Do Firms Game Quality Ratings? Evidence from Mandatory Disclosure of Airline On-Time Performance

Silke J. Forbes
University of California, San Diego

Mara Lederman
University of Toronto, Rotman School of Management

Trevor Tombe
University of Toronto

July 2011

Abstract

Many quality disclosure programs provide consumers with information that is based on whether a product meets a particular threshold. This creates the potential for "gaming" as firms have incentives to improve the quality of specifically those products that can easily be brought above the threshold. We investigate this type of behavior in the context of government-mandated disclosure of airline on-time performance. While this program collects data on the actual minutes of delay incurred, it ranks airlines based only on the fraction of their flights that arrive 15 or more minutes late. This creates the incentive for airlines to selectively reduce delays on flights they expect to arrive with about 15 minutes of delay. We estimate the extent to which airlines engage in this type of gaming and, in particular, whether the occurrence of such gaming depends on whether employees are explicitly incentivized based on the airline's performance in the program. We find little evidence of gaming by airlines that have no incentive programs in place or by airlines that have implemented incentive programs with targets that are unrealistically hard to achieve. On the other hand, we find strong evidence of "gaming" by airlines that have incentive programs with a target level of performance that can realistically be achieved. Specifically, for these airlines, we find that their flights that are predicted to arrive with between 15 and 16 minutes delay have significantly shorter taxi-in times than other flights and are significantly more likely to arrive exactly one minute sooner than predicted. Counterfactual exercises that simulate an airline's distribution of delays in the absence of taxi-time distortions indicate that even small improvements in taxi times can – if applied to the "right" set of flights – result in changes in an airline's ranking.

We thank Severin Borenstein, Bob Gibbons, Matt Mitchell, Steve Puller and seminar participants at Boston University, New York University (Stern), UC Berkeley (Haas), UC San Diego, the University of Maryland, the University of Toronto (Rotman), the U.S. Department of Justice, the AEA Meetings (2011), the Berkeley-Stanford IO Fest (2010) and the USC Conference on Game Theory in Law, Business and Political Economy for helpful comments.

I. Introduction

Disclosure programs exist in many industries in which consumers are imperfectly informed about product quality.[1] While the growing empirical literature on these programs has generally found that they result in improvements in product quality, firms also appear to engage in various types of behavior that attempt to "game" the disclosure scheme. For example, when only some dimensions of quality are reported, firms may substitute effort from unreported to reported margins (see Jacob 2005 and Lu 2009); when quality depends on consumer characteristics, firms may choose to serve only a select set of consumers (see Dranove *et al*. 2003 and Werner and Asch 2005); and when the disclosure program is based on a particular quality threshold, firms may focus on improving the quality of those products that can most easily be brought over the threshold (see Neal and Schanzenbach 2010). The growing body of evidence on gaming implies that, in addition to considering the cost, precision and usefulness of the information being provided, the design of an optimal disclosure program must also consider the potential for firms to game the program. However, anticipating the potential for gaming - as well as the type of gaming - depends not only on the design of the program but also on the characteristics of the product and the internal organization and incentive schemes of the firm. For example, gaming may depend on how product quality can be manipulated and whether those in a position to manipulate it have incentives to influence the information reported.

In this paper, we explore the relationship between gaming of a disclosure program, the design of the program and the incentive schemes in place at the firms covered by the program. Our setting is the U.S. airline industry. Since 1987, airlines have been required to report to the Department of Transportation (DOT) the scheduled and actual arrival times of their domestic flights. Although the DOT collects detailed data about the actual minutes of delay incurred on

[1] See Dranove and Jin (2010) for a review of the literature on disclosure programs.

each flight, it only counts a flight as being "late" if it arrives 15 minutes or more behind schedule. The DOT issues monthly reports that rank airlines based on the percentage of their flights that are late under this definition and excerpts from these rankings are frequently reported in media outlets.[2] The design of this program clearly creates the potential for gaming as airlines have an incentive to reduce delays on specifically those flights that would otherwise arrive just over 15 minutes late. Small reductions in delay on these flights - which can likely be made at low cost - can improve an airline's performance in the DOT rankings even though they may not necessarily improve its overall on-time performance.

Two features of this setting make it a particularly interesting one in which to investigate how gaming behavior is affected by characteristics of the product as well as organizational features of the firm. First, airlines cannot predict in advance which flights will be candidates for gaming. While airlines may be able to anticipate which routes or flights will, on average, have longer delays, they are unlikely to be able to anticipate which flights will arrive with exactly 14 versus 16 minutes of delay. Thus, to the extent that gaming occurs, it occurs in real time. As a result, the effort to game must come from front-line airline employees rather than executives or managers. This makes a consideration of employee-level incentives particularly relevant.

Second, between 1995 and 2009, five different airlines implemented employee bonus programs based explicitly on the airline's performance in the government's ranking of on-time performance. Under these programs, each airline employee would receive a payment of between $65 and $100 in any month in which the airline as a whole placed at or near the top of the DOT ranking. While all of the programs created a free-rider problem by rewarding individuals based on firm-level performance, the programs differed significantly in how easy it was to achieve the target ranking. Thus, this empirical setting – combined with the richness of the flight-level data

[2] These rankings are published in the DOT's "Air Travel Consumer Report", which also contains separate rankings of airlines based on baggage handling, oversales, and customer complaints.

available - allows us to investigate not only the existence of gaming but also explore where and when it occurs and whether it is affected by the incentives provided to the employees most likely to engage in the gaming behavior.

We develop an empirical approach that allows us to estimate whether airlines systematically try to reduce delays on flights which would otherwise arrive slightly above the 15 minute 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 has a fairly precise estimate of the expected delay that the flight will have and can decide whether or not to try to reduce that delay below 15 minutes. We expect that taxi-in times can be reduced in several ways – for example, by preferential allocation of scarce resources such as ground crew, by employees exerting more effort and, in some cases, by simply lying about a plane's actual arrival time. While we cannot observe what actions airline employees take to reduce delays, we devise an empirical strategy to try to distinguish between lying and actual speeding up of planes that exploits the fact that, during our sample, some airlines reported their on-time performance manually while other reported automatically.

Our empirical analysis uses the very data that is collected by the DOT under the mandatory disclosure program. We construct a dataset that includes a random sample of domestic flights operated by the seven largest carriers between 1995 and 2010. We take advantage of the fact that, starting in 1995, the DOT 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). This additional 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 non-

monotonicity right around the 15 minute threshold. We also estimate whether flights that are predicted to be 15 (16) minutes late are systematically more likely than any other flights to arrive exactly one (two) minute(s) earlier than predicted. We estimate these relationships for airlines without incentive programs in place and, separately, for each airline that introduced an incentive program. We focus the analysis, for now, on the seven large network carriers who were initially covered by the reporting requirements.

Our empirical analysis does not show evidence of gaming by airlines without employee bonus programs in place. However, we find strong evidence of gaming by the first two of the five airlines that introduced these types of incentive programs - Continental Airlines (in 1995) and TWA (in 1996). During the first three years of its bonus program, Continental's taxi-in times for flights predicted to be between 15 and 16 minutes late were about 14 percent shorter than its taxi-in times for flights with predicted delays of less than 10 minutes. We see effects of a very similar magnitude when we look at TWA who also introduced a bonus program during this period. Moreover, the estimates for Continental and TWA 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 the costs of delays may be convex), 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. We also find – for both of these carriers – that their flights that we predict to be exactly 15 (16) minutes late are much more likely than any other flights to arrive exactly one (two) minutes sooner than predicted. When we investigate whether this gaming appears to reflect lying or actual reduction in taxi-in times, we find evidence for both. When we carry out the same series of analysis for the three airlines that introduced bonus programs after 2000, we find no evidence of gaming. We suspect that this is due to the much

weaker incentives provided by these programs.[3] While the two early programs rewarded their employees if the airline was among the top five of the 10 airlines that were ranked at the time, the three later programs only rewarded employees if the airline achieved first or, in some cases, second place out of a much larger number of airlines that were, by that time, included in the rankings. Some of these airlines – for example, Hawaiian Airlines – consistently had substantially better on-time performance than any of the large network carriers.

In addition to the literature on gaming of quality disclosure programs, this paper is also related to research on gaming of employee incentive programs, such as Oyer (1998), Courty and Marschke (2004) and Larkin (2007). Finally, this work is related to Knez and Simester (2001) which has studied the effect of one of the airline employee bonus programs on the airline's overall delays. Knez and Simester show that overall departure delays decreased after the introduction of Continental's bonus program, but they do not investigate the gaming of the disclosure program which is the focus of our paper.

The rest of the paper is organized as follows. Section II provides institutional background on the government disclosure program and on the airline bonus programs. Section III describes our data and sample. We outline our empirical approach in Section IV and present our results in Section V. A final section concludes.

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 to the

[3] Some of the differences across firms may also be due to differences in communication. For example, Continental's bonus program was introduced during a time period in which management explicitly communicated on-time performance as an important goal for the organization. We are in the process of investigating the communication strategies of the other carriers.

Department of Transportation 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 to 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.[4]

Airlines can record delays either manually or automatically through technology that is installed in the aircraft. While the automated devices are presumably reliable in recording the actual arrival times, there has been speculation that airlines which record delays manually may not record their arrival times accurately. Indeed, the distribution of arrival delays for manual reporters shows considerable rounding. Our empirical analysis below focuses on the seven largest network carriers. While we believe that most of these airlines reported their delays

[4] The DOT's report also contains a separate ranking based only on the reportable airports, but this ranking is not as highly publicized as the main ranking.

automatically during our sample period, several – including the two which implement the early employee bonus programs – likely used a combination of manual and automatic reporting.[5] This raises the possibility that airline employees who record flight delays manually may report delays of 14 minutes for flights whose actual delays are 15 minutes. While this would appear in our data as shorter taxi-in times for these flights, this would not reflect extra effort or preferential allocation of resources but rather would reflect employees lying about arrival times. Since this represents a different type of gaming, we have developed an approach (described below) for trying to identify the manual aircraft in the data and look separately at gaming behavior on manually and automatically reported flights.

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.[6] 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

[5] Starting in 1998, we know how each carrier reports in each month (automatic, manual or combination). Since our analysis covers the period between 1995 and 1998, we cannot be certain that the manner in which carriers reported in 1998 is the same as how they reported in the earlier years. However, anecdotal evidence and descriptive analysis of their delay distributions suggest they likely did.

[6] In 1994, Continental had the worst average on-time performance ranking among the ten reporting airlines.

performance after the introduction of the bonus program may be the result of a combination of all three changes. While we have no reason to believe that the increased turnaround time would specifically reduce delays on flights near the 15 minute threshold, increased emphasis within the organization on meeting the DOT's on-time target could enhance the effect of the explicit incentives provided by the bonus program.

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.[7] 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. Table 1 summarizes the details of these bonus programs and the airlines' on-time performance one year before and after the introduction of their programs. The table also reports the number of months during the first year after the introduction of the bonus program in which the employees in fact earned bonuses. There are a number of things that are interesting to note

[7] The fourth ranked category, oversales, is a function of the airline's reservation system and not directly related to employee effort.

about this table. First, all of these airlines except American improved their ranking in the first year after the introduction of their bonus program, relative to the year prior. However, the accompanying improvement in on-time performance (i.e.: the percentage of flights delayed less than fifteen minutes) varies quite a bit, from less than one percentage point for TWA to almost ten percentage points for United.

Second, while the two earlier bonus programs by Continental and TWA, made it relatively easy for employees to earn bonuses by rewarding any placement in the top 5 spots of the ranking, the three later programs only rewarded first and (for American and United) second places which made it substantially harder for employees to earn bonuses. In fact, American's and US Airways' employees did not earn a single bonus in the first year after the introduction of their programs, and United's employees had only a single month in which they earned a bonus. In contrast, Continental's and TWA's employees earned bonuses in ten and four months, respectively, during the first year after the introduction of the program.

Another factor which appears to substantially affect the chance that an airline's employees might earn a bonus is the number and type of airlines included in the rankings. Until 2002, there were ten airlines that accounted for more than one percent of domestic passenger revenues and which were therefore included in the DOT's ranking. After 2001, the combination of growth by low-cost and regional carriers and reductions in capacity by the large network carriers led to an increase in the number of carriers that met the DOT reporting requirements. By January 2003, there were 17 airlines included in the ranking. Moreover, Hawaiian Airlines was added as an 18th carrier in November 2003. Since then - in every single month that we have looked at (including all of 2005 and all of 2009) - Hawaiian has always occupied the top spot in the ranking, typically with a substantial lead over the second-ranked airline. We suspect that this may have to do with the fact that Hawaiian operates at relatively uncongested airports with few

weather disruptions. This combination of factors means that the three later bonus programs may have had much more negligible effects on the incentives of employees than the two earlier programs because employees of American, US Airways and United may have been aware that their likelihood of achieving a bonus was quite small.

III. Data

III. A. Data and Sample

Our empirical analysis 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 collected these data for all reporting carriers for every year between 1988 and 2008, inclusive. Our empirical work below focuses on the years 1995 to 1998, 2003 to 2006, and 2008 to 2010 since these are the years during which the airlines introduced their employee bonus programs.[8] Because of the volume of data, we cannot investigate all five bonus programs in a single sample that includes data from the 15 years over which these programs are introduced. As a result, we construct separate samples which include several years of data around the introduction of the programs.

Our regression sample includes domestic flights operated by the following seven airlines: American Airlines, Continental Airlines, Delta Air Lines, Northwest Airlines, TWA, United Airlines, and US Airways. 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

[8] 1995 is also the year in which the DOT began collecting data on wheels-up and wheels-down times and we require this particular data for our empirical analysis.

flight), arrive more than 90 minutes early, depart on what appears to be the following calendar day, have a taxi-out or taxi-in time of more than 60 minutes, have missing values for their scheduled arrival or departure times, have a distance of less than 25 miles, or operate fewer than 20 times during the quarter. Our final sample includes 3, 067,533 flights.

Table 2 presents summary statistics for the main variables in the data.[9] The average arrival delay in the sample is about seven minutes. About 21% of flights arrive 15 or more minutes late and thus are considered "late" under the program's definition. The average air time is 109 minutes, the average taxi-out time is about 15 minutes and the average taxi-in time is 6 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.

III. B. Histograms of Arrival Delays

Figure 1 shows the distribution of arrival delays for the seven network carriers in our regression sample as well as the three other carriers that met the DOT's reporting requirements during our initial sample period. These three additional carriers are Southwest Airlines, America West and Alaska Airlines. We truncate the histogram at -20 on the left and at 60 on the right. The histogram reveals a distribution of delays that peaks at 0. The histogram is fairly smooth but shows discrete increases at certain values. As the next set of histograms will show, these discrete increases appear to reflect rounding by carriers who report their delay data manually. It is interesting to note that the spikes generally occur at five minute intervals (e.g. at -5, 0, 5, 10, etc…); however, instead of there being a spike at 15 minutes, the histogram shows a spike at 14

[9] These are based on the 1995 to 1998 sample.

minutes.[10] This could either reflect rounding (or lying) by carriers who report manually or effort by airlines to systematically reduce delays on flights that would otherwise have delays just above the threshold.

In Figures 2A through 2C, we compare the distribution of arrival delays for carriers who report their delays in different ways. Since we only know an airline's reporting type with certainty beginning in March 1998, these histograms only show delays for flights between March and December 1998. Figure 2A shows the distribution of arrival delays for American Airlines, Northwest Airlines, United Airlines and US Airways – all of which reported fully automatically during this period. Their histogram is smooth with a peak around -5 and no apparent spike at 14 minutes. Figure 2B shows the distribution of arrival delays for Southwest Airlines, Alaska Airlines and American West – all of which reported their on-time data manually during this period. This histogram is much less smooth, has a large spike at zero (with almost 10% of flights arriving with exactly zero minutes delay) and suggests that these airlines are rounding their delays at the five minute intervals (i.e.: 0, 5, 10, etc…). However, rather than a spike at 15 minutes – which would be consistent with the pattern – the histogram shows a spike at 14 minutes. Finally, Figure 2C shows arrival delays for Continental, Delta and TWA – the three airlines that used a combination of manual and automatic reporting. This histogram is quite smooth and looks much more like the histogram of the automatic reporters than the histogram of the manual reports – suggesting these airlines were likely reporting most of their data automatically. The histogram for these carriers - which includes the first two airlines to introduce an employee bonus program based on the DOT ranking - shows a distinct spike at 14 minutes.

[10] Much of this pattern is driven by Southwest Airlines, which schedules its flights to arrive on "the 5s" and appears to report many of its delays in five minute intervals.

In Figures 3A and 3B through Figures 7A and 7B, we compare the distribution of an airline's arrival delays before and after it introduces an employee bonus program. We do this for each of the five airlines that we observe introducing such a program. Figures 3A and 3B show arrival delays for Continental in the two years before and two and a half years after the introduction of its employee bonus program.[11] These histograms suggest a marked increase in the number of flights that arrive exactly 14 minutes late and a decrease in the number of flights that arrive 15 or 16 minutes late after the introduction of the bonus program. Figures 4A and 4B plot analogous histograms for TWA. These figures show a very similar pattern. After the introduction of TWA's program, there is an obvious discontinuity in its distribution right around the relevant threshold, with 14 minute delays being more than twice as likely as 15 minute delays. For both Continental and TWA, the difference in the percentage of flights delayed 14 minutes compared to 15 minutes is much larger after the introduction of the bonus program than before and also much larger than any other difference observed anywhere else in their distributions.

Figures 5A and5B plot the arrival delay distribution for American Airlines one year before and one year after the introduction of its bonus program. The figures show a very small discontinuity around the 15 minute mark which is much less pronounced that the discontinuity in the first two sets of histograms. The analogous figures for US Airways and United Airlines before and after the introduction of their programs show no apparent in the relative heights of the bars at 14 and 15 minutes.

IV. Empirical Approach

IV.A. Overview of Empirical Approach

[11] We add data from 1993 for this histogram so that we can have two years of pre-bonus program data.

We define gaming as a systematic effort by an airline to reduce delays on specifically those flights that it expects to arrive with a delay of just over 15 minutes.[12] To empirically identify gaming, we need to be able to do two things. First, we need to be able to identify flights that an airline expects to be close to the 15 minute threshold. These flights are the most likely candidates for gaming since they are the flights that can presumably be brought below the threshold at the lowest cost. Second, we need to be able to measure whether the airline actually reduces delays on these flights below what they would otherwise have been. This requires a counterfactual measure of what a flight's delay would have been absent any incentive for gaming.

We believe that both of these requirements are met particularly well in our setting. Because our data allow us to observe the various stages of each flight – departure from the gate, take-off from the departure runway, landing on the arrival runway, and arrival at the gate – we can construct a flight's expected delay at each stage and, at any given stage, we can identify those flights whose expected delay is close to 15 minutes. We can then investigate whether – in *subsequent* stages of the flight - airlines attempt to reduce delays on specifically those flights that were expected to be around 15 minutes late. Furthermore, we have several ways of controlling for the counterfactual delay that these flights would have had in the subsequent stages absent the airline's incentive to game. First, we can look at flights just outside the critical threshold. That is, at a given stage of a flight, we can assume that – absent incentives to game – subsequent delays on flights that had expected delays of 15 minutes should be similar to subsequent delays on flights with expected delays of, say, 12 or 18 minutes. Second, we can compare flights with expected delays in the critical range to flights with very long expected delays (which we define

[12] The manipulation we focus on here is on effort spent in real-time (i.e.: once a flight is in progress) to reduce delays. This is distinct from manipulation that may occur in advance through what has been termed "schedule padding" – increasing schedule times for the purpose of appearing to be on-time.

to be delays over 25 minutes). If the costs of delays are convex, then the airline should have the greatest incentive to reduce delays on those flights. If we find that airlines make more effort to reduce delays on flights that they expect to arrive close to the 15 minute threshold than on flights that they expect to arrive with very long delays, this would strongly suggest that there is gaming. It is also worth pointing out that, in our setting, the flights that are candidates for gaming – i.e.: whose predicted delay is right around the critical 15 minute mark – will be identified in real-time and will vary from day to day. This means that airlines cannot engage in ex ante behavior that aims to reduce delays on those flights that it expects to arrive right around 15 minutes late since this is simply not known by the airline in advance. This eliminates selection concerns when comparing flights that are candidates for gaming to their "control groups" of flights outside the threshold.

IV.B. Taxi Time Regressions

Before describing our regression analysis in detail, it is useful to consider at what stages of a flight gaming may take place. Delays can be occurred at any of the stages of a given 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 15 minute threshold could try to reduce delays during any of the phases. However, we expect that airlines 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 therefore can 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, our empirical analysis focuses on estimating 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 – as a function of its *predicted delay* at the time that it touches down at the arrival airport.[13]

It is the richness of the DOT data and, in particular, the fact that in 1995 it began collecting information on wheels up and wheels down times which allow us to construct a fairly precise predicted delay measure. Specifically, to construct each flight's predicted delay at the time that its wheels touchdown, we take the flight's wheels-down time and add to it the median taxi-in time for that flight in the quarter.[14] 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. For example, consider a flight by Delta Air Lines between Boston and Atlanta in March of 1997. Suppose that is has a scheduled arrival time of 4:30 pm. If its wheels-down time is 4:36 pm and Delta's median taxi-in time for this flight in this quarter is 4 minutes, then the flight's predicted arrival time is 4:40 pm and its predicted delay is 10 minutes.

We then construct a series of dummy variables for each level of predicted delay, in one minute increments. For example, we construct a dummy variable that equals one if a flight's predicted delay is greater than or equal to 10 minutes and less than 11 minutes. We construct another dummy variable that is equal to one if a flight's predicted delay is greater than or equal to 11 minutes and less than 12 minutes. And so on. Flights with predicted delays of greater than

[13] In addition, focusing on taxi-in times has the advantage that it minimizes the number of stages of a flight's progression that we need to predict thus eliminating noise from our measure of predicted delay. For example, were we to calculate a flight's predicted delay at the time that it departs from the ground, we would need to estimate both its airborne time as well as its taxi-in time.

[14] We identify a flight as a unique combination of airline, flight number, departure airport and arrival airport.

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. Thus, we define 16 different predicted delay "bins". 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 separately for airlines without bonus programs in place and for each airline with a program in place and, where possible, distinguish between the years before and years after its program was in place. Thus, for the 1995-1998 sample which covers the first two bonus programs, we construct predicted delay bins for four mutually exclusive sets of flights: (1) flights by the five carriers in our data that do not have a bonus program in place during the time period; (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. This means that we have a total of 64 mutually exclusive dummy variables in these models.

We estimate a flight level equation that regresses a flight's taxi-in time, in logs, on these 64 dummy variables, carrier-airport-day fixed effects and a set of control variables which includes a dummy for the departure airport being a hub, controls for two distance categories (500-1500 miles and greater than 1500 miles), and dummies for each (actual) arrival hour. One can think of the model as estimating four vectors of 16 parameters, one for each of the four groups of flights defined above. Within these vectors, each coefficient represents the change in the log(taxi-in time) for flights in a given predicted delay bin relative to the taxi-in time for flights with predicted delay of less than 10 minutes. 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. 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 and in the air. Our primary interest is in testing whether those flights with predicted delay right around the critical 15 minute threshold have systematically shorter taxi times than flights that are 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 non-monotonic pattern.

V. Results

V.A. Taxi-Time Regressions

Our main set of taxi-time results are presented in Tables 3A and 3B. Table 3A shows the results for the two early bonus programs while Table 3B shows the results for the three later programs. Each column of the table represents the coefficients on the 16 predicted delay bins for a particular set of flights. We begin by describing the results in Table 3A. 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. In order to look for evidence of gaming, we perform three hypothesis tests for each group. Specifically, we (separately) test if the coefficient on the 15-16 minute bin is significantly larger in magnitude than the coefficients for the 12-13, 18-19 and 25 and over bins, respectively.

The results in the first column show no evidence of gaming by airlines that do not have bonus programs in place. Flights that are predicted to arrive just above the critical threshold have about 3.5% shorter taxi-in times than flights that are predicted to be less than 10 minutes late; however, flights at every higher level of predicted delay also have taxi-in times that are between 3.5%-5% shorter than those for flights with predicted delays of less than10 minutes. Our hypothesis tests show that the coefficient on the 15-16 minute bin is significantly larger in magnitude than the coefficient on the 12-13 minute bin, but it is significantly smaller in magnitude than the coefficient on the 25 minute and over bin and not significantly different from the coefficient on the 18-19 minute bin.

In contrast, the results for the first two carriers that implemented bonus programs show a different pattern. Looking first at Continental Airlines, its flights with predicted delays of 15 to 16 minutes have taxi-in times that are 14 percent shorter than those of flights with predicted delay of 10 minutes or less. Its flights with predicted delays of 16 to 17 minutes have taxi-in times that are about 14.5 percent shorter. Moreover, the coefficients indicate a non-monotonic relationship between taxi-in times and a flight's predicted delay. Flights with predicted delays above or below the critical range have much smaller coefficients (i.e.: longer taxi-in times) than flights that are within the critical range. All three of our hypothesis tests indicate that the coefficient on the 15-16 minute bin is larger in magnitude than the other coefficients we test it against. Given an average taxi-in time of about 6 minutes, the coefficients we estimate for flights in the critical range translate into average reductions in taxi-in times of about 50 seconds. While this magnitude may appear small, our simulations below reveal that these selective reductions in delay can add up to meaningful changes in on-time performance.

The estimates for TWA after the introduction of its bonus program show a very similar pattern for flights near the 15 minute threshold, with magnitudes that are slightly larger than

those estimated for Continental. While we cannot reject equality of the 15-16 minute and the 18-19 minute coefficients, we find that the 15-16 minute coefficient is significantly larger in magnitude than both the 12-13 and the 25 minute and over coefficients. Since TWA's program was introduced in 1996, we are able to separately estimate the relationship for TWA before and after its program is in place. As the third column of the table indicates, we see no systematic evidence of gaming by TWA prior to the introduction of its program. Figures 8A and 8B contain plots of the coefficients for Continental and TWA after their programs are in place. The non-monotonic relationship is very apparent in these plots.

Table 3B shows the results for the airlines that introduced bonus programs in 2003 and later. In the first two columns we show the results for American Airlines and US Airways after they introduced their bonus programs (estimated on the 2002 to 2006 sample). The third column shows the results for United Airlines after it introduced its program (estimated on the 2008-2010) sample. As above, we also include predicted delay dummy variables for these airlines pre-bonus as well as for the other carriers that did not introduce bonus programs during this period. However, because of space constraints, we only present the post-bonus results in the table. None of the columns show any indication that these programs resulted in gaming as we have defined it. The coefficients on predicted delay bins in the threshold range are very similar in magnitude to or smaller than the coefficients on predicted delay bins above the critical range. In the case of United's program, there is no evidence that taxi-in times for flights in the critical range are any different than taxi-in times for flights that are predicted to be less than 10 minutes late. Thus, while we find strong evidence of gaming following the introduction of Continental's and TWA's bonus programs, we do not find similar evidence of gaming following the introduction of American's, US Airways' and United's programs. As described earlier, we suspect that this is due to the fact that the three later provides provided much weaker incentives to employees as the

programs were structured in such a way that the likelihood of actually earning the bonus was quite low.

V.B. Does it Work?

The results in Table 3A 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. In Tables 4A and 4B, we investigate whether they are successful in doing so.[15] We do this by estimating the probability that flights with predicted delay between 15 and 16 minutes arrive exactly one minute early and compare this to the probability that flights with other levels of predicted delay arrive exactly one minute early. Again, we are looking for a discontinuous relationship right around the relevant threshold. Since our predicted delay measure is not necessarily an integer but the actual delay variable in the data is, we define a flight as arriving exactly one minute earlier than predicted if its actual delay is the integer below its predicted delay (e.g.: a flight that is predicted to have 17.6 minutes of delay would be considered to arrive exactly one minute early if its actual arrival delay was 16 minutes). 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 3A.

The results are presented in Table 4A. As before, each column displays the 16 coefficient estimates for one of the four different groups of flights and we run three separate hypothesis tests for each of these groups to look for evidence of gaming. Consistent with the results presented in Table 3, the estimates in the first column of Table 4A do not suggest gaming by airlines without bonus programs. The results for Continental and TWA in columns 2 and 4, respectively, are again consistent with efforts to systematically reduce delays on flights that

[15] Given that the results Table 3B – as well as the raw data in the histograms presented above - suggest that the later programs did not induce gaming, we restrict our subsequent empirical analyses to Continental and TWA programs.

would otherwise arrive around the threshold for being considered on-time. For Continental and TWA, after the introduction of their bonus programs, their flights with predicted delays between 15 and 16 minutes are 11 percentage points and 9 percentage points, respectively, more likely to arrive exactly one minute earlier than predicted, relative to their flights with less than 10 minutes of predicted delay. For both of these carriers, no other level of predicted delay has a coefficient that is in this range.

In Table 4B, we re-estimate this regression using (as the dependent variable) a dummy variable that equals one if a flight arrives exactly two minutes earlier than expected. The results of this exercise are again consistent with these two airlines attempting to systematically reduce delays on flights that would otherwise arrive just above the threshold for being on-time. For both Continental and TWA, flights that are predicted to be between 16 and 17 minutes late (i.e.: arrive 2 minutes after the cutoff for being considered on-time) are more than 13 percentage points more likely to arrive two minutes sooner than predicted than flights with predicted delay of less than 10 minutes. This effect is again substantially larger than it is for flights with any other level of predicted delay. Note that the results in Tables 4A and 4B are also consistent with what is observed in Continental's and TWA's histograms after they introduce their bonus programs – an increase in the fraction of flights that arrive exactly 14 minutes late.

V.C. Manual vs. Automatic Planes

All of the results presented so far indicate that, after introducing their employee bonus programs, Continental and TWA systematically try to reduce delays on those flights that might otherwise arrive right around the 15 minute threshold. However, as discussed in Section II, we believe that, during our sample period, both of these airlines had some number of aircraft that reported on-time data manually. This raises the possibility that what we are measuring as shorter

taxi-in times are simply airline employees lying about the arrival times of flights that would have arrived 15 or 16 minutes late.[16] This would still represent a form of "gaming" of the incentive program; however, it would represent a different type of gaming than actual reductions in taxi-in times. In addition, the welfare implications would be different.

The fact that the histograms for Continental and TWA look much more similar to the histograms for the automatic reporters than the histograms for the manual reporters suggests that most of these two airlines' planes are likely to be reporting automatically. However, we have also developed an approach that tries to identify specifically which aircraft may be reporting manually. We exploit the fact that we can track planes in our data by tail number. We look for evidence that some of the planes of combination reporters appear to be rounded in a way that is similar to how the manual reporters appear to round their delays at zero. Specifically, for each aircraft in each year of our data, we calculate the fraction of its flights in that year that have a reported arrival delay of zero. We then compare the distribution of this plane-year level variable across airlines which report their on-time data in different ways.

Table 5 shows the distribution of this variable for all 10 airlines who report to the DOT in 1996. The 99th percentile of the distribution of this variable for American Airlines – which we expect reported fully automatically in 1996 – is 0.0509 which indicates that only about 1 percent of American's planes were reported to arrive with a delay of zero minutes more than 5% of the time. In contrast, for America West which was a manual reporter during this time, 50% of its planes landed with a reported delay of zero more than 5% of the time. Southwest is clearly an outlier here with the 50th percentile of its distribution being 11.72%, far higher than any other airline's. If we compare Continental and TWA to the carriers that we expect are fully automatic

[16] In our data, taxi-in times are calculated as the difference between arrival times and wheels down times. As a result, given a plane's wheels down time, if its arrival time at the gate is recorded as one minute earlier than it actually was, this would appear in our data as a one minute shorter taxi-in time.

in 1996, we see that TWA's distribution is very close to the automatic reporters while Continental's planes are more likely than the automatic reporters to have reported delays of zero. Based on this table, we categorize planes that have reported delays of zero more than 5% of the time to be manual planes.

Using this approach for identifying manual aircraft, we re-estimate our earlier regressions with separate predicted delay bins for Continental's and TWA's manual and automatic planes. This allows us to investigate whether the patterns we estimated earlier were being driven by planes that we suspect reported manually and where lying may be taking place. Rather than present the results of this exercise in additional tables, we present plots of the coefficients of interest. The coefficients from the taxi-time regression are presented in Figures 9A and 9B while the coefficients from the "arrive 1 minute early" regression are presented in Figures 10A and 10B. The coefficients in these figures show that the non-monotonic relationship between taxi-in times and predicted delays exists for both manual and automatic planes. However, the pattern is more pronounced and the difference in taxi-in times for threshold flights is greater for manual planes. Our hypothesis tests again suggest evidence of gaming for Continental and TWA after the introduction of their bonus programs. Of the six hypotheses that we test, the only hypothesis test we reject is that the 15-16 minutes coefficient for TWA is greater than its 18-19 minute coefficient.

We have tested the robustness of our definition for identifying manual planes by using an alternative definition which is based on rounding of flight delays throughout the distribution, not just at zero. Specifically, we compute the percentage of a plane's flights during a year that have a reported arrival delay that is either equal to 0 or is equal to a number that falls on the five minute intervals, excluding 15. Based on the distribution of this variable for automatic reporters, we define planes as manual if their flights are reported to arrive with a delay of zero or a multiple

of five more than 20 percent of the time. This alternative definition has a strong overlap with the definition based zero delay and the results are robust to using this alternative definition.

As a final check of our main definition of manual planes, we have tested it on Continental's planes in the period after Continental had switched to fully automatic reporting of delays. We find that our definition identifies about three percent of Continental's automatic planes as "manual" during that time period which is similar to the fraction of planes that arrive with zero delay more than five percent of the time for the automatic reporters on which the definition was based.

V.D. Analysis of Paired Flights

The results in Table 3A clearly suggest that airline employees are systematically shortening taxi-in times for flights that arrive close to the 15 minute threshold. The identification strategy used in those regressions exploits variation in delays incurred prior to arrival across a carrier's flights arriving at the same airport on the same day. While this identification strategy should be fairly convincing given that it is difficult to think of an unobservable factor that would be correlated with predicted delays and generate the particular relationship between predicted delays and taxi-in times that we find, we nonetheless carry out an additional analysis of taxi-in times that controls even more carefully for possible unobservable factors that may lead to differences in taxi-in times across flights. Specifically, we consider pairs of flights by the same airline that land at the same airport at the precisely the same time.[17] 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

[17] The BTS data rounds arrival times to the nearest minute. Thus, we can only be certain that tied arrivals do not deviate in their true arrival times by more than one minute.

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 that 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 earlier regressions) 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 taxi-in time when it arrives, 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.

The results of this exercise are presented in Table 6. Each column again presents the coefficients for one of the four groups of flights that we distinguish. Each coefficient represents the probability that the "late" member of the pair has a shorter taxi time than the "early" member of the pair when the "early" member's expected delay is in the particular bin. The coefficients are relative to the probability that the "late" member has a shorter taxi time when it is paired with a flight with predicted delay less than 10 minutes. The estimates for Continental indicate that, relative to when the late flight lands with a flight that is predicted to be less than 10 minutes late, there is a significant reduction in the probability of the late flight "winning" when it lands at the exact same time as a flight that is predicted to be 14 to 15 or 15 to16 minutes late. While it is reasonable to expect that the probability that the late flight wins falls with the expected delay of its pair, one would expect to observe a monotonic relationship and this is not what the results show. 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. Interestingly, TWA's flights show this pattern prior to the introduction of its bonus program but not after. We are in the process of investigating what may be driving this result for TWA. Perhaps operational changes or changes in scheduling at its

hub (where we are most likely to observe more than flight land at the same time) are influencing taxi-in times.

V.E. Externalities (Preliminary Results)

We have also begun an analysis of whether the selective reduction of delays on threshold flights imposes externalities on other flights. Such externalities will occur if – as a result of gaming – scarce resources are reallocated from other flights to threshold flights. They will not occur if airlines simply lie about the delays of threshold flights – as we suspect may happen with the manual planes – or if gaming is achieved through higher levels of effort from slack resources (e.g., ground crew). We should also point out that any externalities that do occur from reallocation of scarce resources will be inherently difficult to detect. This is because resources are scarce during times when the carrier has many flights arriving at the airport, but any threshold flight may only affect a small number of these flights. Since we do not know which of these flights will be affected (and it is likely that it is not always the same flight that is affected) and since we do not know whether arrival or departure delays are affected – we necessarily have to look for average effects which will be hard to detect.

Based on our analysis of this so far, we have not been able to uncover any externalities beyond the effects on paired flights described above. We have run regressions where we relate a flight's arrival delay to the percentage of other flights by the same carrier that arrive at the same airport during the same 15 minute time block that are threshold flights. We do not find that increases in the fraction of threshold flights that land within the 15 minute window that a flight lands has an effect on that flight's arrival delay.

V.F Additional Results and Robustness Checks

We have explored the robustness of our results to two alternative ways of estimating the taxi-in time that is used to calculate a flight's predicted delay. Specifically, instead of computing the median taxi time for a given flight in a given quarter, we have computed the median taxi-in time for a carrier at a given airport in a given month as well as the median taxi-in time for a carrier at a given airport in a given month during arrival time window. The results are robust to these alternative ways of calculating a flight's expected delay.

We have also re-estimated our regressions on a few subsamples of the data in order to explore whether the results differ across these samples. First, we have created separate samples for flights that arrive at a carrier's hub and flights that do not arrive at a hub. We find evidence of gaming by Continental and TWA in both subsamples. We also find that flights with long expected delays have shorter taxi-in times (relative to flights with expected delays under ten minutes) in the hub sample than in the non-hub sample. This is consistent with the fact that long delays are more costly at hubs, where many more passengers make connections.

Second, we have created subsamples of flights that arrive at times of day where congestion at the arrival airport is above and below the median, respectively. Depending on whether the primary mechanism through which gaming occurs is the reallocation of scarce resources (during congested times) or a higher level of effort from otherwise slack resources, such as ground crew (during uncongested times), gaming may either be more or less prevalent for flights during congested times, compared to flights during uncongested times. We find evidence of gaming in both subsamples for Continental, but only for flights during uncongested times for TWA, suggesting that, for TWA, the primary source of gaming is a higher level of effort from slack resources. Finally, we have explored whether there may be end-of-the-month effects – specifically, whether gaming takes place at the end of months in which the airline is close to achieving the necessary ranking for a bonus payment, but not at the end of months in

which the carrier is far away from achieving that target. Similar types of effects have been found in the prior literature on employee bonus programs. Note that, in order for such effects to occur in our setting, employees would have to be informed not only about their own airline's overall on-time performance in the month so far, but also about the on-time performance of all other carriers. The Department of Transportation only releases this information with a two-month lag, so that the information would have to come from other sources. We find no evidence of end-of-the-month effects, which suggests that airline employees may not have the necessary information to distinguish the months in which the airline is close to achieving the bonus target from months in which it is not.

V.G Simulation of Rankings

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.

We report results from the counterfactual exercises in Tables 7A and 7B. Table 7A shows simulated changes in on-time performance and ranking for Continental in the months after the introduction of its bonus program. Table 7B shows the same thing for TWA. 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' behaviour unchanged, we find that the taxi time distortions result in Continental achieving an improvement in rankings of at least one position in 19 of the 35 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 45 to 55 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.H Are There Any Real Effects of the Bonus Programs?

The results so far suggest that part of the improvements in Continental's and TWA's on-time performance after the introduction of their bonus programs resulted from gaming behaviour. One might question whether these programs - and the other operational and/or managerial changes that accompanied them - resulted in any actual improvements in on-time performance.

Using a very different empirical strategy and different data than us, Knez and Simester (2001) investigate the impact of Continental's program and 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. Since airlines have no incentive to manipulate departure delays, these results would indicate an actual improvement in on-time performance.

In Appendix A we estimate the relationship between the introduction of the bonus programs and several different measures of on-time performance. Using our sample of flights from 1994 to 1998, we estimate a flight-level regression that includes airline and arrival-airport date fixed effects. To estimate whether on-time performance differed after the introduction of the bonus program, we interact the Continental dummy with a variable that equals one in months in which its bonus program is in effect. We do the same with the TWA dummy. Time trends are captured in a very flexible way by the arrival-airport date fixed effects. The results indicate that, in the months after the introduction of its bonus program, Continental's mean arrival was lower by about 2.4 minutes, its likelihood of arriving 15 or more minutes late fell by about 4.8 percentage points, its taxi-in times were on average 0.6 minutes shorter, its departure delays were 1.8 minutes shorter and its taxi-out times were not changed. The results for TWA are roughly similar.

There are a couple of interesting things to note from this table. First, there is evidence of at least some real improvement in on-time performance. Continental's flights are, on average, departing 1.8 minutes less delayed after the introduction of the program. TWA's departure delays are about one minute shorter. Second, the estimates in also suggest the presence of gaming. Specifically, one can take the estimated change in arrival delays – 2.4 minutes – and apply it equally to all of Continental's flights in 1994 (i.e.: reduce each flight's delay by 2.4 minutes). Based on this, one would predict that the fraction of flights with delays of 15 minutes

or more would fall by about 3 percentage points, which is less than the 4.8 percentage points estimated in the second column of the table. The same is true for the estimates on TWA's program. Thus, the findings in these fairly descriptive regressions are consistent with the findings from the more nuanced analysis above.

VI. Conclusion

Prior research has shown that while disclosure programs may induce firms to improve product quality, there is also considerable effort by firms to game the schemes under which they are rated. As a result, those designing disclosure programs must try to anticipate the potential for a given scheme to be gamed. However, the potential for gaming of a disclosure program will depend not only the structure of the program but also on the characteristics of the product being rated and the incentives in place at the firm. In this paper, we have begun to explore these issues in the context of airline reporting of on-time performance. While the structure of this program creates obvious incentives for airline to game by selectively reducing delays on flights that would otherwise arrive with 15 minutes of delay, those flights cannot be identified in advance and so gaming must take plane in real-time. Whether such gaming will take place depends on whether those individuals who are in the position to reduce delays on select flights have incentives to do so.

Our empirical analysis finds no evidence of gaming by airlines without explicit employee bonus programs in place and no evidence of gaming by airlines with bonus programs that set targets that cannot realistically be achieved. On the other hand, our empirical analysis finds very strong evidence of gaming by the two airlines who introduced bonus programs with targets that could plausibly be achieved. We find that those airlines have systematically shorter taxi-in times for their flights that are predicted to arrive with 15 or 16 minutes of delay. These flights are also

much more likely to end up arriving with exactly 14 minutes of delay. Our analysis suggests that some of this represents lying about planes' arrival times while some represents actual reductions in taxi-in times. While the effects we estimate translate into about 50 second shorter taxi-in times, our simulations show that applying this reduction in taxi-in times to the "right" set of flights can result in meaningful changes in the rankings which is what the bonus programs are based on. This paper contributes the growing empirical literature on gaming of disclosure programs by explicitly considering how gaming is affected by the incentives provided to employees who are in a position to carry out the gaming.

References

Courty, P. and G. Marschke (2004), "An Empirical Investigation of Gaming Responses to Explicit Performance Incentives." *Journal of Labor Economics* 22: 23-56.

Dranove, D. and G. Jin (2010), "Quality Disclosure and Certification: Theory and Practice", *Journal of Economic Literature*.

Dranove, D., D. Kessler, M. McClellan, and M. Satterthwaite (2003) "Is More Information Better? The Effects of 'Report Cards' on Health Care Providers." *Journal of Political Economy* 111: 555-88.

Jacob, B. (2005), "Accountability, Incentives and Behavior: Evidence from School Reform in Chicago," *Journal of Public Economics*, 89(5-6): 761-796.

Jacob, B. (2007), "Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments", *NBER Working Paper* 12817.

Kenz, M. and D. Simester (2001), "Firm Wide Incentives and Mutual Monitoring at Continental Airlines," Journal of Labor Economics, 19(4): 743-772.

Larkin, I. (2007), "The Cost of High-Powered Incentives: Employee Gaming in Enterprise Software Sales." *Unpublished manuscript*, Harvard Business School.

Lu, Susan F. (2009), "Multitasking, Information Disclosure and Product Quality: Evidence from Nursing Homes", *Working Paper*, University of Rochester (Simon School of Business).

Neal, D. and D. W. Schanzenbach (2010), "Left Behind by Design: Proficiency Counts and Test-Based Accountability", *Review of Economics and Statistics* 92(2), 263-283.

Oyer, P. (1998), "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality." *Quarterly Journal of Economics* 113:149-85.

Werner, R. and D. Asch (2005), "The Unintended Consequences of Publicly Reporting Quality Information," *Journal of the American Medical Association,* 293(10):1239-44.

Figure 1
Distribution of Arrival Delays
Ten Largest U.S. Carriers, 1994-1998

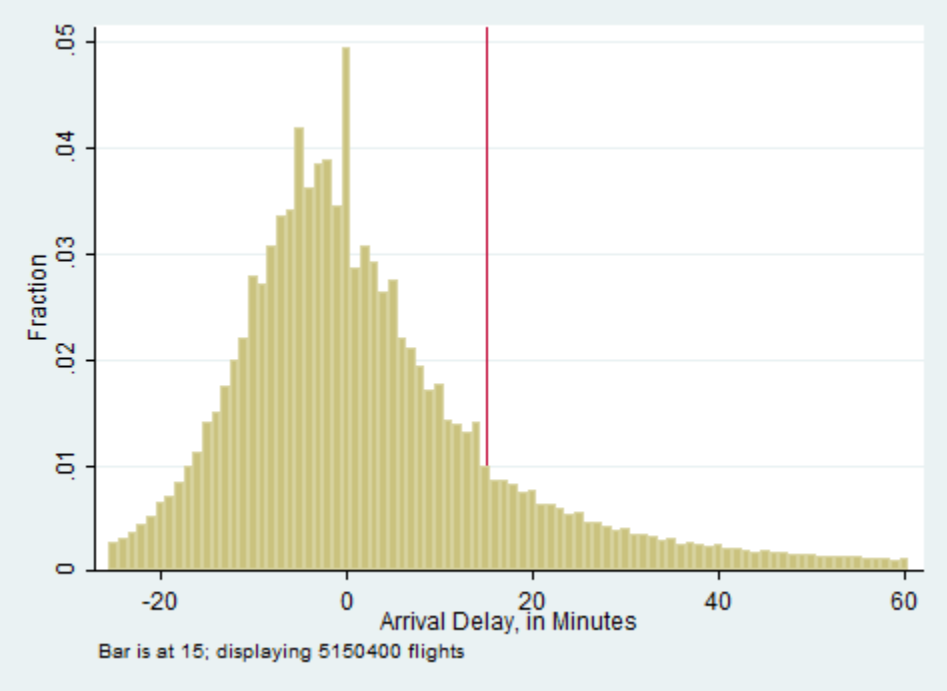

Figure 2A
Distribution of Arrival Delays
Fully Automatic Reporters, March – December 1998

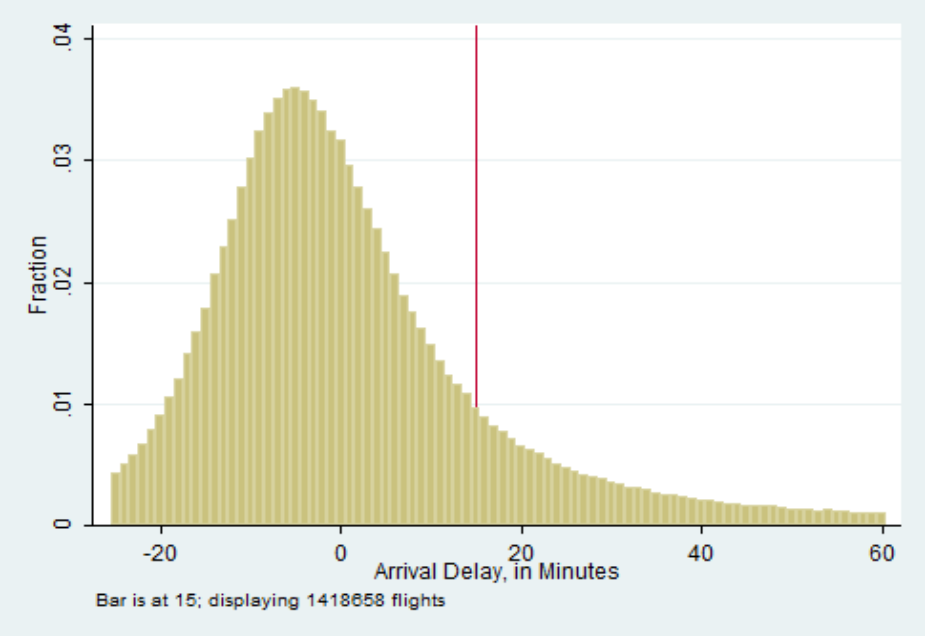

Figure 2B
Distribution of Arrival Delays
Manual Reporters, March – December 1998

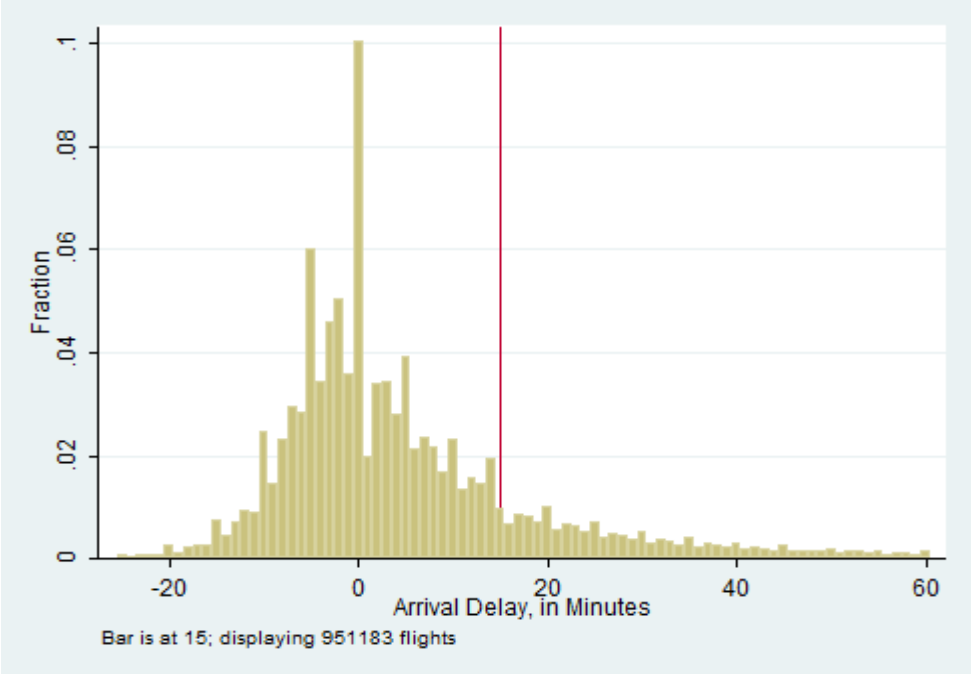

Figure 2C
Distribution of Arrival Delays
Combination Reporters, March – December 1998

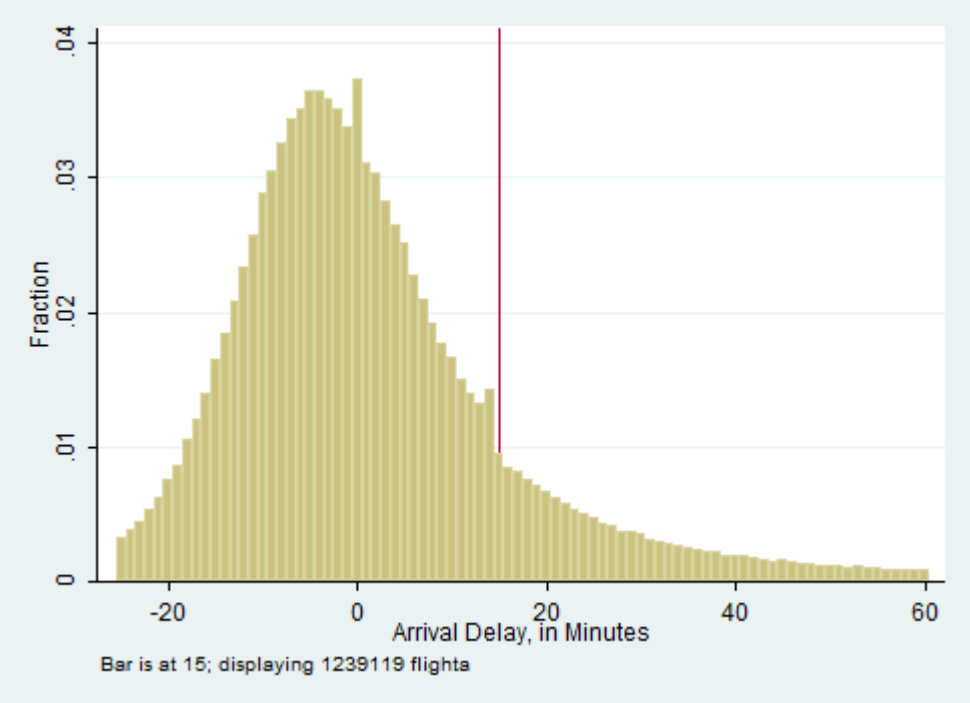

Figure 3A
Distribution of Arrival Delays
Continental Airlines, 1993-1994

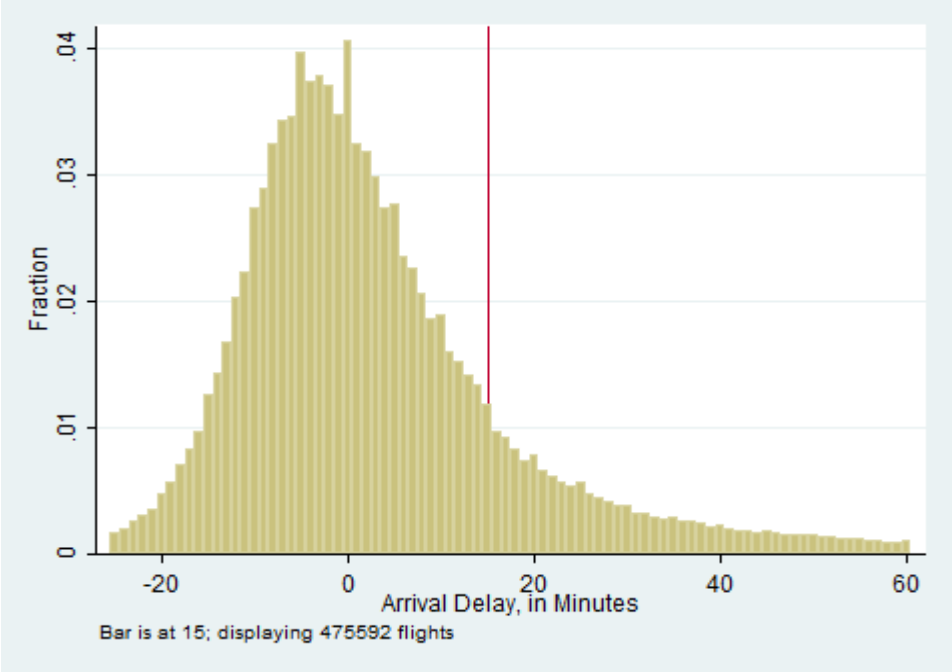

Figure 3B
Distribution of Arrival Delays
Continental Airlines, February 1995-1997

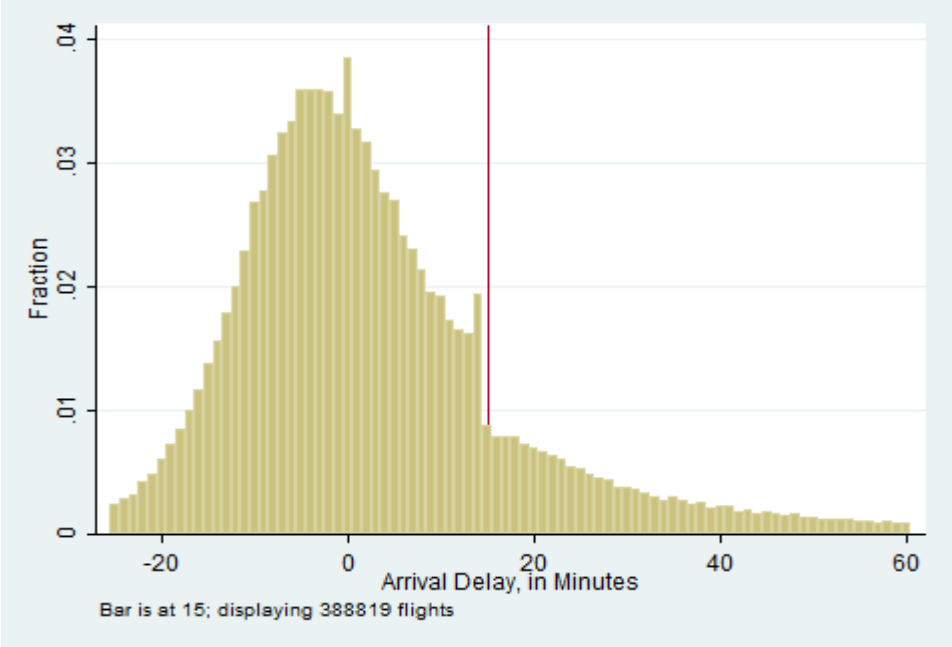

Figure 4A
Distribution of Arrival Delays
TWA, 1994-1995

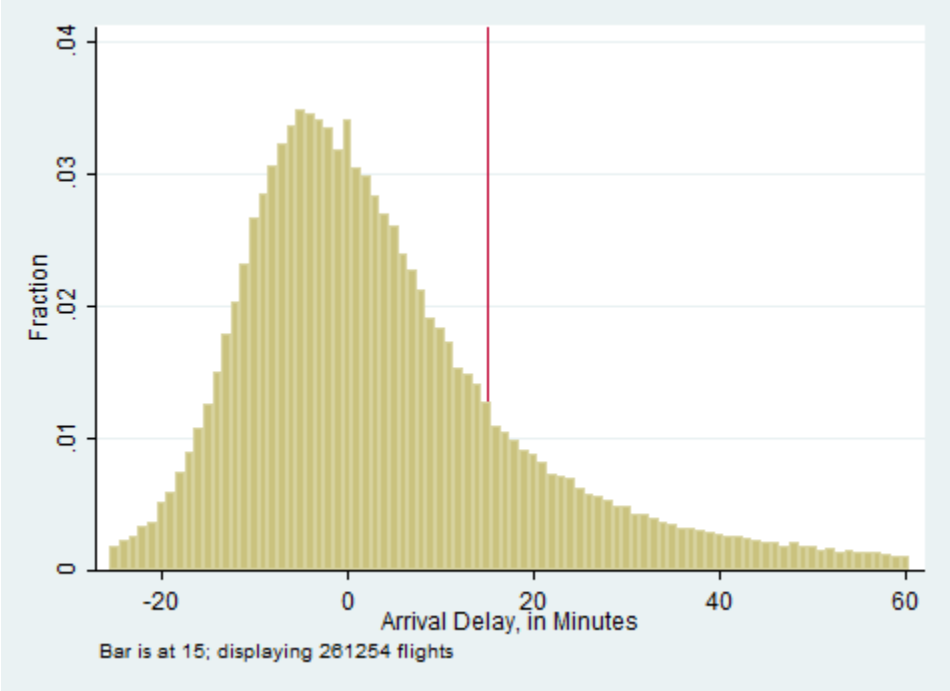

Figure 4B
Distribution of Arrival Delays
TWA, June 1996-1998

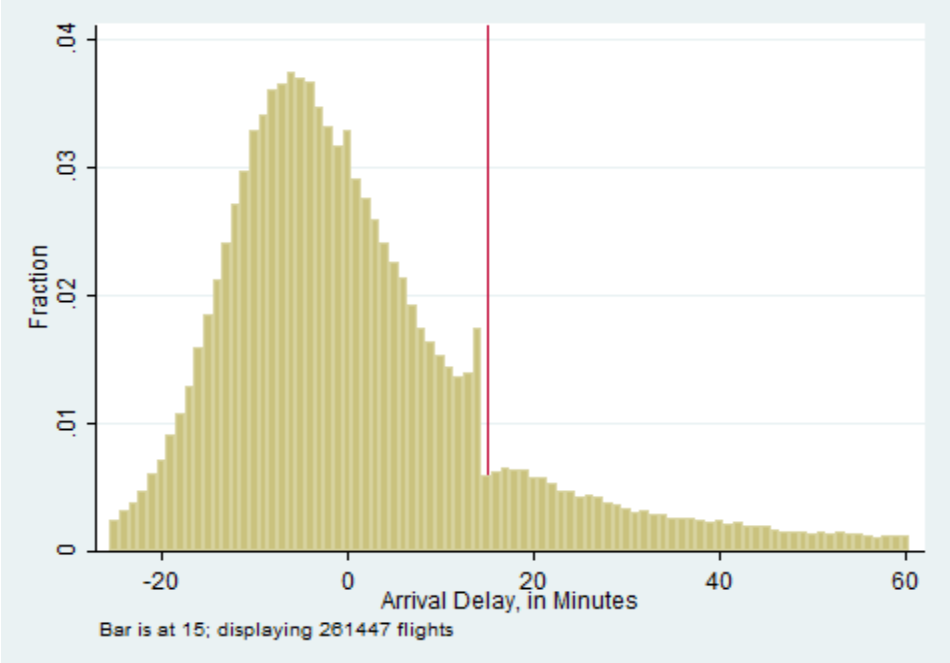

Figure 5A
Distribution of Arrival Delays
American Airlines, 2002

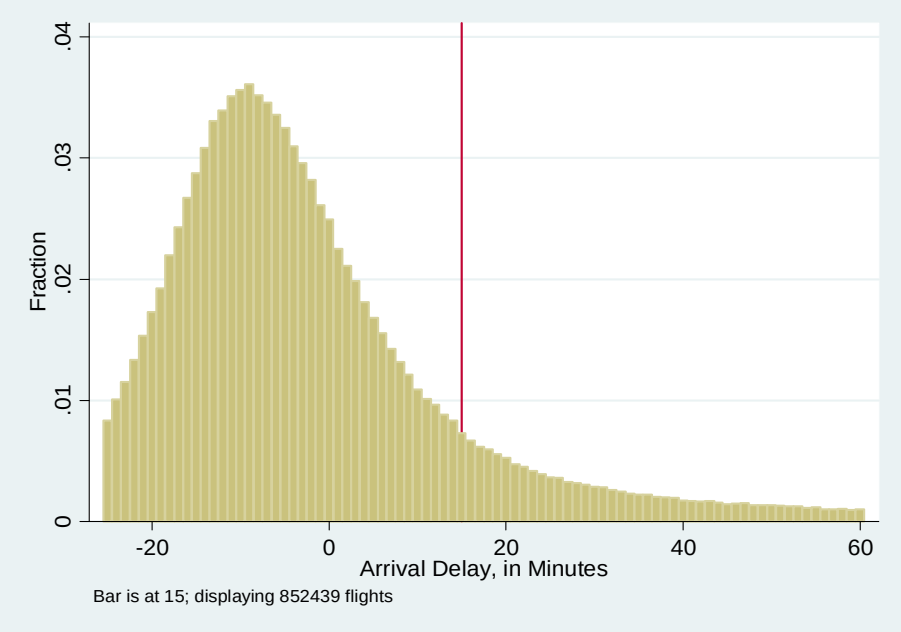

Figure 5B
Distribution of Arrival Delays
American Airlines, 2003

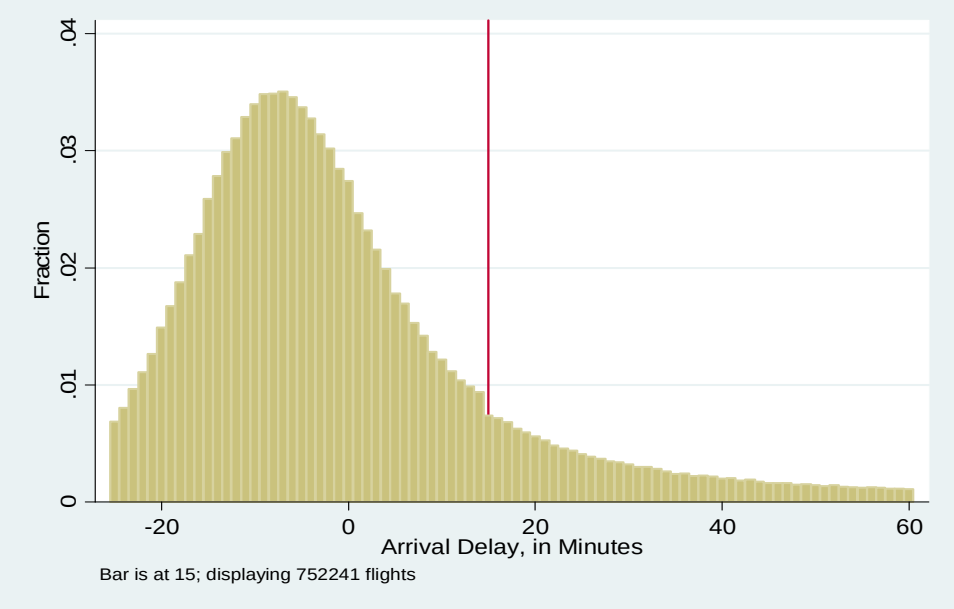

Figure 6A
Distribution of Arrival Delays
US Airways, 2004

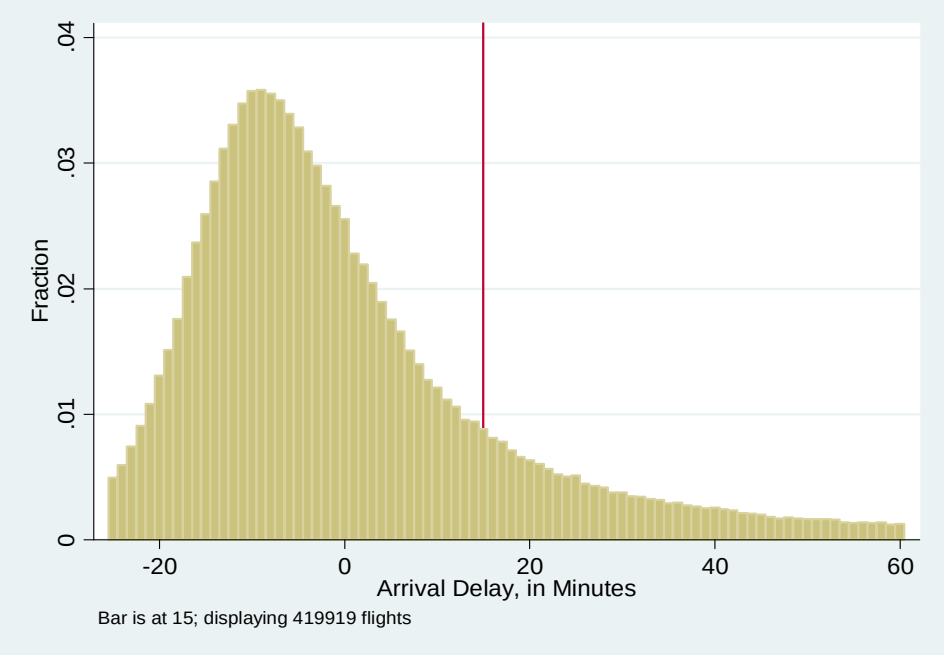

Figure 6B
Distribution of Arrival Delays
US Airways, 2004

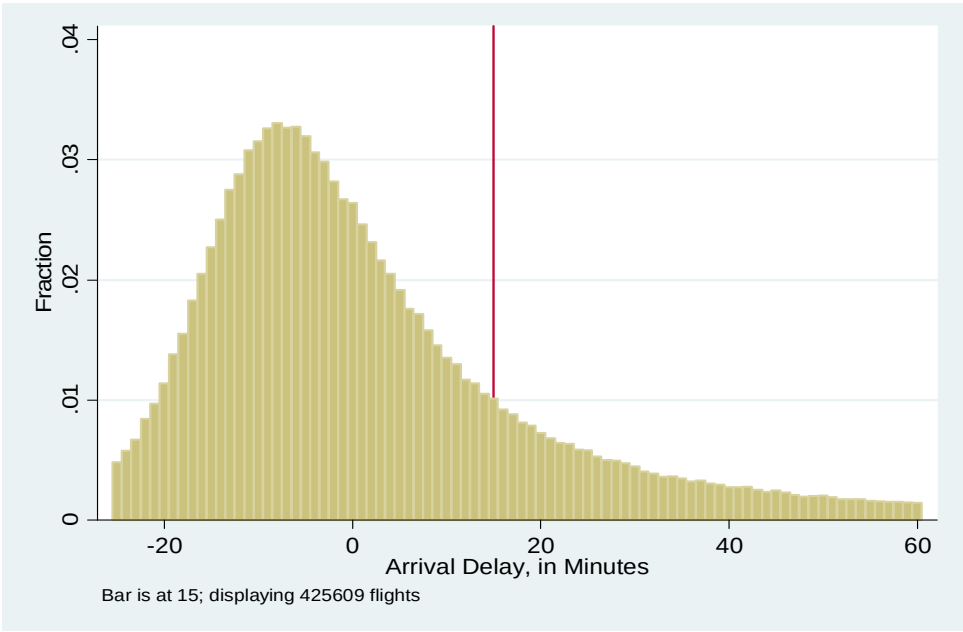

Figure 7A
Distribution of Arrival Delays,
United Airlines, 2008

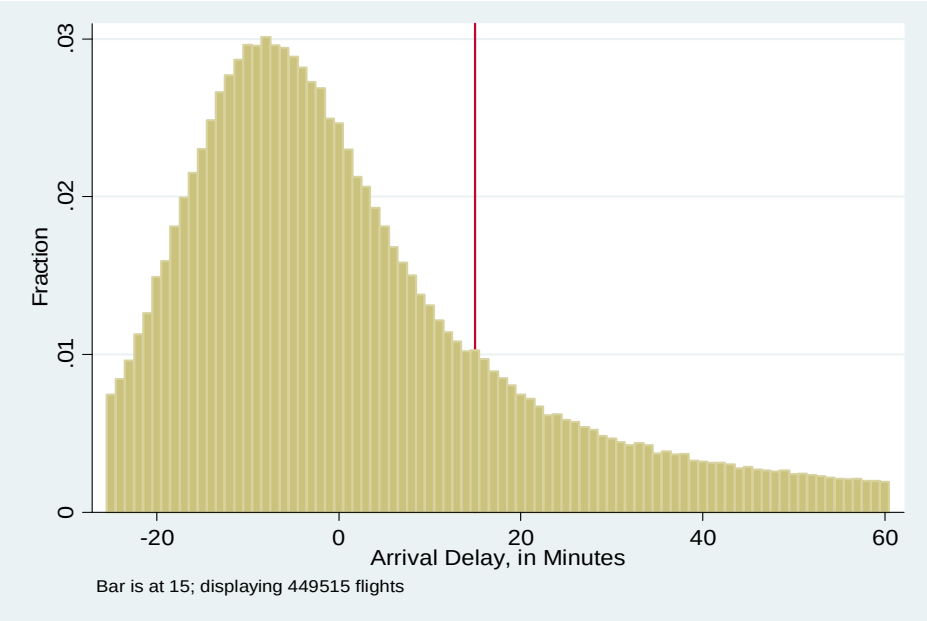

Figure 7B
Distribution of Arrival Delays
United Airlines, 2009

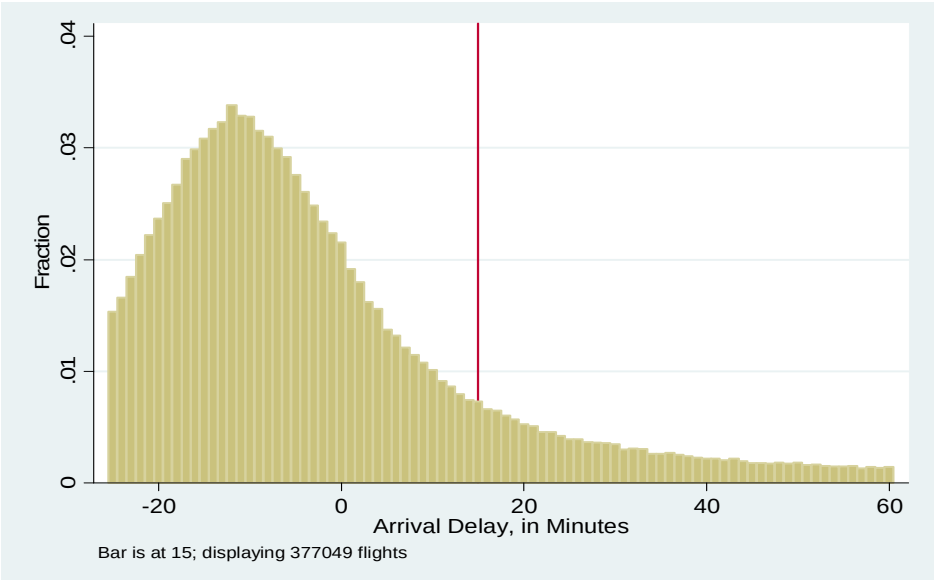

Figure 8A
Coefficients on Continental's Predicted Delay Bins (post-bonus)
(From Table 3)

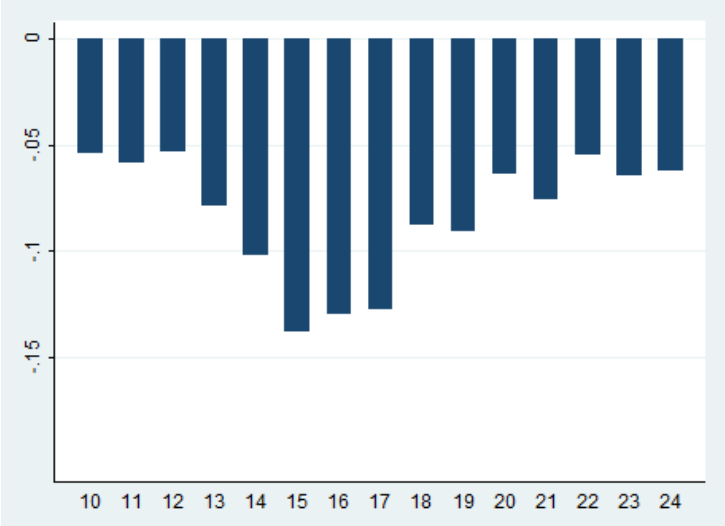

Figure 8B
Coefficients on TWA's Predicted Delay Bins (post-bonus)
(From Table 3)

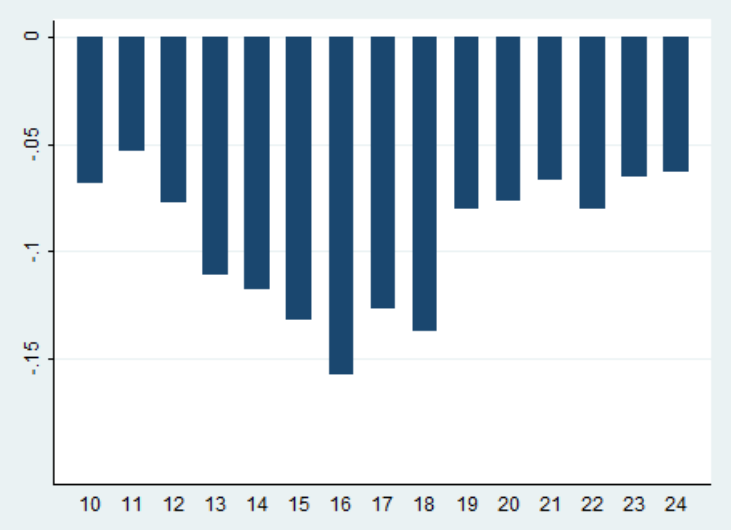

Figure 9A
Coefficients from Taxi-Time Regression
Continental's Predicted Delay Bins – Manual vs. Automatic Planes

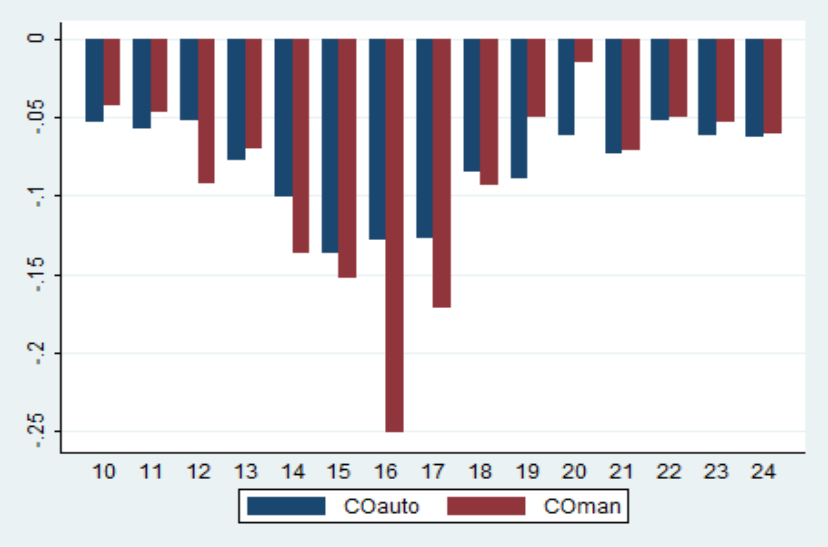

Figure 9B
Coefficients from Taxi-Time Regression
TWA's Predicted Delay Bins – Manual vs. Automatic Planes

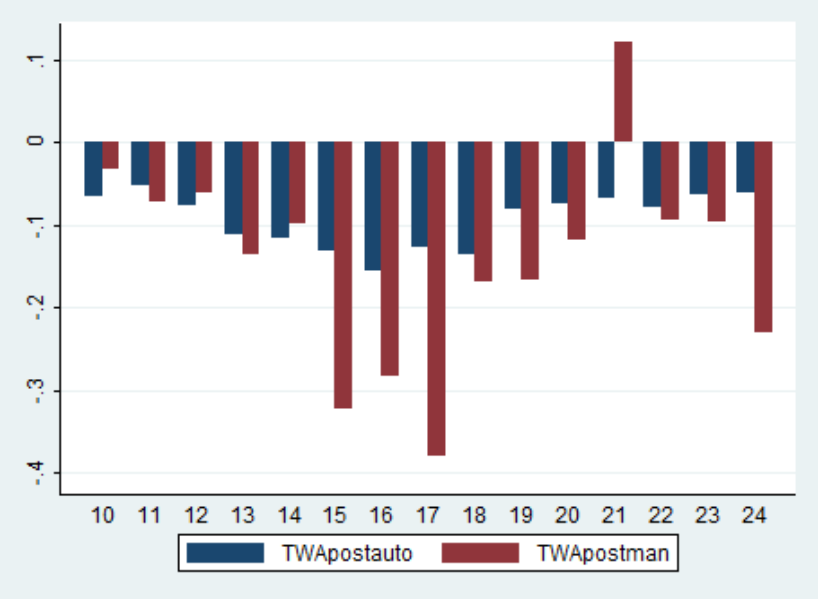

Figure 10A
Coefficients from 1 Minute Early Regression
Continental's Predicted Delay Bins – Manual vs. Automatic Planes

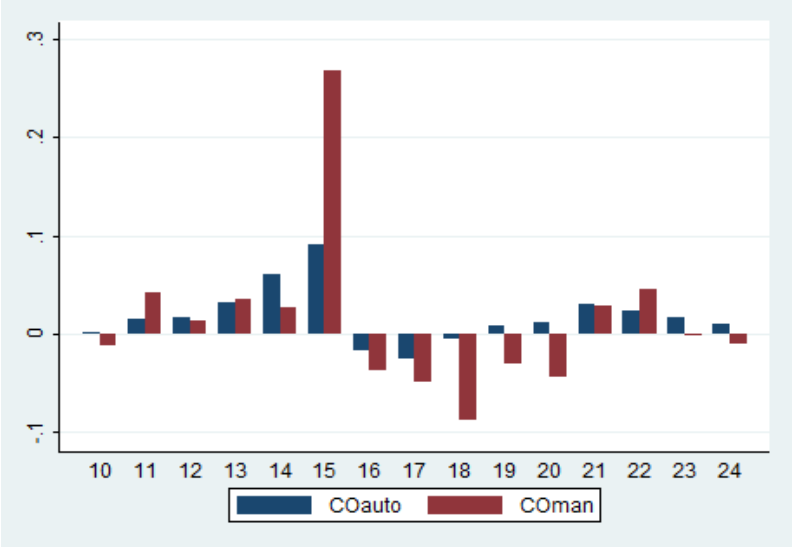

Figure 10B
Coefficients from 1 Minute Early Regression
TWA's Predicted Delay Bins – Manual vs. Automatic Planes

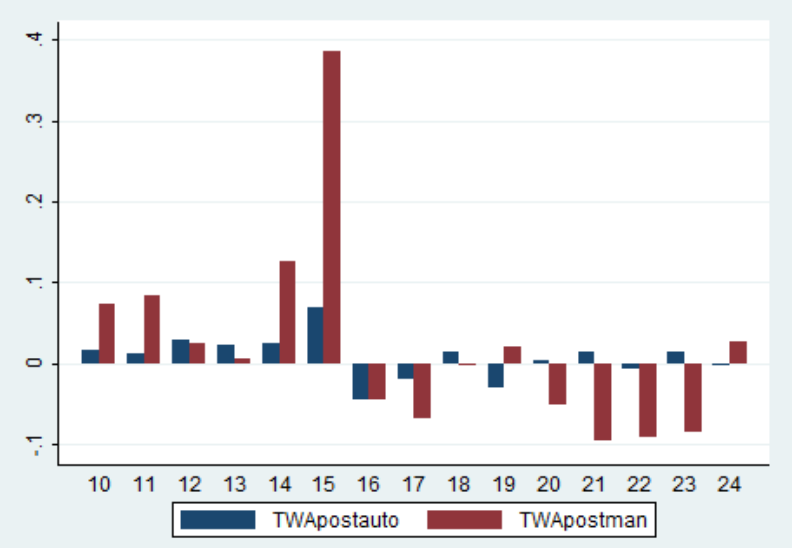

Table 1
Overview of Bonus Programs

Airline	Payment Structure	1 year prior to start of bonus program		1 year after start of bonus program		
		Average Rank	Average On-Time %	Average Rank	Average On-Time %	# Months Bonus Achieved
Continental (Start: Feb 1995)	Initially: $65 per employee in each month that the airline ranked among top 5. Since 1996: $65 for rank 2 and 3; $100 for rank 1.	7.1	80.2	3.4	81.4	10
TWA (Start: Jun 1996)	Initially: $65 per employee in each month that the airline ranked 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.	8.1	74.2	5.7	74.6	4
American (Start: Apr 2003)	Initially: $100 per employee in each month that the airline ranked 1st. $50 in months that the airline ranked 2nd. Since 2009: Bonus based on internal metric that excludes delays that are not under the employees' control.	3.1	81.4	12	79.2	0
US Airways (Start: May 2005)	$75 per employee in each month in which the airline ranks 1st.	9.8	76.1	8.2	79.2	0
United (Start: Jan 2009)	$100 per employee in each month that the airline ranked 1st. $65 in months that the airline ranked 2nd.	14.7	71.6	6.8	81.0	1

Table 2
Summary Statistics for Regression Sample
February1995 - December 1998

	N	Mean	Standard Deviation	Min	Max
Arrival Delay (min)	3,067,533	7.22	27.99	-88	1182
Dummy for Arrive 15 Minutes Late or More	3,067,533	0.21	0.41	0	1
Taxi In Time (min)	3,067,533	6.10	3.92	1	60
Departure Delay (min)	3,067,533	8.43	25.43	-15	1185
Taxi Out Time (min)	3,067,533	14.91	7.44	1	60
Flight Time	3,067,533	108.7	66.50	20	632

***Notes*:** Includes flights by American, Continental, Delta, Northwest, TWA, United, and US Airways.

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

Dependent Variable	***Log(Taxi In)***			
	Coefficient Estimates for:			
	All Other Carriers	**CO post-Bonus**	**TWA pre-Bonus**	**TWA post-Bonus**
Predicted Delay				
[10,11) min	-0.0218***	-0.0522***	-0.0587***	-0.0656***
	(0.00199)	(0.00553)	(0.0123)	(0.0108)
[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)
[20,21) min	-0.0467***	-0.0590***	-0.0609**	-0.0789***
	(0.00293)	(0.00862)	(0.0194)	(0.0171)
[21,22) min	-0.0363***	-0.0728***	-0.0721***	-0.0620***
	(0.00306)	(0.00877)	(0.0175)	(0.0157)
[22,23) min	-0.0411***	-0.0556***	-0.0645**	-0.0811***
	(0.00316)	(0.00892)	(0.0204)	(0.0180)
[23,24) min	-0.0436***	-0.0607***	-0.0938***	-0.0665***
	(0.00331)	(0.00930)	(0.0187)	(0.0183)
[24,25) min	-0.0425***	-0.0615***	-0.0886***	-0.0716***
	(0.00338)	(0.00982)	(0.0207)	(0.0172)
>25 min	-0.0489***	-0.0489***	-0.0841***	-0.0883***
	(0.00145)	(0.00366)	(0.00978)	(0.00846)

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 change in log(taxi time) relative to flights with predicted delay of less than 10 minutes. Calculation of predicted delay is described in the text on page 13. The regression contains 3,067,533 observations.

Table 3B
Taxi Time as a Function of *Predicted* Delay, 2002-2006 and 2008-2010 samples

Dependent Variable	***Log(Taxi In)***		
	Coefficient Estimates for:		
	American Airlines post-Bonus	**US Airways post-Bonus**	**United Airlines post-Bonus**
Predicted Delay			
[10,11) min	-0.0291***	-0.0206*	-0.0124
	(0.00665)	(0.0105)	(0.0143)
[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)
[20,21) min	-0.0545***	-0.0376*	-0.0276
	(0.00994)	(0.0148)	(0.0174)
[21,22) min	-0.0564***	-0.0599***	-0.0428*
	(0.00970)	(0.0144)	(0.0215)
[22,23) min	-0.0601***	-0.0349*	-0.0304
	(0.0103)	(0.0149)	(0.0202)
[23,24) min	-0.0499***	-0.0644***	-0.0352
	(0.0103)	(0.0145)	(0.0201)
[24,25) min	-0.0755***	-0.0618***	-0.0302
	(0.0104)	(0.0158)	(0.0233)
>25 min	-0.0579***	-0.0617***	-0.0470***
	(0.00360)	(0.00512)	(0.00567)

Notes: Standard errors are in parentheses and are clustered at the level of the arrival airport-day. Columns display coefficients from regression of taxi time on mutually exclusive sets of predicted delay "bins" for individual carriers. This table only shows a selected set of coefficients: for carriers with bonus programs, after the introduction of the program. Columns 1 and 2 are based on data from 2002-2006. Column 3 is based on data from 2008-2010. Specifications includes carrier-arrival airport-day fixed effects and arrival hour and hub controls. Coefficients represent the change in log(taxi time) relative to flights with predicted delay of less than 10 minutes.

Table 4A
Probability of Arriving Exactly *One* Minute Earlier than Predicted, 1995-1998

Dependent Variable	*Arrives One Minute Earlier than Predicted*			
	Coefficient Estimates for:			
	All Other Carriers	**CO post-Bonus**	**TWA pre-Bonus**	**TWA post-Bonus**
Predicted Delay				
[10,11) min	0.00520*	0.000474	-0.0204	0.0185
	(0.00209)	(0.00624)	(0.0121)	(0.0101)
[11,12) min	0.00522*	0.0177*	0.00500	0.0160
	(0.00213)	(0.00686)	(0.0124)	(0.00987)
[12,13) min	0.00290	0.0158*	-0.00768	0.0279**
	(0.00224)	(0.00689)	(0.0132)	(0.0108)
[13,14) min	0.00673**	0.0312***	0.00412	0.0228
	(0.00235)	(0.00736)	(0.0144)	(0.0121)
[14,15) min	0.00997***	0.0560***	-0.0145	0.0318**
	(0.00247)	(0.00803)	(0.0148)	(0.0120)
[15,16) min	0.0101***	0.111***	0.0106	0.0888***
	(0.00257)	(0.00852)	(0.0157)	(0.0132)
[16,17) min	0.00769**	-0.0196**	0.00146	-0.0435***
	(0.00261)	(0.00760)	(0.0151)	(0.0118)
[17,18) min	0.00957***	-0.0274***	-0.0125	-0.0223
	(0.00272)	(0.00779)	(0.0155)	(0.0125)
[18,19) min	0.0128***	-0.0131	0.00905	0.0127
	(0.00285)	(0.00870)	(0.0174)	(0.0134)
[19,20) min	0.00896**	0.00288	-0.000275	-0.0292*
	(0.00295)	(0.00924)	(0.0180)	(0.0122)
[20,21) min	0.0127***	0.00856	0.0258	0.000948
	(0.00306)	(0.00998)	(0.0194)	(0.0147)
[21,22) min	0.00504	0.0302**	-0.00486	0.0109
	(0.00323)	(0.0102)	(0.0188)	(0.0153)
[22,23) min	0.0131***	0.0244*	-0.0230	-0.0119
	(0.00325)	(0.0102)	(0.0185)	(0.0150)
[23,24) min	0.00931**	0.0135	-0.0133	0.00964
	(0.00344)	(0.0105)	(0.0183)	(0.0161)
[24,25) min	0.00837*	0.00808	0.0411	-0.00246
	(0.00346)	(0.0108)	(0.0233)	(0.0170)
>25 min	0.00799***	0.00993***	-0.000805	0.00813
	(0.000916)	(0.00264)	(0.00555)	(0.00441)

Notes: Standard errors are in parentheses and are clustered at the level of the arrival airport-day. Columns display coefficients from a single regression 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 change in the probability of a flight arriving exactly one minute earlier than predicted relative to the probability of arriving exactly one minute earlier than predicted 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 3,067,533 observations.

Table 4B

Probability of Arriving Exactly *Two* Minutes Earlier than Predicted, 1995-1998

Dependent Variable	*Arrives Two Minutes Earlier than Predicted*			
		Coefficient Estimates for:		
	All Other Carriers	**CO post-Bonus**	**TWA pre-Bonus**	**TWA post-Bonus**
Predicted Delay				
[10,11) min	0.00876***	0.0249***	0.00725	0.00968
	(0.00151)	(0.00499)	(0.00949)	(0.00760)
[11,12) min	0.00746***	0.0173***	0.00177	0.0171*
	(0.00155)	(0.00479)	(0.00967)	(0.00780)
[12,13) min	0.0107***	0.0193***	-0.00231	-0.00902
	(0.00163)	(0.00521)	(0.00958)	(0.00772)
[13,14) min	0.00969***	0.0267***	-0.00571	0.0289**
	(0.00167)	(0.00544)	(0.0110)	(0.00914)
[14,15) min	0.0147***	0.0291***	0.0140	0.0252**
	(0.00175)	(0.00577)	(0.0114)	(0.00911)
[15,16) min	0.0165***	0.0638***	0.0164	0.0439***
	(0.00186)	(0.00679)	(0.0119)	(0.00962)
[16,17) min	0.0208***	0.139***	0.0110	0.132***
	(0.00201)	(0.00807)	(0.0114)	(0.0131)
[17,18) min	0.0140***	0.0287***	0.0149	-0.0171
	(0.00198)	(0.00659)	(0.0141)	(0.00900)
[18,19) min	0.0118***	0.0212**	-0.0108	0.00496
	(0.00203)	(0.00667)	(0.0123)	(0.0103)
[19,20) min	0.0137***	0.0305***	0.0223	0.0195
	(0.00214)	(0.00748)	(0.0135)	(0.0106)
[20,21) min	0.0147***	0.0287***	0.000792	0.0113
	(0.00227)	(0.00784)	(0.0130)	(0.0110)
[21,22) min	0.0182***	0.0315***	0.0240	0.0389**
	(0.00239)	(0.00738)	(0.0143)	(0.0124)
[22,23) min	0.0155***	0.0120	0.0100	0.0245
	(0.00238)	(0.00743)	(0.0151)	(0.0127)
[23,24) min	0.0170***	0.0187*	0.0276	0.00868
	(0.00258)	(0.00779)	(0.0152)	(0.0122)
[24,25) min	0.0199***	0.0249**	-0.0178	0.0412**
	(0.00265)	(0.00835)	(0.0145)	(0.0142)
>25 min	0.0188***	0.0209***	0.0209***	0.0234***
	(0.000689)	(0.00199)	(0.00427)	(0.00352)

Notes: Standard errors are in parentheses and are clustered at the level of the arrival airport-day. Columns display coefficients from a single regression 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 change in the probability of a flight arriving exactly two minutes earlier than predicted relative to the probability of arriving exactly two minutes earlier than predicted 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 3,067,533 observations.

Table 5
Identification of "Manual" Planes, 1995-1998
Likelihood of a Plane Landing with Exactly Zero Delay, by Reporting Status

	50th percentile	75th Percentile	90th Percentile	95th Percentile	99th Percentile	Reporting Status in 1998
Alaska	0.0577	0.0621	0.0652	0.0671	0.0709	Manual
America West	0.05	0.0552	0.0591	0.0604	0.0653	Manual
American	0.0333	0.0384	0.0429	0.0455	0.0509	Auto
Continental	0.0418	0.0459	0.0521	0.0577	0.0689	Combo
Delta	0.0393	0.0464	0.0537	0.0569	0.0620	Combo
Northwest	0.0356	0.0400	0.0433	0.0455	0.0502	Auto
Southwest	0.1172	0.1230	0.1277	0.1299	0.1335	Manual
TWA	0.0327	0.0360	0.0432	0.0559	0.0613	Combo
United	0.0380	0.0421	0.0466	0.0491	0.0553	Auto
US Airways	0.0385	0.0432	0.0464	0.0483	0.0546	Auto

Notes: Table shows the distribution of a plane-year level variable that equals the probability that the plane is reported to have landed with zero minutes of delay. For example, the fourth entry in the row for American Airlines (third row of table) indicates that only 5% of American's planes in 1996 reportedly landed with zero delay more than 4.5% of the time. The entries in the row for Southwest Airlines (final row of table) indicate that 50% of Southwest's planes in 1996 reportedly landed with zero delay more than 11% of the time. The three shaded rows represent the three carriers that we think were combination reporters in 1996. Their entries show that their planes are slightly more likely than the automatic reporters to land with a reported delay of zero but not nearly as likely as the manual reporters.

Table 6

Probability that "Late" Flight Has Shorter Taxi Time as Function of "Early" Flight's *Predicted* Delay (Flights that Land at the Exact Same Time) , 1995-1998

Dependent Variable	***"Late" Member of Pair Has Shorter Taxi Time***			
	Coefficient estimates for:			
	All Other Carriers	**CO post-Bonus**	**TWA pre-Bonus**	**TWA post-Bonus**
Predicted Delay of "Early" Member of Pair				
[10,11) min	-0.0518***	-0.0702	-0.0159	-0.0274
	[0.0125]	[0.0487]	[0.0504]	[0.0621]
[11,12) min	-0.0523***	-0.0727	0.101***	-0.00537
	[0.0136]	[0.0434]	[0.0197]	[0.0512]
[12,13) min	-0.0297	-0.0679	-0.156***	-0.0907*
	[0.0165]	[0.0585]	[0.0264]	[0.0360]
[13,14) min	-0.0447***	-0.0887**	0.0105	-0.176***
	[0.0111]	[0.0335]	[0.0257]	[0.0396]
[14,15) min	-0.0612***	-0.199***	-0.217***	0.00157
	[0.0136]	[0.0308]	[0.0622]	[0.0343]
[15,16) min	-0.0584***	-0.117*	-0.203***	-0.0188
	[0.0105]	[0.0490]	[0.00323]	[0.104]
[16,17) min	-0.0651***	-0.193***	-0.0581*	0.00669
	[0.0144]	[0.0332]	[0.0231]	[0.0900]
[17,18) min	-0.0597***	-0.133*	-0.0732	-0.165
	[0.0112]	[0.0625]	[0.0647]	[0.112]
[18,19) min	-0.0325**	-0.159*	-0.0341	-0.201***
	[0.0119]	[0.0650]	[0.0360]	[0.0175]
[19,20) min	-0.0529***	-0.0621	0.0115	-0.173*
	[0.0104]	[0.0386]	[0.0292]	[0.0814]
[20,21) min	-0.00644	-0.0737	-0.0953***	-0.0673
	[0.0170]	[0.0500]	[0.0261]	[0.0668]
[21,22) min	-0.0235*	-0.0513	0.0899**	-0.183***
	[0.0110]	[0.0394]	[0.0286]	[0.0304]
[22,23) min	-0.0533***	-0.106	-0.00261	-0.0454
	[0.0118]	[0.0654]	[0.0344]	[0.0281]
[23,24) min	-0.0389*	-0.107*	0.0291	-0.141*
	[0.0151]	[0.0453]	[0.0675]	[0.0708]
[24,25) min	-0.0478***	-0.0772	-0.0243	-0.108**
	[0.0137]	[0.0506]	[0.0446]	[0.0353]
>25 min	-0.0432***	-0.0676***	-0.0533***	-0.0374**
	[0.00474]	[0.0169]	[0.00950]	[0.0142]

Notes: Sample includes carriers' flights that touch-down at the exact same minute. Restricted to two-member pairs. Standard errors are in parentheses. Columns display coefficients from a single regression on four sets of predicted delay "bins" that are defined to be mutually exclusive. Coefficients represent the change in the probability that the "late" member of pair has a shorter taxi time, relative to when "late" member is paired with flight with predicted delay of less than 10 minutes.

Table 7A
Simulated Changes in On-Time Performance and Rankings
Continental, 1995-1997

Year	Month	Actual % On-Time	Simulated % On-Time	Standard Error of Simulated % On-Time	Actual Rank	Simulated Rank
1995	2	0.1704	0.1762	0.0007	4	4
1995	3	0.1507	0.1570	0.0006	1	1
1995	4	0.1451	0.1498	0.0007	2	3
1995	5	0.1963	0.1997	0.0006	9	8
1995	6	0.3313	0.3274	0.0008	10	10
1995	7	0.1691	0.1772	0.0008	2	5
1995	8	0.1286	0.1353	0.0005	1	2
1995	9	0.1037	0.1094	0.0006	2	2
1995	10	0.1324	0.1403	0.0006	3	4
1995	11	0.1709	0.1778	0.0007	4	4
1995	12	0.2111	0.2195	0.0007	1	2
1996	1	0.2370	0.2469	0.0008	2	2
1996	2	0.1901	0.2015	0.0008	2	2
1996	3	0.2011	0.2138	0.0007	5	6
1996	4	0.1800	0.1908	0.0008	4	4
1996	5	0.1334	0.1453	0.0009	2	2
1996	6	0.2441	0.2611	0.0011	6	6
1996	7	0.2170	0.2323	0.0005	5	6
1996	8	0.2358	0.2515	0.0006	5	6
1996	9	0.1960	0.2090	0.0009	4	6
1996	10	0.1797	0.1933	0.0005	3	3
1996	11	0.1653	0.1774	0.0005	1	3
1996	12	0.2421	0.2570	0.0007	1	1
1997	1	0.2434	0.2584	0.0007	2	4
1997	2	0.1869	0.2018	0.0007	2	4
1997	3	0.1941	0.2107	0.0008	5	8
1997	4	0.1785	0.1919	0.0006	6	7
1997	5	0.1698	0.1827	0.0008	8	9
1997	6	0.2131	0.2267	0.0007	8	8
1997	7	0.1723	0.1871	0.0009	4	5
1997	8	0.1720	0.1856	0.0008	4	5
1997	9	0.1367	0.1488	0.0005	5	8
1997	10	0.1728	0.1867	0.0008	7	8
1997	11	0.2050	0.2182	0.0007	6	7
1997	12	0.2270	0.2397	0.0006	3	5

Number of months in which actual rank is **better** than simulated: **19**

Number of months in which actual rank is **same** as simulated: **13**

Number of months in which actual rank is **worse** than simulated (others simulated): **1**

Notes: Based on 20 iterations, standard errors average 300 times smaller than the reported on-time.

Table 7B
Simulated Changes in On-Time Performance and Rankings
TWA, 1996-1998

Year	Month	Actual % On-Time	Simulated % On-Time	Standard Error of Simulated % On-Time	Actual Rank	Simulated Rank
1996	6	0.2845	0.2927	0.0008	9	9
1996	7	0.2995	0.3046	0.0010	8	8
1996	8	0.2836	0.2931	0.0009	8	8
1996	9	0.2106	0.2135	0.0008	6	6
1996	10	0.2146	0.2221	0.0010	5	6
1996	11	0.1861	0.1929	0.0010	5	6
1996	12	0.3302	0.3377	0.0010	6	7
1997	1	0.2833	0.2923	0.0009	6	6
1997	2	0.2081	0.2154	0.0008	5	5
1997	3	0.2041	0.2128	0.0010	8	8
1997	4	0.1402	0.1456	0.0006	1	2
1997	5	0.1040	0.1121	0.0007	1	1
1997	6	0.1372	0.1489	0.0008	1	1
1997	7	0.1275	0.1445	0.0009	1	2
1997	8	0.1515	0.1696	0.0007	2	3
1997	9	0.0848	0.0977	0.0006	1	2
1997	10	0.1175	0.1317	0.0005	1	2
1997	11	0.1872	0.2032	0.0009	3	5
1997	12	0.2756	0.2977	0.0008	8	9
1998	1	0.2259	0.2421	0.0007	5	5
1998	2	0.1906	0.2107	0.0012	4	4
1998	3	0.2571	0.2781	0.0009	9	9
1998	4	0.1891	0.2092	0.0012	6	7
1998	5	0.2093	0.2302	0.0011	6	6
1998	6	0.2985	0.3179	0.0010	7	9
1998	7	0.1836	0.2001	0.0007	6	6
1998	8	0.1392	0.1522	0.0007	1	2
1998	9	0.1081	0.1186	0.0007	1	3
1998	10	0.1046	0.1172	0.0008	1	1
1998	11	0.1075	0.1217	0.0007	1	1
1998	12	0.2080	0.2275	0.0013	4	5

Number of months in which actual rank is **better** than simulated (others simulated): **15**

Number of months in which actual rank is **same** as simulated (others simulated): **16**

Number of months in which actual rank is **worse** than simulated (others simulated): **0**

Based on 20 iterations, standard errors average 300 times smaller than the reported on-time.

Appendix A
Changes in On-Time Performance after Introduction of Employee Bonus Programs
1995-1998

Dependent Variable	*Arrival Delay*	*Arrival Delay≥15 min*	*Taxi In Time*	*Departure Delay*	*Taxi Out Time*
	(1)	(2)	(3)	(4)	(6)
CO*Bonus Period	-2.370***	-0.0476***	-0.585***	-1.797***	-0.227
	(0.177)	(0.00237)	(0.0836)	(0.150)	(0.127)
TW*Bonus Period	-2.609***	-0.0484***	-0.0807**	-0.947***	-0.209***
	(0.207)	(0.00269)	(0.0260)	(0.214)	(0.0482)
Airline Dummies					
CO	2.034***	0.0328***	-0.114	2.482***	0.536***
	(0.171)	(0.00229)	(0.0878)	(0.153)	(0.129)
DL	1.964***	0.0273***	-0.498***	1.170***	0.0290
	(0.0974)	(0.00144)	(0.0352)	(0.112)	(0.0279)
NW	0.289*	0.00992***	-0.0867**	0.372**	0.00792
	(0.113)	(0.00154)	(0.0326)	(0.113)	(0.0311)
TW	1.889***	0.0309***	-0.757***	1.806***	0.276***
	(0.170)	(0.00223)	(0.0352)	(0.179)	(0.0459)
UA	1.742***	0.00991***	-1.266***	3.176***	-1.081***
	(0.107)	(0.00143)	(0.0349)	(0.105)	(0.0291)
US	0.876***	0.0194***	-0.736***	2.193***	-1.873***
	(0.107)	(0.00151)	(0.0314)	(0.106)	(0.0293)
WN	1.040***	0.000669	-2.157***	2.597***	-3.988***
	(0.108)	(0.00159)	(0.0313)	(0.111)	(0.0338)
HP	4.953***	0.0527***	-0.696***	2.940***	-1.150***
	(0.139)	(0.00202)	(0.0325)	(0.144)	(0.0365)
AS	2.818***	0.0284***	-1.535***	0.427*	-1.000***
	(0.209)	(0.00349)	(0.0345)	(0.171)	(0.0426)
N	4,966,448	4,966,448	3,983,280	4,966,448	3,983,280
R-squared					

Notes: Standard errors are in parentheses and are clustered at the arrival airport-day level. All specifications include arrival airport-day fixed effects. All specifications also include departure and arrival hour controls as well as controls airline and airport level controls. Appendix B presents the coefficient estimates on the control variables. Data on taxi time is not available prior to 1995. As a result, columns (3) and (6) have fewer observations.