

Quality Disclosure Programs with Thresholds: Misreporting, Gaming, and Employee Incentives

Silke J. Forbes
University of California, San Diego

Mara Lederman
University of Toronto, Rotman School of Management

Trevor Tombe
Wilfrid Laurier University

March 2012

Abstract

Many quality disclosure programs provide consumers with measures that are based on quality passing a discrete threshold. Such measures invite gaming on the part of firms because they create a discontinuity in the relationship between actual and reported quality right around the relevant threshold. In this paper, we explore whether and why – for a given disclosure program with this type of structure – gaming may vary across firms and within firms over time. In particular, we focus on how gaming may be affected by the technology used for reporting quality as well as by the incentives provided to employees. Our empirical context is 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 less than 15 minutes late. We estimate whether airlines game this program by selectively reducing delays on flights that are expected to land right around the 15 minute threshold. We find strong evidence of gaming by airlines that report their delays manually which we suspect captures airlines misreporting the arrival times of threshold flights so that they appear to land “on-time”. We also find strong evidence of gaming by some of the airlines that explicitly incentivize their employees based on the airline’s performance in the government’s program. Our findings highlight that gaming of a disclosure program will not only depend on the design of the program but also on the extent to which 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, Misreporting, Gaming, Incentives

JEL codes: L2, L5

We thank Severin Borenstein, Florian Ederer, Bob Gibbons, Ginger Jin, Phillip Leslie, Matt Mitchell, Paul Oyer, Steve Puller and seminar participants at Boston University, Case Western Reserve University, Georgetown University, New York University (Stern), UC Berkeley (Haas), UC Los Angeles (Anderson), UC San Diego, the University of Maryland, the University of Rochester (Simon), the University of Toronto (Rotman), the U.S. Department of Justice, the AEA Meetings (2011), the Berkeley-Stanford IO Fest (2010), the NBER IO Summer Institute (2011), the NBER Organizational Economics Working Group Meeting (2011) and the USC Conference on Game Theory in Law, Business and Political Economy 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 When designing such programs, regulators often choose to report quality using simplified measures that are based on discrete thresholds.² While such measures have the advantage that they are easily understood by consumers (see, for example, Hastings and Weinstein, 2008), they also create a discontinuity in the relationship between actual and reported quality right around the relevant threshold. In particular, small improvements in quality that are close to the threshold may have a large impact on reported quality while large improvements in quality that are well above or below the threshold may have no impact at all. As a result, firms subject to such programs face strong incentives to focus on quality improvements that move them over the relevant threshold, potentially at the expense of quality improvements which may be equally or even more highly valued by consumers. Indeed, the empirical literature on quality disclosure has documented numerous examples of “gaming” by individuals and firms subject to disclosure programs with this type of design (see, for example, Figlio and Getzler (2002), Jacob and Levitt (2003), Jacob (2005), Cullen and Reback (2006), Neal and Schanzenbach (2010) and Macartney (2011) on educational accountability and Dranove, Kessler, McClellan and Satterthwaite (2003), Lu (2009) and Snyder (forthcoming) in health care).³

In this paper, we explore whether and why – for a *given* disclosure program – the incidence of gaming may vary across firms and over time. In particular, we focus on two factors

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

² Examples include student test scores, hospital report cards, fuel economy standards, green building standards and restaurant hygiene.

³ There is also a related literature on how firms and employees respond to discontinuous incentives. For example, see Sallee and Slemrod (2010) on the response to “notches” in fuel economy standards, Bergstresser, Desai and Rauh (2006) on earnings manipulation and CEO incentives, and Oyer (1998), Courty and Marschke (2004) and Larkin (2008) on gaming of incentive schemes by employees.

which we expect will affect the extent to which firms engage in gaming. The first factor we consider is the technology used for reporting quality. The second factor is the degree to which the employees who are most likely to carry out the gaming are incentivized to care about the firm's performance in the disclosure program. Our setting is the disclosure of airline on-time performance, which has been mandated by the U.S. Department of Transportation (DOT) since 1987. Under this program, the DOT collects information on the arrival delays incurred on every flight operated by large domestic airlines and produces monthly rankings of these airlines based on the fraction of their flights that arrive less than 15 minutes late. The design of this program results in a clear discontinuity in the relationship between actual and reported on-time performance at the 15 minute threshold and therefore creates incentives for airlines to game the program by reducing delays on specifically those flights that are close to this threshold. This disclosure program is similar in structure to those used in other policy areas such as health care, education or environmental regulation, but the data and sources of variation that we can exploit are unusually rich compared to those other settings. Thus, our particular context is useful because it allows us to develop a very precise way of empirically identifying gaming as well as shed light on how gaming may be affected by differences in technology and employee incentives, both of which are likely relevant in other policy settings as well.

We develop an empirical approach for identifying gaming, which we define as selective reductions in the delays of flights that are close to the 15 minute threshold. We use the data on individual flight delays that are collected by the DOT under the mandatory disclosure program. Our empirical strategy exploits the fact that the data allow us to separately observe each stage of a flight (e.g., departure from the gate, taxi-out time, time in the air, and taxi-in time). Using the observed delays incurred on previous stages of a flight and combining this with estimates for the

timing of subsequent stages of a flight, we can calculate a flight's *expected arrival delay at the gate* at different points in its progression. This allows us to identify flights that – upon touchdown - are expected to arrive at the gate right around 15 minutes late. We then estimate whether airlines try to speed up specifically these flights so that they arrive at the gate just below the 15 minute threshold. Because we observe tens of thousands of flights each year, we can identify those flights that are candidates for gaming very precisely as well as control for possible unobservables that could lead to reductions in delays on the threshold flights.

To explore whether gaming is affected by the technology used for reporting quality, we take advantage of the fact that, during our sample period, there are two different methods for reporting on-time performance and the use of these methods varies both across airlines and within airlines over time. Specifically, some airlines report their delays by manually recording the time at which a flight departs and arrives, while others rely on an automated technology called ACARS. This automated system directly records each stage of the flight, including arrival at the gate. We expect that the manual reporting technology would facilitate gaming as airlines could simply misreport the arrival times of flights that would otherwise appear to have arrived right around 15 minutes late.⁴ In contrast, airlines using ACARS would actually need to speed up a plane's arrival at the gate – for example, by preferentially allocating scarce resources (such as gates and ground crew) or by increasing the plane's taxiing speed. Because the ACARS system has a number of other – and arguably more important - uses for the airlines (such as helping to automate some of the calculations used in determine employees' pay and facilitating

⁴ In 1998, the Office of Inspector General carried out an audit in response to a complaint to the DOT that two air carriers were submitting falsified arrival data. Specifically, the complainant noted an “abnormally high number of flights were reported by the two air carriers as having arrived on time at 14 minutes after scheduled arrival times.” While the audit did not find evidence of misreporting of arrival times by these two carriers (though the audit only examined less than one percent of the flights these two carrier operated), the report did note that planes with manual reporting were more likely than planes with ACARS to arrive exactly 14 minutes late .

communication between the pilots and the ground), it is unlikely that airlines would have strategically delayed the adoption of the ACARS system in order to facilitate gaming of the disclosure program. Moreover, each of the manually reporting airlines adopts ACARS before the end of our sample and we can compare the firms' behavior before and after this transition.

In addition to the variation in reporting technology, our setting also provides variation in the extent to which airline employees are incentivized based on the airline's on-time performance. During our sample period, several airlines introduce employee bonus programs that are based explicitly on the airline's performance in the government's ranking of on-time performance. Under these programs, employees receive a bonus of between \$65 and \$100 in any month in which the airline as a whole placed at or near the top of the DOT ranking.⁵ Such incentives are potentially important in our setting because airlines cannot predict far in advance which flights will be candidates for gaming. While they probably can anticipate which flights will, on average, have longer delays than others, they likely do not learn which flights will have exactly 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.

The results of our empirical analysis indicate that there is substantial heterogeneity across - and even within - airlines in the extent to which they game the DOT disclosure program. First, we find strong evidence of gaming by airlines that report their delays manually. Specifically, we estimate that their flights that are expected to arrive between 15 and 16 minutes late are about twice as likely as an average flight to arrive exactly one minute earlier than expected. We observe no similar pattern for their flights that are expected to arrive between 12 and 13 minutes

⁵ 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 gaming of the disclosure program which is the focus of our paper.

late, between 18 and 19 minutes late, or more than 25 minutes late – all of which could be considered “control groups” for flights in the 15 to 16 minute range. As an additional test of whether these findings represent a response to the design of the DOT program, we investigate whether a similar pattern emerges when we look at the probability of a flight arriving two minutes earlier than expected and we find that it does. When we estimate the same relationships for the airlines that initially report their data manually after they have switched to automatic reporting, we no longer find any evidence that they game. The fact that these airlines stop gaming once they switch to automatic reporting strongly suggests that they had been gaming by misreporting the arrival times of threshold flights.

Second, we find strong evidence of gaming by two of the airlines which introduced employee bonus programs – specifically, Continental and TWA, who were the first to introduce such programs. At the time of the introduction of their incentive programs, these airlines had some planes that recorded their delay data manually and some planes that reported via ACARS. Our results indicate that the introduction of their bonus programs lead to gaming on both types of planes; however, the estimated effects are larger on their manual planes. For the other three airlines that introduced similar employee bonus programs, we find little or no evidence of gaming. While we do not have detailed enough information on the features of the bonus programs to determine conclusively why they had different effects, we note that the later three programs involved harder to achieve targets and may not have been accompanied by an increased organizational focus on on-time performance which, from anecdotal sources, we know accompanied the first two programs. Finally, it is interesting to note that - despite the incentives inherent in the design of the DOT disclosure program – we find no evidence of gaming by airlines that neither report manually nor introduce employee incentive programs.

Gaming of the type we explore here may adversely impact welfare in two ways. First, to the extent that gaming is achieved by reallocating scarce resources to threshold flights, it may come at the expense of longer delays on non-threshold flights. However, such misallocations are difficult to empirically identify because any of a large number of non-threshold flights could be affected. Second, any form of gaming - including misreporting - may distort the information being conveyed to consumers. This is particularly relevant on our setting since – in the rankings produced by the DOT – airlines are often separated by very small differences in absolute on-time performance and so even small improvements that result from misreporting of threshold flights could impact an airline’s rank.⁶ We carry out simulations that show that the gaming we find can indeed improve an airline’s rank by about one spot in an average month, without improving other measures of the airline’s on-time performance such as mean delays. To the extent that the 15 minute cutoff used in the ranking is imperfectly correlated with other dimensions of on-time performance that consumers care about, then changes in rankings due to gaming may lead consumers to believe that an airline has improved on the dimensions they care about when it has not.

We contribute to the existing literature on disclosure programs by explicitly investigating heterogeneity in firms’ propensities to game a given disclosure program. We show that despite the incentives for gaming that may be inherent in a program’s design, gaming does not always incur. In particular, we show that the technology used for reporting quality matters. In our setting, when firm self-report their quality information, they appear to systematically misreport the data for threshold transactions. While gaming is still possible with automatic reporting of data, it is more costly and we find that it is much less common. In addition, we show that, when

⁶ Additionally, the *relative performance* aspect of the ranking may lead consumers to be less focused on changes in absolute performance.

gaming needs to happen in real-time and threshold transactions cannot be identified in advance, gaming is sensitive to the incentives provided to those employees whose effort is required to carry out the gaming. Our findings contribute to the ongoing policy discussion on the use of disclosure programs in various settings with informational asymmetries by highlighting the need to consider how the design of a disclosure program (in particular, the dimensions of quality that are reported) interacts with both the ease of manipulating these quality dimensions and the incentives of the individuals who are in a position to carry out the manipulation.⁷

The rest of the paper is organized as follows. Section II provides institutional background. Section III describes our data and sample. We outline our empirical approach in Section IV. Our results are presented in Section V. A final section concludes.

II. Institutional Background

II.A. Disclosure of Airline On-Time Performance

Since September 1987, all airlines that account for at least one percent of domestic U.S. passenger revenues have been required to submit information about the on-time performance of their flights to the Department of Transportation under Title 14, Part 234 of the Code of Federal Regulations. 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.⁸ 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 and airborne times, as

⁷ See Ederer, Holden and Meyer (2010) for an example of how to design disclosure programs in a way that reduces the incentives for gaming.

⁸ 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 – likely because their performance is better at uncongested airports.

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.

The DOT uses the data it collects to issue monthly reports that rank airlines based on the percentage of their flights that arrive at the gate with less than 15 minutes of delay. 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. National and local media outlets often report these rankings. A typical news story will mention 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.

As mentioned above, airlines can record delays either manually or automatically through technology installed in the aircraft. Table 1 shows each carrier's reporting method in March 1998 – the earliest month for which this information is available – as well as the dates of any subsequent changes in reporting methods. In 1998, three carriers (Alaska, America West and Southwest) were reporting their delays manually, four carriers (American, Northwest, United and US Airways) were reporting automatically, and the remaining three (Continental, Delta and TWA) were using a combination of manual and automatic reporting. Since we are explicitly interested in exploring whether the reporting technology affects gaming, we have developed an approach (which we describe below) for distinguishing the manual and automatic aircraft of the combination reporters. Note that most of the manual and combination reporters adopt fully

automatic reporting in 2002 or 2003, with the exception of Southwest which does not switch until 2007. Our empirical analysis exploits both the cross-airline as well as this within-airline variation in reporting technology.

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.⁹ 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. However, increased emphasis within the organization on meeting the DOT's on-time target may have amplified the effect of the financial incentives that were provided to the employees by communicating that the airline's management cared about on-time performance.

In June 1996, TWA implemented an employee bonus program which closely resembled Continental's. Three other airlines introduced similarly structured bonus programs in subsequent

⁹ In 1994, Continental had the worst average on-time performance ranking among the ten reporting airlines.

years. These were American Airlines in April 2003, US Airways in May 2005, and United Airlines in January 2009. Despite our best efforts, we were unable to obtain systematic and complete information on the specifics of these programs. However, based on our reading of the trade press and annual reports, we have learned that the three later programs only rewarded employees if the airline achieved the first or second place in the rankings.¹⁰ Thus, the incentive effects of the three later bonus program may have been weaker than for the two earlier programs since the target was harder to achieve. We have also learned that American Airlines introduced its bonus program in conjunction with a large negotiated wage cut during the industry's downturn and that US Airways introduced its program around the same time as it merged with America West. Each of these events may have muted the employees' response to the bonus programs.

III. Data

III. A. Data and Sample

Our empirical analysis uses the flight-level data on on-time performance collected by the U.S. Bureau of Transportation Statistics under the DOT's mandatory reporting requirement. Our regression sample includes domestic flights operated by the ten airlines which were large enough to meet the reporting requirement in 1995: Alaska Airlines, America West, American Airlines, Continental Airlines, Delta Air Lines, Northwest Airlines, TWA, Southwest Airlines, United Airlines, and US Airways.¹¹ We estimate our main empirical specifications for the period 1995-2000. 1995 is the year in which the DOT began collecting data on taxi times, which we require for our empirical analysis. We choose not to include post-2000 data in our main specifications

¹⁰ US Airways only rewarded first place, while the other two rewarded first and second place.

¹¹ TWA was acquired by American Airlines in 2001.

for two reasons. First, after September 2001, both the supply-side and the demand-side incentives for on-time performance likely changed substantially (see, for example, Berry and Jia, 2010). Second, the volume of the available data is so large that we are unable to estimate regressions that include all the airlines for a longer time period. However, we use the later data when we explore the bonus programs that are introduced later in the sample as well as when we explore airlines' transitions from manual to automatic reporting which takes place between 2002 and 2007.

Due to the large volume of data, 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 to 2000 period includes over 3.3 million flights.

Table 2 presents some summary statistics for the main regression sample. The average arrival delay in the sample is 7.5 minutes. 21 percent of all flights arrive 15 or more minutes late and thus are considered late under the DOT's definition. The average taxi-in time is six minutes, the average departure delay is nine minutes, the average taxi-out time is 14 minutes and the average air time is 102 minutes.

III. B. Histograms of Arrival Delays

Figures 1A and 1B show the distribution of arrival delays for the ten carriers in our regression sample for the periods 1995-2000 and 2002-2010, respectively. We truncate the

histograms at -25 on the left and at 60 on the right. For the earlier period, Figure 1A reveals a distribution of delays that peaks at 0. The histogram is fairly smooth but shows discrete spikes at certain values. As Figures 2A-2D will show, these discrete spikes appear to reflect rounding by carriers who report their delay data manually, especially Southwest Airlines. 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. In contrast, the histogram for the later period in Figure 1B appears completely smooth and peaks around -10 minutes. During this later period, most carriers have adopted the automated reporting technology.

In Figures 2A-2D, we show the distribution of arrival delays separately by reporting method. Since we only know an airline's reporting technology with certainty beginning in March 1998, we only show delays for flights between March 1998 and December 2000 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 arrival delays for Continental, Delta and TWA – the three airlines that used a combination of manual and automatic reporting during this time period. Their histogram is also quite smooth; however, it shows distinct spikes at 0 and at 14 minutes, but not elsewhere in the distribution. Note that this group includes the first two carriers to introduce employee bonus programs based on the DOT ranking. Figure 2C shows the distribution of arrival delays for two of the three carriers which reported their on-time data manually during this period – Alaska Airlines and American West. This histogram is also reasonably smooth, but has a large spike at zero and a much smaller, but noticeable spike at 14 minutes. Finally, we show the third

manually reporting carrier, Southwest Airlines, separately in Figure 2D. This histogram reveals a distinct pattern for this carrier which has a large spike at zero (with almost 12% of flights arriving with exactly zero minutes delay) and spikes at all of the five minute intervals except 15, and a spike at 14 minutes. This suggests that Southwest Airlines was rounding flight delays to the nearest multiple of five, except at the 15 minute threshold where the rounding appears to have occurred at 14.

In Figures 3-7, we compare the before-and-after distributions of arrival delays for each of the airlines that introduced employee bonus programs in the 18 months before and after the introduction. Figures 3A and 3B show arrival delays for Continental. 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 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 and 5B plot the arrival delay distribution for American. The figures show a small discontinuity around the 15 minute mark which is much less pronounced than 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 discontinuity at 14 minutes.

IV. Empirical Approach

IV.A. Overview of the Empirical Approach

Our objective is to estimate whether airlines systematically reduce delays on specifically those flights that they expect to arrive at the gate with a delay of just over 15 minutes. This is

what we call gaming. 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 arrive close to the 15 minute threshold. These flights are the most likely candidates for gaming since they are presumably the ones that can 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 otherwise would have been. This requires a counterfactual measure of the flight's delay absent any incentive for gaming.

We believe that both of these requirements are met particularly well in our setting. Because our data allow us to observe the various stages of each flight – departure from the gate, take-off from the departure runway, landing on the arrival runway, and arrival at the gate – we can construct a flight's expected delay at each stage and, at any given stage, we can identify those flights whose expected delay is close to 15 minutes. We can then investigate whether – in *subsequent* stages of the flight – airlines attempt to reduce delays on specifically those flights that are 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. 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 and controlling for observable flight characteristics – an airline's behavior with respect to flights that are expected to arrive 15 minutes late should be similar to its behavior with respect to flights that are expected to arrive, say, 12 or 18 minutes late. 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 12 or 18 minutes of delay, or with very long delays, we will interpret this as evidence of gaming. It is also worth pointing out 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.

IV.B. Regression Analysis

Before describing our regression analysis in detail, it is useful to consider when in the flight’s progression gaming may take place. Delays – and systematic effort to reduce delays – can occur at any stage of a flight. However, we expect that airlines that are attempting to game the 15 minute threshold will be more likely to try to do so during the later stages. This is because, as the flight progresses, the airline knows the delay incurred so far and can therefore 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 to selectively reduce the delay of a flight to just below 15 minutes will actually be successful. 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. Focusing on this last stage of the flight also has the advantage that we can predict taxi times with much less noise than airborne time because airborne time depends on wind patterns which change from day to day and on which we do not have any data.

To construct each flight's expected delay at the time that its wheels touchdown, we take the flight's observed wheels-down time and add to it the median taxi-in time for that flight in the quarter.¹² This gives us a predicted arrival time for the flight. The difference between the predicted arrival time and the scheduled arrival time is the flight's predicted delay.¹³ 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 category. Thus, we define 16 different predicted delay "bins".

To investigate whether gaming is affected by an airline's reporting technology, we group airlines by their reporting technology and construct mutually exclusive delay bins for each of these groups. Similarly, to we investigate whether the introduction of an employee bonus program affects gaming, we construct separate predicted delay bins for each airline that introduces a bonus program and, where possible, distinguish between the years before and after the program was in place.

To measure whether flights that are expected to arrive around 15 minutes late are arriving just under 15 minutes late, we construct a dummy variable that equals one if a flight arrives exactly one minute earlier than predicted. We regress this dummy variable on the various sets of predicted delay bins described above and a set of control variables which include carrier-arrival

¹² We identify a flight as a unique combination of airline, flight number, departure airport and arrival airport.

¹³ For example, consider a flight by Delta Air Lines between Boston and Atlanta in March of 1997. Suppose that it 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.

airport-day fixed effects, 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.¹⁴ The coefficients on the delay bins represent the change in the probability that a flight in a given predicted delay bin will arrive exactly one minute earlier than predicted, relative to the probability of arriving one minute early for flights with predicted delay of less than 10 minutes (the excluded category). Our primary interest is in testing whether flights with predicted delay right around the 15 minute threshold are systematically more likely to arrive exactly one minute earlier than predicted, compared to flights that are just above or below the 15 minute threshold. Because we include carrier-airport-day fixed effects, our coefficients are largely identified by 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. 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 15 minute range and that would also result in the flight being more likely than flights just outside this range to arrive one (or two) minute(s) earlier than predicted.

To check that our findings indeed represent a response to the structure of the DOT program, we also estimate the same regressions using the probability that a flight arrives exactly two minutes early as the dependent variable. In these regressions, we look for a discontinuity in the relationship between predicted delay and the probability of arriving two minutes early around the 16 minute threshold. Finally, it is worth pointing out that since airlines report arrival times and not minutes of delay to the DOT, misreporting by a manual reporter would mean that the airline records an earlier arrival time than the one that actually occurred. For example, the airline

¹⁴ We cluster our standard errors at the arrival airport-day level.

would report that the flight arrived at 2:42 pm instead of 2:43 or 2:44. As a result, both misreporting and actual reductions in taxi-in times would appear in the data as a threshold flight arriving earlier than was predicted.

IV.C. Identifying Reporting Technology at the Aircraft Level

As mentioned above, we only have information on the reporting technology that each airline uses beginning in March 1998. We can safely assume that carriers that are manual reporters in 1998 are also manual reporters in the years prior to 1998 because - if their planes do not have ACARS in 1998 - they also would not have ACARS prior to 1998. For the carriers that use a combination of manual and automatic reporting, we do not know which or how many of their aircraft are equipped with ACARS and which are not. Therefore, we develop an approach for identifying specifically which of their aircraft may be reporting manually. We use the same method to verify that carriers that are automatic reporters in 1998 also reported automatically from 1995-1998. Our approach exploits 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 3 shows the distribution of this variable for all 10 airlines who reported to the DOT in 1995. The 99th percentile of the distribution of this variable for American Airlines (an automatic reporter) is 0.056 which indicates that only about 1 percent of American's planes

arrived with a delay of zero minutes more than 5% of the time. Contrast this to America West (a manual reporter): fifty percent 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. Based on an analogous version of this table for March-December 1998, 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.¹⁵

V. Results

Our regression results are presented in Tables 4 through 10. All tables, except Table 9, show the results from a single regression and the different columns of the table present the coefficients for the different sets of predicted delay bins included in that regression. Across tables, we vary the dependent variable, the sample period, and the sets of airlines and/or types of planes distinguished in the predicted delay bins. We begin by presenting a baseline regression for our main sample period (1995-2000) in which we combine all airlines in a single set of predicted delay bins (Table 4). We then explore whether the gaming is affected by reporting technology (Tables 5 and 6). After that, we explore whether gaming is affected by the introduction of an employee bonus program (Tables 7 through 9). Finally, we investigate whether the behavior of manual reporters changes after they switch to fully automatic reporting (Table 10). Our final set of tables presents simulation results that illustrate how the gaming that we identify can affect the carriers' rankings (Tables 11A-C).

¹⁵ Using this approach, we classify 85 of Continental's 441 planes (19%) and 22 of TWA's 241 planes (9%) as manual.

V.A. Overall Effects

In Table 4, we estimate the regression described above with a single set of predicted delay bins for all ten airlines in the sample. The coefficient on the 15-16 minute bin is 0.074 which indicates that flights that are predicted to arrive between 15 and 16 minutes late are 7.4 percentage points more likely than flights in the excluded category to arrive exactly one minute earlier than predicted. The coefficient is statistically and economically significant. Averaging across all flights, the probability of arriving one minute early is 0.21. Thus, our estimate of 0.074 represents about a 35% increase in the likelihood of being one minute early. No other level of predicted delay has a coefficient anywhere near this large. In this and all subsequent specifications, we test whether the coefficient on the 15-16 minute bin is equal to the coefficient on the 12-13 minute bin, the coefficient on the 18-19 minute bin and the coefficient on the 25+ minute bin. The p-values presented in Table 4 indicate that we can reject each of these hypotheses.

V.B. Does Gaming Vary by Reporting Technology?

In Table 5, we estimate the same regression but now include four mutually exclusive groups of predicted delay bins which capture airlines with differences in reporting technologies. The first column shows the coefficients for the group of carriers which report their delays automatically. The second column presents the coefficients for manual reporters. The final two columns show the coefficients for the combination reporters, distinguishing between their manual and automatic planes. In each column, we highlight the coefficient for the flights that are predicted to arrive between 15-16 minutes late.

The estimates in this table indicate that the reporting technology appears to affect the incidence of gaming. The first column reveals that there is no evidence of gaming by automatic reporters. The second column shows striking evidence of gaming by manual reporters. Specifically, we find that their flights that are predicted to arrive between 15-16 minutes late are 22.6 percentage points more likely to arrive exactly one minute early than flights in the omitted group. Since the overall probability of arriving one minute early is 23 percent for flights in this group, their flights in the 15-16 minute range are about twice as likely to arrive one minute early. Most of the other coefficients for this group of manual reporters are an order of magnitude smaller, and many of them are negative. Our hypothesis tests reject the equality of the 15-16 minute coefficient with the 12-13, 18-19 and 25 and over coefficients. Consistent with the appearance of rounding to five-minute intervals which we observed in the histogram, we find positive and reasonably large coefficients on the 11-12 and 21-22 minute bins, although these coefficients are still only about a third of the size of the coefficient on the 15-16 minute bin. In the final two columns of the table, we find positive and significant coefficients on the 15-16 minute bins for both the automatic and manual planes of the combination reporters although the magnitudes are smaller (4.4 percentage points for manual planes and 4.3 percentage points for automatic planes). In our hypothesis tests, we again reject equality of the 15-16 minute coefficient with the 12-13 and 25 and over coefficients. We also reject equality with the 18-19 minute coefficient for automatic planes.

In Table 6, we estimate the same specification using, as the dependent variable, the dummy variable that equals one if a flight arrives exactly two minutes earlier than predicted. We now highlight the coefficients for the flights that are predicted to be between 16 and 17 minutes late. The results in this table confirm the patterns in Table 5. There is again strong evidence of

gaming by manual reporters. Their flights that are predicted to be between 16 and 17 minutes late are 10 percentage points more likely than flights in the excluded category to arrive two minutes early. Relative to the underlying probability of arriving two minutes early in this set of flights (5%), this effect is quite large. As in Table 5, there is evidence of gaming by combination reporters on both of their plane types. There is again little evidence of gaming by automatic reporters. The coefficient on the 16-17 minute bin for automatic reporters is statistically significant but it is small in magnitude and not statistically distinguishable from the coefficients on the 18-19 minute bin or the 25+ minute bin.

V.C. Do Employee Bonus Programs Affect Gaming?

In Table 7, we estimate our main regression using the subset of carriers who introduced employee bonus programs during the 1995-2000 period. These are Continental and TWA. We create separate groups of delay bins for their manual and the automatic planes and also separate TWA's flights before and after the carrier introduced its bonus program. Recall that we observe no data on taxi times prior to the introduction of Continental's bonus program which is why we only estimate the effect during the post-period for this carrier. Since Continental and TWA are both combination reporters, we also include Delta – the one other combination reporter in our sample – in this table. However, Delta did *not* introduce an employee bonus program. Due to space constraints, we only present a subset of the 16 coefficients we estimate for each group.¹⁶

Looking first at Continental, which was the first carrier to introduce an employee bonus program, we find strong evidence of gaming on both its manual and its automatic planes. The estimates in the first two columns imply that Continental's manual planes that are predicted to be 15 to 16 minutes late are 15.5 percentage points more likely to arrive one minute earlier than

¹⁶ The full set of results is available upon request.

predicted. For its automatic planes, the effect is 8.4 percentage points. This is compared to the overall probability of arriving one minute early for these groups of about 20 percent. The coefficients on all of the other predicted delay bins for Continental are substantially smaller in magnitude.

The results for TWA are perhaps even more interesting because we can observe the carrier's behavior before and after the introduction of its bonus program in June 1996. We find some evidence of gaming by TWA on its manual planes prior to the introduction of its bonus program, but very strong evidence of gaming after the introduction. In TWA's post-period, we estimate that its manual planes that are predicted to be between 15 and 16 minutes late are 22.1 percentage points more likely to arrive one minute early than its flights in the excluded category. This is a large effect given that – like Continental's - TWA's flights, on average, arrive one minute earlier than predicted only about 20 percent of the time. For automatic planes, the estimated effect is 7.8 percentage points, which is considerably smaller than the estimate for the manual planes, but still a very sizeable effect given the underlying probability.

The last two columns of the table show the results for Delta, separating its manual and automatic planes. We estimate statistically significant but very small coefficients on the 15-16 bin for Delta's manual and automatic planes (between 1 and 2 percentage points). The much smaller effects we find for Delta suggest to us that the estimates for Continental and TWA reflect the introduction of their bonus program rather than the fact that they are combination reporters.

In Table 8, we estimate the same regression with the dummy for whether a flight arrived exactly two minutes earlier than predicted as the dependent variable. The results are highly consistent with what we find in Table 7. Both Continental and TWA show strong evidence of gaming after introducing their employee bonus programs and the effects are larger for their

manual planes than for their automatic planes. The magnitudes of the effects range from 10 to 20 percentage points and are quite large when compared to the underlying probability of arriving two minutes early which is around 10 percent.

Taken together, the results in Tables 7 and 8 indicate that – for Continental and TWA – the introduction of their bonus programs encouraged gaming some of which was probably misreporting but some of which was likely actual reductions in delays. For TWA, we can observe that the carrier’s behavior changed substantially after the bonus program was introduced. While we do not observe taxi times prior to the bonus program for Continental, we do observe arrival delays. Recall that in the histograms of arrival delays, we observed a very similar change for Continental after the introduction of the bonus program as for TWA and no spike at 14 in Continental’s histogram prior to the introduction of its program.

We now turn to the three later bonus programs which were all introduced by carriers that used fully automatic reporting. We estimate separate regressions for each of these carriers using, for each of them, a time period that covers 12-18 months before and after the introduction of their bonus program. We present the results from these regressions in Table 9. For American Airlines, which introduced its bonus program in April 2003, our estimates suggest that prior to the introduction of its program, its flights in the 15-16 minute bin were 1.2 percentage points more likely to arrive one minute early. After the introduction, the estimated effect increases to 2.2 percentage points and we can reject the equality of this coefficient and those on the control bins. However, the magnitude of the effect is very small compared to the effects we found for Continental and TWA. The remaining results in the table indicate that neither US Airways (which introduced its bonus program in May 2005) nor United (which introduced its program in January 2009) engaged in gaming before or after they introduced their bonus programs.

V.D. Evidence from Carriers that Switch Reporting Technology

In Table 10, we use data from the later sample period to investigate the behavior of carriers that switch from manual to automatic reporting. In particular, we focus on the three carriers which were reporting manually in the earlier sample and for which we found strong evidence of gaming. Two of these carriers, Alaska and America West, switched from manual to combination reporting in February 2002, and America West switched to fully automatic reporting by January 2003, while Alaska remained a combination reporter. Southwest switched from manual to fully automatic reporting in July 2007. The first two columns of Table 10 show that neither Alaska nor America West engaged in gaming during the 2002-2008 period. The coefficients on the 15-16 minute bin are zero. Similarly, while we continue to see strong evidence of gaming by Southwest while it was still reporting manually (the third column of the table), we no longer find any evidence of gaming after it switches to automatic reporting. The coefficient on the 15-16 bin goes from 0.118 in the third column of the table to 0.006 in the fourth column. We see this as further evidence that the reporting technology has a strong effect on gaming, likely through misreporting.

V.E. Simulation of Rankings

For the carriers which report their delays manually, as well as some of the carriers that introduce employee bonus programs, we have found that their threshold flights are significantly more likely to arrive exactly one minute earlier than predicted. We now investigate whether these small but selective reductions in delays – which will have virtually no impact on the airline’s mean delay – impact the airline’s performance in the DOT ranking, which is the primary

source of information about on-time performance for most consumers. To do this, we perform a counterfactual simulation that estimates what arrival delays and rankings would have been absent gaming. We focus here on the three carriers that show the starkest evidence of gaming: Continental and TWA after their bonus programs are introduced, and Southwest which, among the manual reporters, shows the largest increases in flights arriving one minute early.

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. This gives us a simulated arrival time for each flight in our data. We then re-calculate the fraction of flights that are 15 or more minutes delayed to get a counterfactual measure of on-time performance for each airline. We use these to create counterfactual rankings.

We report the results from these simulations in Tables 11A-C, beginning with Continental in Table 11A. We find that, on average, the selective reductions in delays of threshold flights resulted in a 0.82 percentage point increase in the proportion of flights that were recorded as on-time according the DOT measure. This resulted in an average improvement in the rankings of 0.61 ranks. While we find no change in ranking in some months, in others we see improvements of up to three spots. Other measures of on-time performance such as average delays and the proportion of flights delayed more than 30 or 45 minutes, however, are barely changed by our counterfactual simulations. In Table 11B, we find very similar results for TWA. Its reported on-time performance improved a bit more with 1.08 percentage points, but the changes in reported rank were almost identical. Finally, in Table 11C we see an improvement in average on-time performance for Southwest of 0.76, but the average effect on its rankings is

much smaller (about 0.14 spots). This is because the absolute difference between Southwest's on-time performance and that of the airline ranked immediately below it is usually larger than for other carriers. Overall, the results of these simulations indicate that gaming did affect the information relayed to consumers – both via the ranking and via the reported metric of on-time performance – while other dimensions of on-time performance were essentially unaffected. In this sense, gaming distorted the information conveyed to consumers.

V.F . Additional Results and Robustness Checks

We have carried out a number of robustness checks that we briefly describe here. We have replaced our carrier-arrival airport-day fixed effects with flight-quarter fixed effects and find that our main 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 airport in a given month *during the arrival time window*. The results are robust to these alternative ways of calculating a flight's expected delay.

For the carriers that introduced bonus programs, we have explored whether there may be end-of-the-month effects. Specifically, we have tested 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 DOT 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.

We have also investigated whether there is any evidence that airlines appear to systematically reduce airborne times in response to a flight's predicted delay at the time of departure, using an analogous regression procedure to the one presented above. We find no evidence that airborne times 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.

Finally, 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.

VI. Discussion and Conclusion

We have found evidence for selective reductions in the delays of threshold flights for two types of carriers: those which report their delays manually and some of those that have introduced employee bonus programs. For manual reporters, we suspect that the reductions in delay are largely due to misreporting – that is, the arrival times of flights that have actually arrived with 15 or 16 minutes of delay are misreported so that the flights appear to have arrived with 14 minutes of delay. The fact that the threshold effects we find are very strong – approximately doubling the probability of arriving one minute early and tripling the probability of arriving two minutes early – and disappear once the same carriers switch to automatic reporting suggests that accountability systems which rely on self-reporting face substantial challenges because of the inherent incentives to misreport, especially if it is difficult to detect.

Our finding that some of the employee bonus programs encouraged gaming suggests that – at least in some settings – the incidence of gaming will depend on the incentive schemes in place at a firm. Disclosure programs rate firms; yet quality is often improved or manipulated by employees who may or may not care about the firm’s performance in the disclosure programs. Anticipating whether a firm may game a disclosure program therefore requires a nuanced understanding of not only how quality is produced and reported within the firm but also who, within the firm, has explicit or implicit incentives to respond to the disclosure program.

It is also interesting that we find evidence of gaming after the introduction of some of these programs, but not of all of them. Unfortunately, our only sources of information on these programs are the trade press and annual reports, which do not contain as much detail as we would like. However, we suspect that at least some of the heterogeneity in the response to the employee bonus programs may be due to the differing strength of the incentives. In particular,

the earlier programs had target rankings that were easier to achieve than those of the later programs. In addition, anecdotal sources describe the fact that there were other changes implemented at the firms that introduced the early bonus products which may have enhanced employees' response to the financial incentives of the bonus. For example, Bethune and Huler (1999) point out that when Continental introduced its bonus program it also communicated to employees that on-time performance would be an important goal for the company going forward. To the extent that there were differences across firms in their communication strategies or other complementary factors, these might also explain why we find gaming by some of the firms with bonus programs in place but not by others.

Prior research on disclosure programs has shown that there is also considerable evidence of both misreporting and gaming by firms that are rated under schemes like the one we study. As a result, those designing disclosure programs must try to anticipate the potential for misreporting and for a given scheme to be gamed. Our findings indicate that there is heterogeneity across airlines as well as within airlines over time in the extent to which they game a given disclosure program. Moreover, we show that the likelihood of gaming depends not only the structure of the program but also on the scope for quality manipulation given the characteristics of the product that are measured as well as the incentives in place at the firms that are subject to the program.

References

- Bergstresser, Daniel, Mihir Desai and Joshua Rauh (2006), "Earnings Manipulation, Pension Assumptions, and Managerial Investment Decisions", *Quarterly Journal of Economics* 121(1), 157-195.
- Berry, Steven and Panle Jia (2010), "Tracing the Woes: An Empirical Analysis of the Airline Industry", *American Economic Journal: Microeconomics* 2, 1-43.
- Bethune, Gordon and Scott Huler (1999), *From worst to first: behind the scenes of Continental's remarkable comeback*, New York: John Wiley & Sons, Inc.
- 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* 48(4), 935-63.
- 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.
- Ederer, Florian, Richard Holden and Margaret Meyer (2010), "Gaming and Strategic Ambiguity in Incentive Provision", *Working Paper*, University of California, Los Angeles.
- 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.
- Jacob, Brian and Steven Levitt (2003), "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating", *Quarterly Journal of Economics* 118(3), 843-877.
- 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. (forthcoming), "Multitasking, Information Disclosure and Product Quality: Evidence from Nursing Homes", *Journal of Economics & Management Strategy*.
- Macartnety, Hugh (2011), "The Dynamic Effects of Educational Accountability", *Working Paper*, Duke University.
- 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.
- Snyder, Jason (forthcoming), "Gaming the Liver Transplant Market", *Journal of Law, Economics, & Organization*.
- 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 1A
Distribution of Arrival Delays, 1995-2000
Ten U.S. Carriers that Met Original DOT Reporting Requirement

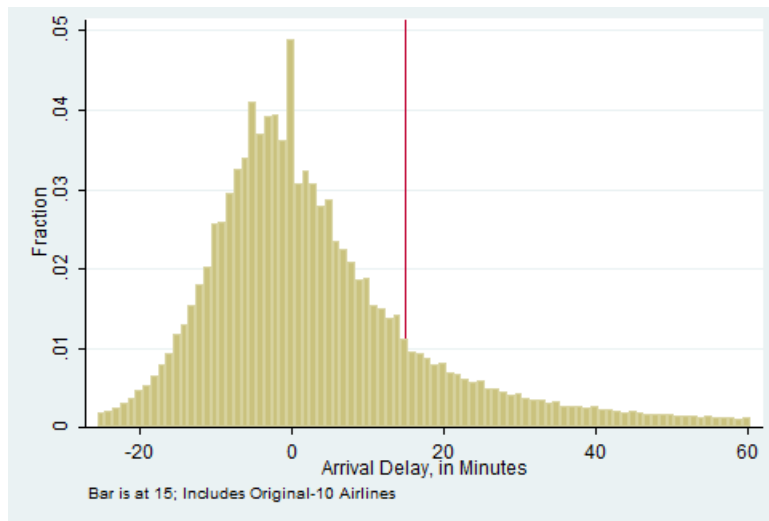
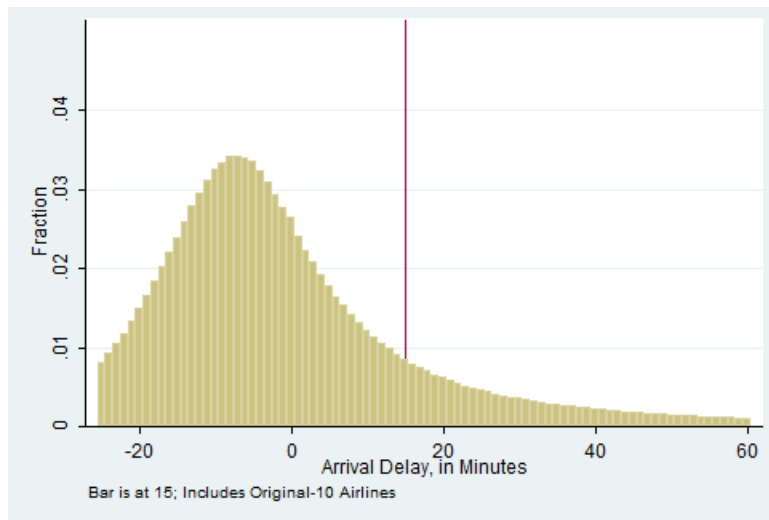


Figure 1B
Distribution of Arrival Delays, 2002-2010
Ten U.S. Carriers that Met Original DOT Reporting Requirement



Figures 2A – 2D
Distribution of Arrival Delays, by Reporting Status
March 1998 - 2000

Figure 2A
Automatic Reporters¹⁷

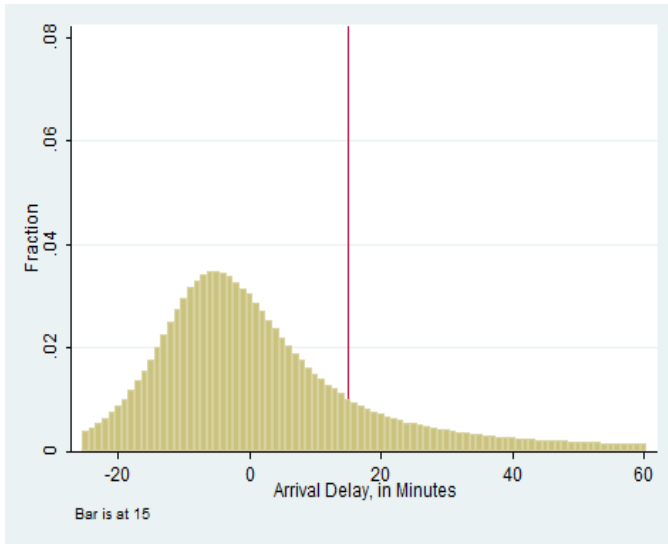


Figure 2B
Combination Reporters¹⁸

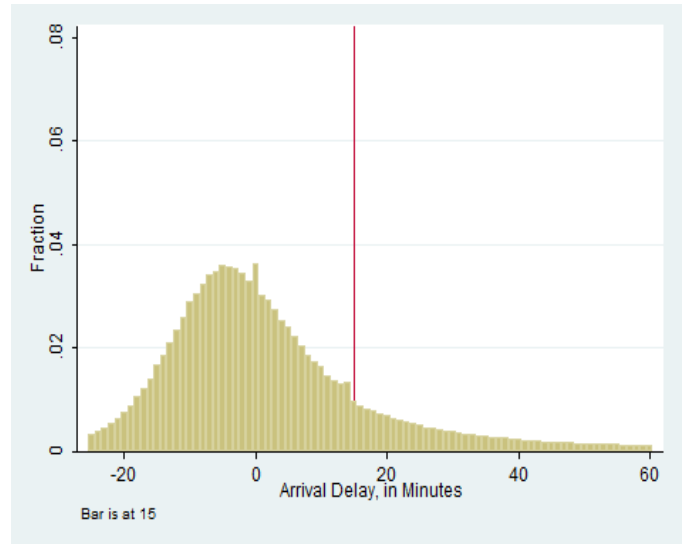


Figure 2C
Manual Reporters, Excluding Southwest¹⁹

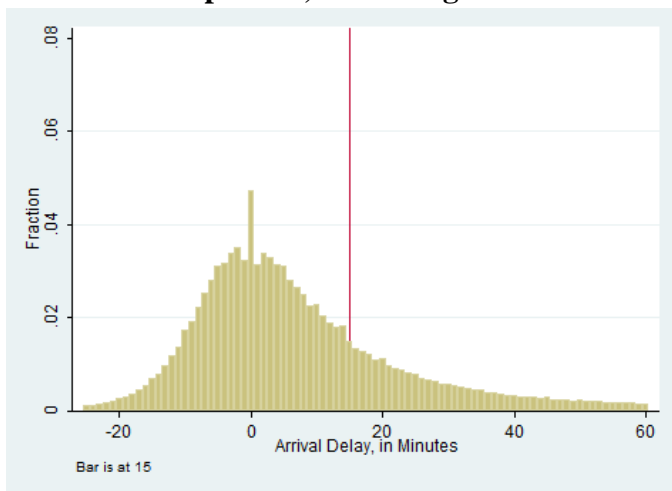
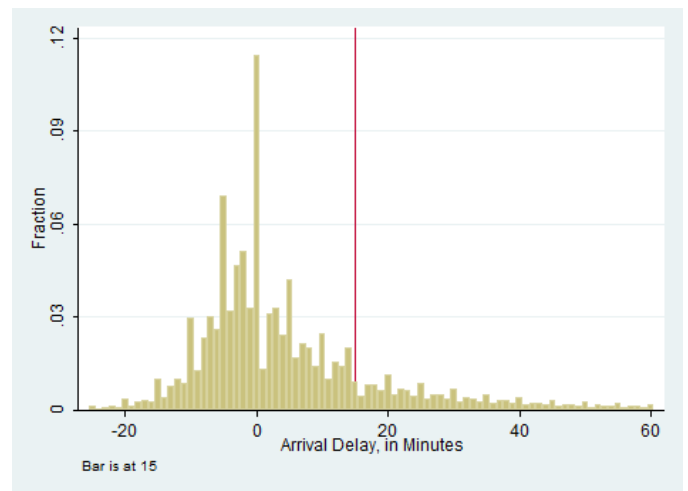


Figure 2D
Southwest Airlines²⁰



¹⁷Automatic reporters include America, Northwest, United and US Airways.

¹⁸Combination reporters include Continental, Delta and TWA.

¹⁹Manual reporters other than Southwest include Alaska and America West.

²⁰Note that the scale on Figure 2D is slightly different than the other three.

Figures 3 – 7
Distribution of Arrival Delays
18 Months Before and After Introduction of Bonus Programs

Figure 3A: Continental Airlines, Before

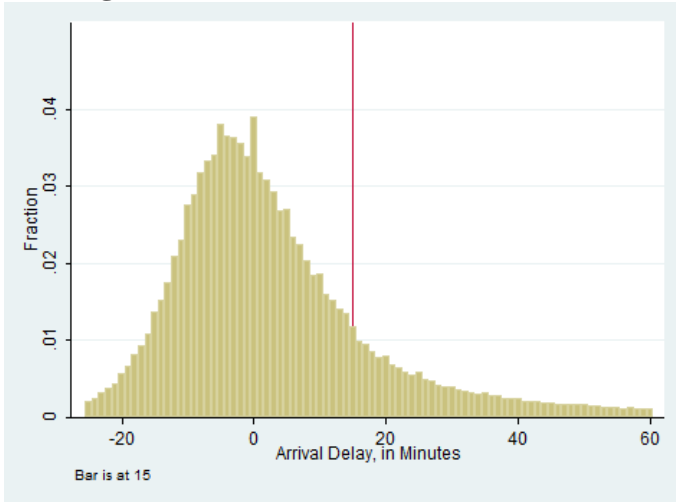


Figure 3B: Continental Airlines, After

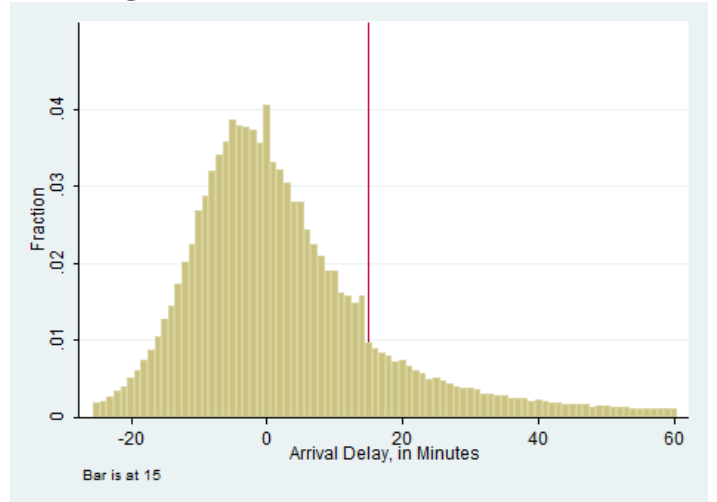


Figure 4A: TWA, Before

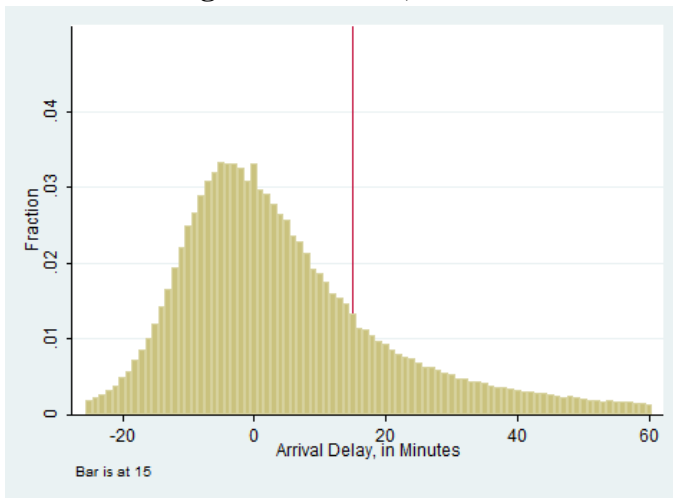


Figure 4B: TWA, After

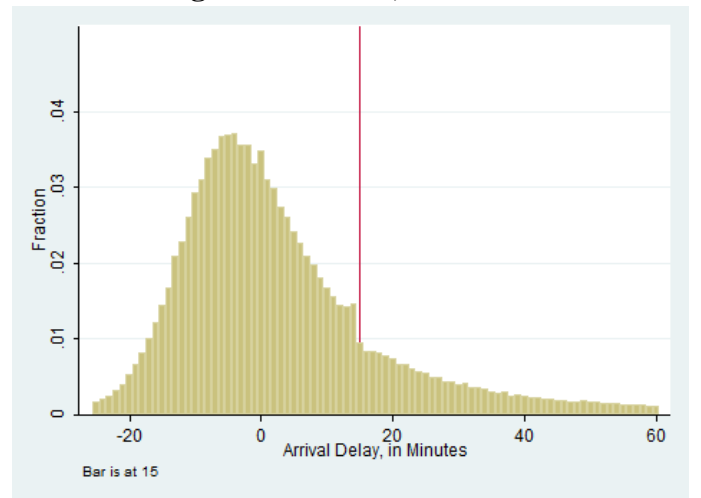


Figure 5A: American Airlines, Before

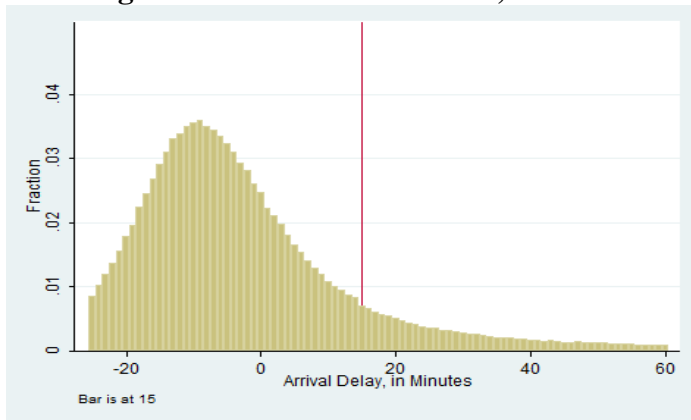


Figure 5B: American Airlines, After

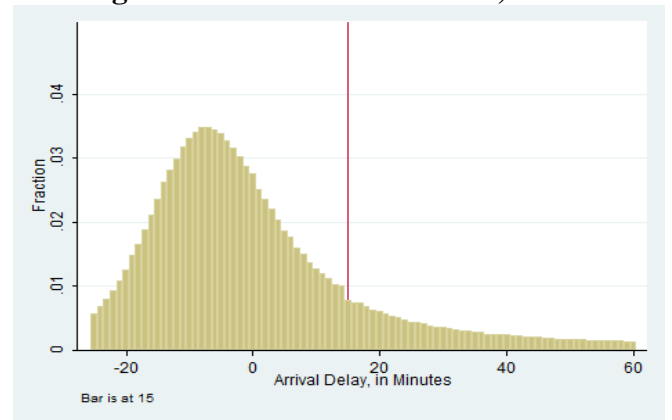


Figure 6A: US Airways, Before

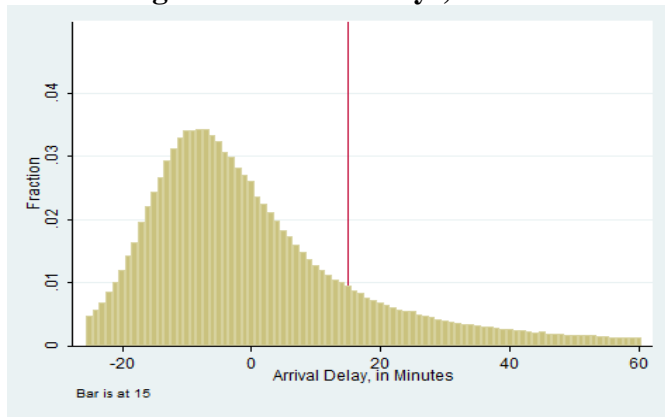


Figure 6B: US Airways, After

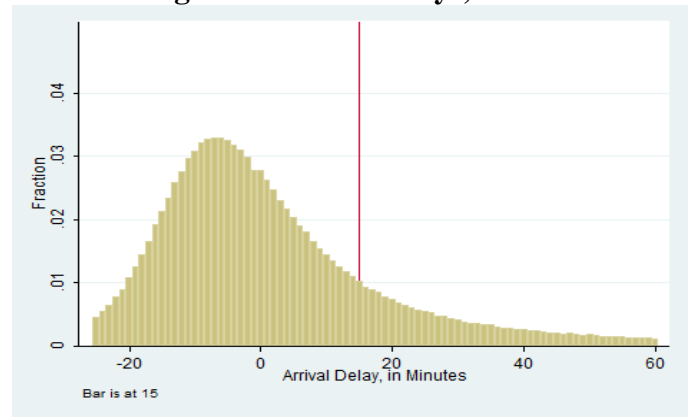


Figure 7A: United Airlines, Before

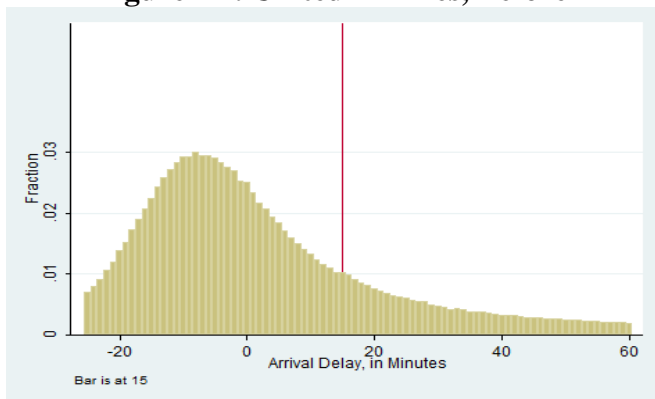


Figure 7B: United Airlines, After

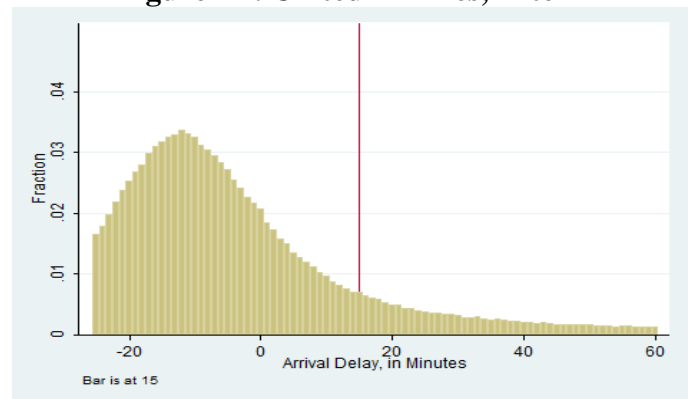


Table 1
Variation in Reporting Technology
By Carrier and Over Time

Carrier	Reporting Technology in March 1998	Date of Switch: Manual →Combination	Date of Switch: Combination →Automatic
Alaska	Manual	Feb. 2002	----
America West	Manual	Feb. 2002	Jan. 2003
American	Automatic	n/a	n/a
Continental	Combo	n/a	Feb. 2002
Delta	Combo	n/a	Nov. 2003
Northwest	Automatic	n/a	n/a
Southwest	Manual	----	July 2007
TWA	Combo	n/a	----
United	Automatic	n/a	n/a
US Airways	Automatic	n/a	n/a

Notes: Based on information provided in the Air Travel Consumer Reports which are issued monthly by the Department of Transportation. Note that Southwest Airlines switches from being a fully manual to being a fully automatic reporter in July of 2007.

Table 2
Summary Statistics for Main Regression Sample

	Mean	Standard Deviation	Min	Max
<u>1995-2000 Sample</u> (3,327,677 observations)				
Arrival Delay (min)	7.50	28.23	-86	1251
Dummy for Arrive 15 Minutes Late or More	0.21	0.41	0	1
Taxi In Time (min)	5.78	3.93	1	60
Departure Delay (min)	8.79	25.91	-15	1246
Taxi Out Time (min)	14.36	7.58	1	60
Air Time (min)	101.74	64.71	20	641

Notes: Includes flights by Alaskan, America West, American, Continental, Delta, Northwest, TWA, United, and US Airways. TWA is acquired by American in 2001. America West merges with US Airways in 2005.

Table 3
Likelihood of a Plane Landing with Exactly Zero Delay
By Carrier, in 1995

	50th Percentile	75th Percentile	90th Percentile	95th Percentile	99th Percentile	Reporting Status in 1998
Alaska	0.064	0.069	0.072	0.074	0.078	Manual
America West	0.061	0.066	0.073	0.075	0.079	Manual
American	0.036	0.042	0.044	0.047	0.056	Auto
Continental	0.041	0.046	0.054	0.062	0.076	Combo
Delta	0.043	0.049	0.055	0.058	0.065	Combo
Northwest	0.038	0.041	0.045	0.048	0.052	Auto
Southwest	0.117	0.123	0.128	0.130	0.134	Manual
TWA	0.034	0.039	0.045	0.056	0.061	Combo
United	0.039	0.042	0.047	0.049	0.053	Auto
US Airways	0.042	0.046	0.050	0.052	0.056	Auto

Notes: We construct an aircraft level variable that equals the fraction of the aircraft's flights in a given year that were recorded as arriving with exactly zero minutes of delay. The table above shows the distribution of this variable for each airline (across its aircraft) in 1995. For example, the fourth entry in the row for American (third row of table) indicates that, in 1995, the 95th percentile of this variable for American was 0.046. This means that 5% of American's planes in 1995 reported landing with zero delay for 4.6% of the plane's flights or more.

Table 4
Probability of Arriving *One Minute Earlier than Predicted*, 1995-2000
All Reporting Carriers

Dependent Variable	<i>=1 if Flight Arrives One Minute Earlier than Predicted</i>
<i>Predicted Delay</i>	
[10,11) min	-0.004* (0.002)
[11,12) min	0.022*** (0.002)
[12,13) min	-0.000 (0.002)
[13,14) min	0.003 (0.002)
[14,15) min	0.005* (0.002)
[15,16) min	0.074*** (0.003)
[16,17) min	0.008** (0.002)
[17,18) min	-0.009*** (0.002)
[18,19) min	0.003 (0.003)
[19,20) min	-0.000 (0.003)
[20,21) min	0.002 (0.003)
[21,22) min	0.029*** (0.003)
[22,23) min	0.000 (0.003)
[23,24) min	0.012*** (0.003)
[24,25) min	0.007* (0.003)
≥25 min	0.006*** (0.001)
<i>P-Values from Hypothesis Tests of Equality of Coefficients (two-tailed)</i>	
Bin15=Bin12	0.00
Bin15=Bin18	0.00
Bin15=Bin25	0.00
Prob(1 min early) in group	0.21

Notes: Standard errors are in parentheses and clustered at the level of the arrival airport-day. Coefficients represent the change in the probability of a flight arriving one minute earlier than predicted relative to flights with predicted delay less than 10 minutes. The regression contains 3,327,677 observations. ** significant at 1%. * significant at 5%. † significant at 10%.

Table 5
Probability of Arriving One Minute Earlier than Predicted, 1995-2000
By Reporting Technology

Dependent Variable	<i>=1 if Flight Arrives One Minute Earlier than Predicted</i>			
	<u>Automatic Reporters</u>	<u>Manual Reporters</u>	<u>Combination Reporters</u>	
			Manual Planes	Automatic Planes
<i>Predicted Delay</i>				
[10,11) min	0.003 (0.003)	-0.033*** (0.004)	0.003 (0.006)	0.010* (0.004)
[11,12) min	-0.002 (0.003)	0.079*** (0.004)	0.015+ (0.006)	0.010* (0.004)
[12,13) min	-0.000 (0.003)	-0.020*** (0.004)	0.003 (0.007)	0.019** (0.004)
[13,14) min	-0.002 (0.003)	-0.006 (0.004)	0.026** (0.007)	0.014* (0.004)
[14,15) min	0.006 (0.003)	-0.008 (0.004)	0.012 (0.007)	0.017** (0.005)
[15,16) min	0.007+ (0.003)	0.226*** (0.005)	0.044** (0.008)	0.043** (0.005)
[16,17) min	0.006 (0.003)	0.021*** (0.005)	0.000 (0.008)	0.004 (0.005)
[17,18) min	0.005 (0.004)	-0.041*** (0.005)	-0.005 (0.008)	-0.007 (0.005)
[18,19) min	0.006 (0.004)	-0.013** (0.005)	0.009 (0.008)	0.013+ (0.005)
[19,20) min	0.006 (0.004)	-0.007 (0.006)	-0.013 (0.009)	-0.001 (0.006)
[20,21) min	0.008 (0.004)	-0.016** (0.005)	-0.007 (0.009)	0.013* (0.006)
[21,22) min	0.006 (0.004)	0.080*** (0.006)	0.013 (0.010)	0.026** (0.006)
[22,23) min	0.001 (0.004)	-0.016** (0.006)	-0.000 (0.010)	0.016* (0.006)
[23,24) min	0.014* (0.004)	0.001 (0.006)	0.024+ (0.011)	0.016* (0.006)
[24,25) min	0.001 (0.004)	0.019** (0.007)	0.010 (0.011)	0.005 (0.007)
≥25 min	0.006** (0.001)	0.002 (0.002)	0.001 (0.002)	0.012** (0.002)
Prob(1 min early)	0.21	0.23	0.20	0.21
<i>P-Values from Hypothesis Tests of Equality of Coefficients (two-tailed)</i>				
Bin15=Bin12	0.12	0.00	0.00	0.00
Bin15=Bin18	0.83	0.00	0.00	0.00
Bin15=Bin25	0.82	0.00	0.00	0.00

Notes: Standard errors are in parentheses and clustered at the level of the arrival airport-day. Columns display coefficients from a single regression that

includes carrier-arrival airport-day fixed effects and arrival hour and hub controls. Coefficients represent the change in the probability of a flight arriving one minute earlier than predicted relative to flights with predicted delay less than 10 minutes. Automatic reporters include American, Northwest, United and US Airways. Combination reporters include Delta, Continental and TWA. Manual reporters include Alaska, America West and Southwest. The regression contains 3,327,677 observations. ** significant at 1%. * significant at 5%. + significant at 10%.

Table 6
Probability of Arriving Two Minutes Earlier than Predicted, 1995-2000
By Reporting Technology

Dependent Variable	<i>=1 if Flight Arrives Two Minutes Earlier than Predicted</i>			
	<u>Automatic Reporters</u>	<u>Manual Reporters</u>	<u>Combination Reporters</u>	
			Manual Planes	Automatic Planes
<i>Predicted Delay</i>				
[10,11) min	0.006** (0.002)	0.004* (0.002)	0.006 (0.004)	0.007* (0.003)
[11,12) min	0.012*** (0.002)	-0.003 (0.002)	0.016*** (0.005)	0.007* (0.003)
[12,13) min	0.006** (0.002)	0.018*** (0.002)	0.026*** (0.005)	0.006 (0.003)
[13,14) min	0.008*** (0.002)	-0.002 (0.002)	0.014** (0.005)	0.016*** (0.003)
[14,15) min	0.007** (0.002)	0.001 (0.002)	0.025*** (0.005)	0.021*** (0.003)
[15,16) min	0.006* (0.002)	0.008** (0.003)	0.023*** (0.006)	0.028*** (0.004)
[16,17) min	0.016*** (0.003)	0.103*** (0.005)	0.074*** (0.007)	0.054*** (0.004)
[17,18) min	0.009** (0.003)	0.019*** (0.003)	0.020** (0.006)	0.010* (0.004)
[18,19) min	0.012*** (0.003)	-0.011*** (0.002)	0.014* (0.006)	0.007 (0.004)
[19,20) min	0.010*** (0.003)	0.015*** (0.003)	0.012 (0.007)	0.014** (0.004)
[20,21) min	0.014*** (0.003)	0.008** (0.003)	0.023*** (0.007)	0.018*** (0.004)
[21,22) min	0.008** (0.003)	0.001 (0.003)	0.018* (0.007)	0.020*** (0.005)
[22,23) min	0.011*** (0.003)	0.020*** (0.004)	0.026*** (0.008)	0.018*** (0.005)
[23,24) min	0.010** (0.003)	-0.010*** (0.003)	0.015 (0.008)	0.009 (0.005)
[24,25) min	0.014*** (0.003)	0.005 (0.004)	0.031*** (0.009)	0.026*** (0.005)
≥25 min	0.013*** (0.001)	0.006*** (0.001)	0.027*** (0.002)	0.022*** (0.001)
Prob(2 min early)	0.09	0.05	0.10	0.10
<i>P-Values from Hypothesis Tests of Equality of Coefficients (two-tailed)</i>				
Bin16=Bin12	0.00	0.00	0.00	0.00
Bin16=Bin18	0.35	0.00	0.00	0.00
Bin16=Bin25	0.37	0.00	0.00	0.00

Notes: Standard errors are in parentheses and clustered at the level of the arrival airport-day. Columns display coefficients from a single regression that

includes carrier-arrival airport-day fixed effects and arrival hour and hub controls. Coefficients represent the change in the probability of a flight arriving one minute earlier than predicted relative to flights with predicted delay less than 10 minutes. Automatic reporters include American, Northwest, United and US Airways. Combination reporters include Delta, Continental and TWA. Manual reporters include Alaska, America West and Southwest. The regression contains 3,320,612 observations. ** significant at 1%. * significant at 5%. + significant at 10%.

Table 7
Probability of Arriving *One Minute Earlier* than Predicted, 1995-2000
Combination Reporters, Variation in Bonus Programs

Dependent Variable	<i>=1 if Flight Arrives One Minute Earlier than Predicted</i>							
	<u>Continental</u>		<u>TWA pre-Bonus</u>		<u>TWA post-Bonus</u>		<u>Delta</u>	
	Manual	Automatic	Manual	Automatic	Manual	Automatic	Manual	Automatic
<i>Predicted Delay</i>								
[12,13) min	0.017 ⁺ (0.008)	0.016** (0.004)	0.010 (0.028)	-0.005 (0.009)	0.044 ⁺ (0.019)	0.019** (0.006)	-0.005 (0.003)	0.010** (0.003)
[13,14) min	0.047** (0.009)	0.036** (0.004)	-0.017 (0.029)	-0.002 (0.009)	0.024 (0.018)	0.039** (0.006)	-0.000 (0.004)	0.017** (0.003)
[14,15) min	0.034** (0.009)	0.049** (0.005)	0.037 (0.031)	0.001 (0.010)	0.049 ⁺ (0.020)	0.039** (0.006)	0.007 (0.004)	0.015** (0.003)
[15,16) min	0.155** (0.010)	0.084** (0.005)	0.080 ⁺ (0.036)	0.017 (0.010)	0.221** (0.024)	0.078** (0.007)	0.013** (0.004)	0.021** (0.003)
[16,17) min	-0.018 ⁺ (0.009)	-0.009 ⁺ (0.005)	0.042 (0.032)	0.001 (0.010)	-0.048* (0.018)	-0.025** (0.006)	0.004 (0.004)	0.013** (0.003)
[17,18) min	-0.006 (0.009)	-0.004 (0.005)	-0.016 (0.030)	-0.011 (0.010)	-0.047 ⁺ (0.020)	-0.013 ⁺ (0.006)	-0.003 (0.004)	0.010* (0.003)
[18,19) min	-0.014 (0.010)	0.005 (0.005)	-0.004 (0.036)	-0.004 (0.011)	0.006 (0.024)	-0.001 (0.007)	-0.003 (0.004)	0.019** (0.003)
.....								
≥25 min	0.011** (0.003)	0.012** (0.002)	0.004 (0.009)	0.007* (0.003)	0.003 (0.006)	0.009** (0.002)	0.002 (0.001)	-0.005 (0.027)
<i>P-Values from Hypothesis Tests of Equality of Coefficients (two tailed)</i>								
Bin15=Bin12	0.00	0.00	0.13	0.10	0.00	0.00	0.00	0.00
Bin15=Bin18	0.00	0.00	0.10	0.14	0.00	0.00	0.01	0.55
Bin15=Bin25	0.00	0.00	0.04	0.31	0.00	0.00	0.01	0.04
Prob(1 min early)	0.22	0.21	0.21	0.20	0.20	0.20	0.20	0.21

Notes: Standard errors are in parentheses and clustered at the level of the arrival airport-day. Columns display select coefficients from a single regression that includes carrier-arrival airport-day fixed effects and arrival hour and hub controls. Coefficients represent the change in the probability of a flight arriving one minute earlier than predicted relative to flights with predicted delay of less than 10 minutes. The regression includes flights by Delta, Continental and TWA on every other day between 1995 and 2000 for a total of 4,485,758 observations. ** significant at 1%. * significant at 5%. ⁺ significant at 10%.

Table 8
Probability of Arriving *Two* Minutes Earlier than Predicted, 1995-2000
Combination Reporters, Variation in Bonus Programs

Dependent Variable	<i>=1 if Flight Arrives Two Minutes Earlier than Predicted</i>							
	<u>Continental</u>		<u>TWA pre-Bonus</u>		<u>TWA post-Bonus</u>		<u>Delta</u>	
	Manual	Automatic	Manual	Automatic	Manual	Automatic	Manual	Automatic
<i>Predicted Delay</i>								
[12,13) min	0.035** (0.006)	0.025** (0.003)	-0.008 (0.020)	-0.001 (0.006)	0.017 (0.013)	0.017** (0.004)	0.022** (0.003)	0.005* (0.002)
[13,14) min	0.024** (0.006)	0.029** (0.003)	0.023 (0.022)	-0.001 (0.007)	0.008 (0.014)	0.027** (0.004)	0.018** (0.003)	0.005* (0.002)
[14,15) min	0.036** (0.007)	0.042** (0.004)	-0.020 (0.021)	0.006 (0.008)	0.037 ⁺ (0.016)	0.037** (0.005)	0.020** (0.003)	0.010** (0.002)
[15,16) min	0.060** (0.007)	0.059** (0.004)	0.014 (0.027)	-0.004 (0.007)	0.080** (0.018)	0.064** (0.006)	0.018** (0.003)	0.013** (0.002)
[16,17) min	0.164** (0.010)	0.103** (0.005)	0.051 (0.029)	0.022* (0.008)	0.229** (0.023)	0.119** (0.006)	0.045** (0.003)	0.018** (0.002)
[17,18) min	0.021* (0.007)	0.024** (0.004)	0.039 (0.027)	0.013 (0.008)	-0.009 (0.016)	0.006 (0.005)	0.024** (0.003)	0.010** (0.002)
[18,19) min	0.005 (0.007)	0.019** (0.004)	-0.000 (0.027)	0.001 (0.008)	-0.010 (0.017)	0.001 (0.005)	0.012** (0.003)	0.009** (0.002)
.....								
≥25 min	0.029** (0.002)	0.022** (0.001)	0.027* (0.009)	0.019** (0.003)	0.025** (0.005)	0.023** (0.002)	0.029** (0.001)	0.021** (0.001)
Prob(2 min early)	0.10	0.09	0.10	0.10	0.11	0.10	0.10	0.10
<i>P-Values from Hypothesis Tests of Equality of Coefficients (two tailed)</i>								
Bin16=Bin12	0.00	0.00	0.10	0.03	0.00	0.00	0.00	0.00
Bin16=Bin18	0.00	0.00	0.20	0.06	0.00	0.00	0.00	0.01
Bin16=Bin25	0.00	0.00	0.44	0.77	0.00	0.00	0.00	0.17

Notes: Standard errors are in parentheses and clustered at the level of the arrival airport-day. Columns display select coefficients from a single regression that includes carrier-arrival airport-day fixed effects and arrival hour and hub controls. Coefficients represent the change in the probability of a flight arriving two minutes earlier than predicted relative to flights with predicted delay of less than 10 minutes. The regression includes flights by Delta, Continental and TWA on every other day between 1995 and 2000 for a total of 4,485,758 observations. ** significant at 1%. * significant at 5%. + significant at 10%.

Table 9
Probability of Arriving Exactly *One* Minute Earlier than Predicted
Introduction of Later Bonus Programs

Dependent Variable	<i>=1 if Flight Arrives One Minute Earlier than Predicted</i>					
	<u>American Airlines</u>		<u>US Airways</u>		<u>United Airlines</u>	
	Pre-Bonus	Post-Bonus	Pre-Bonus	Post-Bonus	Pre-Bonus	Post Bonus
<i>Predicted Delay</i>						
[12,13) min	-0.001 (0.005)	0.005 (0.003)	-0.001 (0.006)	0.004 (0.003)	-0.005 (0.006)	0.011 (0.006)
[13,14) min	0.008 (0.005)	0.009* (0.003)	0.002 (0.006)	0.003 (0.003)	-0.013 ⁺ (0.006)	0.002 (0.007)
[14,15) min	0.014* (0.005)	0.016** (0.003)	0.003 (0.006)	0.000 (0.003)	-0.008 (0.007)	0.005 (0.007)
[15,16) min	0.012 ⁺ (0.005)	0.022** (0.004)	0.001 (0.006)	-0.001 (0.003)	-0.004 (0.007)	0.002 (0.007)
[16,17) min	-0.001 (0.005)	-0.003 (0.004)	0.004 (0.006)	0.004 (0.004)	-0.003 (0.007)	0.008 (0.007)
[17,18) min	-0.002 (0.005)	-0.007 (0.004)	0.007 (0.007)	0.003 (0.004)	0.018 ⁺ (0.007)	0.007 (0.007)
[18,19) min	0.001 (0.006)	0.011* (0.004)	0.010 (0.007)	-0.000 (0.004)	0.005 (0.008)	0.006 (0.008)
≥25 min	-0.000 (0.002)	0.002 (0.001)	0.012** (0.002)	0.007** (0.001)	-0.000 (0.002)	0.000 (0.002)
N (Years included in sample)	2,777,448 (2002-2005)		2,101,260 (2004-2007)		1,087,605 (2008-2010)	
Prob(1 min early)	0.18		0.21		0.19	
<i>P-Values from Hypothesis Tests of Equality of Coefficients (two tailed tests)</i>						
Bin15=Bin12	0.05	0.00	0.79	0.26	0.94	0.36
Bin15=Bin18	0.13	0.03	0.34	0.79	0.40	0.70
Bin15=Bin25	0.02	0.00	0.06	0.02	0.59	0.80

Notes: Standard errors are in parentheses and clustered at the level of the arrival airport-day. Separate regressions are run for each of the carriers in the table. The regressions include arrival airport-day fixed effects and arrival hour and hub controls. Coefficients represent the change in the probability of a flight arriving one minute earlier than predicted, relative to flights with predicted delay less than 10 minutes. ** significant at 1%. * significant at 5%. ⁺ significant at 10%.

Table 10
Probability of Arriving Exactly *One* Minute Earlier than Predicted, 2002-2008
Carriers that Switch Reporting Technology

Dependent Variable	<i>=1 if Flight Arrives One Minute Earlier than Predicted</i>			
	<u>America West</u>	<u>Alaska</u>	<u>Southwest</u> Manual Reporting Period	<u>Southwest</u> Automatic Reporting Period
<i>Predicted Delay</i>				
[12,13) min	0.013 (0.011)	-0.004 (0.009)	-0.060** (0.004)	-0.001 (0.007)
[13,14) min	0.006 (0.011)	-0.016 (0.009)	-0.010* (0.004)	0.001 (0.007)
[14,15) min	0.007 (0.011)	0.003 (0.009)	0.007 (0.004)	0.022* (0.007)
[15,16) min	-0.007 (0.012)	-0.003 (0.010)	0.118** (0.005)	0.006 (0.007)
[16,17) min	0.007 (0.012)	-0.027* (0.009)	0.128** (0.006)	0.024* (0.008)
[17,18) min	0.018 (0.013)	-0.008 (0.010)	-0.064** (0.005)	0.017+ (0.008)
[18,19) min	0.014 (0.013)	0.009 (0.011)	-0.029** (0.005)	0.000 (0.008)
.....				
≥25 min	0.010* (0.004)	0.008* (0.003)	0.003 (0.002)	0.001 (0.002)
<i>P-Values from Hypothesis Tests of Equality of Coefficients (two tailed tests)</i>				
Bin15=Bin12	0.19	0.95	0.00	0.46
Bin15=Bin18	0.25	0.41	0.00	0.60
Bin15=Bin25	0.15	0.25	0.00	0.49

Notes: Standard errors are in parentheses and clustered at the level of the arrival airport-day. Columns display select coefficients from a single regression that includes carrier-arrival airport-day fixed effects and arrival hour and hub controls. Coefficients represent the change in the probability of a flight arriving one minute earlier than predicted, relative to flights with predicted delay less than 10 minutes. The regression contains 2,904,668 observations. ** significant at 1%. * significant at 5%. + significant at 10%.

Table 11A
Changes in DOT On-Time Performance and Rank Due to Taxi-Time Distortions
Continental Airlines, February 1995 – December 2000 (71 months)

Measure of On-Time Performance	Mean	Min	Max	St. Dev.
Change in Fraction < 15 Minutes Late (percentage points)	0.82	-0.69	1.63	0.37
Change in BTS Rank	-0.61	-3.00	0.00	0.73
Change in Monthly Mean Delay (minutes)	-0.05	-0.10	0.00	0.02
Change in Fraction < 30 Minutes Late (percentage points)	0.03	-0.18	0.34	0.08
Change in Fraction < 45 Minutes Late (percentage points)	-0.01	-0.12	0.08	0.04
# Months Rank Improves: 34				
# Months Rank Worsens: 0				
# Months Rank Unchanged: 37				

Table 11B
Changes in DOT On-Time Performance and Rank Due to Taxi-Time Distortions
TWA, June 1996 – December 2000 (55 months)

Measure of On-Time Performance	Mean	Min	Max	St. Dev.
Change in Fraction < 15 Minutes Late (percentage points)	1.08	0.01	2.01	0.50
Change in BTS Rank	-0.60	-3.00	0.00	0.78
Change in Monthly Mean Delay (minutes)	0.25	-0.96	0.74	0.25
Change in Fraction < 30 Minutes Late (percentage points)	0.13	-0.06	0.38	0.08
Change in Fraction < 45 Minutes Late (percentage points)	0.04	-0.29	0.21	0.08
# Months Rank Improves: 24				
# Months Rank Worsens: 0				
# Months Rank Unchanged: 31				

Notes: Illustrates Continental's and TWA's improvements in various measures of on-time performance as a result of taxi-time distortions. Based on the simulations described in the text on page 25. For example, the first row in Table 7A indicates that, averaging over the 71 months following the introduction of Continental's bonus program, taxi-time distortions resulted in Continental improving its on-time performance (as measured by the fraction of flights less than 15 minutes late) by 0.82 percentage points.

Table 11C
Changes in DOT On-Time Performance and Rank Due to Misreporting
Southwest Airlines, 1995-2000 (72 months)

Measure of On-Time Performance	Mean	Min	Max	St. Dev.
Change in Fraction < 15 Minutes Late (percentage points)	0.76	0.42	1.31	0.19
Change in BTS Rank	-0.14	-1.00	0.00	0.35
Change in Monthly Mean Delay (minutes)	-0.13	-0.21	-0.06	0.03
Change in Fraction < 30 Minutes Late (percentage points)	-0.22	-1.21	0.00	0.23
Change in Fraction < 45 Minutes Late (percentage points)	-0.09	-0.49	0.02	0.09

Months Rank Improves: 10

Months Rank Worsens: 0

Months Rank Unchanged: 62

Notes: Illustrates Southwest's improvement in various measures of on-time performance as a result of misreporting. Counterfactual arrival delays based on smoothing of actual reported arrival delay with month-specific 4th-degree polynomial kernel smoother across arrival delays between 25 minutes early and 60 minutes late. For example, the first row indicates that, averaging over the 72 months between 1995 and 2000, misreporting resulted in Southwest improving its on-time performance (as measured by the fraction of flights less than 15 minutes late) by 0.76 percentage points.