Quality Disclosure and Gaming: Do Employee Incentives Matter?

Silke J. Forbes
University of California, San Diego

Mara Lederman
University of Toronto, Rotman School of Management

Trevor Tombe
Wilfrid Laurier University

October 2011

Abstract

We investigate gaming of a public disclosure program and, in particular, whether gaming depends on the incentives provided to the employees who are most likely to carry out the gaming. We do this in the context of the government-mandated disclosure of airline on-time performance. While this program collects data on the actual minutes of delay incurred on each flight, it ranks airlines based only on the fraction of their flights that arrive 15 or more minutes late. This creates incentives for airlines to game the program by reducing delays on specifically those flights they expect to arrive with about 15 minutes of delay. In addition, several airlines have introduced employee incentive programs based explicitly on the airline's performance in the government program. Our empirical analysis finds no evidence of gaming by airlines without incentive programs or with incentive programs with targets that are unrealistically hard to achieve. On the other hand, we find strong evidence of gaming by airlines that implemented incentive programs with targets that could be – and were - achieved. Specifically, we find that their flights that are predicted to arrive with about 15 minutes of delay have significantly shorter taxi-in times and are significantly more likely to arrive exactly one minute sooner than predicted. Our findings highlight that gaming of a disclosure program will not only depend on the design of the program but will also depend on if and how the measured quality dimensions can be manipulated and whether those who are in a position to manipulate them have incentives to do so.

Key words: Disclosure; Gaming; Incentives

JEL codes: L2, L5

We thank Severin Borenstein, Bob Gibbons, Ginger Jin, 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), the USC Conference on Game Theory in Law, Business and Political Economy and the NBER Summer Institute (2011) for helpful comments. Forbes gratefully acknowledges financial support from NSF grant SES-1124154. Lederman gratefully acknowledges financial support from the Social Science and Humanities Research Council of Canada.

I. Introduction

Disclosure programs exist in many industries in which consumers are imperfectly informed about product quality.[1] While the empirical literature on these programs has generally found that they result in improvements in product quality, there is also considerable evidence that firms make targeted efforts to improve their *reported* quality, potentially at the expense of other quality dimensions (see, for example, Dranove *et al*., 2003, Jacob, 2005, Werner and Asch, 2005, Lu, 2009 and Neal and Schanzenbach 2010). Such gaming may both distort the information being conveyed to consumers as well as lead firms to inefficiently allocate resources.[2] The existing evidence 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 anticipate the ability of firms to game the program.[3] However, the potential for gaming will depend not only on features of the program but also on the characteristics of the product as well the organizational structure and incentives in place at the firm. For example, whether gaming takes place will depend on which dimensions of product quality are measured, if and how these dimensions can be manipulated and whether those who are in a position to manipulate them have incentives to do so.

This paper begins to explore this problem by investigating the relationship between gaming and the incentives provided to the employees who are most likely to implement the actions required for gaming to occur. While we focus on a particular

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

[2] In designing disclosure programs, policy makers face a trade-off between providing information that is comprehensive on all dimensions of quality versus information that is sufficiently easy for consumers to process and understand (e.g. Hastings and Weinstein, 2008). This is particularly important when consumers are heterogeneous in their valuation of different quality dimensions.

[3] A related literature on "notches" points out that similar issues exist in the design of taxes and subsidies (see, e.g., Sallee and Slemrod, 2010).

empirical context, the issues we consider are relevant in other settings in which disclosure programs exist or are being contemplated, including public education, health care and environmental regulation. The specific context we consider is the reporting of airline on-time performance which has been mandated by the Department of Transportation (DOT) since 1987. This setting has several advantageous features. First, the design of the disclosure program clearly encourages gaming. Although the DOT collects detailed data about the actual minutes of delay incurred on each flight, it only considers a flight to be late if it arrives 15 minutes or more behind schedule. Since this is the metric that is used by the DOT and the media when producing monthly rankings of airline on-time performance, airlines have an incentive to game the program by trying to reduce delays on specifically those flights that would otherwise arrive close to the 15 minute cutoff.

Second, airlines cannot predict far in advance which flights will be candidates for gaming. While they may be able to anticipate which routes or flights will, on average, have longer delays, they likely do not learn which flights will have 14 versus 16 minutes of delay until shortly before the plane's arrival. Thus, to the extent that gaming occurs, it must occur in real-time and 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. Third, between 1995 and 2009, five different airlines implemented employee bonus programs based explicitly on the airline's performance in the government's ranking.[4] Under these programs, each airline employee would receive a payment of between $65 and $100 in any month in which the airline as a

[4] Knez and Simester (2001) study the effect of one of the airline employee bonus programs (Continental's) on the airline's overall delays. They show that overall departure delays decreased after the introduction of the bonus program, but they do not investigate the gaming of the disclosure program which is the focus of our paper.

whole placed at or near the top of the DOT ranking. While all of the programs potentially faced a free-rider problem, the programs differed significantly in how easy it was to achieve the target ranking and thus in the strength of the incentives provided to employees.

Finally, the richness of the data available allows us to identify gaming in a very precise way. Because we observe each stage of a flight, we can calculate an estimate of a flight's *expected* delay at various points in its progression. This allows us to identify flights that are expected to arrive right around the 15 minute cutoff. We can then estimate whether delays on *subsequent* stages of the flight are systematically different for those flights that are close to the cutoff. Moreover, because we observe tens of thousands of flights each year, we can construct quite precise counterfactuals of what these flights' delays would have been absent the incentive to game.

Our empirical analysis uses the very data that is collected by the DOT under the mandatory disclosure program. We focus on the final stage of a flight and construct a measure of every flight's expected delay at the time it touches down at the arrival airport. We then estimate the relationship between a flight's taxi-in time (i.e., the time between touch-down on the runway and arrival at the gate) and the flight's expected delay upon touch-down. We look for evidence of a non-monotonicity right around the 15 minute threshold. That is, we look for evidence that flights that are expected to arrive about 15 minutes late have systematically shorter taxi-in times than flights that are expected to arrive with slightly shorter or slightly longer delays. We estimate these relationships for airlines that do not have incentive programs in place and, separately, for each airline that introduced an incentive program.

Perhaps surprisingly, we do not find 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 expected to be between 15 and 16 minutes late were about 13 percent shorter than its taxi-in times for flights with expected delays of less than 10 minutes. We see effects of a very similar magnitude when we look at TWA. Moreover, the estimates for Continental and TWA reveal a discontinuous relationship between taxi-in times and expected 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 much longer predicted delays. We also find that both of these carriers' flights which we are predicted to be 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 misreporting or actual reduction in taxi-in times, we find evidence for both.

When we carry out the same sets 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. In particular, 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. Moreover, by the time the later programs had been introduced, the DOT rankings had expanded to include between 17 and 20 - rather than 10 – airlines. At least one of these airlines – Hawaiian Airlines – consistently had much better on-time performance than any of the large network carriers.

The gaming we document may impact welfare in two ways. First, it may distort the information being conveyed to consumers. We carry out simulations that show that airlines' selective reduction in taxi-in times of threshold flights can result in an improvement in their DOT rank of at least one place. To the extent that the 15 minute cutoff used in the ranking is imperfectly correlated with the dimensions of on-time performance that consumers care about, then changes in rankings that are simply due to gaming may cause consumers to believe that an airline has improved on the dimensions they care about when they have not. Second, if the reductions in delay for the threshold flights are achieved by reallocating scarce resources, then there could be negative externalities on other flights in the form of longer delays. In our empirical setting, identifying such externalities without knowledge of when and from where resources are reallocated is difficult because, at the times when resources are scarce, any one of a very large number of flights could potentially be affected. We have not found evidence of externalities in the data, but we also cannot rule out that they exist.

We believe that this paper makes an important contribution to the existing literature in this area. To our knowledge, it is the first large-scale empirical analysis of gaming to explicitly investigate the link between gaming by firms and changes in the incentives in place inside those firms.[5] Our results show that despite the incentives for gaming that are inherent in the design of the DOT disclosure program, gaming only takes

[5] There is a related literature on gaming of employee incentive programs, including Oyer (1998), Courty and Marschke (2004) and Larkin (2007).

place when the employees who are in the position to improve the relevant dimension of quality are explicitly incentivized to do so. More generally, we believe this paper highlights the importance of considering interactions between the design of a disclosure program design - specifically, the dimensions of quality that are being measured - how, when and by whom these dimensions can be manipulated and the incentive schemes in place at a firm. Finally, we also see this paper as contributing to the ongoing policy discussion on the use of disclosure programs to resolve informational asymmetries in areas such as public education, health care and environmental regulation. For example, our results suggest that recent efforts to financially reward public school teachers based on the percentage of their students who pass standardized tests may exacerbate the teachers' incentives to focus their efforts on students who are near the threshold for passing, at the expense of other students.[6]

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 Department of Transportation under Title 14, Part 234 of the Code of

[6] The results in Neal and Schanzenbach (2010) suggest that the introduction of accountability programs has already shifted teachers' attention to students near the threshold.

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.[7] A January 1995 amendment expanded the requirements to include flights that were delayed or cancelled because of mechanical problems. The same 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 report delay causes 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.

Airlines can record delays either manually or automatically through technology 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. Starting in 1998, we know how each carrier reports in each month. Prior to that, we cannot be certain which method was used for reporting. While we believe that most of the planes in our sample reported their delays automatically, we suspect that the two airlines which implemented bonus programs in the mid-1990s likely had a combination of manual and automatic planes at the time. 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. Since such misreporting represents a different type of gaming, we have developed an approach (described below) for identifying the manual aircraft in the data.

[7] The legislation only requires flights to and from 29 of the most congested airports to be included, but all airlines voluntarily report the on-time performance of all of their flights.

The DOT uses the data it collect to issue monthly reports that rank airlines based on the percentage of their flights that are late under the 15 minute definition. 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. Firms only have an incentive to game the disclosure program if consumers in fact care respond to the disclosed information. Forbes (2008) shows that consumers' willingness-to-pay falls in response to longer flight delays. Similarly, the fact that several airlines refer to their placement in the DOT rankings in their advertising campaigns suggests that at least the airlines perceive that consumers care about on-time performance. Finally, the DOT rankings are often picked up in national or local media outlets. A typical news story will report the percentage of on-time flights for all airlines and may point out which airlines have improved or deteriorated relative to the others, often highlighting which carriers are consistently near the top or the bottom. Local media outlets tend to focus on carriers that have a big market share in the local city. It is not uncommon for the media reports to simply refer to flights being "on-time", without explaining the DOT's definition of "on-time".

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.[8] 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. While overall improvement in on-time performance after the introduction of the bonus program may be the result of a combination of all three changes, the other components should not differentially affect flights close to the 15 minute threshold.[9]

In June 1996, TWA implemented an employee bonus program which closely resembled Continental's. The program was later amended to reward employees if high rankings were sustained for an entire quarter and, in 1999, was changed to reward absolute measures of on-time performance rather than relative rankings. 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. With the exception of American Airlines, all of these carriers introduced their programs after periods of poor performance. A notable difference between these later programs and the earlier programs, however, is that the later programs would only reward employees in months that the airline ranks first or second, even though by this time the number of carriers that participated in the rankings had increased.[10]

[8] In 1994, Continental had the worst average on-time performance ranking among the ten reporting airlines.
[9] However, 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.
[10] US Airways only rewarded a first place in the rankings.

Table 1 summarizes the details of these bonus programs and shows the number of months during the first year after the introduction of the bonus program in which the employees in fact earned bonuses. The table reveals that Continental's employees earned bonuses in 10 of the first 12 months after the introduction of the program while TWA's employees earned bonuses in four of the first 12 months. In contrast, 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. We suspect these differences are related to the changes in the number and types of airlines included in the rankings over time. As the final column of Table 1 shows, when Continental and TWA introduced their programs, only 10 carriers were included in the rankings. In addition, the group was generally more homogenous in size, network structure and geographic coverage.

A combination of growth by low-cost and regional carriers and reductions in capacity by the large network carriers increased the number of carriers that met the DOT reporting requirements from 10 to 17 carriers in 2003, peaking at 20 carriers between 2005 and 2007. The group also became more heterogeneous as low-cost carriers operating point-to-point networks as well as regional carriers qualified for the ranking. Moreover, two of the newly added carriers (Hawaiian Airlines and Aloha Airlines) occupied the top spot in the ranking in 98 percent of the months between November 2003 and the end of 2010, typically with substantially better on-time performance than the next highest ranked airline.[11] This combination of factors means that the three later bonus

[11] This is likely due to the fact that these two airlines operate at relatively uncongested airports that face few weather disruptions.

programs created much weaker incentives for gaming because the probability of the airline ranking high enough to achieve the bonus was extremely low.

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 requirement. We have collected these data for all reporting carriers for every year between 1988 and 2010, inclusive. However, our empirical work below utilizes three separate samples covering the different time periods during which the bonus programs are introduced: 1995 to 1998, 2002 to 2006, and 2008 to 2010.[12] We do this for several reasons. First, the volume of data is such that we cannot estimate regressions using all of the flights of all of the large carriers over a 15 year period in a single sample. Second, as we explain below, our identification strategy exploits variation across an airline's flights arriving at a given airport on a given day. Thus, changing the length of the sample does not substantially affect how our estimates are identified. Finally, given the aggregate changes that have impacted the industry over this 15 year period (e.g., fluctuations in aggregate demand, increases and decreases in congestion), we prefer to estimate our effects over shorter periods of time.

All of our regression samples include domestic flights operated by the following seven airlines: American Airlines, Continental Airlines, Delta Air Lines, Northwest

[12] 1995 is also the year in which the DOT began collecting data on wheels-off and wheels-on times and we require this particular data for our empirical analysis.

Airlines, TWA, United Airlines, and US Airways.[13] Even focusing on the three separate subsamples, the datasets are very large and so 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 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 for the 1995-1998 period includes 3,067,533 flights. The 2002-2006 and 2008-2010 samples include 2,942,492 and 1,349,666 observations respectively.

Table 2 presents summary statistics for the main variables in the data. The different panels of the table show summary statistics for our different regression samples. The average arrival delay in the early sample is about seven minutes and, during this period, 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. In the later samples, average arrival delays are shorter (with an average of about 4 minutes) though the fraction of flights more than 15 minutes late is similar (about 20%). Taxi-in and taxi-out times are slightly longer in the later samples and the average airtime of a flight is longer by 20 to 30 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

[13] The two later samples do not include TWA as it was acquired by American Airlines in 2001.

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 spikes at certain values. As the next set of histograms will show, these discrete spikes 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 minutes.[14]

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, we only show delays for flights between March and December 1998 in these histograms. 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

[14] 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.

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. Finally, Figure 2C shows arrival delays for Continental, Delta and TWA – the three airlines that used a combination of manual and automatic reporting during this time period. This histogram is quite smooth and looks much more like the histogram of the automatic reporters than the histogram of the manual reporters – 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.

In Figures 3A and 3B through Figures 7A and 7B, we compare the before-and-after distributions of arrival delays for each of the airlines that introduced employee bonus programs. 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. 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 and show a very similar pattern. 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 elsewhere 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 difference in the relative heights of the bars at 14 and 15 minutes.

IV. Empirical Approach

IV.A. Overview of Empirical Approach

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.[15] 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 ones 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

[15] 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. Schedule padding is potentially a costly strategy because it decreases aircraft utilization and increases labor costs which, in a typical airline contract, are based on the maximum of the scheduled and the actual flight time (see also Mayer and Sinai, 2003).

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 determining 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. 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 15 minute range to flights with very long expected delays. If the costs of delays are convex, then airlines should have the greatest incentives 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 specifically on those flights that they expect to arrive right around 15 minutes late since they simply do not know in advance which flights these will be. This eliminates selection concerns when comparing flights that are candidates for gaming to their "control groups" of flights outside the threshold range. It is also what makes an analysis of the employee bonus programs particularly relevant and interesting.

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 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 expected delay at the time that it touches down at the arrival airport.[16]

To construct each flight's expected delay at the time that its wheels touch down, we take the flight's wheels-on time and add to it the median taxi-in time for that flight in the quarter.[17] 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

[16] 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.

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

delay.[18] 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. 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 gaming is affected by the introduction of an employee bonus 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 the data that do not have a bonus program in place during the sample 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.

[18] 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 four minutes, then the flight's predicted arrival time is 4:40 pm and its predicted delay is 10 minutes.

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 of the 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-in 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 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.

Perhaps surprisingly, 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 13 percent shorter than the taxi-in times of its flights that are predicted to arrive less than 10 minutes late. Its flights with predicted delays of 16 to 17 minutes also have taxi-in times that are about 13 percent shorter. Moreover, the coefficients indicate a non-monotonic relationship between taxi-in times and a flight's predicted delay. While flights with predicted delays above or below the critical range also have negative coefficients – indicating they have shorter taxi-in times than the flights in the excluded category – their coefficients are smaller in absolute value indicating that the relative reduction in taxi-in times for these flights is not as great as for flights in the 15 minute 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 47 seconds. While this magnitude may appear small, our simulations below reveal that these selective reductions in delay can add up to meaningful changes in the on-time performance metric that is reported by the DOT.

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 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 these programs effectively provided much weaker incentives to employees because they only rewarded first or second rank at a time when the rankings included a large number of airlines some of which persistently outperformed the network carriers.

V.B. Does it Work?

The results in Table 3A suggest that – after the introduction of their bonus programs - Continental and TWA tried 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 were successful in doing so.[19] 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.

[19] 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.

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 the second and fourth columns, respectively, are again consistent with efforts to systematically reduce delays on flights that 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. To provide a sense of the magnitude of these estimates, consider the fact that, for both of these carriers, their flights that are predicted to be less than 10 minutes late arrive one minute earlier than predicted about 20 percent of the time. Thus, the coefficient estimates imply these airlines' flights in the critical 15 minute range are 50 percent more likely than their other flights to arrive exactly one minute earlier than predicted. For both of these carriers no other level of predicted delay has a coefficient that is in this range and almost none of the coefficients on any of the other predicted delay levels is statistically distinguishable from zero.

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 two minutes after the cutoff for being considered on-time) are 13 to 14 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 and is quite large in magnitude given that their flights in the excluded category arrive two minutes earlier than predicted only about 10 percent of the time. 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 misreporting the arrival times of flights that would have arrived 15 or 16 minutes late.[20] This would still represent a form of gaming of the incentive program; however, it would be a different type of gaming than actual reductions in taxi-in times. In addition, the welfare implications would be different.

[20] In our data, taxi-in times are calculated as the difference between arrival times and wheels-on times. As a result, given a plane's wheels-on 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.

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 have their delays 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 reported 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 arrived 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 in 1996, we see that TWA's distribution is very similar 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 any plane that has reported delays of zero for more than 5% of its flights in *any* year as a manual plane for *every* year of our sample. We see this as a conservative approach to identifying manual planes since it classifies a plane as manual based on it meeting the criteria described above in only a single year. However, since our goal is to construct a sample of planes that we are fairly confident are not manual, we prefer to erroneously categorize some automatic planes as manual rather than erroneously categorize some manual planes as automatic. Using this approach, we classify 85 of Continental's 441 planes (19%) and 22 of TWA's 241 planes (9%) as manual.

We then re-estimate our earlier regressions with separate predicted delay bins for Continental's and TWA's manual and automatic planes. 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 one 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 for manual planes and the difference in taxi-in times and in the probability of arriving one minute early for threshold flights is larger for manual planes. Since the manual planes only account for a small fraction of the airline's overall flights, the coefficient estimates for the automatic planes is only slightly smaller in magnitude than the estimates in Tables 3A and 4A for the full sample of planes. Our hypothesis tests for automatic planes 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 interpret this as evidence that at least some of the gaming we are measuring represents actual reductions in taxi-in times.

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 an additional 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. In addition, we have estimated our regressions for Continental in the years after it becomes a fully automatic reporter (which takes place in February 2002). We find similar patterns of gaming by Continental in 2002 though the magnitudes are smaller but do not find evidence of gaming in 2003 or later years. However, as described earlier, in 2003 the number of airlines included in the rankings increased substantially and included several small carriers who consistently outperformed the large carriers on on-time performance. As a result, the incentives inherent in Continental's program – like those introduced by American and US Airways in the later time period – were much weaker relative to the earlier time period.

V.D. Analysis of Paired Flights

The identification strategy used in all of our earlier analyses exploits variation in delays incurred prior to arrival across a carrier's flights arriving at the same airport on the same day. While 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 an airline that land at the same airport on the same day during the same minute. 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 and we relate this variable to a measure of the predicted delay of the early member of the pair. 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 threshold for being considered on-time. The benefit of this is that if there is an unobservable that is correlated with both a flight's arrival time and its taxi-in time, this unobservable should equally affect the threshold flight and the flight with which it is paired.

This empirical exercise requires several changes to the sample and specification. First, because we are only using pairs of flights that land at the exact same time and that have one member of the pair that is predicted to be more than 25 minutes late, we no longer restrict to a random sample of every fifth day of the year. Even utilizing the full sample, we only have about 179,000 pairs (as compared to over 3 million flights in the earlier regressions). Second, we do not have enough pairs by a given airline at a given airport on a given day to include airline-destination-day fixed effects as we do before. Instead, we include airline-destination-month fixed effects and the following control variables: a measure of airport congestion at the arrival time of the pair, dummies for each arrival hour of the day, separate dummies for whether the early flight and late flight departed from the airline's hub and measures of the distance of each flight in the pair. Third, since we still want to separately estimate the effects for four groups of flights (Continental, TWA pre- and post-bonus, and all other carriers) but do not want to have predicted delay bins that consist of a very small number of pairs, we replace the minute-by-minute predicted delay bins that we have used so far with a smaller number of wider predicted delay bins. Specifically, we distinguish four levels of predicted delay for the early member of a pair: predicted delay of less than 10 minutes (which is used as the excluded category as before), predicted delay between 10 and 13 minutes, predicted delay between 14 and 17 minutes, and predicted delay of greater than 17 minutes.

The results of this empirical exercise are presented in Table 6. Each column presents the coefficients for one of the four groups of flights that we distinguish. Each coefficient represents the change in 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 has predicted delay in the particular range relative to when the "late" member is paired with a flight with predicted delay less than 10 minutes. The first column shows the estimates for all non-bonus carriers. We find no evidence that the probability of the late flight having a shorter taxi-in time is affected by the predicted delay of its paired flight. On the other hand, the estimates for Continental indicate that when a late flight lands with a flight that is predicted to be15 to 17 minutes late, it is almost 13 percentage points less likely to have a shorter taxi-in time than when it lands with a flight that is predicted to be less than 10 minutes late. While it is reasonable to expect that the probability that the late flight wins falls with the expected delay of the other flight in its pair, one would expect to observe a monotonic relationship and this is not what the results for Continental show as the magnitude of the coefficient on the next predicted delay bin is significantly smaller. 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, while TWA's flights exhibit this pattern both before and after the introduction of its bonus program, the pattern is more pronounced before. Since airlines typically only have pairs of flights that land at the same time at their hubs and since TWA only has a single hub (at St. Louis), the results for TWA may be sensitive to other changes TWA made at its lone hub around the time it introduced its bonus program.[21]

V.E. Externalities

All of our results indicate that, after the introduction of their bonus programs, both Continental and TWA selectively reduced delays on flights that would otherwise

[21] We have also estimated these paired models for American, US Airways and United when they introduce their bonus programs and, consistent with our earlier analyses, find no evidence of gaming.

have been likely to arrive just above the cut-off for being considered on-time. While some of this may be misreporting, given the small number of manual planes we identify, much of what we are measuring is likely actual reductions in flights' taxi-in times. If the reductions in the taxi-in times of threshold flights are driven by the reallocation of scarce resources, negative externalities on other flights may result. Furthermore, if resources are reallocated from flights where the cost of an additional minute of delay is greater than on threshold flights, then this behaviour will be welfare-reducing. On the other hand, if the shorter taxi-in times on threshold flights are a result of lying or of higher levels of effort from slack resources (e.g., ground crew), then gaming will not impose externalities on other flights.

Empirically uncovering externalities that may result from a reallocation of scarce resources is difficult for a number of reasons. First, it requires us to identify those periods of time when resources are, in fact, scarce. This will depend on how airlines match their demand for and supply of airport and personnel resources over the course of the day. In addition, it will depend on the extent to which actual schedules deviate from anticipated schedules. Second, even if we could identify periods when resources are likely to be scarce and speeding up a threshold flight would require resources to be reallocated, we have no way of knowing which flights will be affected and what way (e.g.: departure or arrival delays). As a result, we are at risk of either missing the effects (if we focus on a very small set of flights) or diluting the effects (if we include many flights and estimate averages).

We have carried out a number of different empirical analyses that explore the existence of externalities and have not found evidence that the gaming behaviour that we have documented imposes negative externalities on other flights. At the same time, we cannot rule out that externalities may exist. While the paired analysis described above finds that late flights that land with a threshold flight are less likely to have a shorter taxi-in time than the flight they land with, we do not find that those same flights have longer than expected taxi-in times. This suggests that the threshold flight is being sped up but not at the expense of the late flight with which it lands. However, the threshold flight may of course be sped up at the expense of other flights which are not members of the pair. In addition, we have estimated a series of regressions in which we relate the probability that a particular flight lands later than predicted as a function of the number or fraction of threshold flights landing within five minutes of the flight. We do not find that flights that land close to one or more threshold flights are systematically more likely to arrive later than we predict but, again, we cannot be certain that the flights which may be affected are contained within any specific time window we select.

V.F Additional Results and Robustness Checks

The richness of our data allows us to carry out a large number of supplemental analyses and robustness checks that we briefly describe here.[22] We have replaced our carrier-arrival airport-day fixed effects with flight-quarter fixed effects and find that our results are robust to this modification. We have also 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

[22] The results for any of these additional analyses are available from the authors upon request.

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.

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.

Finally, we have investigated whether there is any evidence that airlines appear to systematically reduce airtimes in response to a flight's predicted delay at the time of departure. To do this, we have estimated regressions analogous to the taxi-time regressions but with a flight's airtime on the left-hand side and using predicted delay bins that are based on a flight's predicted delay at the time that its wheels leave the ground. We find no evidence that airtimes are systematically shorter for flights that – upon departure – are predicted to be about 15 minutes late. A likely explanation for this is that the delay prediction at the time of departure is quite noisy; thus the airline may not want to devote resources to specific flights based on this prediction.

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 three years 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 both Continental and TWA is about one percentage point. Put differently, these airlines' selective reduction of taxi-in times results in their fraction of flights delayed 15 minutes or more falling, on average, by one percentage point. For both Continental and TWA, these effects are slightly larger (about 1.2 percentage points) in the second and third year after they introduce their bonus programs. Given that positions in the DOT ranking are often determined by very small differences in absolute on-time performance between airlines, these changes in the fraction of flights delayed more than 15 minutes map into changes in Continental's and TWA's rankings. For example, we find that the taxi time distortions result in Continental achieving an improvement in rankings of at least one position in 15 of the 35 months following the introduction of their program. The simulations indicate that, in 1997, gaming improved Continental's rank in 10 of the 12 months of that year with its rank improving by two or more positions in three of those months. When we simulate TWA's taxi-in times after the introduction of its bonus program, we find its rank improved in 11 of the 31 months we look at it. 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, this selective reduction of delays can impact the reporting rankings and the information conveyed to consumers.

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 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 place in real-time by front-line employees who may not have the incentives to manipulate delay in the necessary way.

Our empirical analysis finds no evidence of gaming by airlines who 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 be – and often were – achieved. We find that those airlines have systematically shorter taxi-in times for their flights that are predicted to arrive close to the 15 minute cut-off for being considered on-time. 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 45 to 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 the main source of information communicated to consumers.

This paper contributes the growing empirical literature on gaming of disclosure programs by explicitly considering the interaction between the dimensions of quality that a program reports, the scope for measured quality to be manipulated and the incentives of those individuals who are the best position to manipulate the relevant dimensions of quality. We believe that considering these interactions in other disclosure settings will help shed light on the potential for gaming as well as explain possible variation in whether and when gaming takes place.

References

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

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

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

Forbes, Silke J. (2008), "The Effect of Air Traffic Delays on Airline Prices", *International Journal of Industrial Organization* 26(5), 1218-1232.

Hastings, Justine and Jeffrey Weinstein (2008), "Information, School Choice, and Academic Achievement: Evidence from Two Experiments", *Quarterly Journal of Economics* 123(4), 1373-1414.

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

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

Larkin, Ian (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).

Mayer, Chris and Todd Sinai (2003), "Why Do Airlines Systematically Schedule Their Flights to Arrive Late?", *Working Paper*, University of Pennsylvania (Wharton School of Business).

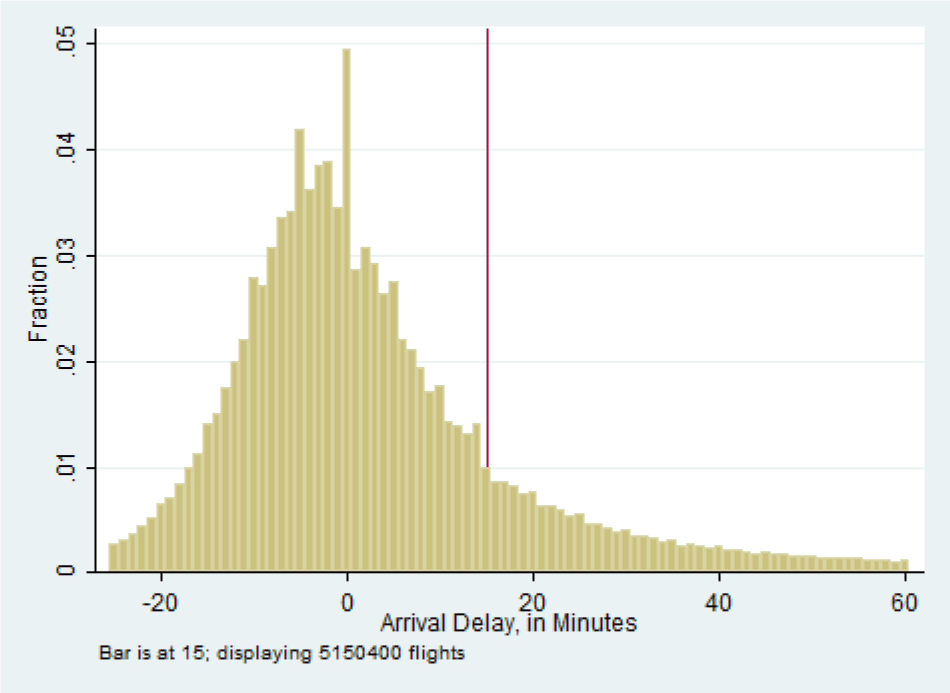
Neal, Derek and Diane W. Schanzenbach (2010), "Left Behind by Design: Proficiency Counts and Test-Based Accountability", *Review of Economics and Statistics* 92(2), 263-283.

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

Sallee, James and Joel Slemrod (2010), "Car Notches: Strategic Automaker Responses to Fuel Economy Policy", *NBER Working Paper* 16604.

Werner, Rachel and David 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



Bar is at 15; displaying 5150400 flights

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

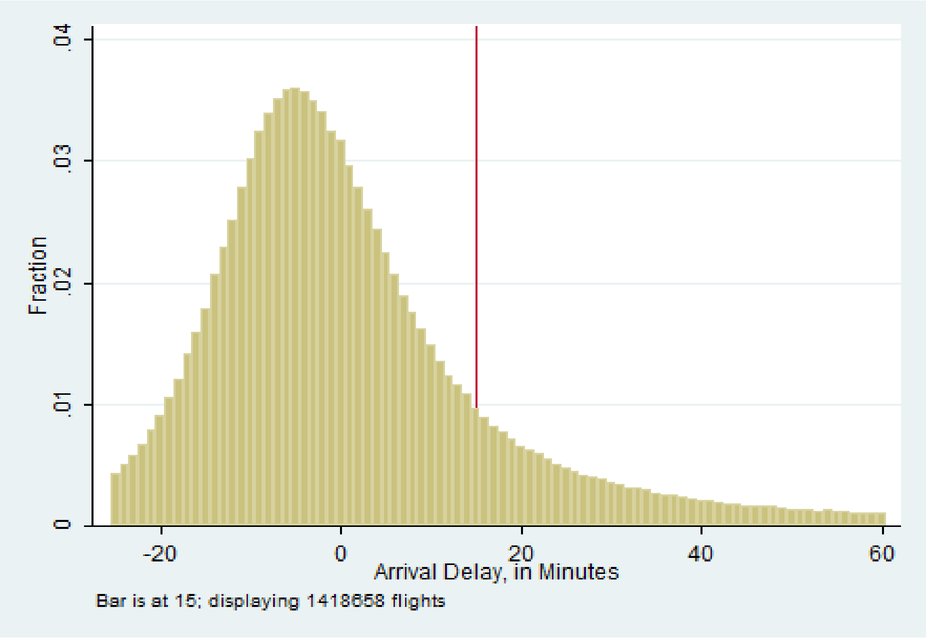


Figure 2B
Distribution of Arrival Delays
Manual Reporters, March – December 1998

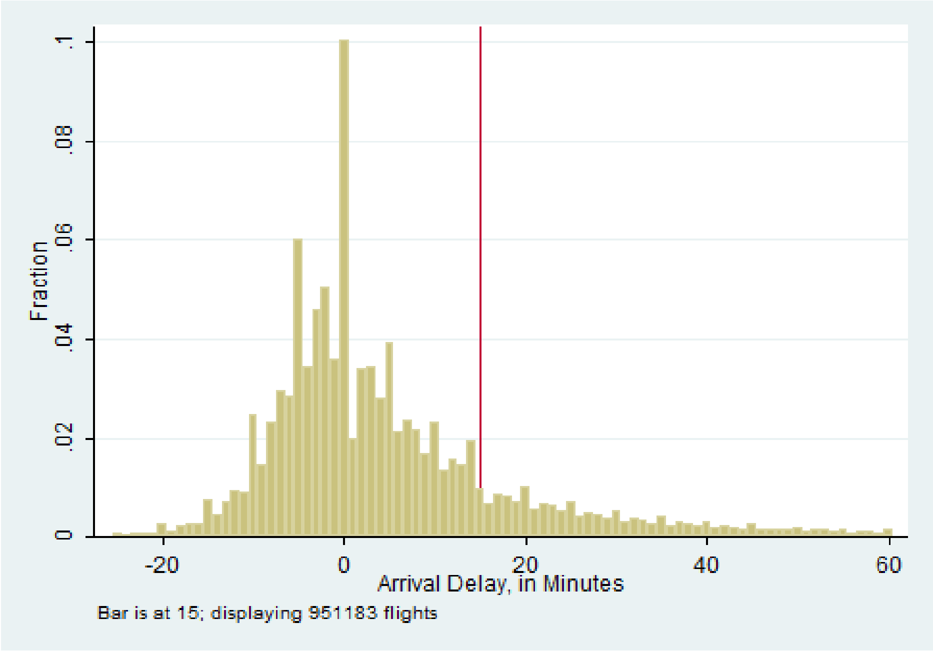


Figure 2C
Distribution of Arrival Delays
Combination Reporters, March – December 1998

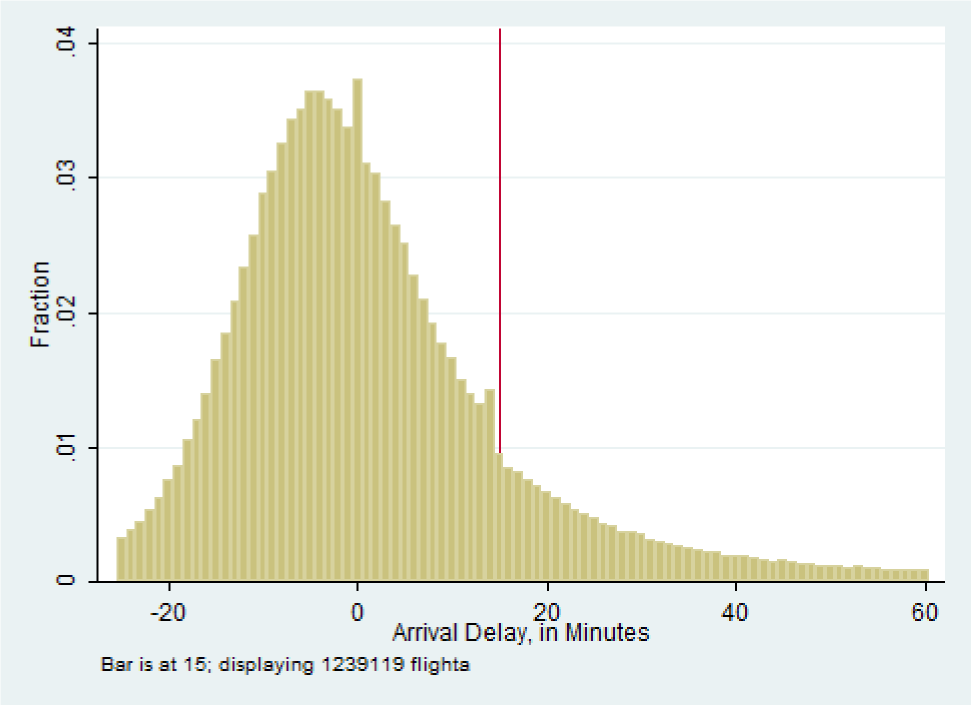


Figure 3A
Distribution of Arrival Delays
Continental Airlines, 1993-1994

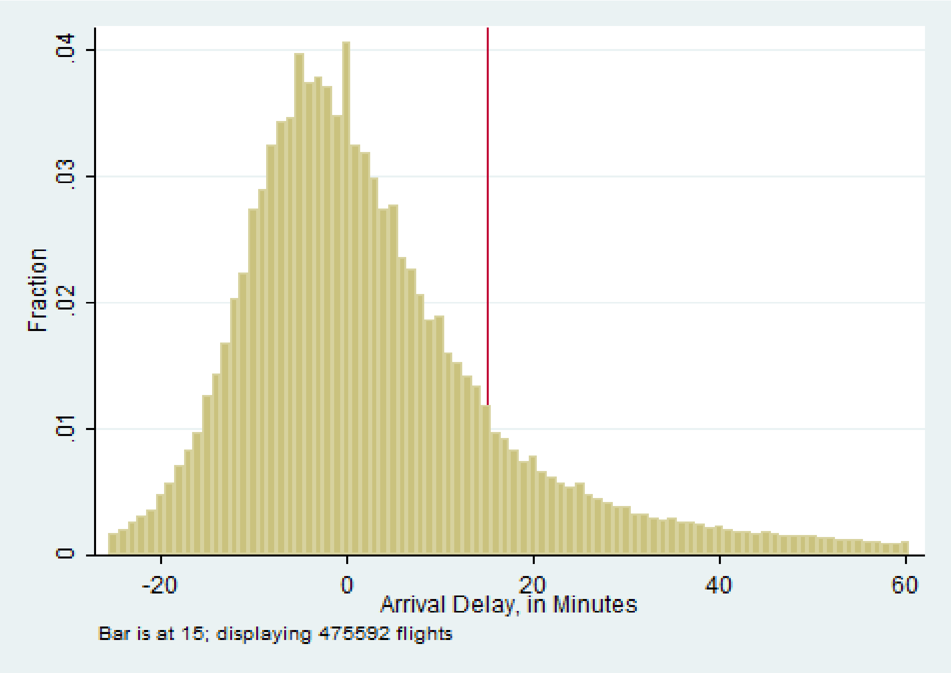


Figure 3B
Distribution of Arrival Delays
Continental Airlines, February 1995-1997

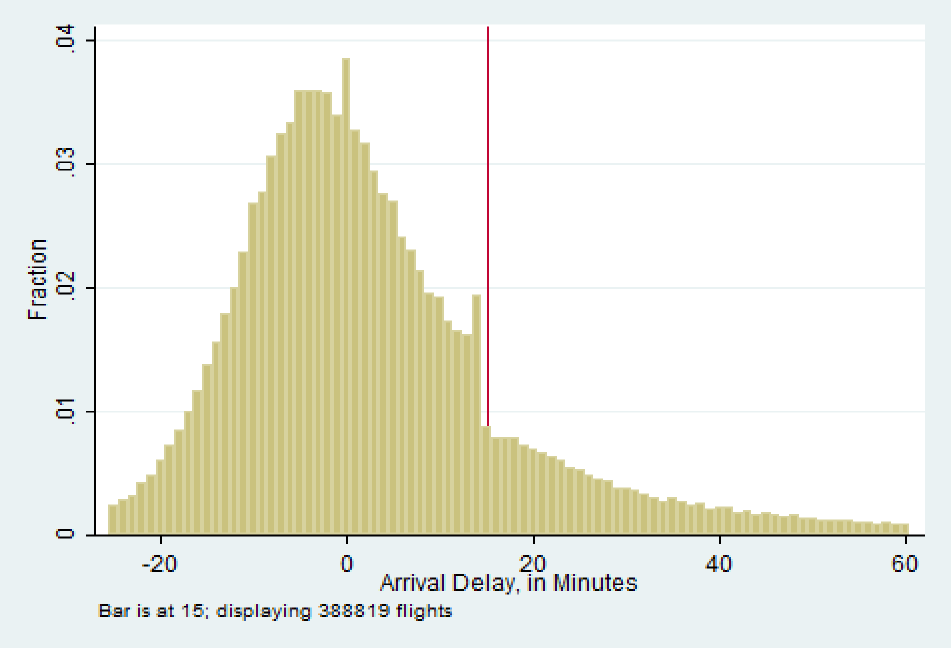


Figure 4A
Distribution of Arrival Delays
TWA, 1994-1995

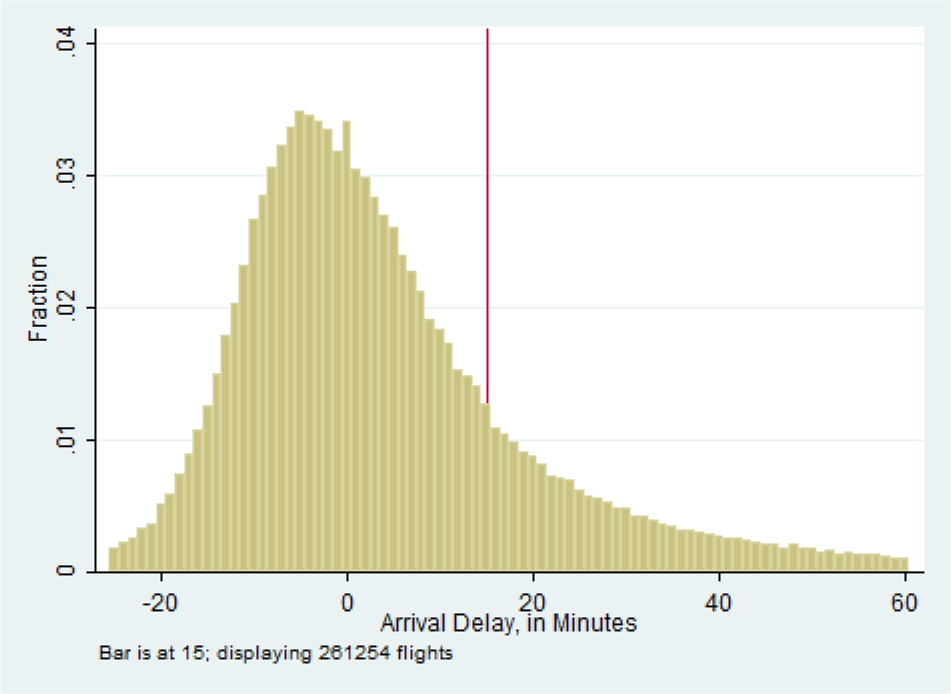


Figure 4B
Distribution of Arrival Delays
TWA, June 1996-1998

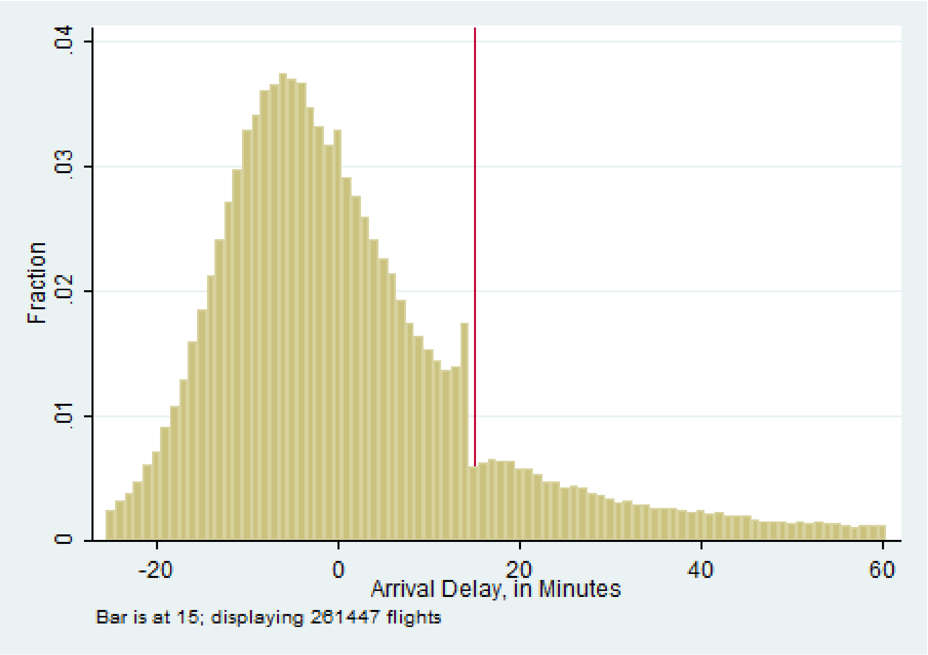


Figure 5A
Distribution of Arrival Delays
American Airlines, 2002

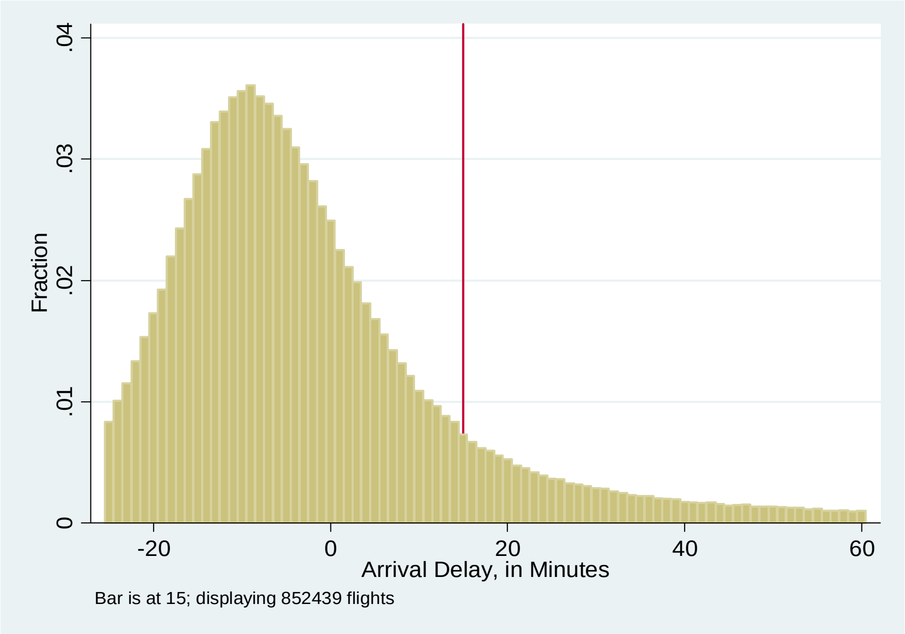


Figure 5B
Distribution of Arrival Delays
American Airlines, 2003

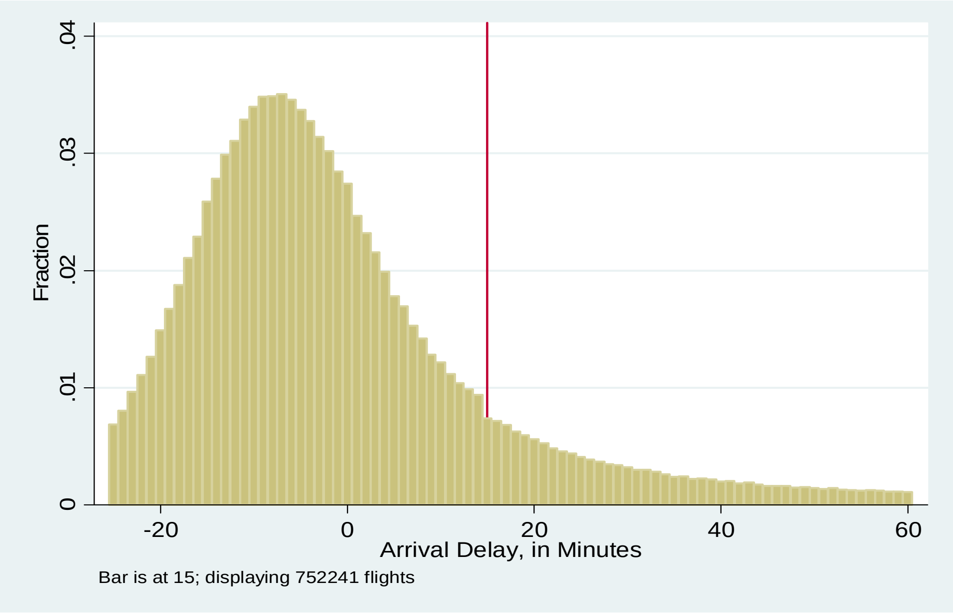


Figure 6A
Distribution of Arrival Delays
US Airways, 2004

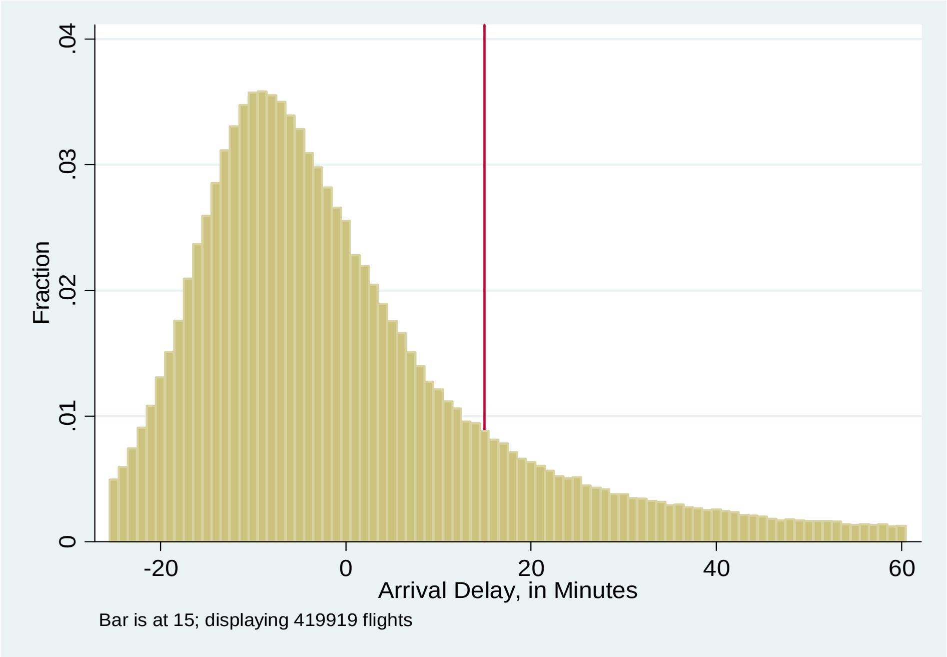


Figure 6B
Distribution of Arrival Delays
US Airways, 2004

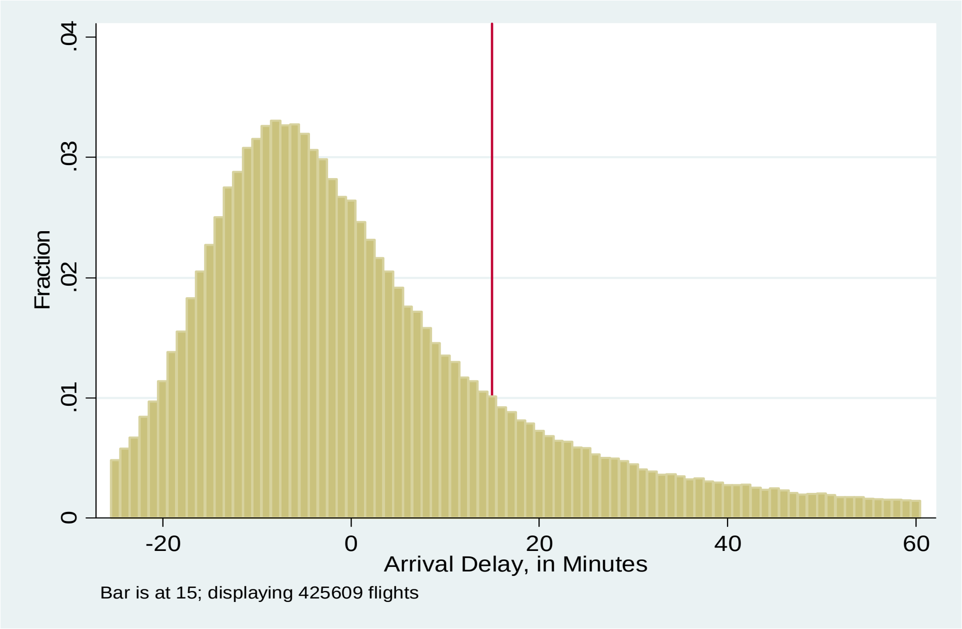


Figure 7A
Distribution of Arrival Delays,
United Airlines, 2008

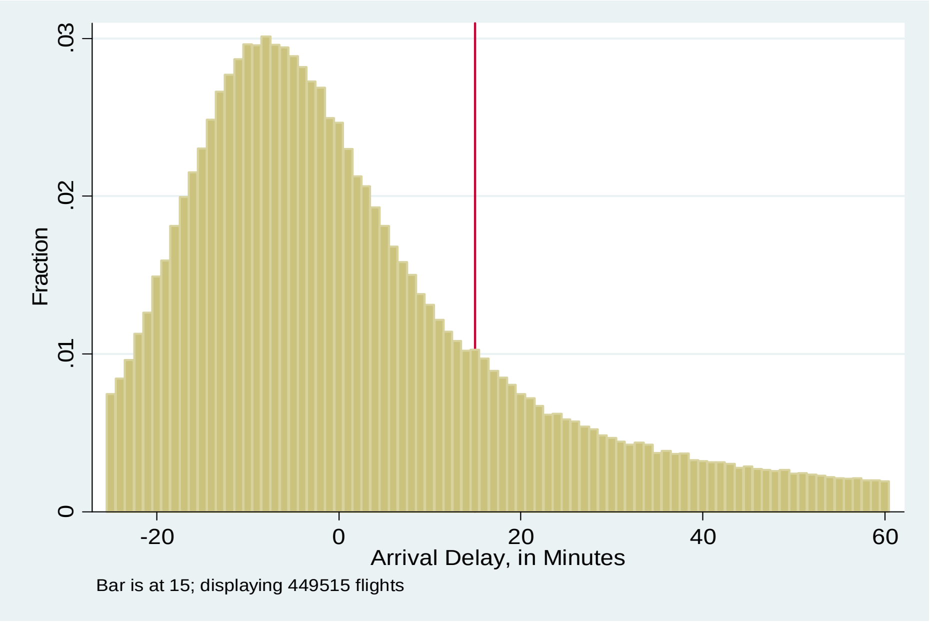


Figure 7B
Distribution of Arrival Delays
United Airlines, 2009

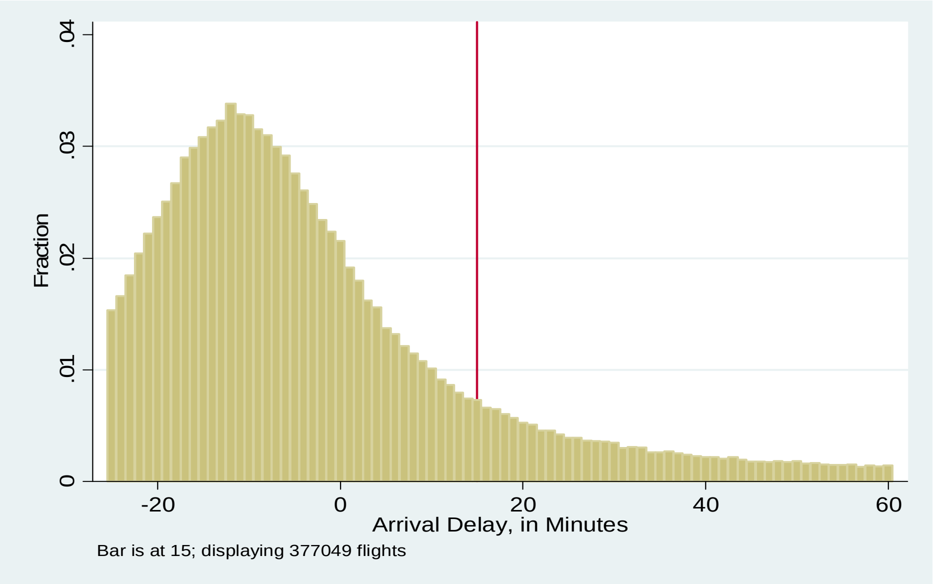


Figure 8A
Coefficients on Continental's Predicted Delay Bins (post-bonus)
(From Table 3A)

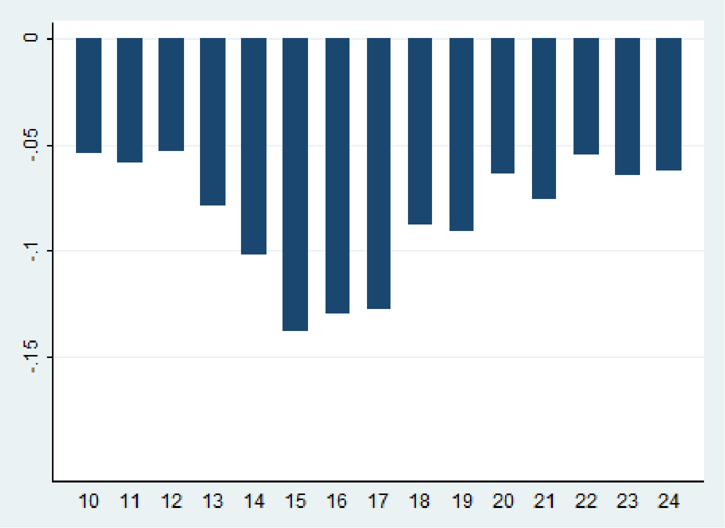


Figure 8B
Coefficients on TWA's Predicted Delay Bins (post-bonus)
(From Table 3A)

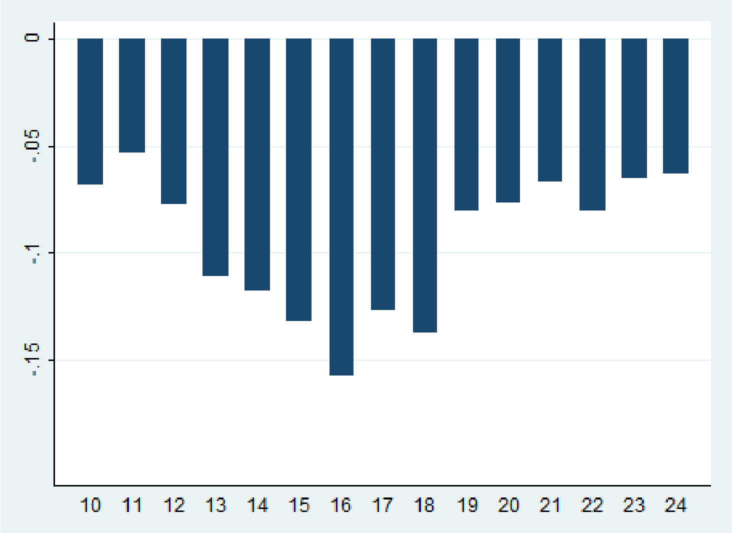


Figure 9A
Coefficients from Taxi-Time Regression
Continental's Predicted Delay Bins – Manual vs. Automatic Planes

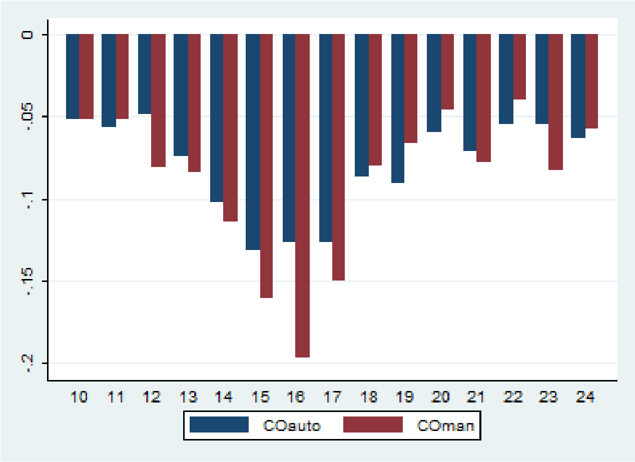
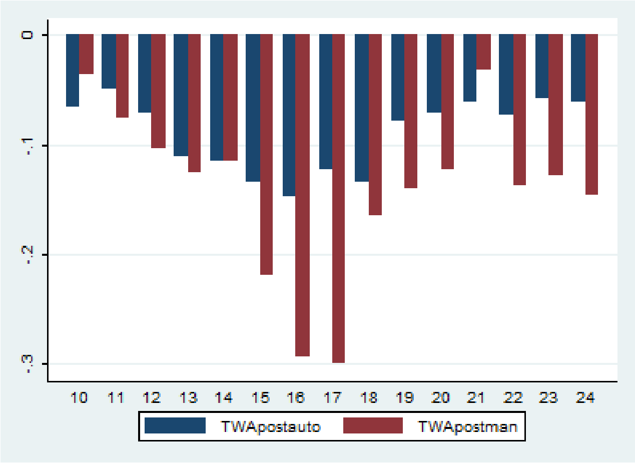


Figure 9B
Coefficients from Taxi-Time Regression
TWA's Predicted Delay Bins – Manual vs. Automatic Planes (post-Bonus)



Notes: Blue bars are for automatic planes, red bars are for manual planes. Both types of planes exhibit similar patterns, but the gaming around the threshold is more pronounced for manual planes.

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

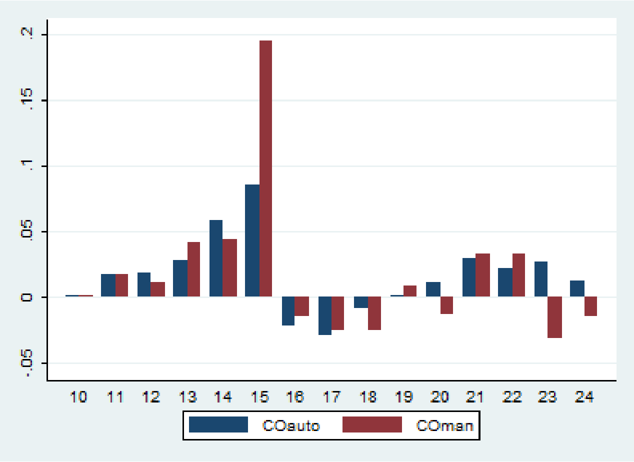
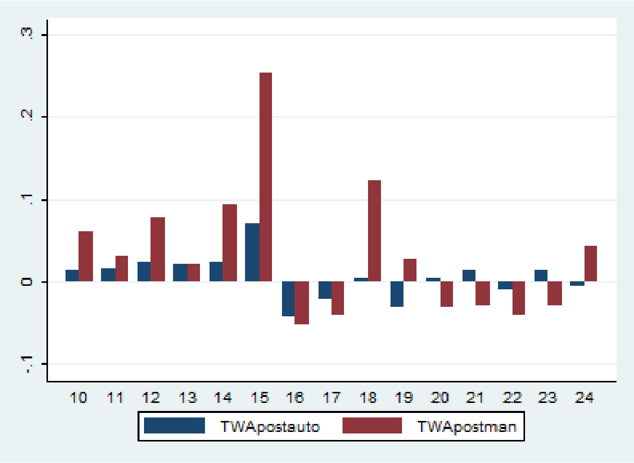


Figure 10B
Coefficients from 1 Minute Early Regression
TWA's Predicted Delay Bins – Manual vs. Automatic Planes (post-Bonus)



Notes: Blue bars are for automatic planes, red bars are for manual planes. Both types of planes exhibit similar patterns, but the gaming around the threshold is more pronounced for manual planes.

Table 1
Overview of Bonus Programs

Airline	Payment Structure	# Months Bonus Achieved in First Year After Introduction	# Airlines in Ranking when Bonus Introduced
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.	10	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.	4	10
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.	0	17
US Airways (Start: May 2005)	$75 per employee in each month in which the airline ranks 1st.	0	19
United (Start: Jan 2009)	$100 per employee in each month that the airline ranked 1st. $65 in months that the airline ranked 2nd.	1	20

Table 2
Summary Statistics for Regression Sample
February1995 - December 1998

	Mean	Standard Deviation	Min	Max
1995-1998 Sample (3,067,533 observations)				
Arrival Delay (min)	7.22	27.99	-88	1182
Dummy for Arrive 15 Minutes Late or More	0.21	0.41	0	1
Taxi In Time (min)	6.10	3.92	1	60
Departure Delay (min)	8.43	25.43	-15	1185
Taxi Out Time (min)	14.91	7.44	1	60
Flight Time	108.7	66.50	20	632
2002-2006 Sample (2,942,493 observations)				
Arrival Delay (min)	4.03	29.74	-89	1925
Dummy for Arrive 15 Minutes Late or More	0.18	0.38	0	1
Taxi In Time (min)	7.02	4.54	1	60
Departure Delay (min)	5.76	27.21	-15	1930
Taxi Out Time (min)	16.92	7.96	1	60
Flight Time	125.04	74	20	713
2008-2010 Sample (1,340,666 observations)				
Arrival Delay (min)	4.45	34.77	-90	1632
Dummy for Arrive 15 Minutes Late or More	0.20	0.40	0	1
Taxi In Time (min)	7.69	5.06	1	60
Departure Delay (min)	8.31	32.29	-15	1626
Taxi Out Time (min)	17.59	8.29	1	60
Flight Time	135.61	78.08	20	677

***Notes*:** Includes flights by American, Continental, Delta, Northwest, TWA, United, and US Airways. TWA acquired by American in 2001.

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. 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 clustered at 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 coefficients for carriers with bonus programs, after its introduction. Columns 1 and 2 are based on data from 2002-2006 (2,942,493 observations). Column 3 is based on data from 2008-2010 (1,340,666 observations). Specifications include 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	***=1 if Flight 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* (0.00209)	0.000474 (0.00624)	-0.0204 (0.0121)	0.0185 (0.0101)
[11,12) min	0.00522* (0.00213)	0.0177* (0.00686)	0.00500 (0.0124)	0.0160 (0.00987)
[12,13) min	0.00290 (0.00224)	0.0158* (0.00689)	-0.00768 (0.0132)	0.0279** (0.0108)
[13,14) min	0.00673** (0.00235)	0.0312*** (0.00736)	0.00412 (0.0144)	0.0228 (0.0121)
[14,15) min	0.00997*** (0.00247)	0.0560*** (0.00803)	-0.0145 (0.0148)	0.0318** (0.0120)
[15,16) min	0.0101*** (0.00257)	0.111*** (0.00852)	0.0106 (0.0157)	0.0888*** (0.0132)
[16,17) min	0.00769** (0.00261)	-0.0196** (0.00760)	0.00146 (0.0151)	-0.0435*** (0.0118)
[17,18) min	0.00957*** (0.00272)	-0.0274*** (0.00779)	-0.0125 (0.0155)	-0.0223 (0.0125)
[18,19) min	0.0128*** (0.00285)	-0.0131 (0.00870)	0.00905 (0.0174)	0.0127 (0.0134)
[19,20) min	0.00896** (0.00295)	0.00288 (0.00924)	-0.000275 (0.0180)	-0.0292* (0.0122)
[20,21) min	0.0127*** (0.00306)	0.00856 (0.00998)	0.0258 (0.0194)	0.000948 (0.0147)
[21,22) min	0.00504 (0.00323)	0.0302** (0.0102)	-0.00486 (0.0188)	0.0109 (0.0153)
[22,23) min	0.0131*** (0.00325)	0.0244* (0.0102)	-0.0230 (0.0185)	-0.0119 (0.0150)
[23,24) min	0.00931** (0.00344)	0.0135 (0.0105)	-0.0133 (0.0183)	0.00964 (0.0161)
[24,25) min	0.00837* (0.00346)	0.00808 (0.0108)	0.0411 (0.0233)	-0.00246 (0.0170)
≥25 min	0.00799*** (0.000916)	0.00993*** (0.00264)	-0.000805 (0.00555)	0.00813 (0.00441)

Notes: Standard errors are in parentheses and clustered at 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 flights with predicted delay of less than 10 minutes. The regression contains 3,067,533 observations.

Table 4B
Probability of Arriving Exactly *Two* Minutes Earlier than Predicted, 1995-1998

Dependent Variable	***=1 if Flight 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 clustered at 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 flights with predicted delay of less than 10 minutes. The regression contains 3,067,533 observations.

Table 5
Identification of "Manual" Planes, 1996
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.

Table 6
Analysis of Pairs of Flights that Land at the Exact Same Time

Dependent Variable	***=1 if "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				
[10,14) min	-0.0220*** (0.00512)	-0.0154 (0.0236)	0.0477 (0.0335)	-0.0264 (0.0375)
[14,18) min	-0.0290*** (0.00643)	-0.129*** (0.0207)	-0.128*** (0.0384)	-0.0653*** (0.0160)
≥18 min	-0.0279*** (0.00328]	-0.0305** (0.0100)	-0.0166 (0.00927)	-0.0473*** (0.00814)
# of pairs in 14-18 minute range	8492	617	158	327

Notes: Sample includes carriers' flights that touch down at the exact same minute. Restricted to two-member pairs in which the "late" member of the pair is predicted to be 25 or more minutes late. Standard errors are in parentheses. Regression includes carrier-destination-month fixed effects and the controls described in the text on page 30. 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. The regression contains 179,557 observations.

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

Year	Month	Actual Fraction Delayed	Simulated Fraction Delayed	Difference in Fraction Delayed	Actual Rank	Simulated Rank
1995	2	0.1728	0.1767	0.0039	4	4
1995	3	0.1521	0.1565	0.0043	1	1
1995	4	0.1468	0.1506	0.0039	1	2
1995	5	0.1984	0.2005	0.0021	8	8
1995	6	0.3355	0.3276	-0.0079	10	10
1995	7	0.1733	0.1777	0.0044	2	3
1995	8	0.1304	0.1358	0.0053	1	2
1995	9	0.1051	0.1095	0.0044	2	2
1995	10	0.1341	0.1412	0.0071	3	3
1995	11	0.1730	0.1785	0.0056	3	4
1995	12	0.2152	0.2208	0.0056	1	1
1996	1	0.2408	0.2491	0.0083	2	2
1996	2	0.1931	0.2020	0.0090	2	2
1996	3	0.2054	0.2153	0.0099	4	6
1996	4	0.1827	0.1943	0.0117	4	4
1996	5	0.1359	0.1472	0.0113	2	2
1996	6	0.2502	0.2657	0.0154	6	6
1996	7	0.2209	0.2334	0.0125	5	5
1996	8	0.2399	0.2544	0.0145	5	5
1996	9	0.1999	0.2128	0.0129	4	4
1996	10	0.1828	0.1935	0.0108	3	3
1996	11	0.1692	0.1790	0.0098	1	1
1996	12	0.2455	0.2586	0.0130	1	1
1997	1	0.2482	0.2610	0.0127	2	3
1997	2	0.1893	0.2039	0.0146	2	3
1997	3	0.1965	0.2122	0.0157	5	8
1997	4	0.1819	0.1938	0.0119	6	6
1997	5	0.1742	0.1851	0.0109	8	9
1997	6	0.2175	0.2279	0.0105	7	8
1997	7	0.1772	0.1896	0.0123	3	4
1997	8	0.1762	0.1885	0.0123	4	5
1997	9	0.1402	0.1518	0.0116	5	7
1997	10	0.1747	0.1884	0.0137	6	8
1997	11	0.2081	0.2195	0.0115	6	6
1997	12	0.2304	0.2418	0.0113	2	4

Number of months in which actual rank is **better** than simulated: **15**
Number of months in which actual rank is **same** as simulated: **20**
Number of months in which actual rank is **worse** than simulated (others simulated): **0**

Notes: Variation in the simulated fraction delayed is minor. Based on 20 iterations, t-statistics for the simulated fraction delayed are typically over 300, with no month having a t-statistic below 100. The differences between the simulated and actual fractions are highly significant.

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

Year	Month	Actual Fraction Delayed	Simulated Fraction Delayed	Difference in Fraction Delayed	Actual Rank	Simulated Rank
1996	6	0.2908	0.2915	0.0007	8	8
1996	7	0.3039	0.3058	0.0018	8	8
1996	8	0.2903	0.2951	0.0048	8	8
1996	9	0.2130	0.2145	0.0015	5	5
1996	10	0.2184	0.2217	0.0034	4	5
1996	11	0.1901	0.1913	0.0011	5	5
1996	12	0.3345	0.3348	0.0004	7	7
1997	1	0.2891	0.2928	0.0037	6	6
1997	2	0.2117	0.2158	0.0041	5	5
1997	3	0.2064	0.2113	0.0048	7	8
1997	4	0.1423	0.1478	0.0055	1	1
1997	5	0.1059	0.1118	0.0059	1	1
1997	6	0.1410	0.1508	0.0098	1	1
1997	7	0.1315	0.1467	0.0152	1	2
1997	8	0.1531	0.1713	0.0182	2	3
1997	9	0.0863	0.0997	0.0134	1	1
1997	10	0.1186	0.1314	0.0128	1	2
1997	11	0.1901	0.2051	0.0149	3	5
1997	12	0.2795	0.2992	0.0197	8	9
1998	1	0.2280	0.2418	0.0139	5	5
1998	2	0.1916	0.2098	0.0182	4	4
1998	3	0.2605	0.2780	0.0175	9	9
1998	4	0.1921	0.2092	0.0171	5	6
1998	5	0.2139	0.2313	0.0174	5	6
1998	6	0.3056	0.3187	0.0131	6	7
1998	7	0.1878	0.2032	0.0154	6	6
1998	8	0.1432	0.1521	0.0089	1	1
1998	9	0.1097	0.1196	0.0099	1	3
1998	10	0.1068	0.1180	0.0112	1	1
1998	11	0.1084	0.1203	0.0119	1	1
1998	12	0.2174	0.2301	0.0127	5	5

Number of months in which actual rank is **better** than simulated (others simulated): **11**
Number of months in which actual rank is **same** as simulated (others simulated): **20**
Number of months in which actual rank is **worse** than simulated (others simulated): **0**

Notes: Variation in the simulated fraction delayed is minor. Based on 20 iterations, t-statistics for the simulated fraction delayed are typically over 300, with no month having a t-statistic below 100. The differences between the simulated and actual fractions are highly significant.