README for Itai Ater., Eugene Orlov, “The Effect of the Internet on Performance and Quality: Evidence from the Airline Industry”

This document provides a description of the programs and datasets/data sources necessary to replicate results in

Itai Ater., Eugene Orlov, “The Effect of the Internet on Performance and Quality: Evidence from the Airline Industry”

****

The table below outlines the order of the programs, data sources and the names of the intermediate datasets.[1] In the “Files provided” column, we note which of the datasets are provided. For some source datasets, we attempted to provide “sample” datasets (i.e. a dataset for one quarter or for one month). The notes below the table provide additional information about the datasets we use.

With respect to not provided intermediate datasets, we would be happy to fully cooperate with investigators seeking to conduct a replication who request them.

	Stata Program Name	Input file(s)	Output file(s)	Description	Files Provided
1.	dotdb1b data read in	Origin_and_Destination_Survey_DB1BCoupon_`YR'_`QTR'.csv	DOT DB1B Coupon_`YR'_`QTR'	Reads in raw DB1B data	DOT DB1B Coupon_1997_1 (zipped)
		Origin_and_Destination_Survey_DB1BMarket_`YR'_`QTR'.csv	DOT DB1B Data Market_`YR'_`QTR'		DOT DB1B Data Market_1997_1 (zipped)
		Origin_and_Destination_Survey_DB1BTicket_`YR'_`QTR'.csv	DOT DB1B Data Ticket_`YR'_`QTR'		DOT DB1B Data Ticket_1997_1 (zipped)

[1] All programs were run in Stata, version 12.1, on a Unix server.

2.	dotdb1b_merge	DOT DB1B Coupon_`YR'_`QTR' DOT DB1B Data Market_`YR'_`QTR' DOT DB1B Data Ticket_`YR'_`QTR'	DOT DB1B Data ALL_`YR'_`QTR'	Combines DB1B data into one file	DOT DB1B Data ALL_1997_1 (zipped)
3.	Create_distance	DOT DB1B Coupon_`YR'_`QTR'	DistNS	Creates a file with non-stop distances between airports	DistNS
4.	OAG_Prepare_Variables.do	oag97v1-oag03v1 y2007_08 OAG2007_equip T_T100_SEGMENT_US_CARRIER_1997.csv _1998, _1999, _2000, _2001, _2002, _2003, _2007 DistNS	oag domestic_airlines	1. Read in and process OAG data for 1997-2003, 2007 a. "Expand" sample to individual days and keep only data from Jan, Apr, Jul and Oct (only for 1997-2003) b. Selected Thursday in each month. c. Create competition variables: i. # carriers on route - more than 5 RTs. ii. HHI on route using #flights	T_T100_SEGMENT_US_CARRIER_2000.csv domestic_airlines

				d. Create congestion variables: i. # flights/day at origin and destination ii. # flights/hour at origin and destination e. HHI at origin/destination using #flights f. Hubs at origin/destination - (Share >= 0.5) g. #flights on route h. Presence of LCCs i. Netwok variables i. # flights arriving to origin within windows of time ii. # flights departing from the destination within windows of time
5.	OAG_Prepare_Variables.do	oag97v1-oag03v1 OAG_2007_access04152011 domestic_airlines DistNS	oag_04192011	Same as OAG_Prepare_Variables.do but using alternative data source for OAG 2007 data

6.	delays97_03.do	On_Time_On_Time_Performance_1997_1-On_Time_On_Time_Performance_2003_12	delays1997_2003	Read in and append on-time performance data for 1997/2003	On_Time_On_Time_Performance_1997_1-On_Time_On_Time_Performance_1997_12 (zipped)
7.	delays04_09.do	On_Time_On_Time_Performance_2004_1-On_Time_On_Time_Performance_2009_12	delays2004_2009	Read in and append on-time performance data for 2004/2009	
8.	Delays_Step2.do	delays1997_2003, delays2004_2009 Dates_of_Interest	delays1997_2007_help	Create three variables: a. Flight in Day - a given aircraft's consecutive flight number within a day b. Late Arrival - arrivaly delay (in minutes) of the aircraft on previous flight c. Scheduled buffer - how many minutes airline schedules between given flight and previous flight on the same aircraft	Dates_of_Interest delays1997_2007_help

9.	Create_T100_Dataset	T_T100_SEGMENT_US_CARRIER_1997.csv _1998, _2000, _2001, _2003, _2007	Load_Factors.dta	1. Read in original T100 monthly **segment** data a. Domestic region b. Keep airline configuration = 1 (i.e. passenger configuration), and drop 2 (freight) and 3 (freight/passenger combo); 1 makes up 97% of observations, 3 flies mainly in AK c. Drop if seats = 0 (this also means departures performed = 0) d. Keep Class = F (scheduled passenger service) 2. Create load factors	Load_Factors.dta
10.	Create_origin_state	T_T100_SEGMENT_US_CARRIER_1997.csv _1998, _2000, _2001, _2003, _2007	airport_state	Creates a file with airports and its state; eliminates overseas territories (i.e.keeps only 50 states and DC)	airport_state
11.	OD_data_clean	DOT DB1B Data ALL_`YR'_`QTR'	db1b_processed_`YR'_q`QTR'	Clean OD data	cpi

where `YR' = 97, 98, 00, 01, 03, 07

cpi

sifl

airport_state

foreign_airlines

Steps:

1. Keep itineraries with at most one stop
2. Drop itineraries involving bulk and non-credible fares
3. Drop itineraries involving carriers with missing or unknow names
4. Delete itineraries with foreign carriers
5. Create a variable for itineraries involving trips to/from AK and HI (delete itineraries to/from overseas territories)
6. Create variable class = Coach / First
7. Determine jaw itineraries
8. Deflate fares using CPI
9. Identify fares with unlikely fares:
   a. Fares "outside" SIFL

sifl

foreign_airlines

				b. Coach fare < $25, First class fare < $75 10. Identify fares with distance restrictions: a. Non-stop miles < 100 b. Non-stop miles > 1000, and market miles flown > 2 * non-stop miles 11. Identify a variable **itineraries**	
12.	OD_fare	db1b_processed_`YR'_q`QTR' airport_state DistNS	db1b_complete_`YR'_q`QTR' Fare_1_0_0_0_1 weights_0	Calculate average fares on each segment as well as passenger weights for each segment-origin airport	
13.	Create_Weighted_Variables	airpAllan CPS demog weights_0	Weighted_Variables	Creates weighted Internet on a segment by origin airport, % of passengers traveling on duopoly, monopoly and competitive route, and weighted demographic variables.	airpAllan CPS demog
14.	Instruments	db1b_complete_`YR'_q`QTR airpAllan	instruments1 instruments2	Creates some of the instruments	hub instruments1

		demog hub			
15.	Merge_Data_04192011.do	delays1997_2007_help oag_04192011 load_factors.dta Weighted_Variables Fare_1_0_0_0_1 Dates_of_Interest	Sample_04192011	Merges OTP and OAG data with weighted variables, load factors and fares	Sample_04192011
16.	Merge_Data	delays1997_2007_help oag load_factors.dta Weighted_Variables Fare_1_0_0_0_1 Dates_of_Interest	Sample	Merges OTP and OAG data with weighted variables, load factors and fares	Sample
17.	Merge_Data_04192011_OAG.do	oag_04192011 load_factors.dta Weighted_Variables	Sample_04192011_OAG	Same as Merge_Data_04192011.do but this keeps all carriers from OAG	Sample_04192011_OAG

		Fare_1_0_0_0_1 Dates_of_Interest			
18.	Regressions_Final_03272013_nrf_areg_try.do	Sample_04192011	arrdelay_segment_large100_fnl_segment_nrf_try.xls	Prepare Tables 1, 2, 3, 4, 8 and 9	arrdelay_segment_large100_fnl_segment_nrf_try.xls
		DistNS	total_actual_time_segment_large100_fnl_segment_nrf_try.xls schelapsedtime_segment_large100_fnl_segment_nrf_try.xls		total_actual_time_segment_large100_fnl_segment_nrf_try.xls schelapsedtime_segment_large100_fnl_segment_nrf_try.xls
19.	Table5.do	Sample DistNS instruments1	Table5.xls	Prepare Table 5	Table5.xls
20.	Table6.do	Sample_04192011	schelapsedtime_segment_large100_final_BUS_0419_nrf.xls	Prepare Table 6	schelapsedtime_segment_large100_final_BUS_0419_nrf.xls
		DistNS			

		instruments1			
21.	Regressions_help_nw_inv1.do	Sample_04192011_OAG	min_flt_time__segment_large100_final0925b_wnw_sob_inv1.xls	Prepare Table 7	min_flt_time__segment_large100_final0925b_wnw_sob_inv1.xls
22.	make_90_07_scheduled_data.do	On_Time_On_Time_Performance_1990_4-On_Time_On_Time_Performance_2007_4 DistNS	90_07_scheduled_data	Prepared data for Figure 1	90_07_scheduled_data
23.	1990-2007_graph_for_paper_restat.do	90_07_scheduled_data	Figure 1.gph	Prepare Figure 1	Figure 1.gph
24.	fast_time_analysis1506.do	oag	Figure 2.gph Figure 3.gph	Prepare Figures 2 and 3	Figure 2.gph Figure 3.gph

Note: When replicating the results, the standard errors in IV regressions should be manually adjusted to account for the loss of degrees of freedom that happens when data are manually demeaned in the programs.

****

Additional notes on data sources:

I. On-Time Performance Data

The on-time performance data were downloaded from the Bureau of Transportation ("BTS") website at http://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time. This database reports scheduled and actual flight information for all flights on each airline with at least one percent of US domestic passenger revenues. In particular, the data include measures of scheduled and actual departure and arrival times as well as identifying information for each aircraft and flight.

Data for the period 1997-2009 were downloaded and data for 1997-2003 and 2007 were included in our regression analysis. We eliminate flights, for which data do not report aircraft's tail number, as well as cancelled and diverted flights.

II. BTS T-100 Data – Domestic Segment, Airline Level

T-100 data were downloaded from the BTS website at http://www.transtats.bts.gov/TableInfo.asp?Table_ID=311&DB_Short_Name=Air%20Carriers&Info_Only=0. These data are monthly and represent domestic non-stop segment data reported by both U.S. and foreign air carriers, including carrier, origin, destination, aircraft type and service class for transported passengers, freight and mail, available capacity, scheduled departures, departures performed, aircraft hours, and load factor when both origin and destination airports are located within the boundaries of the United States and its territories.

Data for the period 1997-2003 and 2007, except for Sep. 2001 – Dec. 2001, were downloaded and included in our regression analysis. In each month, we kept only domestic flights (using variable "region"); flights corresponding to the following service classes (variable "class"): Scheduled Passenger/ Cargo Service, and Non Scheduled Civilian Passenger/ Cargo Service; flights with positive number of seats (variable "seats"); and flights that correspond to aircrafts with passenger configuration (variable "aircraft_config"). We then summed up the number of seats and the number of passengers by the directional segment (origin and destination airport), airline, year and quarter. Load factor for each observation is defined as the ratio of the number of passengers to the number of seats.

III. Internet Penetration Data

Internet adoption data were created using a series of the Internet and Computer Use Supplements for the Current Population Survey.[2] The data for these supplements were collected in a single month in the following years: 1994, 1997, 1998, 2000, 2001, 2003 and 2007, of which we use all years but the 1994.[3] The data were downloaded from the CPS website using their DataFerrett tool (see

[2] The Current Population Survey (CPS) is a joint effort between the Bureau of Labor Statistics and the Census Bureau. It is a "monthly survey of about 50,000 households". Its "sample is scientifically selected to represent the civilian noninstitutional population".

[3] In 1997, the supplement was called "Internet Usage Data".

http://dataferrett.census.gov/index.html).

To calculate Internet penetration in 1997, we used respondents' answers to the following questions:
– Computer-home use connect to internet/on-line service,y/n (variable PESCU6I)
– Computer-internet use at home,y/n (variable PESCU12A)
– Computer-internet use at school,y/n (variable PESCU12C)
– Computer-internet use at work,y/n (variable PESCU12B).

If the answer to either of these is 'yes', then we assumed that a respondent has access to the Internet. We then used one of the CPS supplied weights (variable PWSSWGT) to calculate Internet penetration by MSA (variable GTMSA).[4]

To calculate Internet penetration in 1998 and 2000, we used respondents' answers to the following questions:
– Internet home use, computer or WebTV owners, y/n (variable HESIU3)
– Internet use (home) person recode - y/n (variable PRS11)
– Internet use (outside home) - y/n (variable PES14)

If the answer to either of these is 'yes', then we assumed that a respondent has access to the Internet. We then used one of the CPS supplied weights (variable PWCMPWGT, composite final weight) to calculate Internet penetration by MSA (variable GTMSA).

To calculate Internet penetration in 2001, 2003 and 2007, we used respondents' answers to the following questions:
– Internet home use - y/n (variable HESINT1)
– Internet use (anywhere) person recode (variable PRNET1)
– Internet use (home) person recode (variable PRNET2)
– Internet use (outside home) person recode (variable PRNET3)
– Interne use (variable PESNET)

If the answer to either of these is 'yes', then we assumed that a respondent has access to the Internet. We then used one of the CPS supplied weights (variable PWCMPWGT, composite final weight) to calculate Internet penetration by MSA (variable GTMSA). We then interpolated Internet penetration, by MSA, for years 1999 and 2002, using linear regression of Internet penetration on trend and MSA dummy variables.[5]

[4] Note that there are also other supplied weights, e.g. weight-family (variable PWFMWGT) or weight-household (HWHHWGT). All three weights are highly correlated and similar in magnitude.

To calculate adjusted Internet penetration, we followed the following procedure:
1. Using O&D Data (with the same restrictions on observations outlined in part IV below), calculate the number of passengers for each year, quarter, first segment, second segment, carrier on segment 1, carrier on segment 2 of the flight combination. (Note that we only have observations with no more than two segments/coupons, i.e. the sample includes all one-way, round-trip and open-jaw with two segments flights).
2. Sum up passengers by year, quarter, segment 1, and carrier on segment 1. For these data create segment 2 equal to segment 1, and carrier on segment 2 equal to segment 1.
3. Get the data from (1) but keep only two-segment flights; and append the data from (2).
4. Merge data on Internet penetration by each observation's *origin* airport (via its MSA) and year.
5. Calculate weighted average of the Internet penetration, by year, quarter, segment two and the carrier on segment 2 (!), with weights being the number of passengers.

IV. O&D Data

O&D Data were downloaded from the BTS website at http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=247&DB_Short_Name=Origin%20and%20Destination%20Survey . This is a market-level dataset (a market is defined as origin-destination pair, not as a segment). The data are quarterly and represent a 10% random sample of all domestic flights reported by both U.S. and foreign air carriers, including the reporting carrier, origin and destination airport, prorated market fare, number of market coupons, market miles flown, number of passengers, and carrier change indicators.

Data for the period 1997-2003 and 2007, except for the fourth quarter of 2001, were downloaded. In each quarter, we kept only itineraries with no more than two coupons, thus essentially leaving only one-way, round-trip and openjaw (with two coupons) tickets (using variable MktCoupons). We dropped itineraries where at least one carrier on either leg of the trip is unknown, itineraries with top-coded fares (equal 5000 or 9999), and itineraries with fares, in 2000 dollars, below $25 or higher than five times SIFL (standard industry fare level that is determined by the Department of Transportation) for long-haul trips and six times SIFL for short-haul trips.[6] (Long-haul trip is defined as total travel distance is greater than 500 miles, short-haul trip – otherwise). We then dropped very short

[5] Note that an alternative interpolation method is to run a hedonic regression at an individual level of the Internet penetration on demographic characteristics, and then interpolate using CPS respondents in years 1999 and 2002, and then calculate Internet adoption. We don't think this method is more beneficial to the one we used.
[6] See http://ostpxweb.dot.gov/aviation/domfares/siflb.pdf.

trips, with travel distance less than 100 miles, and itineraries with unusual patterns: where market distance is more than twice larger than the non-stop mileage between the origin and destination airports, and the non-stop mileage is more than 1000 miles.

For round-trip itineraries, we divided the total itinerary fare between legs proportionally to the leg distances (use variable AirportGroup to determine the stop-over airports). We then created a duplicate for each observation number of times equal the number of coupons (i.e. we duplicated each round-trip observation) and created "segment" level observations, with corresponding origin and destination airports, a fare and the number of passengers. For instance, an A->B->C with fair X and number of passengers N was transformed into two observations, A->B and B->C with the same number of passengers N and fares proportional to the distances on each leg. Each of the segments has their original carrier.

Using this sample of segments, we then calculated weighted across observations (by the number of segment passengers) average fare and total number of passengers for each year, quarter, segment and carrier combination.
This number of passengers was then used to weigh Internet penetration. Note that the number of passengers does not reflect "real origin" for each itinerary.

V. Official Airline Guide ("OAG") Data

The OAG data are proprietary and available for purchase at www.oag.com. (A contact person: Mark Mattio, T: +1 630 515 3388 M: +1 630 965 4390, Mark.Mattio@oag.com.)

OAG publishes a complete set of scheduled flights for all airlines between all U.S. airports. We use the OAG data to construct the measures of competition for each flight segment in addition to several measures of aircraft operations in each airport and at different times throughout the day. In addition, we use these data to identify flights operated by low-cost carriers as well as to investigate the effect of Internet access on the shortest scheduled flight times.

VI. Demographic and Economic Measures

Demographic and economic measures at the metropolitan-area level were downloaded at http://www.bea.gov/regional/index.htm.

****

Instructions for the replication of robustness checks

README for Itai Ater., Eugene Orlov, “The Effect of the Internet on Performance and Quality: Evidence from the Airline Industry”

1. Footnote 10: “These and other results hold if instead of Thursdays in each quarter in each year in our sample we consider Mondays or Tuesdays.”

   Instructions: See enclosed programs in folders /Programs/Reg4/MN and /Programs/Reg4/TU, and the corresponding output in folders /Data/Reg4/MN and /Data/Reg4/TU. These programs are based on programs Regressions_Final_03272013_nrf_areg_try.do and Table6.do, but instead of dataset Sample_04192011 use (enclosed) datasets Sample_04192011Mn and Sample_04192011Tu, respectively. (These datasets are created in exactly the same way as Sample_04192011 but instead of relying on dataset Dates_of_Interest use datasets Dates_of_Interest_Mn and Dates_of_Interest_Tu, respectively.)

2. Footnote 10: “Furthermore, these results also hold if we add airport/carrier fixed effects to the regression analysis.”

   Instructions: Take program Table6.do, and on the lines 337, 356 and 370, change “cym* arcft*” to “cym* arcft* oc*”.

3. Footnote 11: “The regression results remain qualitatively the same if we use the On-Time Performance data instead of the OAG data.”

   Instructions: See the enclosed program in folder /Programs/Reg12, and the corresponding output in folder /Data/Reg12.

4. Section IV.B: “We also performed similar regression analyses for other measures of flight times on a given segment: 25th percentile, median, 75th percentile and maximum flight times.”

   Instructions: In program “Regressions_help_nw_inv1.do”, on line 118, change min_flt_time to median_flt_time for median, to p25_flt_time for the 25the percentile, to mean_flt_time for the mean, to p75_flt_time for the 75^{th} percentile, and to max_flt_time for maximum flight times.

5. Footnote 12: “Regressing the Internet variable on segment and aircraft fixed effects, we get that the standard deviation of the residuals of the Internet penetration is 0.16. When we add date fixed effects to this regression, the standard deviation of the residuals of Internet penetration reduces to 0.02.”

   Instructions: Run enclosed program “Internet_Residual_Deviation.do”.

6. Footnote 13: “We also explored whether the increased actual time is driven by longer departure delays, measured as the elapsed time between the actual gate pushback time and the scheduled departure time… The regression results, presented in the online appendix, suggest that departure delays are significantly higher for flights on segments in which larger shares of passengers have Internet access. This difference is greater in competitive markets.”

   Instructions: In program “Regressions_Final_03272013_nrf_areg_try.do”, on line 297, change “arrdelay total_actual_time” to depdelay. See enclosed program “Regressions_Final_03272013_nrf_areg.do” in folder /Programs/Reg910, and the corresponding output in folder /Data/Reg910.

7. Section IV.D: “We experimented with several non-linear functions of prices and verified that the results are unaffected,” and Footnote 7: “… we verified that including non-linear price terms does not affect our results.”

   Instructions: See enclosed programs “Regressions_Final_03272013_nrf_areg.do” in folder /Programs/Reg11, and “Regressions_Final_03272013_nrf_areg.do” in folder /Programs/Reg11/Ptile, and the corresponding output in folders /Data/Reg11, and /Data/Reg11/Ptile.

8. Section IV.D: “We also found that in most specifications, the Internet variable has a positive and statistically significant coefficient even if we exclude the interaction term between the Internet and competition variable… The results we obtained were qualitatively the same.”

   Instructions: Rerun the regressions excluding the interaction term between the Internet and competition variable.

9. Section IV.D: “We also reran our regressions to include all airlines that reported on-time performance statistics to the U.S. Department of Transportation (instead of including only airlines that reported on-time performance throughout the entire period studied)… The results we obtained were qualitatively the same.”

   Instructions: See enclosed programs in folder /Programs/Reg1, and the corresponding output in folder /Data/Reg1.

10. Section IV.D: “We also ran the regressions separately for earlier and later time periods. These regressions suggest that most of the effect of the Internet on scheduled elapsed flight times took place between the years 1998 and 2007.”

    Instructions: See enclosed programs in folder /Programs/Reg3, and the corresponding output in folder /Data/Reg3.

11. Section IV.D: “We also verified that our results remained unchanged when we used flights from other days of the week or for flights performed between the largest 75 airports instead of the top 100 airports.”

    Instructions: See the enclosed program in folder /Programs/Reg2, and the corresponding output in folder /Data/Reg2.

12. Section IV.D: “…we fixed the weights based on the total number of passengers in each route between 1997 and 2007. We then reran the regressions with the fixed weights. The results remained qualitatively and quantitatively the same.”

    Instructions: See the enclosed program in folder /Programs/Reg5, and the corresponding output in folder /Data/Reg5. This program uses (enclosed) dataset Sample_04192011_Fixed, which is constructed similarly to dataset Sample_04192011 but instead of relying on dataset “Weighted_Variables” it uses dataset “Weighted_Variables_Fixed” (which is generated in the enclosed program Create_Weighted_Variables_Fixed.do).

13. Section IV.D: “An additional concern is that the HHI variable could be endogenous. Therefore, we also ran instrumental variable regressions where we used as an instrument the lagged HHI values on the same segment. Again, the regression results remained qualitatively the same.”

    Instructions: See enclosed programs in folder /Programs/Reg6, and the corresponding output in folder /Data/Reg6.

14. Section IV.D: “We also verified that our results are not sensitive to the exclusion of the late arrival and the scheduled buffer variables,” and Section III.C.: “… omitting any or all of these additional variables [aircraft’s arrival delay, the scheduled buffer, number of a given aircraft’s prior flights in a given day] does not qualitatively affect our results.”

    Instructions: See enclosed programs in folder /Programs/Reg7, and the corresponding output in folder /Data/Reg7.

15. Section IV.D: “…we verified that the significance of coefficients is not affected when we cluster the standard errors at alternative levels, such as at the origin/date level.”

    Instructions: See enclosed programs in folder /Programs/Reg8, and the corresponding output in folder /Data/Reg8.

16. Section V.C: “…in a separate analysis reported in the online appendix, we measure delays using the definition used by the Department of Transportation to compute airlines’ on-time performance statistics. In other words, we use as a dependent variable a

dummy variable which equals one if the flight is late more than 15 minutes and zero otherwise. In this analysis we do not find that the Internet is associated with fewer delays.”

Instructions: In program “Regressions_Final_03272013_nrf_areg_try.do”, on line 297, change “arrdelay total_actual_time” to arrdelayDum. See enclosed program “Regressions_Final_03272013_nrf_areg.do” in folder /Programs/Reg910, and the corresponding output in folder /Data/Reg910.