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

November 2010

Preliminary draft. Please do not circulate without authors' permission.

Abstract

Many quality disclosure programs provide consumers with binary quality measures that are based on whether or not a product meets a particular threshold. This creates incentives for firms to improve the quality of specifically those products that can easily be brought above the threshold. We investigate this type of "gaming" behavior in the context of government-mandated disclosure of airline on-time performance. While this program collects detailed data on the delay of each flight, it ranks airlines based only on the fraction of flights that arrive 15 or more minutes after their scheduled arrival time. We estimate whether airlines systematically reduce delays on precisely those flights that they expect to arrive with delays of about 15 minutes. Our results show that flights with predicted delays between 15 and 16 minutes have significantly shorter taxi-in times than flights with shorter or longer predicted delays. We also find that these particular flights are significantly more likely to arrive exactly one minute sooner than predicted. The results are even stronger when airlines explicitly incentivize their employees based on the airline's ranking in the government program. 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 Bob Gibbons, Matt Mitchell and seminar participants at the Rotman School of Management and the USC Conference on Game Theory in Law, Business and Political Economy for helpful comments.

I. Introduction

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

In this paper, we investigate these issues in the context of a government mandated program that requires airlines to disclose the on-time performance of their flights. Since 1987, airlines have been required to report to the Department of Transportation (DOT) the scheduled and actual arrival times for their domestic flights. Although the DOT collects detailed data about the actual minutes of delay incurred on each flight, it counts a flight as being "late" if it arrives 15 minutes or more behind schedule. The DOT issues monthly reports to consumers that rank airlines based on the percentage of their flights that are late as defined by the 15 minute cutoff.[1] These rankings - or excerpts from these rankings - are frequently reported in media outlets, which is likely the primary source of this information for consumers.

Because flights are counted as late if they arrive 15 or more minutes later scheduled, this program gives airlines an incentive to exert effort to reduce delays on specifically those flights

[1] 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.

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

We develop an empirical approach that allows us to estimate whether airlines systematically exert more effort to reduce delays on flights which would otherwise arrive slightly above the 15 minute threshold, relative to the effort they exert on flights that are expected to arrive well above or below the threshold. Much of our empirical analysis focuses on differences in flights' taxi-in times. We focus on taxi-in times because this represents the final stage of a flight and thus the final point at which delays may be incurred or reduced. By the time a flight has touched down at the arrival airport of its route, an airline should have a fairly precise estimate of the expected delay that the flight will have and can decide whether or not to exert effort to try to reduce that delay below 15 minutes. Furthermore, we as researchers can also predict fairly precisely the typical delay of a flight based on the time that the plane lands on the runway. This allows us to identify flights that are likely candidates for gaming. While we cannot observe precisely what actions airline employees undertake to reduce arrival delays, we expect that taxi-in times can be reduced through several channels. For example, airlines can preferentially allocate resources – such as ground crew and even gates - across flights. In addition, it may be possible for ground crew members to reduce taxi-in times by exerting more effort.

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 10 largest carriers between 1994 and 1998.[2] We exploit the fact that, starting in 1995, the DOT also began collecting information about each flight's wheels-off and wheels-on times (i.e.: the times at which it leaves the runway and touches down on the runway). These additional pieces of information allow us to construct a measure of every flight's *predicted delay* at the time that it touches down at the arrival airport. Our main set of regressions relate a flight's taxi-in time to its predicted delay and look for evidence of a reduced taxi-in times right around the 15 minute threshold. We also construct pairs of flights that land at the exact same time and investigate whether a flight's own taxi-in time depends on whether it lands at the same time as a flight with a predicted delay around the 15 minute threshold. We estimate these relationships separately for airlines with and without employee bonus programs based on the DOT rankings.

Our results provide clear evidence that airlines do indeed try to game these rankings and that this behavior is enhanced when employee bonus programs based on these rakings are in place. For airlines that do *not* have an employee bonus program in place, taxi-in times of flights that are predicted to be between 15 and 16 minutes late are almost 8 percent shorter than taxi-in times for flights that are predicted to be less than 10 minutes late. Moreover, the estimates reveal a discontinuous relationship between taxi-in times and predicted delay right around the 15

[2] We plan to extend the analysis to later years in a future draft of the paper.

minute threshold. While one might have thought that airlines have the greatest incentive to reduce very long delays (because there may be convex costs of delay), we find that taxi-in times for the flights with predicted delays in the critical 15 minute range are significantly shorter than taxi-in times for flights with longer predicted delays.

When we estimate these relationships for two of the airlines that introduced employee bonus programs between 1994 and 1998, we find the same patterns but the magnitudes are much larger. During the time that Continental Airlines had a bonus program in place, its taxi-in times for flights with predicted delays between 15 and 16 minutes were 13 percent shorter than its taxi-in times for flights with predicted delays of less than 10 minutes. Its flights with predicted delays between 16 and 17 minutes had taxi-times that were 15 percent shorter. We see effects of a similar magnitude when we look at TWA who also introduced a program during this period.

We have begun to explore the "costs" of this gaming behavior on two dimensions – whether it leads airlines to inefficiently allocate resources across flights and whether it distorts the information presented to consumers. With the respect to the first of these, our results so far indicate that more effort is allocated to flights that are near the threshold at 15 minutes than to flights that have very long delays. However, when we look at pairs of flights that land at the exact same time, we do not find that a shorter taxi-time for the threshold member of the pair results in a longer taxi-time for the other member. Thus, the extra effort allocated to the threshold flight does not appear to be coming from the "paired" flight but may be coming from other flights that arrive or depart around the same time. We are in the process of trying to measure this. With respect to the second potential cost of gaming in this context, we carry out a series of simulations that illustrate the extent to which these distortions in taxi-in times can affect an airline's rankings. Our results so far indicate that they can. For example, we find that

Continental's gaming behavior improved its rankings by at least one position in more than half of the months following the introduction of its employee bonus program.

This paper is related to a number of other papers which have studied incentives to game quality disclosure programs in a variety of contexts. Dranove, Kessler, McClellan and Satterthwaite (2003) and Werner and Asch (2005) present evidence that hospitals select lower-risk patients when they are subject to hospital report card programs. Haney (2000), Deere and Strayer (2001), Jacob (2005), Cullen and Reback (2006) and Figlio and Getzler (2006) find that schools increase the rate at which students are placed in special education – and thus not counted in a school's score – when schools become subject to accountability programs. Similarly, Jacob (2007) provides evidence of "teaching to the test" after the introduction of state-level school accountability programs. The results in Neal and Schanzenbach (forthcoming) suggest that teachers pay special attention to students who are near an accountability threshold. Our work is also related to research on gaming of employee incentive programs by Oyer (1998), Courty and Marschke (2004) and Larkin (2007). Finally, this work is related to Knez and Simester (2001) which is the one paper we know of that has looked at one of the airline employee bonus programs that we consider.

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 explain our empirical analysis and present our results in Section IV. 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 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.[3]

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 through an automatic device inside the aircraft. Many airlines use a combination of manual and automated recording. While the automated devices are presumably reliable in recording the actual arrival times, there is a

[3] Airlines had previously not been required to report taxi times for these flights.

[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.

possibility that airline employees who record flight delays manually report delays of 14 minutes for flights whose actual delays are 15 minutes. This raises the concern that what we interpret as airlines systematically exerting more effort to improve the on-time performance of threshold flights may just reflect employees lying about the arrival times of those flights. While airlines indicate each month whether they record delays manually, automatic or with a combination of the two, for those carriers that use a combination, we do not have information on which planes report automatically and which report manually. However, we can track planes by tail number and thus we can investigate whether any gaming behavior that we identify for the carriers using combination reporting is concentrated in a particular subset of aircraft (which we would presume were those without the automatic device). The two airlines whose bonus programs we investigate in the current draft of the paper (Continental and TWA) do use combination reporting. Our early analysis of the distribution of gaming behavior across tail numbers does not suggest that the gaming is concentrated on a particular set of aircraft.

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.[5] 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

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

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

In June 1996, TWA implemented an employee bonus program which closely resembled Continental's. TWA would pay $65 to each employee in every month in which the airline ranked top five in the following three rankings published by the DOT: on-time performance, baggage handling and customer complaints.[6] 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. [Our current empirical analysis does not include these programs. They will be added in future versions of the paper.] Table 1 summarizes the details of these bonus programs

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

and the airlines' on-time performance in the two years before and after the introduction of their programs.

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 (see the discussion in Section II.A. above). We have collected these data for all reporting carriers for every year between 1988 and 2008, inclusive. Our empirical work below only uses data for 1994 through 1998 since this is the period during which both Continental Airlines and TWA introduced their employee bonus programs. Our primary sample includes domestic flights operated by the following 10 airlines: American Airlines, Continental Airlines, Delta Air Lines, Northwest Airlines, TWA, United Airlines, US Airways, Southwest Airlines, America West and Alaskan Airlines. Because this dataset is very large, we only include their flights between the 29 airports for which the airlines are required to report their on-time performance. To further reduce the size of the dataset, we take a random sample of flights by restricting to every fifth day of the year. In addition, we drop flights that meet any of the following conditions: depart more than 15 minutes early (since we suspect this may represent a rescheduled flight), arrive more than 90 minutes early, depart on what appears to be the following calendar day, have a taxi-out time of more than 55 minutes, have a taxi-in time of more than 25 minutes or have distance of less than 25 miles. Our final sample includes 5,165,322 flights.

Table 2 presents summary statistics for the main variables in the data. The average arrival delay in our sample is about seven minutes. 20% of flights in our sample arrive 15 minutes late or more and thus are considered "late" under the program's definition. The average air time is 100 minutes, the average taxi-out time is about 14 minutes and the average taxi-in time is 5.5 minutes.[7] 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 in our sample. 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. We suspect that these discrete increases reflect rounding by carriers who report their delay data manually.[8] It is interesting to note that the spikes appear to 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 minutes. 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.

Figures 2A and 2B show the distribution of Continental's arrival delays in the two years before and two and a half years after the introduction of its employee bonus program.[9] 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. Figure 3A and 3B plot analogous histograms for TWA before and after the introduction of its employee bonus program and show a very similar pattern. After the

[7] Recall from the discussion in Section II that the airborne, taxi-in and taxi-out times were only collected beginning in 1995. That is why there are fewer observations for these variables.

[8] One of the carriers reporting manually during this period is Southwest, which only reports delays in 5-minute invervals.

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

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 height of the bars at 14 and 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.

IV. Empirical Approach and Results

IV.A. Overview of Empirical Approach

To evaluate the impact that the government's mandatory reporting program has on the overall distribution of airline delays, we would need data on airline on-time performance in both the absence and presence of this program. Since the data on on-time performance are only available *because* of the government program, we have no way of observing or estimating what the distribution of delays would look like if airlines were not subject to mandatory reporting. Instead, what we do is develop an empirical approach that allows us to estimate the extent to which airlines manipulate delays around the 15 minute threshold. That is, we look for evidence of gaming, which we define as an airline systematically exerting more effort to reduce delays on flights which would otherwise arrive slightly above the threshold to be considered on-time.[10] Our empirical approach exploits the fact that the incentive to game changes discontinuously with a flight's expected delay. Assuming that it is on average less costly for an airline to push a flight that it expects to be 16 minutes late below the 15 minute threshold than to try to do the same for a flight that it expects to be 26 minutes late, the design of the disclosure program creates an

[10] 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. We plan to add an analysis of schedule padding to the analysis of real-time manipulation.

incentive to make a greater effort to reduce delays for those flights that are near the threshold.[11] We also exploit the fact that the employee bonus programs should augment the incentives to game that are inherent in the government program and thus we are able to use the introduction of the bonus programs as an additional source of identification.

Before describing how we empirically identify gaming, it is useful to consider when it may take place. Delays can be occurred – and reduced - at several different points in a flight's progression: at the departure gate, while taxiing out, in the air, or while taxiing in. The delay that is recorded upon a plane's arrival at the gate - and which is used as the basis for classification of a flight as on-time or late - represents the sum of delays incurred during all phases of a flight. In theory, an airline that is trying to systematically improve the on-time performance of a flight that it expects to arrive just above the threshold could try to reduce delays during any of the phases. However, we expect that airlines that are trying to game will be more likely to try to reduce delays during the later stages of a flight. This is because, as the flight progresses, the airline knows the delay that has been incurred so far and can therefore more precisely predict the total delay the flight will have. For example, when a flight is airborne, the airline knows how delayed the plane was leaving the ground but must predict both how delayed it will be in the air and how delayed it will be while taxing in. However, once a flight has touched down at the arrival airport, the airline knows how delayed the plane was leaving the ground and while in the air and must only predict how delayed it will be while taxiing in. For any given predicted level of delay, reducing the amount of noise associated with that prediction increases the likelihood that the airline's effort at reducing a flight's delay will actually result in the flight having a shorter delay. Based on this logic, much of our empirical analysis of gaming focuses on measuring an airline's

[11] Note that convex costs of delays give the airline an opposing incentive to exert more effort on flights with longer delays.

effort to reduce delays during the final phase of the flight – i.e.: when it is taxiing in to its arrival gate.

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

IV.B. Taxi Time Analysis

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

We construct a series of dummy variables for each level of predicted delay, in one minute increments. 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 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 for four mutually exclusive sets of flights: (1) flights by carriers that do not have a bonus program in place; (2) flights by Continental after the introduction of its bonus program (which is introduced in the second month for which we have taxi-time data); (3) flights by TWA before the introduction of its bonus program; and (4) flights by TWA after the introduction of its bonus program. This means that we have a total of 64 mutually exclusive dummy variables. Every flight in our sample will have at most one of these dummy variables equal to one and may have none equal to one if the flight's predicted delay is less than 10 minutes (because then it is in the excluded category).

We estimate a flight level equation that regresses a flight's taxi-in time on these 64 dummy variables, carrier-airport-day fixed effects and a set of control variables. 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 percentage change in 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. As mentioned above, this variation results from differences in the delays that flights incur prior to arrival which will largely be driven by factors at the flights' respective departure airports. Our primary interest is in testing whether those flights with predicted delay right around the critical threshold have systematically shorter taxi times than flights that are well above or below the threshold and whether this relationship is affected by the introduction of an employee bonus program. The key identifying assumption of the model is that there are no observable factors that are correlated with a flight having a predicted delay in the threshold range and that affect the flight's taxi-in time. Because evidence of gaming would come from a non-monotonic relationship between predicted delay and taxi time, we can rule out most other possible sources of correlation between predicted delay and taxi time since these are not likely to result in the same type of pattern.

The results of this analysis are presented in Table 3. Each column of the table represents the coefficients on the 16 predicted delay bins for one of the four sets of flights described above. The first column represents the coefficients for airlines without bonus programs, the second column represents the coefficients for Continental, the third column represents the coefficients for TWA prior to the introduction of its bonus program and the final column represents the coefficients for TWA after the introduction of its bonus program. The results clearly indicate that flights that are predicted to arrive just above the critical threshold have systematically shorter taxi-in times than flights with shorter or longer predicted delays. Looking first at the results for the carriers *without* bonus programs, the coefficient estimates imply that taxi-in times for flights predicted to arrive between 15 and 16 minutes late (16 and 17 minutes late) are 8 percent (almost 6 percent) shorter than taxi-in times for flights predicted arrive less than 10 minutes late. Taxi times for flights with predicted delays in this critical range are also statistically significantly shorter than taxi times for flights on either side of this range. For example, flights with predicted delay between 17 and 18 minutes have taxi-in times that are only 4 percent shorter than the omitted group. The same is true for flights with predicted delays between 13 and 14 minutes. In fact, the coefficients on virtually all of the predicted delay bins outside of the critical range are quite similar to each other and are all between 3 and 4 percent. Taxi-in times for flights near the 15 threshold are shorter than even the taxi-in times for flights with very long delays.

The results for the carriers that implemented bonus programs based on the DOT rankings show a very similar pattern but the magnitudes are much larger. For Continental, flights with predicted delays of 15 to 16 (16 to 17 minutes) minutes have taxi-in times that are 13 (15) percent shorter than those of flights with predicted delay of 10 minutes or less. Given an average taxi-in time of about 5.5 minutes, a 13 percent reduction reflects about 42 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 coefficients again show a non-monotonic relationship with the flights with predicted delays on either side of the critical range having significantly longer taxi-in times. As in the first case, even flight with delays that are predicted to be very long do not have taxi-in times that are as short as those for flights in the critical range. The estimates for TWA after the introduction of its program show a very similar pattern which is much weaker prior to the introduction of the program. Overall, the results in Table 3 indicate that airlines appear to be systematically shortening the taxi-in times of flights that they expect to arrive just above the critical threshold for being considered on-time.

IV.C. Does it Work?

The results in Table 3 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 Table 4, we investigate whether they are successful in doing so. 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 3. 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. The estimates in the first column show that, for carriers without bonus programs, flights that are predicted to be 15 to 16 minutes late are 6.9 percentage points more likely to arrive one minute earlier than predicted than flights that are predicted to be less than 10 minutes late. They are actually much more likely to arrive one minute early than flights with any other level of predicted delay. Few of the other predicted delay bins have coefficients that are statistically different from zero and, when they are, they are much smaller than that on the 15 to 16 minute bin. As in Table 3, these effects are even stronger for carriers that have bonus programs in place. For both Continental and TWA, after the introduction of their bonus programs, their flights with predicted delays between 15 and 16 minutes are almost 10 percentage points more likely to arrive exactly one minute earlier than predicted, relative to their flights with less than 10 minutes of predicted delay.

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 airlines attempting to systematically reduce delays on flights that would otherwise arrive just above the threshold for being on-time. 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) have a higher probability of arriving two minutes earlier than expected, relative to both flights in the omitted category and flights in every other predicted delay bin. The effects here are again much larger for Continental and TWA. For both of these carriers, flights with predicted delay between 16 and 17 minutes are more than 13 percentage points more likely to arrive two minutes sooner than predicted.

IV.D. Analysis of Paired Flights

The results in Table 3 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.[12] 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

[12] 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 that equals one if the “late” flight (i.e.: the one that lands with predicted delay of more than 25 minutes) has a shorter taxi-in time than the “early” member of the pair. We relate this variable to the predicted delay of the early member of the pair by regressing it on the same expected delay bins used in the analysis above. Intuitively, what we are doing is estimating whether the probability that a very late flight has a shorter taxi-in time 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 regression in Table 3) 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 5A. 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 carriers without a bonus program 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. The results for

Continental show a very similar pattern. 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.

IV.E. Implications/Costs of Manipulation

Having established that gaming takes place, we are now interesting in investigating whether there are welfare costs to this behaviour. As mentioned in the Introduction, there are two possible costs that may result from airlines improving the on-time performance of specifically those flights with delays close to the 15 minute threshold. First, airlines may be inefficiently allocating resources across flights. Second, airlines may be distorting the information conveyed to consumers in the DOT rankings. In particular, as a result of their gaming, small changes in underlying delays may result in significant changes in rankings. We have begun to explore both of these issues.

i. Misallocation of Effort Across Flights

Assuming that there are convex costs of delays (which there likely are for both consumers and airlines), the effort allocated to reducing delays on threshold flights could be more efficiently allocated to flights with long delays. The results in Table 3 and Table 5A indicate that threshold flights are indeed being allocated effort and/or resources that could be allocated to other flights. However, they do not indicate whether the resources allocated to those flights are diverted from other flights or whether these resources would otherwise be slack. For example, one of the ways in which taxi-in times could be reduced for threshold flights is by speeding up the process by which the plane is directed to the gate by ground crew. Intuitively, we would like to be able to distinguish whether a flight's taxi-in time is reduced because the ground crew that was allocated a different flight is allocated to the threshold flight or because the crew originally allocated to the threshold flights works harder/faster.

To begin to explore these issues, we look at the same sample of paired flights that we use for Table 5A. We construct a new dependent variable that equals the sum of their taxi-in times. We regress this variable on the same set of predicted delay bins. This allows us to get a sense of whether threshold flights impose costs – in the form of longer taxi-in times – on other flights by the same airline that arrive at the exact same time. The results of this regression are presented in Table 5B. They do not indicate that threshold flights divert resources from the flights with which they are paired. The combined taxi time for pairs of flights that involve a "late" flight and an early flight with predicted delay in the critical range is statistically significantly shorter than the combined taxi time for pairs that involve a "late" flight and an "early" flight with predicted delay less than 10 minutes. It is also shorter than the combined taxi time for almost all other pairs of flights. These results are, of course, limited because they only consider one other flight from which resources may be diverted. We are in the process of expanding this analysis to look for costs on other flights that land close in time to a threshold flight.

ii. Distortion of Information Conveyed to Consumers

To investigate whether the distortions in taxi-in times that we find in our regression analysis can actually impact airlines' overall on-time performance and DOT rankings, we perform a counterfactual simulation that estimates what arrival delays and rankings would be absent gaming. To do this, we take the following approach. Our data suggest that taxi-in times are distributed approximately log-normal. We calculate the mean and variance of the log taxi-in time for each carrier-airport-month. Then, for each flight in our data, we replace the actual taxi-in time in the data with a random draw from a log-normal distribution with the mean and variance for the appropriate carrier-airport-month. The idea behind this exercise is to replace a flight's taxi-in time with the taxi-in time it would likely have absent any incentive for the airline to systematically reduce taxi-in times on threshold flights. After doing this exercise for every flight in our data, we can recalculate the fraction of flights that are 15 or more minutes delayed. This leads to counterfactual measures of on-time performance for each airline and these can be used to create counterfactual rankings of airlines. Repeating the simulation a number of times yields standard errors for our simulated on-time performance measures.

We report results from the counterfactual exercises in Tables 6A and 6B. Table 6A shows simulated changes in on-time performance and ranking for Continental in the months after the introduction of its bonus program. Table 6B shows a similar 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 22 of the 36 months following the introduction of their program. When we simulate Continental as well as all other airlines' taxi-in times, we find that the taxi-time distortions result in Continental achieving an improvement in rankings in 8 of the 36 months. Thus, the results of the simulation exercise indicate that while a 40 second reduction in delay may be small in absolute value (and in terms of the disutility to consumers), when applied to flights that are close to the relevant threshold, the impact on reported rankings can be significant.

iii. 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 Table 7A, 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 Table 7A 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.

V. **Discussion/Conclusion [not yet completed]**

References

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

Cullen, J. and R. Reback (2006), “Tinkering Toward Accolades: School Gaming Under a Performance Accountability System,” NBER Working Paper W12286.

Deere, D. and W. Strayer (2001), “Putting Schools to the Test: School Accountability, Incentives, and Behavior,” Working Paper 113, Private Enterprise Research Center, Texas A&M University.

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.

Figlio, D. and L. Getzler. (2006), “Accountability, Ability and Disability: Gaming the System?” in: *Advances in Microeconomics*, T. Gronberg, ed., Amsterdam: Elsevier.

Haney, W. (2000), “The Myth of the Texas Miracle in Education,” *Education Policy Analysis Archives* 8(41).

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.

Neal, D. and D. W. Schanzenbach (forthcoming), “Left Behind by Design: Proficiency Counts and Test-Based Accountability”, *Review of Economics and Statistics*.

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

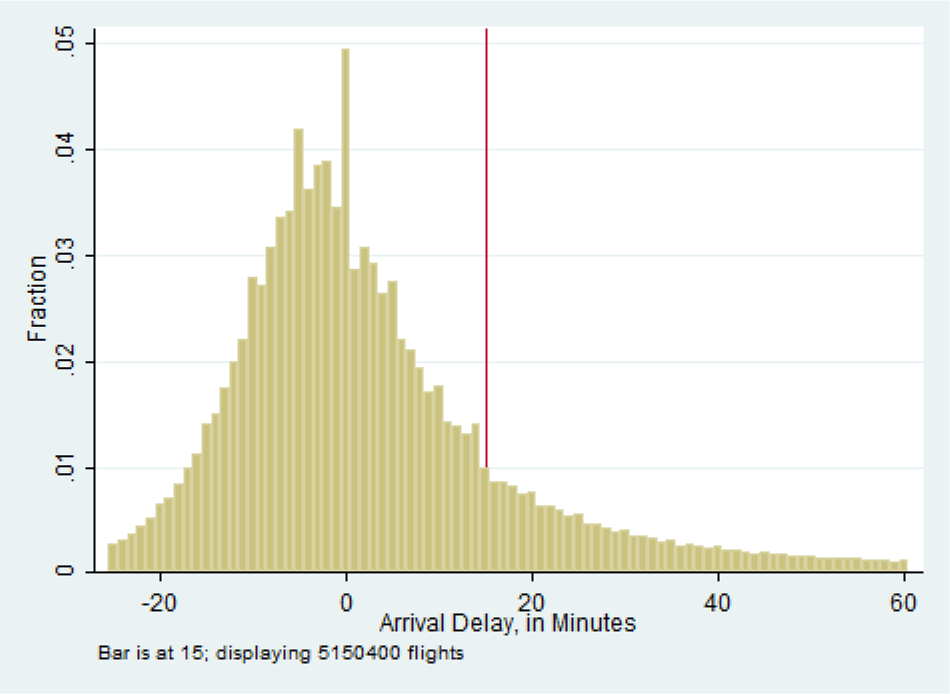

Figure 2A
Distribution of Arrival Delays
Continental Airlines, 1993-1994

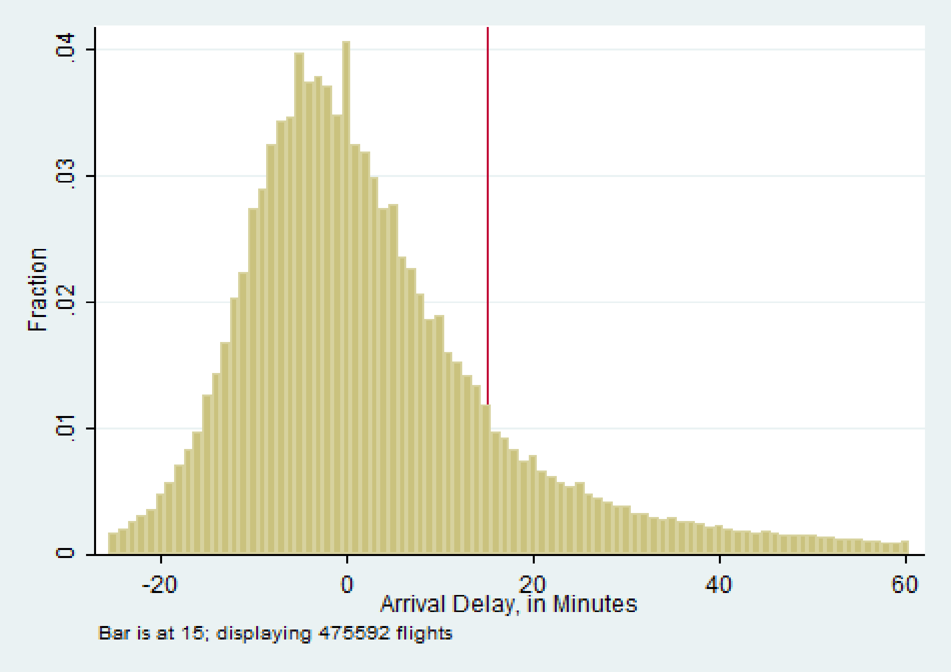

Figure 2B
Distribution of Arrival Delays
Continental Airlines, February 1995-1997

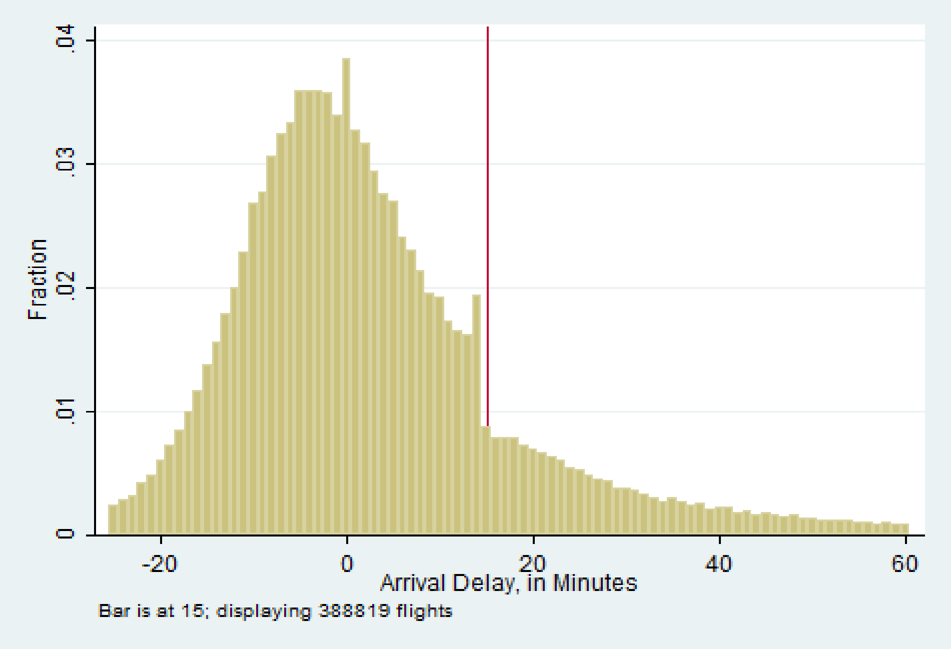

Figure 3A
Distribution of Arrival Delays
TWA, 1994-1995

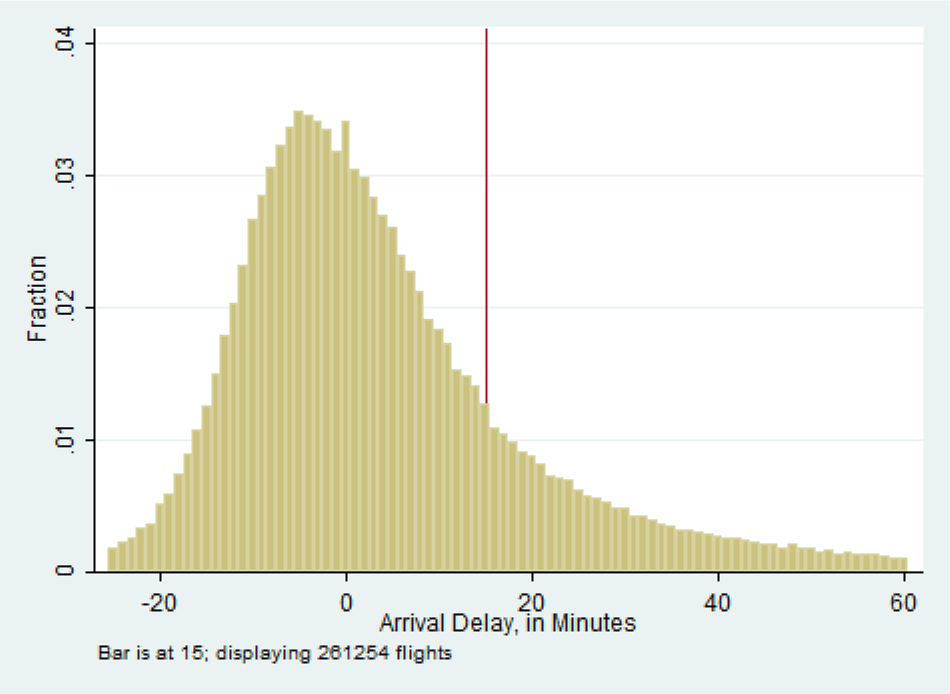

Figure 3B
Distribution of Arrival Delays
TWA, June 1996-1998

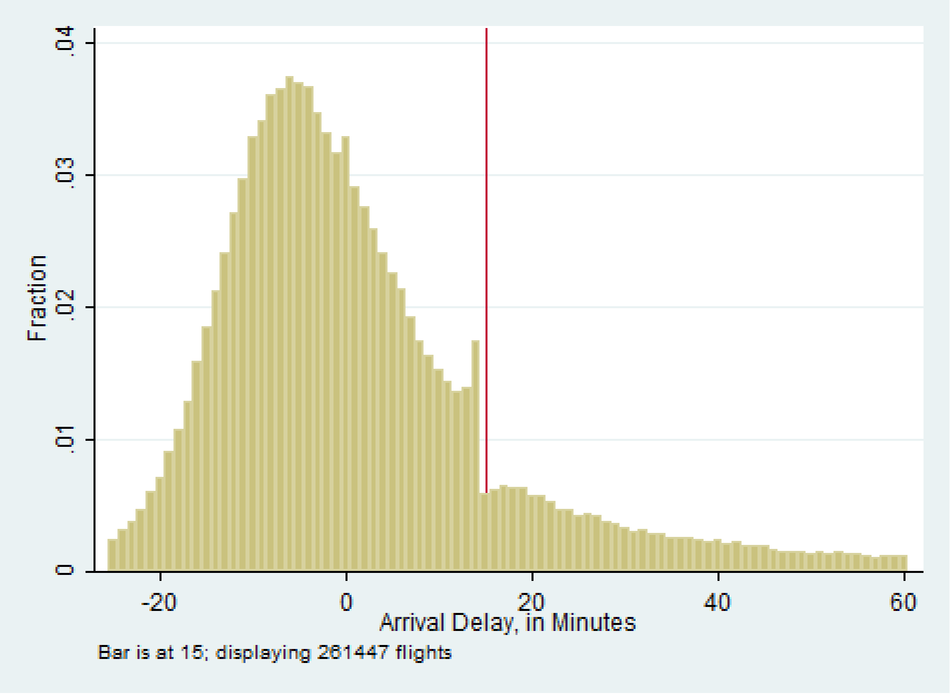

Table 1
Overview of Bonus Programs

Airline	**Start Period**	**Payment Structure**	**Average rank in on-time performance prior to start of bonus program**		**Average rank in on-time performance after start of bonus program**	
			2 years	1 year	1 year	2 years
Continental	February 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.	8.5	7.5	3.5	3.4
TWA	June 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.	7.3	8	6.8	3.7
American	April 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.				
US Airways	May 2005	$75 per employee in each month in which the airline ranks 1st.				
United	January 2009	$100 per employee in each month that the airline ranked 1st. $65 in months that the airline ranked 2nd.				

Table 2
Summary Statistics, 1994-1998

	N	Mean	Standard Deviation	Min	Max
Arrival Delay (min)	5,165,322	6.96	26.63	-90	1609
Dummy for Arrive 15 Minutes Late or More	5,165,322	0.20	0.40	0	1
Taxi In Time (min)	4,143,402	5.51	3.26	1	25
Departure Delay (min)	5,165,322	8.16	24.05	-15	1618
Taxi Out Time (min)	4,143,402	13.90	7.17	1	55
Flight Time	4,143,402	100.01	63.95	1	626

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

Table 3
Taxi Time as a Function of *Predicted* Delay

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	-0.0333***	-0.0577***	-0.0538***	-0.0705***
	(0.00159)	(0.00482)	(0.0119)	(0.0104)
[11,12) min	-0.0376***	-0.0702***	-0.0332**	-0.0615***
	(0.00167)	(0.00517)	(0.0126)	(0.0102)
[12,13) min	-0.0318***	-0.0643**6*	-0.0301*	-0.111***
	(0.00175)	(0.00536)	(0.0130)	(0.0108)
[13,14) min	-0.0359***	-0.0781***	-0.0246	-0.108***
	(0.00178)	(0.00579)	(0.0151)	(0.0114)
[14,15) min	-0.0511***	-0.101***	-0.0457**	-0.135***
	(0.00181)	(0.00580)	(0.0141)	(0.0131)
[15,16) min	-0.0795***	-0.133***	-0.0469**	-0.146***
	(0.00202)	(0.00656)	(0.0160)	(0.0127)
[16,17) min	-0.0583***	-0.151***	-0.0730***	-0.160***
	(0.00220)	(0.00723)	(0.0157)	(0.0149)
[17,18) min	-0.0401***	-0.147***	-0.0667***	-0.186***
	(0.00225)	(0.00825)	(0.0182)	(0.0159)
[18,19) min	-0.0380***	-0.123***	-0.0661***	-0.106***
	(0.00233)	(0.00856)	(0.0163)	(0.0169)
[19,20) min	-0.0395***	-0.103***	-0.0717***	-0.0775***
	(0.00238)	(0.00878)	(0.0165)	(0.0155)
[20,21) min	-0.0425***	-0.0780***	-0.0382*	-0.0604***
	(0.00241)	(0.00871)	(0.0175)	(0.0149)
[21,22) min	-0.0470***	-0.0619***	-0.0602**	-0.0757***
	(0.00251)	(0.00833)	(0.0191)	(0.0156)
[22,23) min	-0.0414***	-0.0689***	-0.0511*	-0.0742***
	(0.00262)	(0.00862)	(0.0202)	(0.0154)
[23,24) min	-0.0365***	-0.0801***	-0.0460*	-0.0851***
	(0.00271)	(0.00870)	(0.0188)	(0.0153)
[24,25) min	-0.0412***	-0.0569***	-0.0450	-0.0783***
	(0.00278)	(0.00954)	(0.0242)	(0.0185)
>25 min	-0.0448***	-0.0534***	-0.0757***	-0.0841***
	(0.00108)	(0.00281)	(0.00916)	(0.00697)

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

Table 4A
Probability of Arriving Exactly *One* Minute Earlier than Predicted

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	-0.00252	0.0142*	0.0117	0.00374
	(0.00171)	(0.00588)	(0.0118)	(0.00919)
[11,12) min	0.0191***	0.0352***	0.00953	0.0102
	(0.00184)	(0.00647)	(0.0120)	(0.00997)
[12,13) min	0.000373	0.0201**	-0.0184	0.0359***
	(0.00186)	(0.00665)	(0.0128)	(0.0105)
[13,14) min	0.00233	0.0256***	0.000105	0.0240*
	(0.00192)	(0.00699)	(0.0137)	(0.0118)
[14,15) min	-0.000210	0.0591***	0.0153	0.0512***
	(0.00197)	(0.00747)	(0.0136)	(0.0119)
[15,16) min	0.0692***	0.0978***	-0.0130	0.0988***
	(0.00238)	(0.00817)	(0.0133)	(0.0140)
[16,17) min	0.00552*	-0.0186*	-0.00617	-0.0414***
	(0.00223)	(0.00734)	(0.0141)	(0.0110)
[17,18) min	-0.00333	0.00101	-0.0103	-0.00158
	(0.00227)	(0.00794)	(0.0146)	(0.0116)
[18,19) min	0.00131	0.0180*	-0.00667	-0.00978
	(0.00237)	(0.00839)	(0.0158)	(0.0124)
[19,20) min	0.00635*	0.0174	0.0150	0.0221
	(0.00251)	(0.00888)	(0.0157)	(0.0134)
[20,21) min	0.00263	0.0103	0.0000666	0.0345*
	(0.00251)	(0.00886)	(0.0163)	(0.0149)
[21,22) min	0.0207***	0.0269**	0.0268	0.0453**
	(0.00268)	(0.00940)	(0.0183)	(0.0160)
[22,23) min	0.00535	0.0231*	-0.00925	0.0293
	(0.00273)	(0.0103)	(0.0165)	(0.0150)
[23,24) min	0.00692*	0.0369***	0.00290	0.0102
	(0.00280)	(0.0105)	(0.0190)	(0.0155)
[24,25) min	0.00452	0.00821	0.0298	0.0130
	(0.00293)	(0.0106)	(0.0201)	(0.0169)
>25 min	0.00804***	0.0176***	0.0149**	0.0234***
	(0.000765)	(0.00271)	(0.00507)	(0.00410)

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 might 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 4,143,402 observations.

Table 4B
Probability of Arriving Exactly *Two* Minutes Earlier than Predicted

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	0.00509***	0.0215***	0.00291	0.0189*
	(0.00120)	(0.00458)	(0.00890)	(0.00847)
[11,12) min	0.00519***	0.0184***	-0.00636	0.00781
	(0.00127)	(0.00475)	(0.0104)	(0.00844)
[12,13) min	0.00884***	0.0266***	0.00108	0.0282**
	(0.00135)	(0.00508)	(0.0111)	(0.00916)
[13,14) min	0.00930***	0.0258***	0.000461	0.0465***
	(0.00136)	(0.00542)	(0.0106)	(0.00984)
[14,15) min	0.00839***	0.0385***	0.00328	0.0411***
	(0.00140)	(0.00563)	(0.00997)	(0.00970)
[15,16) min	0.0118***	0.0533***	0.0173	0.0246*
	(0.00150)	(0.00605)	(0.0117)	(0.0104)
[16,17) min	0.0350***	0.136***	0.0230	0.136***
	(0.00183)	(0.00773)	(0.0131)	(0.0141)
[17,18) min	0.00967***	0.0146*	0.00227	-0.00624
	(0.00169)	(0.00618)	(0.0128)	(0.00977)
[18,19) min	0.00809***	0.0282***	0.0173	-0.00832
	(0.00163)	(0.00695)	(0.0128)	(0.0108)
[19,20) min	0.00371*	0.0239***	0.0101	0.0274*
	(0.00174)	(0.00696)	(0.0125)	(0.0112)
[20,21) min	0.0108***	0.0309***	0.0107	-0.00170
	(0.00183)	(0.00731)	(0.0141)	(0.0112)
[21,22) min	0.00993***	0.0186*	0.00109	0.0212
	(0.00190)	(0.00732)	(0.0137)	(0.0129)
[22,23) min	0.0138***	0.0262***	0.0214	0.0205
	(0.00198)	(0.00781)	(0.0132)	(0.0123)
[23,24) min	0.0100***	0.0318***	0.00837	0.0281
	(0.00192)	(0.00837)	(0.0150)	(0.0145)
[24,25) min	0.00765***	0.0269**	-0.00452	0.0297*
	(0.00211)	(0.00876)	(0.0148)	(0.0140)
>25 min	0.0135***	0.0216***	0.0184***	0.0194***
	(0.000587)	(0.00212)	(0.00454)	(0.00350)

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 4,134,402 observations.

Table 5A
Probability that "Late" Flight Has Shorter Taxi Time as Function of "Early" Flight's *Predicted* Delay (Flights that Land at the Exact Same Time)

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	-0.0406***	-0.115*	-0.0545	0.0206
	(0.0114)	(0.0535)	(0.0923)	(0.0676)
[11,12) min	-0.0407***	-0.0541	0.0747	-0.0792
	(0.0116)	(0.0491)	(0.0894)	(0.0805)
[12,13) min	-0.0231*	-0.0566	-0.0212	-0.0302
	(0.0117)	(0.0523)	(0.0923)	(0.0625)
[13,14) min	-0.0274*	-0.118*	0.0264	-0.190*
	(0.0115)	(0.0566)	(0.0857)	(0.0768)
[14,15) min	-0.0743***	-0.197***	-0.202*	-0.0351
	(0.0118)	(0.0532)	(0.0857)	(0.0714)
[15,16) min	-0.0675***	-0.172**	-0.188*	0.0168
	(0.0116)	(0.0553)	(0.0923)	(0.0736)
[16,17) min	-0.0460***	-0.197**	-0.136	-0.0151
	(0.0122)	(0.0601)	(0.0834)	(0.0837)
[17,18) min	-0.0512***	-0.115*	0.0122	-0.0706
	(0.0124)	(0.0535)	(0.0923)	(0.0751)
[18,19) min	-0.0307*	-0.134*	-0.0519	-0.209**
	(0.0123)	(0.0574)	(0.0814)	(0.0721)
[19,20) min	-0.0324*	-0.0416	0.055	-0.158*
	(0.0129)	(0.0609)	(0.0857)	(0.0777)
[20,21) min	-0.0105	-0.0819	-0.0878	-0.0378
	(0.0125)	(0.0574)	(0.0923)	(0.0759)
[21,22) min	-0.00994	-0.0231	0.137	-0.154
	(0.0130)	(0.0570)	(0.103)	(0.0837)
[22,23) min	-0.0443***	-0.154*	-0.00135	-0.0151
	(0.0130)	(0.0614)	(0.0834)	(0.0815)
[23,24) min	-0.0504***	-0.0796	-0.0172	-0.164*
	(0.0130)	(0.0614)	(0.0869)	(0.0826)
[24,25) min	-0.0284*	-0.0808	0.0836	-0.126
	(0.0135)	(0.0648)	(0.0954)	(0.0837)
>25 min	-0.0296***	-0.0747***	-0.0311	-0.023
	(0.00410)	(0.0187)	(0.0283)	(0.0224)

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 5B

Total Taxi Time for Pair as a Function of "Early" Flight's *Predicted* Delay
Flights that Land at the Exact Same Time

Dependent Variable	*Total Taxi Time for Pair*			
	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	-0.529***	-1.181	-1.24	-0.194
	(0.142)	(0.668)	(1.152)	(0.844)
[11,12) min	-0.386**	-1.529*	-1.904	0.612
	(0.145)	(0.614)	(1.117)	(1.005)
[12,13) min	-0.392**	-1.024	-2.14	-0.879
	(0.146)	[0.653)	[1.152)	[0.780)
[13,14) min	-0.710***	-0.749	-1.702	-0.516
	[0.144)	[0.707)	[1.070)	[0.959)
[14,15) min	-0.729***	-1.838**	-0.502	-2.296**
	[0.147)	(0.664)	(1.070)	(0.891)
[15,16) min	-0.786***	-1.402*	-2.973**	-1.643
	(0.145)	(0.690)	(1.152)	(0.918)
[16,17) min	-0.347*	-2.822***	-1.333	-1.182
	(0.152)	(0.750)	(1.042)	(1.045)
[17,18) min	-0.226	-2.080**	-1.807	-1.271
	(0.155)	(0.668)	(1.152)	(0.938)
[18,19) min	-0.454**	-2.033**	-1.437	-1.557
	(0.153)	(0.716)	(1.016)	(0.900)
[19,20) min	0.014	-1.146	-1.816	-1.206
	(0.161)	(0.761)	(1.070)	(0.970)
[20,21) min	-0.343*	-0.696	-0.973	-0.811
	(0.156)	(0.716)	(1.152)	(0.948)
[21,22) min	-0.168	-0.537	-0.832	-0.266
	(0.162)	(0.712)	(1.283)	(1.045)
[22,23) min	0.0181	-0.771	-0.657	-0.621
	(0.162)	(0.766)	(1.042)	(1.018)
[23,24) min	-0.339*	-0.398	-0.0791	-1.164
	(0.162)	(0.766)	(1.085)	(1.031)
[24,25) min	-0.0663	-2.077*	-1.552	-1.405
	(0.169)	(0.809)	(1.191)	(1.045)
>25 min	0.943***	-0.312	-0.625	0.448
	(0.0512)	(0.234)	(0.353)	(0.280)

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.

Table 6A
Simulated Changes in On-Time Performance and Rankings
Continental

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

Number of months in which actual rank is **better** than simulated (others simulated): **8 (22.9%)**

Number of months in which actual rank is **same** as simulated (others simulated): **24 (68.6%)**

Number of months in which actual rank is **worse** than simulated (others simulated): **3 (8.6%)**

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

Table 6B
Simulated Changes in On-Time Performance and Rankings
TWA

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

Number of months in which actual rank is **better** than simulated (others simulated): **6 (19.4%)**

Number of months in which actual rank is **same** as simulated (others simulated): **25 (80.6%)**

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

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

Table 7A
Changes in On-Time Performance after Introduction of Employee Bonus Programs

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.

Table 7B
Changes in On-Time Performance after Introduction of Employee Bonus Programs
Year-by-Year Effects

Dependent Variable	***Arrival Delay***	***Arrival Delay≥15 min***	***Taxi In Time***
	(1)	(2)	(3)
CO*Bonus*1995	-2.289***	-0.0404***	-0.259**
	(0.245)	(0.00339)	(0.0865)
CO*Bonus*1996	-3.190***	-0.0645***	-0.580***
	(0.226)	(0.00319)	(0.0862)
CO*Bonus*1997	-1.969***	-0.0526***	-0.829***
	(0.235)	(0.00307)	(0.0864)
CO*Bonus*1998	-2.005***	-0.0321***	-0.674***
	(0.246)	(0.00308)	(0.0876)
TW*Bonus*1996	-0.442	0.00112	-0.0362
	(0.345)	(0.00442)	(0.0405)
TW*Bonus*1997	-4.214***	-0.0720***	-0.150***
	(0.247)	(0.00337)	(0.0322)
TW*Bonus*1998	-2.272***	-0.0550***	-0.101**
	(0.292)	(0.00360)	(0.0367)
N	4,966,448	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 airline and airport level controls. Data on taxi time is not available prior to 1995. As a result, column (3) has fewer observations.

Appendix A
Air Time as a Function of Predicted Delay When Wheels Leave the Ground

Dependent Variable	*Log(Air Time)*			
	Coefficient Estimates for:			
	All Other Carriers	**CO post-Bonus**	**TWA pre-Bonus**	**TWA post-Bonus**
Predicted Delay				
[10,11) min	-0.00390***	-0.000924	-0.00179	0.0101***
	(0.000393)	(0.00138)	(0.00257)	(0.00248)
[11,12) min	-0.00340***	-0.00400**	0.00212	0.00137
	(0.000415)	(0.00141)	(0.00265)	(0.00229)
[12,13) min	-0.00313***	-0.00325*	0.00540	0.00231
	(0.000438)	(0.00151)	(0.00301)	(0.00256)
[13,14) min	-0.00337***	-0.00353*	0.00376	0.00604*
	(0.000447)	(0.00154)	(0.00290)	(0.00306)
[14,15) min	-0.00418***	-0.00211	0.00856*	0.00155
	(0.000476)	(0.00167)	(0.00347)	(0.00287)
[15,16) min	-0.00394***	-0.00373	0.00449	0.00613
	(0.000510)	(0.00197)	(0.00329)	(0.00313)
[16,17) min	-0.00333***	-0.00679***	0.00613	0.00629
	(0.000530)	(0.00192)	(0.00394)	(0.00327)
[17,18) min	-0.00485***	-0.00698***	-0.00118	0.00555
	(0.000551)	(0.00196)	(0.00346)	(0.00346)
[18,19) min	-0.00431***	-0.00532**	0.00557	0.00401
	(0.000582)	(0.00204)	(0.00350)	(0.00342)
[19,20) min	-0.00450***	-0.00525*	0.00514	0.00213
	(0.000615)	(0.00222)	(0.00356)	(0.00348)
[20,21) min	-0.00586***	-0.00678**	0.00619	0.00153
	(0.000624)	(0.00218)	(0.00421)	(0.00352)
[21,22) min	-0.00451***	-0.00958***	0.000866	0.00442
	(0.000658)	(0.00216)	(0.00453)	(0.00407)
[22,23) min	-0.00388***	-0.00791***	0.00535	0.00637
	(0.000684)	(0.00229)	(0.00497)	(0.00424)
[23,24) min	-0.00394***	-0.00631*	0.00898	0.00164
	(0.000719)	(0.00262)	(0.00486)	(0.00390)
[24,25) min	-0.00299***	-0.00553	0.00274	0.00230
	(0.000739)	(0.00289)	(0.00421)	(0.00392)
>25 min	-0.00192***	-0.00241**	0.00800***	0.00807***
	(0.000257)	(0.000855)	(0.00152)	(0.00124)

Notes: Standard errors are in parentheses and are clustered at the level of the route-year-month. Columns display coefficients from a single regression of air time on four sets of predicted delay “bins” that are defined to be mutually exclusive. Specification includes route-year-month fixed effects and arrival hour, departure hour and hub controls. Calculation of predicted delay at the time that a flight’s wheels leave the ground is described in the text on page 13. The regression contains 4,143,402 observations.

Appendix B

Full Set of Control Variables Included in Table 7A

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	-0.687***	0.00692***	0.146***	0.908***	0.770***
	(0.0510)	(0.000699)	(0.00470)	(0.0336)	(0.00985)
DL	-1.876***	0.0243***	0.475***	1.890***	2.321***
	(0.138)	(0.00181)	(0.0103)	(0.0857)	(0.0250)
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)
Control Variables					
500≤Distance<1500	-0.687***	0.00692***	0.146***	0.908***	0.770***
	(0.0510)	(0.000699)	(0.00470)	(0.0336)	(0.00985)
1500≤Distance	-1.876***	0.0243***	0.475***	1.890***	2.321***
	(0.138)	(0.00181)	(0.0103)	(0.0857)	(0.0250)
Log(#Daily Flights at Departure Airport)	0.245***	0.00969***	0.0340***		
	(0.0205)	(0.000284)	(0.00232)		
Dep Flights/Runway	-0.00711***	0.0000672***	-0.00217***	0.00346	0.0188***
	(0.00129)	(0.0000178)	(0.000164)	(0.00283)	(0.00178)
Departs Airline's Hub	2.267***	0.0374***	-0.0777***	1.739***	1.797***
	(0.0488)	(0.000715)	(0.00657)	(0.0736)	(0.0211)
Arr Flights/Runway	-0.0197***	-0.000137***	0.0113***	-0.0145***	0.00738***
	(0.00263)	(0.0000307)	(0.000905)	(0.00136)	(0.000354)
Arrives Airline's Hub	-0.0434	-0.00257*	0.888***	-0.289***	-0.278***
	(0.0813)	(0.00115)	(0.0131)	(0.0384)	(0.0121)
				0.571***	0.0184**
Log(#Daily Flights Arrival Airport)				(0.0162)	(0.00561)
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. Coefficients on departure and arrival hour dummies not reported.