus pay based on AYP or other student test score measures. Results of this study can inform policy makers about potential unintended consequences of such policies, such as substantially increased gaming of the program when employees receive financial incentives that are based on the program's rating scheme.