The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains

Greer K. Gosnell, John A. List, and Robert D. Metcalfe*

October 22, 2019

Abstract

Increasing evidence indicates the importance of management in determining firms' productivity. Yet, causal evidence regarding the effectiveness of management practices is scarce, especially for high-skilled workers in the developed world. In an eight-month field experiment measuring the productivity of captains in the commercial aviation sector, we test four distinct management practices: (i) performance monitoring; (ii) performance feedback; (iii) target setting; and (iv) prosocial incentives. We find that these management practices—particularly performance monitoring and target setting—significantly increase captains' productivity with respect to the targeted fuel-saving dimensions. We identify positive spillovers of the tested management practices on job satisfaction and carbon dioxide emissions, and captains overwhelmingly express desire for deeper managerial engagement. Both the implementation and the results of the study reveal an uncharted opportunity for management researchers to delve into the black box of firms and rigorously examine the determinants of productivity amongst skilled labor.

JEL Classification: D01, J3, Q5, R4.

---

*Gosnell: London School of Economics and Political Science, List: University of Chicago, Metcalfe: Boston University.
Acknowledgments: We thank participants at the 2015 EEE and PPE NBER Summer Institute sessions for excellent remarks that considerably improved the research, and seminar participants at various universities. Omar Al-Ubaydli, Scott Barrett, Steve Cicala, Diane Coyle, Paul Dolan, Robert Dur, Samuel Fankhauser, Roger Fouquet, Robert Hahn, Glenn Harrison, Justine Hastings, David Jimenez-Gomez, Matthew Kahn, Kory Kroft, Edward Lazear, Steve Levitt, Bentley MacLeod, Jonathan Meer, Kyle Meng, Michael Norton, Jim Rebitzer, Jonah Rockoff, Sally Sadoff, Laura Schechter, Jessie Shapiro, Kathryn Shaw, Kerry Smith, Alex Teytelboym, Christian Vossler, Gernot Wagner, and Catherine Wolfram provided remarks that helped to sharpen our thoughts. In particular we thank Harald Uhlig for his excellent editorial comments that allowed us to sharpen and improve the paper. Four referees also provided astute remarks that improved the manuscript. Thanks to The Templeton Foundation and the Science of Philanthropy Initiative at the University of Chicago for providing the generous funds to make this experiment possible, and to the ESRC Centre for Climate Change Economics and Policy through the Centre for Climate Change Economics and Policy [grant number ES/K006576/1], and the Grantham Foundation for the Protection of the Environment for funding the research of Greer Gosnell. Further thanks to the UK Civil Aviation Authority and the pilots' unions who took the time to review and approve of the study objectives and material. A special thanks to those individuals at Virgin Atlantic Airways—especially to Paul Morris, Claire Lambert, Dr. Emma Harvey, and Captain David Kistruck—and Rolls Royce (especially Mark Goodhind and Simon Mayes) for their essential roles in the implementation of this experiment. These parties are in no way responsible for the analyses and interpretations presented in this paper. We thank Florian Rundhammer and Andrew Simon for their excellent research assistance. This paper was formerly titled "A New Approach to an Age-Old Problem: Solving Externalities by Incenting Workers Directly." We gained IRB approval from the University of Chicago under IRB13-1272.

1. Introduction

One of the longstanding puzzles in economics relates to the striking observed differences in firm-level productivity across space and time. For example, total factor productivity ratios of 3:1 or more are not unusual across 90th percentile to 10th percentile producers within major manufacturing industries (Foster et al., 2008). Understanding the sources of such differences remains key to deepening our knowledge of the causes of economic growth and the nature of prosperity. Syverson (2011) provides a discussion of the determinants of and underlying differences in observed productivity at the micro-level, but what remains rare are causal tests of an ingredient that behavioral and management economists deem as first order: the role of management practices.

Recently, a rich literature has developed that provides key evidence of a robust relationship between management and firm-level performance. Specifically, it points to the import of operations management, performance monitoring, target setting, and people management for improving productivity (Bloom and Van Reenen, 2007, 2011; Bloom et al., 2013, 2015; Tsai et al., 2015; McKenzie and Woodruff, 2017; Bruhn et al., 2018). In general, evidence is scant concerning the causal relationship between management practices and the productivity of skilled labor, particularly in the developed world. On the one hand, the correlations produced in the literature might suggest that skilled workers select into organizations that manage effectively, or that particular management strategies may causally improve workers' productivity. On the other, (excessive) management could backfire due to perceptions of control—and therefore reduced choice autonomy—and distrust on the part of the principal that such management may elicit in the agent (Akerlof and Kranton, 2005; Falk and Kosfeld, 2006; Ellingsen and Johannesson, 2008). While several studies have attempted to delve into the "black box" of firm-level operations to observe the effects of human resource management on productivity (Ichniowski et al., 1997; Lazear, 2000; Shaw, 2009; Bloom et al., 2013, 2016), a lack of robust causal evidence renders the effectiveness of distinct management practices well-grounded in principal-agent models unsettled.

We aim to narrow this gap in understanding by reporting results from a large-scale field experiment conducted in partnership with a major international airline. Our primary goal is to identify rigorously the impact of pertinent management practices—with increasing degrees of intensity—on measured productivity of skilled labor, particularly in the developed world where such research is especially lacking. Our secondary goal is to pressure test theoretically and empirically supported indications of the importance of prosocial motivation in determining productivity in a real-world labor setting. We focus on commercial airline captains' productivity, where productivity is defined as a function of fuel use, time delays, and safety. Several features render the commercial aviation context—and airline captains in particular—ideal for investigating the impact of management practices on productivity.[1]

First, like much skilled labor, captains in the commercial aviation industry possess strong

[1]Two observational approaches have assessed the historical impact of management on productivity. Giorcelli (2016) uses a quasi-experiment on management training and technology adoption in Italy under the post-war Marshall Plan and finds positive impacts on productivity. Bender et al. (2018) match the Bloom and Van Reenen (2007) survey data with German employee administrative data (from 1975 onwards) and find that about half of the TFP-management relationship is related to managerial ability.

professional identities and a sense of social obligation and organizational mission. The extent to which various management practices affect the productivity of such identity- and mission-driven personnel is largely unknown. Second, and relatedly, captains embody significant human capital, operate in a high-stakes environment, and receive a considerable professional wage as a result.[2] One might posit that management practices should have little to no effect on such high-wage and -ability types, in which case direct application of findings from the emerging literature on the productivity impacts of management would be a futile exercise. Finally, captains' decisions play an integral part in determining their firm's bottom line, so discerning which management practices motivate these flagship employees has significant implications for the short-term profitability and long-run financial success of the firm.

Various combinations of these attributes characterize a number of professional occupations, such as architects, civil servants, consultants, engineers, lawyers, medical doctors, military personnel, researchers, and tech workers. While some prominent management practices may be implemented in such professional settings, little evidence exists to support their effectiveness to increase productivity. For instance, with respect to employee performance monitoring, the few relevant field studies that exist typically focus on low-wage workers with little or no professional identity, and thus monitoring can be posited to strengthen motivation by simply increasing the potential cost of poor performance (see Nagin et al., 2002). In high-skilled labor contexts, however, it is plausible that monitoring is irrelevant to motivation since workers are typically well-trained and intrinsically motivated. In fact, monitoring may even demotivate these types if they perceive it to question or undermine their occupational proficiency or status (Falk and Kosfeld, 2006). Commercial airline captains train meticulously to earn their seat in the cockpit and are profoundly motivated to excel in their work. Such professionalism combined with captains' considerable accountability for discretionary input costs within airlines renders them prime candidates for discernment of the effects of monitoring, and other management practices, in consequential professional settings.

We make use of a rare opportunity to isolate the responses of high-skilled labor to specific management strategies to which they have no prior exposure. Since airline captains' unions are traditionally reluctant to instill managerial changes—including incentive provision—airlines have faced a fundamental incapacity to alter wage structures and management practices to boost captains' productivity in line with principal-agent models (Holmström, 1979). By focusing on behaviors embedded within the airline's standard operating procedures and holding captains' private financial incentives constant, we elicited union-level authorization to conduct the study, thereby providing a unique opportunity to "look under the hood" of an organization that critically relies on the performance of skilled employees.

We observe more than 110,000 binary behavioral outcomes across 40,000 unique flights over a 27-month period for the entire population of captains within Virgin Atlantic Airways ("VAA"; N=335) who were eligible to fly during the full time period under investigation (January 2013 to March 2015). Captains were randomly allocated to one of four study groups subject to: (i) performance monitoring (i.e., our control group), (ii) informational

[2]The average salary of a captain in our study is roughly $175,000-$225,000 (based on information updated in June 2015: http://www.pilotjobsnetwork.com/jobs/Virgin_Atlantic).

performance feedback, (iii) target setting, and (iv) prosocial incentives.[3] In this form, the experiment provides an opportunity to measure the incremental effects of each distinct mechanism on measured aspects of productivity while holding personnel fixed. Such a design is in contrast to the previous field experimental literature that has focused on the Bloom and Van Reenen (2007) management practices in a bundled manner (Bloom et al., 2013; Fryer Jr., 2017; Bruhn et al., 2018) and thus has difficulties identifying the marginal impact of each practice on productivity.

While prosocial incentives and motivations are not explicitly considered within the traditional management toolkit (e.g., Bloom and Van Reenen, 2007), they have recently received considerable attention in the literature both theoretically (Ellingsen and Johannesson, 2008; Bénabou and Tirole, 2010) and empirically (Tonin and Vlassopoulos, 2010; Anik et al., 2013; Imas, 2014; Charness et al., 2016; Hedblom et al., 2016). Moreover, according to a recent survey, 67% of CEOs believe prosocial considerations are increasingly essential for acquiring and motivating high-skilled labor (PricewaterhouseCoopers, 2016). We randomize prosocial incentives in our field experiment to investigate whether their inclusion into the management practice survey enhances its applicability to skilled and/or mission-driven labor contexts.

Our field experiment importantly leverages recent developments in aircraft data processing that capture precise flight-level measures of fuel-related productivity across three distinct phases—pre-flight, in-flight, and post-flight—which we package in a binary "hit or miss" fashion. The pre-flight measure, denoted ***Fuel Load***, assesses the accuracy with which captains implement final adjustments to aircraft fuel load prior to takeoff given all relevant factors (e.g., weather and aircraft weight). The in-flight measure, denoted ***Efficient Flight***, captures the fuel efficiency of captains' decisions between takeoff and landing. The post-flight measure, denoted ***Efficient Taxi***, indicates whether the captain turns off at least one engine during taxi-in. Since captains maintain ultimate authority over these decisions, airlines generally encourage but do not mandate these or similar measures to optimize firm efficiency.

We report three primary results. First, within-subject analysis of behavior before versus during the experiment—akin to the design of Bandiera et al. (2007, 2009)—strongly suggests that simple performance monitoring can significantly induce readily attainable improvements in labor productivity.[4] Within-subject analysis indicates that the proportion of flights in the control group—who were aware that we began monitoring behaviors during the study—on which captains successfully performed the Efficient Flight metric increased by nearly 50 percent compared to the pre-experimental period. Moreover, these captains successfully increased implementation of Fuel Load and Efficient Taxi by around 10 percent. In short, a

[3]These practices could be construed as falling within the *performance monitoring*, *target setting*, and *people management* concepts of workplace management by Bloom and Van Reenen (2007, 2011) and Bloom et al. (2017). While a vast economic literature demonstrates theoretical and empirical robustness of financial incentive provision (e.g., see Lazear, 2000), strong unionization of airline captains precluded incorporation of direct financial incentives in our experimental design.

[4]In an ideal setting, we would compare a genuine business-as-usual 'control' group that remained unaware of the experiment to our aware 'control' group. Such a design was precluded by the VAA- and union-dictated stipulation that we be fully transparent with all captains about the project's undertaking. Even so, had such a design been possible, the high likelihood of contamination to unaware captains in this idealized control group would have only allowed for estimation of a lower bound monitoring effect that diminished with time (i.e., as information spread through the captain population). As a result, and given the granular (pre-)experimental data Virgin had captured at the captain level, the before-and-after approach may nonetheless be preferred in this context.

simple lesson arises: what is measured is improved, even among skilled professionals.

Second, despite the sizable monitoring effects, we find a significant role for the additional management practices contained in our experimental treatments. While all three practices led to statistically significant increases in at least one of the measured behaviors, we observe significant differences between the performance of captains who received personalized performance targets (with or without incentives) versus those who did not. The experimental findings indicate that captains who received performance targets implemented the measured behaviors by up to 28 percent above the monitoring group's implementation.[5] Including a conditional prosocial incentive in the form of a donation to charity did not further improve productivity beyond the effects of providing a personal target.[6] Overall, these interventions were quite cost effective: they resulted in a reduction in fuel use of more than 7,700 tons (i.e., $6.1 million in 2014 prices) over the eight-month experimental period.[7]

Third, performance monitoring leads to productivity gains for the firm beyond the experimental window. The six-month post-experiment baseline (i.e., performance in the control group) remains considerably improved from the pre-experiment baseline, indicating either that captains believe that they are still being monitored or that monitoring for a fixed period of time induces captains to make low-effort efficiency improvements that are quickly habituated. Conversely, treated captains' productivity reverts to post-experiment baseline levels for Fuel Load and Efficient Flight, while the treatment effects remain but attenuate for Efficient Taxi. Such attenuation suggests that the productivity improvement induced by the experimental treatments depend on recurrent administration.

While the management strategies led to reductions in fuel costs for the firm, it is possible that their implementation induces spillovers on other important internal metrics—such as delays, safety, and captains' job satisfaction—as well as external costs, such as carbon dioxide emissions. With regards to on-time performance, we find that our experimental groups did not increase delays. With respect to safety, all behaviors targeted within the study are well within VAA's stringent safety standards.[8] All communications with the captains were vetted by senior decision-makers (i.e., unions, select senior captains, and VAA management). Each highlighted safety as the airline's utmost priority and emphasized the targeted behaviors' situation well within the airline's standard operating procedures. Moreover, all targets were capped at 90% of a captain's flights to alleviate pressure and allow flexibility in decision-

[5]These findings are consistent with two studies in the developed world that show that management practices may causally impact on productivity (Fryer Jr., 2017; Bloom et al., 2018).

[6]To provide a comparison to existing organizational practices, simulator trainings do not have any identifiable impacts on the targeted behaviors in our study. Ensuring that employees are up-to-date in their training is an important management practice across numerous sectors, and we are fortunate to observe the incidence of randomly timed sessions in which the captains in our sample completed simulator trainings. We find that attending a simulator session has no impact on productivity as measured by the three metrics highlighted in the study. These training sessions however provide other benefits not observed in our data.

[7]Additionally, we find that the environmental benefits associated with the fuel savings from this study are in the range of $0.5 to $2 million (depending on the social cost of carbon considered) – equivalent to 8 to 33% of the total fuel savings. In this manner, our approach provides a new way to combat firm-level externalities: target workers rather than the firm as a whole.

[8]These standards comply with those set by the UN International Civil Aviation Organization, the European Aviation Safety Agency, and the U.K. Civil Aviation Authority. Moreover, as in our case, many airlines maintain stricter safety standards than those imposed upon them by these external bodies.

making on any given flight. No fuel-related safety incidents were reported during the study.[9]

Finally, survey results indicate that management practices can improve employees' well-being, and captains express demand for further engagement. The experimental treatments appear to have positively affected captains' reported job satisfaction relative to the control group with the largest gains coming from the prosocial group captains, whose average job satisfaction exceeded that of the control group by 6.5%. Additionally, of the 60% of captains who responded to the survey, 79% indicated a desire for the continuation of the management strategies embodied in the experimental treatments, while only 6% expressed a preference for the pre-study status quo.

Empirical results from our field experiment hold implications for the design of management practices within firms comprised of skilled workers in advanced economies, suggesting a compelling role for: (i) managerial oversight in the form of performance monitoring of defined productivity outcomes; and (ii) comparison of employees' own productivity against personalized targets set by management personnel. Performance feedback on its own provides little enhancement to employees' productivity beyond the effect of monitoring on its own, and prosocial incentives do not boost productivity beyond the effect of personalized targets in this context. As emerging and developed economies continue to advance, economists should pursue a more thorough and nuanced understanding of the generalizability and effectiveness of distinct management practices in increasing the productivity of skilled labor. As our results suggest, the potential to unlock greater economic growth is ubiquitous.

The remainder of the paper is structured as follows. Section 2 provides a detailed contextual background and outlines the experimental design. Section 3 reports the results of the field experiment, focusing on productivity change, resultant efficiency gains, and measurable spillover effects. Section 4 concludes. The online appendix includes additional theoretical and empirical analysis.

2. Background and Experimental Design

In this section, we highlight the three productivity-related behaviors under investigation (section 2.1), the field experimental design (section 2.2), and the details of implementation (section 2.3), including a description of the sample and the method of randomization.

2.1 Captains' Behavior and Fuel Efficiency

While many of the decisions of airline captains' are important in determining fuel use, we worked with VAA to identify three measurable and non-overlapping levers to improve productivity for the purpose of this study: Fuel Load, Efficient Flight, and Efficient Taxi.

[9]While we do not have access to granular safety data that allows for robust analysis on this front, VAA has assured us that, according to their careful and meticulous analyses, there were no fuel-related safety incidents during or following the experiment. However, the lack of means to robustly assess potential nuanced changes in flight safety or captains' risk-taking is a shortcoming of our data. In addition, we are unable to investigate any unobserved consequences from multitasking (a la Holmström and Milgrom, 1991) or unintended impacts on customer satisfaction due to lack of data on these outcomes.

Behavior 1: Fuel Load

The first lever concerns a pre-flight procedure known in the aviation industry as the Zero Fuel Weight (ZFW) adjustment. Approximately 90 minutes prior to each flight, captains utilize flight plan information (e.g., expected fuel usage, weather, and aircraft weight) in conjunction with their own professional judgment to determine initial fuel uptake, which usually corresponds to approximately 90% of the anticipated fuel necessary for the flight. This amount is fueled into the aircraft simultaneous to the loading of passengers and cargo. Near to completion of passenger boarding and cargo/baggage loading, the pilots—now on the flight deck—receive updated information regarding the final weight of the aircraft and may adjust their fuel uptake accordingly. The information they receive from Flight Operations includes a ZFW measure, which indicates the weight of the aircraft with passengers and cargo onboard, as well as the Takeoff Weight (TOW), which additionally includes fuel.

Captains then perform a ZFW calculation in which they first determine the amount by which they should increase or decrease planned fuel load based on the final ZFW using a mathematical formula that is standard across the airline industry. Should they decide to increase the fuel load, they subsequently compute a second iteration to account for the additional fuel necessary to carry the increased fuel load. If the fuel previously loaded onto the aircraft is sufficient according to the calculations, the captain may choose not to load additional fuel. At present, the captain makes this calculation at the beginning of each flight (the process is not automated).

We denote this binary outcome variable as *Fuel Load*. Fuel Load indicates whether the double iteration calculation has been performed and the fuel level adjusted accordingly. We deem the captains' behavior successful if their final fuel load is less than or equal to 200 kg above the "correct" amount of fuel as dictated by the calculation.[10] This allowance prevents penalizing captains for rounding and slight over-fueling on the part of the fueler while providing measurable targets for captains in two treatment groups. According to VAA, accurate Fuel Load adjustment should be performed on every flight, corresponding to 100% attainment of the metric provided. In the thirteen months prior to the experiment, this behavior was performed correctly on just 42% of the flights (see Table 3).

Behavior 2: Efficient Flight

The second behavior is an in-flight consideration, *Efficient Flight*, which captures whether captains (in conjunction with their co-pilots) use less fuel during flight than is projected in the updated flight plan. An original conservatively cost-optimized flight plan is drawn up several hours prior to departure based on flight-specific information and performance data that is particular to the type of aircraft to be flown. Inputs to the flight plan are updated subsequent to decisions made on Fuel Load so that decisions regarding the first metric do not affect one's ability to meet this in-flight metric. Efficient Flight captures whether

[10] Using data from a major U.S. airline, Ryerson et al. (2015) estimate that 4.5% of fuel burned on an average flight is attributable to carrying unused fuel, and that more than 1% of fuel burned on an average flight is due to addition of contingency fuel "above a reasonable buffer". Virgin Atlantic deemed 200 kg—equivalent to allowing for 0.5% error in the calculation—a reasonable buffer to allow for rounding and fueler error. Our results in Section 3 are robust to upward and downward adjustments of this buffer by 50 kg.

captains have actively engaged in fuel-efficient practices between takeoff and landing, such as requesting and executing optimal altitudes and shortcuts from air traffic control, maintaining ideal speeds, performing continuous climb and descent approaches, optimally adjusting to en route weather updates, and ensuring efficient aerodynamic arrangements with respect to flap settings as well as takeoff and landing gear. Captains may approximately predict the fuel savings of, for example, changing speeds or altitudes using computers on board the aircraft.

This in-flight metric is designed to capture various available risk-free and fuel-optimizing behaviors that require effort and are not always implemented. Furthermore, by focusing on fuel use rather than the execution of specific behaviors, the metric affords captains the flexibility to achieve the target while using professional judgment to ensure that safety remains the utmost priority. Under some uncommon circumstances, operational requirements dictate that captains sacrifice fuel efficiency (and VAA accepts the captains' decisions as final), so we would not expect even a "model" captain to perform this metric on 100% of flights, though the metric should be attainable on a vast majority of flights (contrasted with 31% pre-experimental attainment). In our analysis, the Efficient Flight indicator variable is 1 if the actual in-flight fuel use does not exceed the projected fuel use (adjusted for actual TOW), and 0 if the in-flight fuel use is more than projected.[11]

Behavior 3: Efficient Taxi

The final behavior, *Efficient Taxi*, occurs post-flight. Once the aircraft has landed and the engines have cooled, captains may choose to shut down one (or two, in a four-engine aircraft) of their engines while they taxi to the gate, thereby decreasing fuel burn per minute spent taxiing. Captains meet the criteria for this metric if they shut down one or more engines during taxi-in.[12] As with Efficient Flight, there are circumstances characterized by technical or operational restrictions under which the airline would not expect or prescribe captains to undertake Efficient Taxi.[13] Obstacles include geographical constraints (e.g., the placement or layout of the runway), route complexity (e.g., number of stops, turns, or cul-de-sacs), short taxi-in times, weather conditions, (e.g., ice or snow, or heat on asphalt surfaces), and low visibility, all of which are uncorrelated with treatment. Nevertheless, the metric should be attainable on a vast majority of flights, but in the 13-month pre-experimental period, there was a relatively low attainment (roughly 34%) for this metric.

[11]Note that it was essential to create binary metrics for Fuel Load and Efficient Flight so we could assign targets to captains in the targets and prosocial incentives group.

[12]Fuel savings from Efficient Taxi depend on scheduling and delays as savings are accrued on a per-minute basis. Fuel savings also depend on aircraft type and only begin to accrue after engines have cooled, which takes 2-5 minutes from touch down. Savings per minute for aircraft operated within the study are as follows: 12.5 kg (Boeing 747-400, Airbus 330-300), 8.75 kg (Airbus 340-600), and 6.25 kg (Airbus 340-300). Efficient Taxiing data is physically stored on QAR cards inside the aircraft, which are removed every 2-4 days to pull data. These cards can corrupt or overwrite themselves, and also can reach full memory capacity before being removed. Therefore, data capture for Efficient Taxi is not complete—exactly 37% of flights are missing data for this metric. The reason for the missing data is purely technical and cannot be influenced by captains. We regress an indicator variable of missing Efficient Taxi data on treatment indicators and find no statistically significant relationship at any meaningful level of confidence (individual and joint $p > 0.4$). Consequently, this phenomenon should not affect results beyond reducing the power of estimates.

[13]An international survey of aircraft captains highlighted potential issues associated with excessive thrust, maneuverability, and extensive workload that may preclude them from undertaking this behavior in particular circumstances (Balakrishnan et al., 2011).

2.2 Experimental Design

In accordance with captains' optimization problem as proposed in our theoretical model (see Appendix A), the eight-month field experiment focuses on four management practices targeting productivity: monitoring (control), performance information, performance targets, and prosocial incentives. Our goal is to maximize productivity in relation to Fuel Load, Efficient Flight, and Efficient Taxi. Respectively, these behaviors allow us to measure captains' effort before takeoff, during the flight, and after landing. The captains did not receive detailed information relating their decision-making to their fuel efficiency prior to this experiment, consistent with both airline and industry standards.[14]

Importantly, *all* eligible VAA captains were included in the experimental sample.[15] Hence, captains did not select into the experiment, and as a result we can estimate the average treatment effect for the entire roster of eligible captains in VAA. As such, our behavioral parameter of interest shares much in common with that estimated in a natural field experiment (see Al-Ubaydli and List, 2015). Yet, all captains knew that they were part of an experiment, and therefore our study shares features with both framed and natural field experiments (Harrison and List, 2004).

We observe captains' behavior from January 2013 through March 2015, and the experimental window was from February through September of 2014. During this period, monthly branded feedback reports pertaining to the previous month's flights were sent to the home addresses of treated captains, who received their first feedback report in mid-March 2014 and their final feedback report in mid-October 2014. The experimental treatment groups can be summarized as follows:

Control Group: Monitoring. All captains included in the study were aware that they were part of an experiment; that is, the monitoring (control) group did not receive any feedback but was aware that their productivity was being monitored.[16] Two weeks prior to the study start date of February 1, 2014, all captains were informed that VAA would be undertaking a study on fuel efficiency as part of its "Change is in the Air" sustainability initiative. The initial letter outlined the three performance-related behaviors to be measured

[14]One might question why VAA did not perform these management practices prior to the study. Claire Lambert, Fuel Efficiency Manager at VAA during the study, explains, "*There are a number of reasons why Virgin Atlantic was not undertaking this type of initiative in earlier years. Firstly, we had not had much exposure to behavioral science before we established our partnership with the universities. Secondly the granularity of data required to implement the study was a development for us. For airlines, the introduction of the EU Emissions Trading Scheme in 2010 really drove the need for better data, and the emergence of data service providers' software systems at around the same time enabled us access to the data. Thirdly, undertaking this study, even having outsourced a large portion of the experimental planning and implementation to the academic team, was quite labor and time-intensive, from the early-stage engagement with our pilot union and captains through to the frequent data processing and collaboration with the academic team to conduct the experiment properly. We were already tackling most of the other fuel efficiency margins, and with access to the newly available data, we found ourselves in a position to advance our continuously evolving fuel efficiency strategy by providing nuanced data to captains to facilitate fuel-efficient decision making in the flight deck. The academic research partnership provided an opportunity to do so in an innovative way and to test how such a strategy might be optimized going forward.*"

[15]Additionally, all routes were included in the study apart from within U.K. flights; Appendix B contains a map of all VAA destinations during the study period.

[16]A pure monitoring effect aligns with agency theory (e.g., Alchian and Demsetz, 1972; Stiglitz, 1975), as well as with experimental results such as those in Boly (2011), Nagin et al. (2002), and observational study results from Hubbard (2000, 2003) and Pierce et al. (2015).

and the possible study groups to which the captains may be randomly assigned.[17] Captains in treatment groups were to receive letters the following week to inform them of what to expect in the coming months, and the monitoring group would receive no additional information.

In the final week of January 2014, VAA sent letters to all treated captains informing them of the intervention to which they had been assigned. The letters included a sample feedback report, which contained the individuals' targets if they had been assigned to either the targets or prosocial group.[18]

Treatment Group 1: Information. Each feedback report details the captain's performance of Fuel Load, Efficient Flight, and Efficient Taxi for the prior month (see Figure A1 in Appendix C). Specifically, the feedback presents the percentage of flights flown during the preceding month on which the captain successfully implemented each of these metrics. For instance, if a captain flew four times in the prior month, successfully performing Fuel Load and Efficient Taxi on one of the flights and Efficient Flight on two of them, his feedback report would indicate 25% attainment for the former behaviors and 50% attainment for the latter. This treatment aligns closely with the "Performance Tracking" and "Performance Review" management practices outlined in Bloom and Van Reenen (2007, 2011).

Treatment Group 2: Targets. Captains in this treatment group received the same information outlined above but were additionally encouraged to achieve personalized targets of 25 percentage points above their pre-experimental baseline attainment levels for each metric (capped at 90%; see Figure A2 in Appendix C). The targets were communicated to these captains prior to the start of the experiment. An additional box is included in the feedback report to provide a summary of performance (i.e., total number of targets met). Captains were not rewarded or recognized in any public or material fashion for their achievements. This intervention is in line with the management practices called "Target Balance", "Target Connectedness", "Time Horizon of Targets", "Target Stretch", and "Clarity and Comparability of Targets" (Bloom and Van Reenen, 2007, 2011).[19]

Treatment Group 3: Prosocial Incentives. In addition to the information and targets outlined above, those in the prosocial incentive treatment group were informed that achieving their targets would result in donations to charity (see Figure A3 in Appendix C). Specifically, for each target achieved in a given month, £10 was donated on behalf of the captain to a chosen charity. When captains in this group were informed of their assignment to treatment, they were offered the opportunity to choose one of five diverse charities to support with their prosocial incentives: Free the Children, MyClimate, Help for Heroes,

[17] Given that all captains were aware of the start date of monitoring, we additionally derive estimates of the effects of monitoring on captains' performance in the manner of Bandiera et al. (2007). The identified effects for monitoring are therefore not experimental per se, but are based on careful analysis controlling for relevant observables and trends.

[18] Captains were encouraged to engage with the material and send any questions to an email address created specifically for study inquiries. Once the experiment was complete, we sent treated captains a debrief letter informing them of their overall monthly results with respect to their targets (if in the targets or prosocial treatment groups) and their total charitable donations (if in the prosocial incentives treatment group). All (treatment and control) captains were informed that a follow-up survey would be sent to their company email addresses in early 2015. The follow-up survey was designed and administered by the academic researchers alone. Again, captains were assured that data from their responses would be used for research purposes only, that their responses would remain anonymous, and that VAA would not be privy to individual-level information provided by survey respondents.

[19] Such target-setting has its roots in industrial organization psychology through SMART (specific, measurable, attainable, relevant and timebound) targets (Locke and Latham, 2006)

Make A Wish UK, and Cancer Research UK.[20] Therefore, captains in this group each had the opportunity to donate £30 ($49) per month for a total of £240 ($389) to their chosen charity over the course of the eight-month trial. Captains were reminded each month of the remaining potential donations that could result from realizing their targets in the future.

While incentives are a cornerstone of management in Bloom and Van Reenen (2007) (and beyond), *prosocial* incentives are not explicitly included in typical management surveys to date. Our research contributes to the conversation surrounding whether and how personnel economics might broaden the notion of people management to incorporate such incentives. Evidence appears to suggest that prosocial considerations may be quite important to employee productivity, particularly on the extensive margin.[21] That said, many of the existing studies focus on low-stakes occupations, or tasks that do not require high human capital.[22] Our goal is to understand how conditional prosocial incentives causally change productivity of high-skilled employees in high-stakes work situations.

The overarching goal of our design is to identify the marginal effects of management practices, with each component of conditional incentive provision—monitoring, information, targets, incentives—considered in isolation (see Table 1).[23] This design is in contrast to the previous literature that has focused on applying the Bloom and Van Reenen (2007) management practices in a bundled manner (Bloom et al., 2013; Fryer Jr., 2014; Tsai et al., 2015; Bloom et al., 2015; Fryer Jr., 2017; Bruhn et al., 2018). Such studies have difficulties identifying the marginal effect of each management practice on productivity. In other words, to provide conditional incentives to an employee, a firm needs to put in place an appropriate target. In order to provide a meaningful target, a firm needs to share information on the employee's performance. In providing such information, the employee becomes aware that the manager can monitor her performance. Accordingly, each additional component is layered to isolate the real behavioral motivator behind conditional incentives, a strategy that most personnel economists would support first and foremost to motivate employee performance (see Holmström's (1979) seminal theoretical work and Lazear's (2000) seminal empirical work).

[20] Eighteen captains selected a charity by emailing the designated project email address, and 67 captains who did not actively select a charity were defaulted to donate to Free the Children. Captains could choose to remain anonymous, otherwise exact donations were attributed to each individual (identified by their first initial and last name).

[21] On the supply side, workers have revealed a preference for being employed by a company with strong CSR practices, which appear to attract higher ability types and increase productivity. This literature started in the 1990s with the observational datasets in Turban and Greening (1997) and Greening and Turban (2000), and recent field experiments highlights that CSR can motivate high-ability types to apply for job openings (Hedblom et al., 2016).

[22] Online or lab experiments have assessed the effect of charitable incentives on productivity in low-effort tasks (i.e., the intensive margin)—see Tonin and Vlassopoulos (2014), Imas (2014), and Charness et al. (2016). Anik et al. (2013) use a field study to estimate the impact of unconditional charitable bonuses on productivity, and Tonin and Vlassopoulos (2010) recruit university students for a field experiment using charitable incentives for a data entry task to measure pure and impure altruism. Elfenbein et al. (2012) show that sellers who tie products to a charitable donation may be deemed more trustworthy by consumers. Relatedly, field experimental research into unconditional gifts is a burgeoning area of research—see Gneezy and List (2006); Bellemare and Shearer (2009); Hennig-Schmidt et al. (2010); Englmaier and Leider (2012); Kube et al. (2012); and Cohn et al. (2015).

[23] We acknowledge that management practices may extend beyond monitoring, information provision, target setting, and incentive provision. For example, the Bloom and Van Reenen (2007) survey additionally captures more intangible and complex management dynamics, such as the means by which problems are addressed within the organization and the processes behind decisions to hire, promote, retain, and fire employees.

2.3 Further Experimental Details

2.3.1 Sample

Our data consists of the entire eligible universe of VAA captains in 2013 and 2014 (N = 335), of which 329 are male and 6 are female.[24] Of the debrief survey respondents (N = 202), 97 classified their training as military and 102 as civilian (the remaining declined to state). Eleven captains are "trusted pilots" selected for pre-study consultation regarding study feasibility and communications[25], and 62 captains are "trainers" who are responsible for regularly updating and training their colleagues in the latest flight techniques. Captains range from 37 to 64 years of age, where the average captain is 52 years old and had been an employee of the airline for over 17 years when the study initiated. Captains in the sample flew five flights per month on average, where the captain flying most averaged almost eight flights per month and the captain flying least averaged just over two flights per month.

The resulting dataset consists of 42,012 flights and 110,489 observations of the three fuel-related behaviors. Among other variables, we observe fuel (kg) onboard the aircraft at four discrete points in time: departure from the outbound gate, takeoff, landing, and arrival at the inbound gate. In addition, we observe fuel passing through each of the aircraft's engines during taxi, which provides a precise measure of fuel burned on the ground. We use such data to understand how the management practices ultimately affect fuel use in Section 3.2. We also observe flight duration, flight plan variables (i.e., expected fuel use, flight duration, departure and arrival destinations), and aircraft type. We control for several flight-level variables—e.g., ports of departure and arrival, weather on departure and arrival, whether the aircraft had just received maintenance (belly wash, engine change), and aircraft type—and individual fixed effects and captain-level time-varying observables, such as current contracted work hours and whether the captain had attended VAA's annual training.

Four months after the study's completion, we elicited captains' job satisfaction and preferences over the various management practices through an online survey (response rate = 60%). We report these respective analyses in Sections 3.3.3 and 3.3.4. Through the survey, we sought to gain a rigorous understanding of the relevance of management to subjective reports of job satisfaction, an important outcome of particular relevance in this context as mental health concerns have gained prominence in the aviation industry.[26] Similarly, ascertaining captains' demand for various management practices allows us to glean more in-depth insights into the effects of these practices on employees' choices and well-being.

[24] While we understand that there may be partner selection bias inherent in our (or any) firm-level study (Allcott, 2015), our experience with many other international commercial airlines suggests they are no more (and sometimes considerably less) advanced in their management of captains' fuel efficiency. The most advanced airlines—including VAA—purchase software that allows management to visualize some flight-relevant information ex post. The aviation sector appears similar to other sectors on this dimension. For instance, Bruhn et al. (2018) find that small- to medium-sized enterprises in Mexico do not use particular management consulting services because they lack the funds, do not have knowledge of potential benefits, or simply have not considered the possibility.

[25] We run the data analysis both including and excluding trusted pilots and the results do not change.

[26] For instance, Wu et al. (2016) surveyed a random set of 1,866 captains and found that 13% had clinical depression and 4% had suicidal thoughts in the last two weeks. Moreover, the recent U.S. Federal Aviation Administration has made a recommendation for greater focus on pilot mental health in aviation policy and practice (Federal Aviation Administration, 2015).

2.3.2 Randomization

To randomize captains into treatment, we first blocked three months of pre-experimental data (September through November, 2013) on five dummy variables that captured whether subjects were above or below average for the: i) number of engines on aircraft flown, ii) number of flights executed per month, and iii) attainment of the three selected fuel-relevant behaviors, which are our primary dependent variables. Number of engines and monthly flights proved significant in predicting the selected outcome behaviors in preliminary regressions. Once blocked, captains were randomly allocated to one of the four study groups through a matched quadruplet design (for further details, see Appendix D). To ensure that individual-specific observable characteristics are balanced across groups, we performed subsequent balance tests for seniority, age, trainer status, and trusted pilot status as well as flight plan fuel use (as a proxy for average flight distance), actual fuel use, average number of engines on aircraft flown, flying frequency, and the three targeted behaviors (see Table 2).

Table 3 and Figure 1 provide a summary description of captains' performance before and during the experimental period within each experimental group. In accordance with the balance checks above—which focus on just three months of pre-experimental data—the summary statistics from January 2013 through January 2014 (i.e., Table 3, 'Before Experiment') provide assurance that the pre-experimental behavioral outcomes are balanced across various study groups. None of the differences across groups are statistically discernible. In short, an exploration of all available aspects of captain and flight data reveals that the randomization was successful in that the observables are balanced across the four experimental conditions.[27]

3. Results

We summarize our main results in four steps. First, we estimate the impacts of the four management interventions on the selected behaviors. Second, we assess the consequences for overall fuel usage. Third, we consider whether the study affected the airline's reported delays. Fourth, we assess spillovers with respect to delays, safety, greenhouse gas emissions, and captains' well-being.

3.1 The Effects of Management Practices

Figure 2 presents aggregated data for each of the targeted fuel-related behaviors for the 21-month period for which we have pre-experimental and experimental data. This period includes 13 months of data prior to the announcement of the experiment (i.e., monitoring), and 8 months of within-experiment data. The dashed vertical line indicates the beginning of experimental monitoring. The pooled data in Figure 2a indicates that the implementation of Efficient Flight and Efficient Taxi substantially increases after monitoring is announced.

[27] In Tables A1 and A2, we additionally check for balance in pre-experimental trends for each behavior-group pairing for differences in pre-experimental fuel use trends, respectively. We find no major differences in these trends across conditions.

For both behaviors the increases are approximately 10 percentage points in the first month—equivalent to a 25-33% treatment effect—with a vast majority of captains experiencing improvements (see Figure 3). Concerning Fuel Load, captains increase implementation by approximately 4 percentage points in the month following the announcement of monitoring, equivalent to a 10% treatment effect. Removing treatment effects, Figure 2b also suggests an increase in implementation due to monitoring, albeit less pronounced due the exclusion of treatment effects. The difference in behaviors before and during the experiment—including that of the control captains—leads to our first formal result:

Result 1. *The performance of captains in the control group improves considerably upon announcement of behavioral monitoring.*

While the aforementioned summary statistics are certainly consistent with Result 1, they do not account for the data dependencies that arise from each captain's provision of more than one data point, nor any trends in the pre-experimental period. To accommodate the panel nature of the data set, we estimate a regression model of the form:

$$\text{EfficientBehavior}_{it} = \alpha + \text{Exp}_{it} \cdot T_{it}\beta + \text{Exp}_{it}\gamma + T_{it}\delta + X_{it}\zeta + \tau_t + \omega_i + e_{it}$$

where $\text{EfficientBehavior}_{it}$ equals one if captain i performed the fuel-efficient behavior on flight t, and equals zero otherwise; Exp_{it} is an indicator variable that turns on during the experimental period; T_{it} represents a vector with indicator variables for the three treatments; X_{it} is a vector of control variables; τ_t is a linear monthly time trend; and ω_i is a captain fixed effect. We include all available and relevant flight-level variables as controls, which include weather (temperature and condition) on departure and arrival, number of engines on the aircraft, airports of departure and arrival, engine washes and changes, and airframe washes. Additionally, we control for captains' contracted flying hours and whether the captain has completed an annual training.[28]

We estimate the above difference-in-difference model specification for each of the behaviors treating the first day of the experiment as February 1, 2014, when monitoring of captains begins. Three different empirical approaches yield qualitatively similar results: a linear probability model (LPM), a probit model, and a logit model. For ease of interpretation, we focus on the results of the LPM in Table 4.[29]

Given that we do not have a group of captains lacking knowledge of experimental monitoring, we perform an investigation of pre-experimental trends to ensure that our econometric

[28]There are various training channels, foremost of which is time spent in the simulator in which captains must pass assessments; we do not have accurate data on these trainings. We control for attendance at the two-day "Ops Day" seminar, a gathering of small groups of pilots (approximately 20 per training) that includes discussion of the airline's goals and directions, with some informal training for pilots.

[29]Robust standard errors are clustered at the captain level. We also present Newey-West standard errors that are robust to heteroskedasticity and arbitrary autocorrelations within each captain. We perform two robustness checks to control for attrition and different lag lengths in the Newey-West errors. To ensure attrition does not influence these results, we include Table A3 in Appendix D, which performs the same specification excluding quadruplets in our randomization within which captains attrited (all five of which did so prior to the announcement of the study) - we find no differences in our results. Furthermore, we estimate the Newey-West errors with lags of m=1 and m=4 to test the robustness of our results to the underlying model (Newey and West, 1987). Our results are identical under each model.

estimates of the effects of monitoring do not merely represent ongoing shifts in behavior that would have taken place despite the study. Figures 4a-4c demonstrate pre-experimental trends (i.e., from January 2013 through January 2014) and provide a visual representation of the differences in implementation of the prescribed metrics before and during the experiment. Across Fuel Load and Efficient Flight, it is clear that there is no upward trend for any group pre-experiment. For Efficient Taxi, we do observe an upward trend, though there remains a substantial increase in the level of implementation during the experimental period across all groups. To control for this trend, we estimate the specification controlling for a linear time trend.[30] As expected, including a linear trend attenuates the monitoring effect on Efficient Taxi, where the metric drops from 12.5 percentage points to 3.8 percentage points.[31]

Our main regressions in Table 4 therefore reports the results of the difference-in-difference specification controlling for linear time trends. We first note that the coefficient estimate of the experimental period ("Expt"), which provides a point estimate of the extent to which the control group improves their performance once monitoring begins. The influence of monitoring is apparent: the control group increases their implementation of Fuel Load by 3.3 percentage points (7.8% effect, 0.07 standard deviations (σ), $p < 0.05$), Efficient Flight by 13.2 percentage points (42.4% effect, 0.29σ, $p < 0.01$), and of Efficient Taxi by 3.8 percentage points (10.8% effect, 0.08σ, $p < 0.05$).

The above insights lend evidence in favor of a strong monitoring effect, a result consistent with the importance of social pressure in our theoretical structure. They do not, however, shed light on the incremental effectiveness of the treatments in stimulating fuel-efficient behaviors. Results 2-4 address this central question:

Result 2. *Providing captains with information on recent performance moderately improves their fuel efficiency (particularly with respect to Efficient Taxi).*

Result 3. *The inclusion of personalized targets significantly increases captains' implementation of all three measured behaviors.*

Result 4. *Adding a charitable component to the personalized targets intervention does not induce greater effort than providing targets alone.*

Evidence of Result 2 can be found in Table 3 and Figures 1-4, which suggest that—despite increased performance in Fuel Load and Efficient Flight—the differences between the control and information groups are rather slight. Yet, there is a considerable change in Efficient Taxi implementation between the information and control groups (58.8% versus 50.7%). The standard difference-in-difference model estimates in Table 4 complement the raw data in Table 3, indicating that the information treatment induces captains to engage in more fuel-efficient taxiing behavior. The coefficient estimate suggests that the percentage of flights for which captains receiving the information treatment turned off at least one engine while taxiing to the gate increases by 7.9 percentage points ($p < 0.01$) relative to the improvement exhibited in the control group.

When considering the behavior of captains who receive personalized targets, we observe consistent treatment effects across all three performance metrics. From Tables 3 and 4 and

[30] See Table A4 in Appendix E for specifications without controlling for a linear time trend and without controls.

[31] This effect holds because the linear trend accounts for a growing share of gains during the experimental window.

Figures 1-4, it is apparent that the targets treatment pushed each measured behavior in the fuel-saving direction and the effects also appear to be in the fuel-saving direction for captains receiving prosocial incentives. Table 4 reveals positive and statistically significant effects of the intervention for a majority of behavior-treatment combinations, even beyond the sizable effect of monitoring. Most striking is the effect of the interventions on the implementation of Efficient Taxi, which captains in the targets group undertook on almost 10 percentage points more flights (19.1% effect, 0.19σ, $p < 0.01$).

We now isolate the incremental productivity impacts of each management practice in turn. The coefficients associated with the targets and prosocial treatments in Table 4 are very similar ($\beta_{targets} = 0.025$ vs. $\beta_{prosocial} = 0.022$ for Fuel Load; $\beta_{targets} = 0.047^{***}$ vs. $\beta_{prosocial} = 0.037^{***}$ for Efficient Flight; and $\beta_{targets} = 0.088^{***}$ vs. $\beta_{prosocial} = 0.096^{***}$ for Efficient Taxi).[32] However, these two groups of captains appear to outperform captains lacking performance targets. To investigate statistically this claim, we pool captains receiving personalized targets (i.e., targets and prosocial treatment groups) and compare outcomes to a pooled information and control group in an additional regression. We find that captains who receive targets significantly outperform those who do not on all three dimensions: Fuel Load ($\beta = 0.020^{*}$), Efficient Flight ($\beta = 0.034^{***}$), and Efficient Taxi ($\beta = 0.052^{***}$). A similar exercise confirms that prosocial incentives do not significantly improve behavior compared to targets alone. Thus, while information is an important mechanism in encouraging fuel-efficient behavior change, targets augment its effect in a manner that prosocial incentives do not appear to boost further.[33]

We supplement this analysis by investigating whether various captains are motivated to increase implementation of just one of the behaviors, or whether the effects are driven by some captains' improving on multiple behaviors relative to their own implementation prior to the study (see Appendix F). On average, (some) captains are more likely to increase implementation of both Fuel Load and Efficient Flight, but these captains did not necessarily also improve on Efficient Taxi. Similarly, captains who respond most strongly on the taxiing dimension may not have been more likely to fuel and fly efficiently. We therefore infer that the effects are not solely driven by a small subset of captains improving on all three dimensions. Rather, many captains are increasing their efficiency in various phases of flight.

Importantly, our data provide the ability to go beyond short-run substitution effects and explore treatment effects in the longer run. We therefore conduct a more nuanced investigation of the treatment effects by exploring their persistence after the experimental

[32] These treatment effects are extremely similar to those identified using experimental data alone (see Table A5 in Appendix E).

[33] Since each treatment builds on the last, we can "control" for the contents of previous treatments and are therefore able to make distinct comparisons across treatments. If we solely examine the comparison between the information group and the targets group, we find three positive coefficients: Fuel Load ($\beta = 0.015, p = 0.33$), Efficient Flight ($\beta = 0.020, p = 0.17$), and Efficient Taxi ($\beta = 0.016, p = 0.36$). For reference, the estimates of the effects of prosocial incentives relative to targets are relatively attenuated: Fuel Load ($\beta = 0.003, p = 0.84$), Efficient Flight ($\beta = 0.010, p = 0.52$), and Efficient Taxi ($\beta = -0.008, p = 0.67$). That said, we do not have enough statistical power (partially due to the decreased sample size when comparing just two experimental groups) to argue that the coefficients are statistically different from zero, though the effect sizes are noteworthy.

window.[34] Inspection of these data yields a fifth result:

Result 5. *Treatment effects attenuate or disappear after treatment is removed, though the monitoring effect remains for Fuel Load and Efficient Flight.*

In the six months following the experiment, control captains continue to outperform their pre-experimental baseline implementation of Fuel Load ($\beta = 0.043$, $p < 0.05$) and Efficient Flight ($\beta = 0.239$, $p < 0.01$; see columns 1-3 of Table 5).[35] Furthermore, the monitoring and information effects on Efficient Taxi effectively disappear once the experiment stops, while the treatment effects of targets and prosocial incentives remain quite strong ($\beta = 0.078$, $p < 0.05$ and $\beta = 0.062$, $p < 0.05$, respectively). Discontinuation of the feedback letters leads to a reduction in implementation of Efficient Flight for all treated captains, suggesting a benefit of repeated performance feedback for this outcome metric (see column 5).

3.2 Fuel Savings

Given the substantial behavioral change observed during the experimental period of the study, we report economically significant fuel and cost savings for our final formal result:

Result 6. *Largely due to fuel savings from the monitoring effect, we estimate the total overall savings to be 7,769 metric tons ($6,106,434) from the study throughout the eight-month experimental period. The three experimental treatments alone led to an estimated 1,355 metric tons in fuel savings and $553,000 in cost savings for Virgin Atlantic.*

To determine fuel savings from the study, we estimate within-captain differences in the disparity between flight plan (planned) fuel use and actual fuel use from the pre-study period to the study period (Table 6). We calculate fuel savings using an intent-to-treat approach captured in the following OLS specification, where the difference-in-difference regression coefficient provides the fuel savings per flight for each respective group:

$$F_{it} = \alpha + \text{Exp}_{it} \cdot T_{it}\beta + \text{Exp}_{it}\gamma + T_{it}\delta + X_{it}\zeta + \omega_i + e_{it}$$

where F_{it} is the fuel saved per flight (i.e., the difference between planned and actual fuel use) for captain i at time t. We sum the per-flight savings for each treatment group with the average per-flight monitoring effect to estimate the average flight-level savings for each group, which we then multiply by the number of flights flown by captains in the respective groups during the experimental period (see Notes of Table 6).

For the control group, we estimate the total fuel savings to be 1,648 tons (496.1 kg per flight saved $\times$ 3321 flights). For the treatment groups, we estimate additional fuel savings beyond the control group of 500 tons in the information group (150.19 kg x 3330 flights),

[34] We also investigate the dynamics of captains' responses within each treatment month with respect to temporal distance from the previous feedback report received. We do not find consistent evidence of a 'salience effect' (see Table A6 in Appendix E).

[35] It is important to note here that this effect may be due to captains' belief that monitoring continued after the experiment ended. Since we do not elicit such beliefs, we cannot distinguish between a persistent effect of having been monitored versus a continued effect of monitoring.

498 tons in the targets group (165.1 kg × 3016 flights), and 357 tons in the prosocial group (109.7 kg × 3258 flights). Summing these marginal effects with the monitoring effect (which affects all groups), we estimate the total savings for the information, targets, and prosocial incentives groups to be 2,152 metric tons ("ton" hereafter), 1,994 tons, and 1,974 tons, respectively. Taken together, the three treatments led to a marginal 1,355 ton decrease in fuel use in comparison to the control group, and incorporating the monitoring effects into the calculation, fuel savings sum to 7,769 tons for an overall value savings of $6.1 million (in 2014 fuel prices, where a ton of fuel cost $786). [36,37] For comparison, using a regression discontinuity design with randomly timed simulator trainings as input variables, we find no effect of simulator trainings on fuel efficiency, even when interacted with experimental monitoring (see Appendix I).

3.3 Additional Spillover Effects

In this section, we examine four potential spillovers from our experimental treatment groups. These include the impact on delays, greenhouse gas emissions, captain welfare (measured through job satisfaction), and the demand for further management practices.

3.3.1 Delays

As demand for air travel grows, on-time departure becomes an increasingly important aspect of operational efficiency for commercial airlines. Airlines may incur direct financial sanctions for departure delays due to regulations on airport slot misuse, and may also experience additional fuel-related costs in their attempts to recover time to remain on schedule for arrival. It is quite conceivable that departure delays could increase in frequency during the experimental period, since we are encouraging captains to make a more deliberate fuel calculation and to consider the fuel efficiency of in-flight decisions. Alternatively, there are a number of reasons why we might expect the number of delays to decline, such as the case when captains' anticipate monitoring extends to other important outcomes.[38]

Table A7 summarizes results of three fixed effects regressions specified as in Section 3.1, but with delay-related dependent variables. The first dependent variable captures whether

[36]Similar data-driven fuel savings estimates can be calculated separately for each of the three discrete behaviors measured in a given flight (see Table A12 in the online appendix). We find significant fuel savings across all of the groups during the experimental period compared to the pre-experimental period, and the highest fuel savings comes from Efficient Flight followed by Fuel Load, as expected. We find that the targets and prosocial groups exhibit the largest fuel savings, consistent with the greatest changes in observed behavior.

[37]Comparable reductions in fuel demand would require fuel prices to increase by between 2.3% and 17.5%, according to price elasticity of jet fuel demand estimates from the literature. See Appendix H.

[38]Beyond possible anticipation that delays would receive additional scrutiny during the study, captains' drive to implement both the Fuel Load and Efficient Flight behaviors could have encouraged punctuality in their pre-flight procedures. In the case of the former, by narrowing their focus on performing the calculation properly—thereby reducing attention paid to other discretionary professional judgment calls—the time taken to implement this pre-flight behavior may be reduced. With respect to the latter, captains are much less likely to achieve Efficient Flight if they arrive late to the port of arrival due to having missed a landing slot, so on-time takeoff is important to avoid such delays on arrival. That most of the effect of the study comes in preventing short delays (see Table A7) appears to justify the plausibility of the above considerations, as even a savings of a few minutes would, on the margin, trigger a decrease in recorded delays during the study period.

the flight was delayed ('Delays'). The second dependent variable indicates whether the flight was between 1 and 15 minutes delayed ('Short Delays'), and the third indicates whether the flight was delayed more than 15 minutes ('Long Delays').

Empirically speaking, there does not appear to be evidence in favor of an increase in delays during the experimental period. The *Expt* coefficient in Column 1 demonstrates that delays actually decreased by 4.3% ($p < 0.01$) during the experiment. The *Prosocial* treatment induces an additional reduction in delays of 2.6%, which appears to be largely driven by a decrease in short delays (see Column 2). Thus, we can be confident that delays did not increase due to the management strategies implemented in the experiment, and indeed there is some evidence that the number of delays actually declined.

3.3.2 Greenhouse Gas Emissions

Improving fuel efficiency also reduces environmental costs. A fixed emissions factor of 3.15 tons of CO_2 per ton of fuel allows for straightforward calculation of the emissions savings. Deriving from the fuel savings estimates in Section 3.2.1, our study prevented 24,472 tons of CO_2 from entering the atmosphere (excluding savings from persistent behavior change). Given the lack of consensus on the social cost of carbon (SCC), we use two measures of the SCC to derive monetized estimates of the corresponding environmental savings: the 2010 SCC of \$21/ton derived in Greenstone et al. (2013) and the 2020 SCC range of \$40-\$80 from Stiglitz et al. (2017). The resulting environmental savings amount to \$0.51 million using the Greenstone et al. SCC, and \$0.98-\$1.96 million using the Stiglitz et al. SCC range. Thus, in industries where employees' behavior is a determinant of fuel or energy use (e.g., shipping, trucking, aviation, retail, and manufacturing), strategies surrounding behavior change may present a cost-effective means to more closely align private marginal benefits with social marginal costs, particularly in the absence of carbon prices. As such, our study highlights a new way to combat firm-level externalities: target workers rather than the firm as a whole. More work in this area would be welcome.

3.3.3 Captain welfare

Captains' wellbeing is central to airline operations and passenger safety. It is therefore worthwhile asking how captains themselves are affected by the various forms of managerial oversight. We take a first step down this important path by considering captains' job satisfaction. Table A8 presents the intent-to-treat estimates for the effects of each treatment on job satisfaction relative to the control group. The coefficient estimates are positive for all treatments. The largest estimate indicates a positive and significant effect of prosocial incentives, where captains reported a 0.37-point (6.5%) higher job satisfaction rating than captains in the control group ($p < 0.10$). For context, this difference in self-reported job satisfaction is equivalent to that between an employee with poor health compared to one with excellent health (see Clark and Oswald, 1996).

Furthermore, among captains who received personalized targets (i.e., those in the targets and prosocial groups), those who met more targets over the course of the experiment reported greater job satisfaction (see Table A9). More nuanced investigation reveals that performance

on Efficient Taxi drives this result, increasing job satisfaction by 0.12 points (on an eleven-point scale) per monthly target met. In other words, a captain who met all Efficient Taxi targets (out of a possible eight) had a job satisfaction rating 0.96 points higher than a captain who did not meet any Efficient Taxi targets, assuming a linear effect.[39] Thus, airlines may wish to seek means in which to assist captains in reaching fuel efficiency targets for reasons pertaining not only to cost minimization, but also to employee well-being.

3.3.4 Demand for managerial oversight

Finally, in the study debrief survey we assess captains' appetite for continued managerial oversight.[40] Having provided full descriptions of each treatment, we elicited captains' feedback regarding the receipt of similar interventions in the future. Of the 60% of captains who responded to the survey, 79% indicated a desire for the continuation of the management strategies embodied in the study treatments, while only 6% expressed a preference for the pre-study status quo. This qualitative insight supports the notion that captains' welfare as a result of the study appears to have improved.

4. Conclusion

Economists are increasingly confirming that management within firms is a core contributor to the relatively high dispersion of productivity within and between sectors (see e.g., Syverson, 2011; Bloom et al., 2013). This microeconomic facet has earned its place amongst better-understood economy-wide factors—such as the flexibility of capital and labor markets and the regulation of these markets—in explaining sector- and firm-level productivity. In a large-scale field experiment, we lend insights into the causal effects of management practices on productivity in the context of skilled labor in a professional setting in the developed world.

Robust evidence indicates that monitoring of defined behaviors that are directly related to productivity provides gains for the firm, here in the form of reduced input costs. Furthermore, performance targets lead to productivity gains beyond those motivated by monitoring alone. We find that prosocial incentives do not increase productivity beyond targets in this setting, though they do lead to higher job satisfaction, which might have longer term benefits on the extensive margin. In a practical sense, these findings have implications for any corporation aiming to increase labor productivity. For academics, our work highlights the potential of field experimental partnerships to inform productivity models—and provide empirical content to those models—by examining highly-skilled professionals.

Our research speaks to multiple fields within economics. For example, in labor economics, how best to incent workers to motivate effort in the workplace has been a principal topic of

[39] One should take care not to provide a structural interpretation of this result since it is garnered from non-experimental variation.

[40] Each captain received an email on January 29, 2015 with a link to the study debrief survey, and the survey closed three weeks later. A total of 202 captains at least partially completed the survey and 187 completed it, which represents a 60% (56%) response rate. This response rate was achieved after sending each captain up to three emails within four weeks offering incentives up to £105. We find that there are no statistically significant differences in survey taking across experimental conditions (joint F-tests feature $p = 0.69$ and $p = 0.68$ for participation and completion indicators, respectively).

inquiry for decades. The imperfect relationship between employees' effort and productivity renders firms incapable of directly rewarding effort with precision (Miller, 1992; Lazear, 1999; Malcomson, 1999; Prendergast, 1999). A burgeoning field experimental literature on incentives and workplace initiatives attempts to understand the employee-employer relationship and effective means by which employers may increase effort and productivity (see List and Rasul, 2011; Levitt and Neckermann, 2014). Our research aims to advance this literature by identifying the distinct impacts of management practices on defined measures of workplace performance in a developed world labor context.

We augment the management literature in a number of ways. First, we marginally test isolated management practices and demonstrate their differing effects on productivity. This methodology is in contrast to the field experimental management literature at large, which generally tests the effects of a broad array of management practices implemented simultaneously. While these studies (e.g., Bloom et al., 2013; Fryer Jr., 2017; and Bruhn et al., 2018) have brought the importance of management in explaining productivity to the fore, we provide more nuanced understanding of the precise mechanisms underlying the effectiveness of prevalent practices. Broad application of this methodology in future research will lend insights into the complementarity and substitutability of the various components that comprise each management strategy, and in which markets.

Second, we show that these well-studied practices—particularly monitoring and target provision—improve productivity even for high-skilled employees in the developed world, a context that has been largely neglected in the empirical management literature. As the quality of productivity data continues to improve, opportunities to investigate management practices rigorously in a wider variety of contexts will allow researchers to shed further light on means by which to optimize the productivity of skilled labor.

Third, our study highlights the potential for additional benefits of such targeted management initiatives in terms of employee welfare and greenhouse gas emissions. With respect to the former, we find that injecting prosocial elements into employees' incentive structures boosts reported job satisfaction relative to a control group, providing an inexpensive opportunity for employers to improve employee welfare. With regard to the latter, previous research explores how the principal-agent model can be applied to motivate the adoption of energy-efficient technologies in the residential sector (Gillingham and Palmer, 2014); research of this kind remains to be conducted in commercial sectors despite existing evidence of a correlation between management and firm-level emissions (see Bloom et al., 2010; Martin et al., 2012). We provide a well-identified confirmation that the principal-agent problem exists in a relevant commercial setting, and that resulting inefficiencies extend beyond industry profits to the welfare of the worker and of the global environment. The existence of principal-agent problems in polluting industries additionally suggests that implementation of an otherwise optimal Pigouvian carbon tax will result in suboptimal abatement levels.

Overall, our hope is that the research will inspire economists to consider means by which to rigorously and ambitiously pursue knowledge of the drivers of productivity in contexts of considerable relevance and applicability to policy makers and business practitioners in decades to come. Understanding mechanisms through careful partnership of models and field experimentation represents a unique path forward that holds much promise.

References

Akerlof, G. A. and R. E. Kranton (2005). Identity and the economics of organizations. *Journal of Economic perspectives 19*(1), 9–32.

Al-Ubaydli, O. and J. A. List (2015). Do natural field experiments afford researchers more or less control than laboratory experiments? *American Economic Review: Papers & Proceedings 105*(5), 462–66.

Alchian, A. A. and H. Demsetz (1972). Production, information costs, and economic organization. *American Economic Review 62*(5), 777–795.

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics 130*(3), 1117–1165.

Anik, L., L. B. Aknin, M. I. Norton, E. W. Dunn, and J. Quoidbach (2013). Prosocial bonuses increase employee satisfaction and team performance. *PloS one 8*(9), e75509.

Balakrishnan, H., I. Deonandan, and I. Simaiakis (2011). Opportunities for reducing surface emissions through airport surface movement optimization. Technical report.

Bandiera, O., I. Barankay, and I. Rasul (2007). Incentives for managers and inequality among workers: Evidence from a firm-level experiment. *Quarterly Journal of Economics 122*(2), 729–773.

Bandiera, O., I. Barankay, and I. Rasul (2009). Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica 77*(4), 1047–1094.

Bellemare, C. and B. Shearer (2009). Gift giving and worker productivity: Evidence from a firm-level experiment. *Games and Economic Behavior 67*(1), 233–244.

Bénabou, R. and J. Tirole (2010). Individual and corporate social responsibility. *Economica 77*(305), 1–19.

Bender, S., N. Bloom, D. Card, J. Van Reenen, and S. Wolter (2018). Management practices, workforce selection, and productivity. *Journal of Labor Economics 36*(S1), S371–S409.

Bloom, N., E. Brynjolfsson, L. Foster, R. Jarmin, M. Patnaik, I. Saporta-Eksten, and J. Van Reenen (2018). What drives differences in management practices? Technical report, Working Paper, Stanford University.

Bloom, N., E. Brynjolfsson, L. Foster, R. S. Jarmin, M. Patnaik, I. Saporta-Eksten, and J. Van Reenen (2017). What drives differences in management? Technical report, National Bureau of Economic Research.

Bloom, N., B. Eifert, A. Mahajan, D. McKenzie, and J. Roberts (2013). Does management matter? Evidence from India. *The Quarterly Journal of Economics 128*(1), 1–51.

Bloom, N., C. Genakos, R. Martin, and R. Sadun (2010). Modern management: Good for the environment or just hot air? *Economic Journal 120*(544), 551–572.

Bloom, N., R. Lemos, R. Sadun, and J. Van Reenen (2015). Does management matter in schools? *The Economic Journal 125*(584), 647–674.

Bloom, N., R. Sadun, and J. Van Reenen (2016). Management as a technology? Technical report, National Bureau of Economic Research.

Bloom, N. and J. Van Reenen (2007). Measuring and explaining management practices across firms and countries. *Quarterly Journal of Economics 122*(4), 1351–1408.

Bloom, N. and J. Van Reenen (2011). Human resource management and productivity. *Handbook of labor economics 4*, 1697–1767.

Boly, A. (2011). On the incentive effects of monitoring: Evidence from the lab and the field. *Experimental Economics 14*(2), 241–253.

Bruhn, M., D. Karlan, and A. Schoar (2018). The impact of consulting services on small and medium enterprises: Evidence from a randomized trial in mexico. *Journal of Political Economy 126*(2), 635–687.

Charness, G., R. Cobo-Reyes, and Á. Sánchez (2016). The effect of charitable giving on workers' performance: Experimental evidence. *Journal of Economic Behavior & Organization 131*, 61–74.

Clark, A. E. and A. J. Oswald (1996). Satisfaction and comparison income. *Journal of Public Economics 61*(3), 359–381.

Cohn, A., E. Fehr, and L. Goette (2015). Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science 61*(8), 1777–1794.

Elfenbein, D. W., R. Fisman, and B. McManus (2012). Charity as a substitute for reputation: Evidence from an online marketplace. *The Review of Economic Studies 79*(4), 1441–1468.

Ellingsen, T. and M. Johannesson (2008). Pride and prejudice: The human side of incentive theory. *American economic review 98*(3), 990–1008.

Englmaier, F. and S. G. Leider (2012). Managerial payoff and gift exchange in the field. CESifo Working Paper: Behavioral Economics 3707.

Falk, A. and M. Kosfeld (2006). The hidden costs of control. *American Economic Review 96*(5), 1611–1630.

Federal Aviation Administration (2015). Aviation emissions, impacts and mitigation: A primer.

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review 98*(1), 394–425.

Fryer Jr., R. G. (2014). Injecting charter school best practices into traditional public schools: Evidence from field experiments. *The Quarterly Journal of Economics 129*(3), 1355–1407.

Fryer Jr., R. G. (2017). Management and student achievement: Evidence from a randomized field experiment. Technical report, National Bureau of Economic Research.

Gillingham, K. and K. Palmer (2014). Bridging the energy efficiency gap: Policy insights from economic theory and empirical evidence. *Review of Environmental Economics and Policy 8*(1), 18–38.

Giorcelli, M. (2016). The long-term effects of management and technology transfer: Evidence from the us productivity program. *Stanford Institute for Economic Policy Research, Discussion Paper* (16-010).

Gneezy, U. and J. A. List (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica 74*(5), 1365–1384.

Greening, D. W. and D. B. Turban (2000). Corporate social performance as a competitive advantage in attracting a quality workforce. *Business & Society 39*(3), 254–280.

Greenstone, M., E. Kopits, and A. Wolverton (2013). Developing a social cost of carbon for us regulatory analysis: A methodology and interpretation. *Review of Environmental Economics and Policy 7*(1), 23–46.

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic Literature*, 1009–1055.

Hedblom, D., B. R. Hickman, and J. A. List (2016). Toward an understanding of corporate social responsibility: Theory and field experimental evidence.

Hennig-Schmidt, H., A. Sadrieh, and B. Rockenbach (2010). In search of workers' real effort reciprocity: A field and a laboratory experiment. *Journal of the European Economic Association 8*(4), 817–837.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, 74–91.

Holmström, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 24–52.

Hubbard, T. N. (2000). The demand for monitoring technologies: The case of trucking. *Quarterly Journal of Economics 115*(2), 533–560.

Hubbard, T. N. (2003). Information, decisions, and productivity: On-board computers and capacity utilization in trucking. *American Economic Review 93*(4), 1328–1353.

Ichniowski, C., K. Shaw, and G. Prennushi (1997). The effects of human resource management practices on productivity: A study of steel finishing lines. *American Economic Review 87*(2), 768–809.

Imas, A. (2014). Working for the "warm glow": On the benefits and limits of prosocial incentives. *Journal of Public Economics 114*, 14–18.

Kube, S., M. A. Maréchal, and C. Puppe (2012). The currency of reciprocity: Gift exchange in the workplace. *American Economic Review*, 1644–1662.

Lazear, E. P. (1999). Personnel economics: Past lessons and future directions. *Journal of Labor Economics 17*(2), 199–236.

Lazear, E. P. (2000). Performance pay and productivity. *The American Economic Review 90*(5), 1346–1361.

Levitt, S. D. and S. Neckermann (2014). What field experiments have and have not taught us about managing workers. *Oxford Review of Economic Policy 30*(4), 639–657.

List, J. A. and I. Rasul (2011). *Field Experiments in Labor Economics*, Volume 4 of *Handbook of Labor Economics*, Chapter 2, pp. 103–228. Elsevier.

Locke, E. A. and G. P. Latham (2006). New directions in goal-setting theory. *Current Directions in Psychological Science 15*(5), 265–268.

Malcomson, J. M. (1999). Individual employment contracts. In O. C. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*. Amsterdam: North Holland. Vol. III.

Martin, R., M. Muûls, L. B. de Preux, and U. J. Wagner (2012). Anatomy of a paradox: Management practices, organizational structure and energy efficiency. *Journal of Environmental Economics and Management 63*(2), 208–223.

McKenzie, D. and C. Woodruff (2017). Business practices in small firms in developing countries. *Management Science 63*(9), 2967–2981.

Miller, G. J. (1992). *Managerial Dilemmas: The Political Economy of Hierarchy*. Cambridge University Press.

Nagin, D. S., J. B. Rebitzer, S. Sanders, and L. J. Taylor (2002). Monitoring, motivation, and management: The determinants of opportunistic behavior in a field experiment. *American Economic Review 92*(4), 850–873.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Pierce, J. L., D. Snow, and A. McAfee (2015). Cleaning house: The impact of information technology on employee corruption and performance. *Management Science 61*(10), 2299–2319.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature 37*(1), 7–63.

PricewaterhouseCoopers (2016). Redefining business success in a changing world: Ceo survey. Report, PricewaterhouseCoopers, London.

Ryerson, M. S., M. Hansen, L. Hao, and M. Seelhorst (2015). Landing on empty: Estimating the benefits from reducing fuel uplift in us civil aviation. *Environmental Research Letters 10*(9), 094002.

Shaw, K. (2009). Insider econometrics: A roadmap with stops along the way. *Labour Economics 16*(6), 607–617.

Stiglitz, J. E. (1975). Incentives, risk, and information: Notes towards a theory of hierarchy. *The Bell Journal of Economics 6*(2), 552–579.

Stiglitz, J. E., N. Stern, M. Duan, O. Edenhofer, G. Giraud, G. Heal, E. La Rovere, A. Morris, E. Moyer, M. Pangestu, et al. (2017). Report of the high-level commission on carbon prices. *Carbon Pricing Leadership Coalition 29*.

Syverson, C. (2011). What determines productivity? *Journal of Economic literature 49*(2), 326–365.

Tonin, M. and M. Vlassopoulos (2010). Disentangling the sources of pro-socially motivated effort: A field experiment. *Journal of Public Economics 94*(11), 1086–1092.

Tonin, M. and M. Vlassopoulos (2014). Corporate philanthropy and productivity: Evidence from an online real effort experiment. *Management Science 61*(8), 1795–1811.

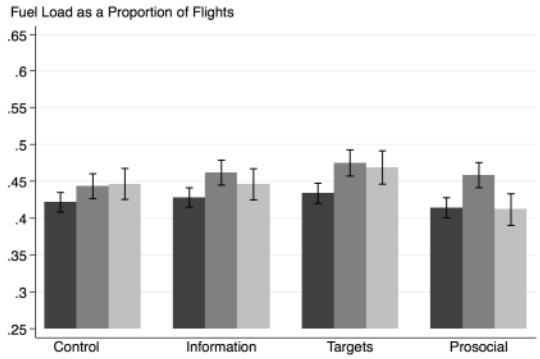
Tsai, T. C., A. K. Jha, A. A. Gawande, R. S. Huckman, N. Bloom, and R. Sadun (2015). Hospital board and management practices are strongly related to hospital performance on clinical quality metrics. *Health Affairs 34*(8), 1304–1311.

Turban, D. B. and D. W. Greening (1997). Corporate social performance and organizational attractiveness to prospective employees. *Academy of management journal 40*(3), 658–672.

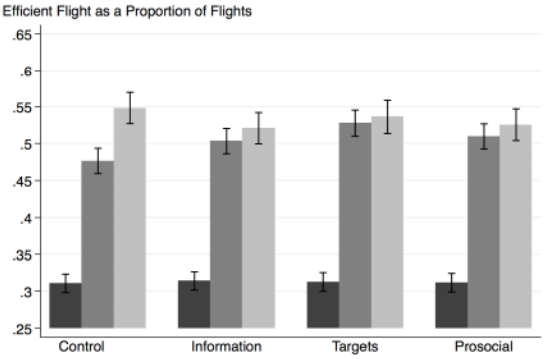
Wu, A. C., D. Donnelly-McLay, M. G. Weisskopf, E. McNeely, T. S. Betancourt, and J. G. Allen (2016). *Environmental health 15*(1), 121.

Figure 1
Behavioral Implementation by Study Group Before, During, and After the Experiment

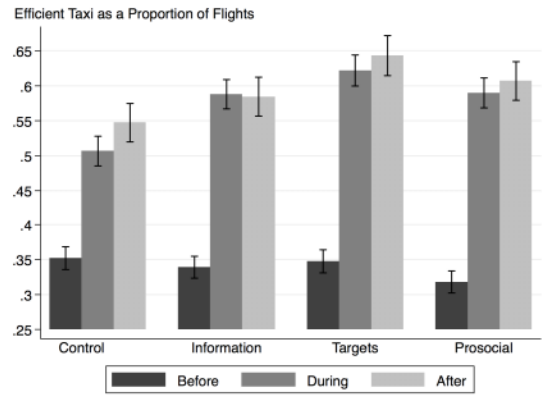
(a) Fuel Load



(b) Efficient Flight



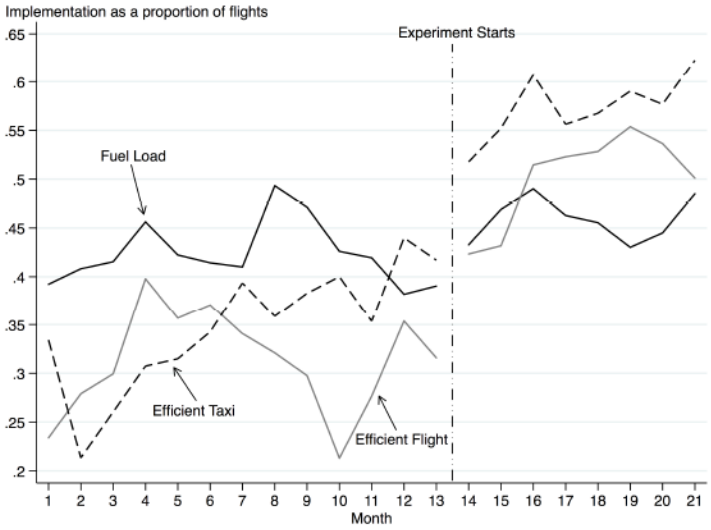
(c) Efficient Taxi



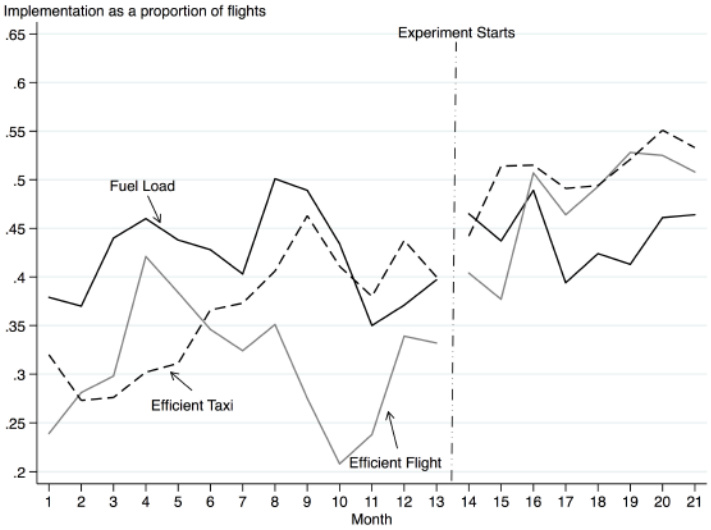
Notes: The y-axis for each of the above graphs represents the proportion of flights for which a fuel-related behavior has been implemented before (dark gray), during (medium gray), and after (light gray) the experiment for each experimental condition in our study (x-axis). These behaviors are averaged at the study group level for the months in our dataset preceding (January 2013-January 2014), during (February 2014-September 2014), and following (October 2014-March 2015) the study period. The error bars represent standard errors.

Figure 2
Behavioral Implementation Over Time

(a) Pooled



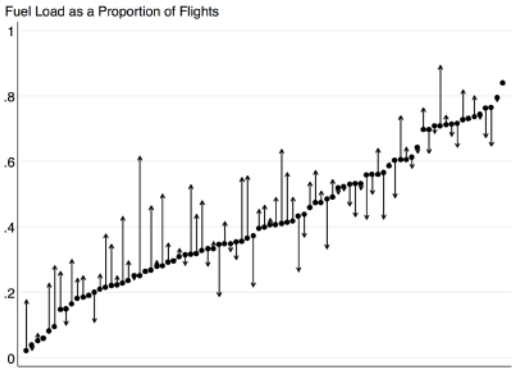
(b) Monitoring Group Only



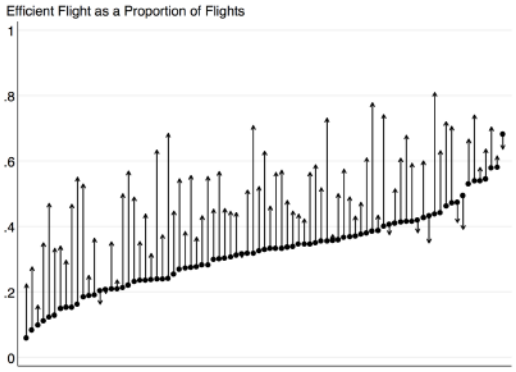
Notes: The y-axis represents the proportion of flights for which each fuel-related behavior has been implemented, while the x-axis indicates the corresponding month in our dataset, where January 2013 is month 1 and October 2014 is month 21, and the experiment took place during months 14 to 21.

Figure 3
Within-Subject Implementation Changes in Control Group after Monitoring is Announced

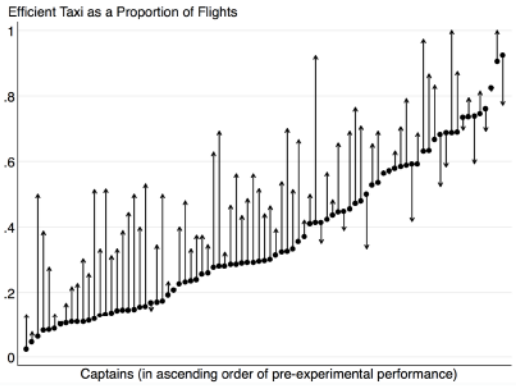
(a) Fuel Load



(b) Efficient Flight



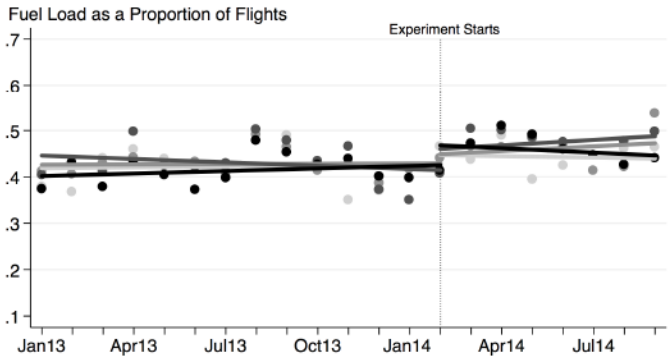
(c) Efficient Taxi



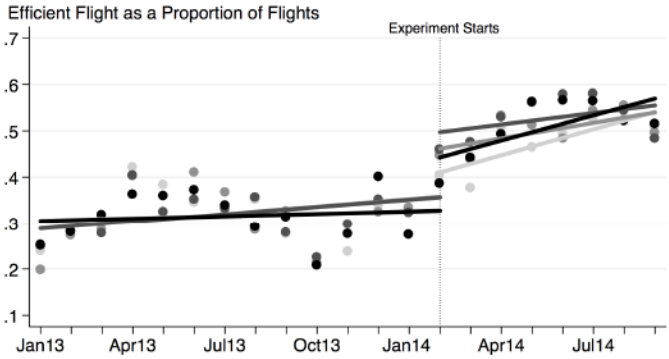
Notes: The data points in the graph represent the proportion of flights for which each captain in the control group implemented the fuel-related behaviors on average before the experiment (January 2013 - January 2014), in ascending order of pre-experimental performance. The vertical arrows represent the same proportion during the experimental period (February 2014 - September 2014). An upward arrow indicates an improvement in implementation (as a proportion of total flights) of the behavior, while a downward arrow indicates a decline in implementation.

Figure 4
Linear Trend of Behavioral Implementation Over Time, by Study Group

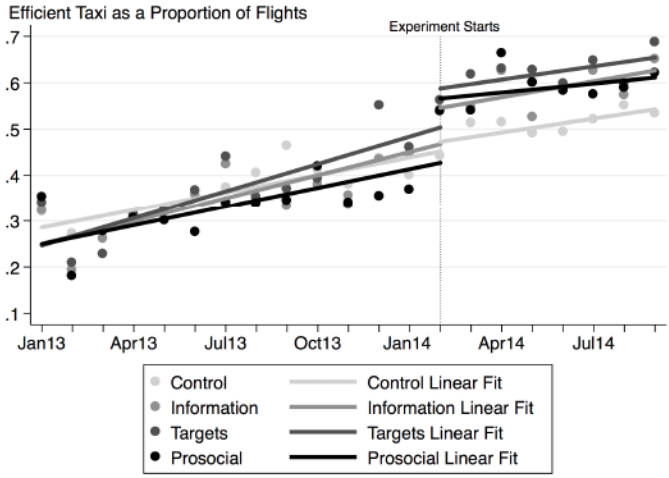
(a) Fuel Load



(b) Efficient Flight



(c) Efficient Taxi



Notes: The y-axis represents the proportion of flights for which a fuel-related behavior has been implemented. The x-axis represents time in months, and the dashed line indicates the start of the experiment. We provide a linear fit of implementation for each of the four experimental groups before and after the start of the experiment, represented by the solid lines.

Table 1
Treatment Group Design

	Monitoring	Information	Targets	Prosocial
Control	✓			
Treatment Group 1	✓	✓		
Treatment Group 2	✓	✓	✓	
Treatment Group 3	✓	✓	✓	✓

Table 2
Balance on Captain and Flight Characteristics

	C: Control	T1: Information	Test of Equality C=T1	T2: Targets	Test of Equality: C=T2	Test of Equality: T1=T2	T3: Prosocial	Test of Equality: C=T3	Test of Equality: T1=T3	Test of Equality: T2=T3
Captain characteristics:										
Seniority	177.93	157.16	p=0.161	174.56	p=0.825	p=0.263	171.87	p=0.682	p=0.327	p=0.863
	(94.68)	(97.38)		(102.00)			(97.17)			
Age	52.23	51.93	p=0.707	51.20	p=0.232	p=0.387	52.31	p=0.926	p=0.633	p=0.193
	(5.34)	(5.10)		(5.73)			(5.15)			
Trainer	0.165	0.188	p=0.687	0.185	p=0.728	p=0.960	0.202	p=0.527	p=0.817	p=0.780
	(0.373)	(0.393)		(0.391)			(0.404)			
Trusted	0.035	0.047	p=0.700	0.025	p=0.690	p=0.440	0.024	p=0.660	p=0.414	p=0.971
Pilot	(0.186)	(0.213)		(0.156)			(0.153)			
Flight characteristics:										
Plan Ramp	76,750	78,559	p=0.440	76,666	p=0.971	p=0.764	76,042	p=0.442	p=0.294	p=0.793
	(14,993)	(15,467)		(14,815)			(15,587)			
Actual Fuel	67,851	69,497	p=0.448	67,763	p=0.967	p=0.770	67,216	p=0.426	p=0.302	p=0.802
	(13,861)	(14,327)		(13,629)			(14,326)			
Engines	3.439	3.483	p=0.648	3.419	p=0.840	p=0.515	3.392	p=0.633	p=0.354	p=0.786
	(0.629)	(0.615)		(0.640)			(0.658)			
Flights/Month	5.182	5.150	p=0.877	5.305	p=0.571	p=0.465	5.261	p=0.706	p=0.586	p=0.837
	(1.372)	(1.310)		(1.406)			(1.328)			
Fuel Load	0.417	0.422	p=0.866	0.424	p=0.831	p=0.769	0.408	p=0.956	p=0.616	p=0.589
	(0.208)	(0.175)		(0.180)			(0.185)			
Eff Flight	0.322	0.322	p=0.979	0.327	p=0.778	p=0.835	0.326	p=0.789	p=0.849	p=0.942
	(0.124)	(0.114)		(0.130)			(0.130)			
Eff Taxi	0.365	0.359	p=0.874	0.367	p=0.947	p=0.460	0.339	p=0.821	p=0.561	p=0.418
	(0.230)	(0.229)		(0.222)			(0.226)			
Sample	n=85	n=85		n=81			n=84			

Notes: The table reports means and standard deviations (in parentheses) for captains in the four experimental conditions in the pre-experimental data (January 2013-January 2014), in addition to tests of equality for each pair of groups (t-test for continuous variables, χ^2 test for indicator variables). *Seniority* and *age* are continuous variables, while *trainer* and *trusted pilot* are indicator variables. *Seniority* captures the captain's ranking amongst VAA captains. *Age* is the captain's age in years (in 2014). *Trainer* captures whether the captain trains other captains in the latest flight techniques, and *trusted pilot* indicates whether the captain was included in pre-study focus groups. *Plan Ramp* measures the amount of fuel anticipated for the entire flight (including taxi-out and taxi-in)—which therefore acts as a proxy for distance flown—and *Actual Fuel* is the actual amount of per-flight fuel realized. *Engines* is the average number of engines on aircraft flown. *Flights/Month* is the average number of flights a captain flew in a given month in the thirteen months leading up to the study. *Fuel Load*, *Eff Flight*, and *Eff Taxi* represent the proportion of each captain's flights on which each of the three fuel-efficient behaviors targeted by the study were met in the pre-experimental period.

Table 3
Summary Statistics: Average Attainment of Fuel Load, Efficient Flight, and Efficient Taxi in all Time Periods

	Control	Treatment 1: Information	Treatment 2: Targets	Treatment 3: Prosocial	All Captains
Fuel Load					
Before Experiment	0.421 (0.494) 5,258 obs	0.428 (0.495) 5,429 obs	0.434 (0.496) 5,070 obs	0.414 (0.493) 5,140 obs	0.424 (0.494) 20,897 obs
During Experiment	0.443 (0.497) 3,321 obs	0.462 (0.499) 3,330 obs	0.475 (0.499) 3,016 obs	0.458 (0.498) 3,258 obs	0.459 (0.498) 12,925 obs
After Experiment	0.446 (0.497) 2,140 obs	0.446 (0.497) 2,120 obs	0.469 (0.499) 1,867 obs	0.412 (0.492) 2,063 obs	0.442 (0.497) 8,190 obs
Efficient Flight					
Before Experiment	0.311 (0.463) 5,258 obs	0.314 (0.464) 5,429 obs	0.313 (0.464) 5,070 obs	0.312 (0.463) 5,140 obs	0.312 (0.463) 20,897 obs
During Experiment	0.476 (0.500) 3,321 obs	0.503 (0.500) 3,330 obs	0.528 (0.499) 3,016 obs	0.510 (0.499) 3,258 obs	0.504 (0.500) 12,925 obs
After Experiment	0.548 (0.498) 2,140 obs	0.521 (0.500) 2,120 obs	0.536 (0.499) 1,867 obs	0.525 (0.499) 2,063 obs	0.533 (0.499) 8,190 obs
Efficient Taxi					
Before Experiment	0.352 (0.478) 3,380 obs	0.339 (0.473) 3,596 obs	0.348 (0.476) 3,260 obs	0.318 (0.466) 3,341 obs	0.339 (0.473) 13,577 obs
During Experiment	0.507 (0.500) 2,117 obs	0.588 (0.492) 2,109 obs	0.622 (0.485) 1,864 obs	0.590 (0.492) 2,014 obs	0.575 (0.494) 8,104 obs
After Experiment	0.547 (0.498) 1,277 obs	0.585 (0.493) 1,201 obs	0.643 (0.479) 1,090 obs	0.607 (0.489) 1,218 obs	0.594 (0.489) 4,786 obs

Notes: The table reports the proportion of flights for which captains in a given group performed each of the three selected behaviors. Due to random memory errors, Efficient Taxi data is unavailable for 37.0% of flights in our dataset. This missing data is in no way systematic and therefore does not bias the results, though it moderately reduces the power of the Efficient Taxi estimates in the subsequent analysis. Standard deviations are reported in parentheses, which is followed by the total number of observations (flights) from which the summary statistics are calculated.

Table 4
Treatment Effect Identification using Difference-in-Difference Regression

	(1) Fuel Load	(2) Eff Flight	(3) Eff Taxi	(4) Fuel Load	(5) Eff Flight	(6) Eff Taxi
Expt	0.033**	0.132***	0.038*	0.033**	0.132***	0.038**
	(0.014)	(0.014)	(0.020)	(0.013)	(0.013)	(0.016)
Expt · Information	0.007	0.017	0.079***	0.007	0.017	0.079***
	(0.017)	(0.016)	(0.025)	(0.015)	(0.014)	(0.017)
Expt · Targets	0.022	0.037**	0.096***	0.022	0.037**	0.096***
	(0.018)	(0.018)	(0.026)	(0.015)	(0.015)	(0.018)
Expt · Prosocial	0.025	0.047***	0.088***	0.025*	0.047***	0.088***
	(0.016)	(0.017)	(0.026)	(0.015)	(0.014)	(0.018)
Observations	33,822	33,822	21,681	33,822	33,822	21,681
# of Captains	335	335	335	335	335	335
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Standard Errors:						
Clustered	Yes	Yes	Yes			
Newey-West				Yes	Yes	Yes

Notes: The table shows the results of a panel difference-in-difference regression specification with captain fixed effects and both clustered and Newey-West standard errors (lag=1), controlling for linear trends in the data. The regressions compare pre-experiment behavior (January 2013-January 2014) to behavior during the experiment ("Expt"; February 2014-September 2014). The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior is performed, and since predicted values are not constrained between 0 and 1, we do not report a constant and instead focus on treatment effects. As such, the coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest is successfully performed. Robust errors are clustered at the captain level. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed an annual training. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

Table 5
Persistence: Treatment Effect Identification Post Experiment

	Pre- vs. Post-experiment			During vs. Post-experiment		
	(1)	(2)	(3)	(4)	(5)	(6)
	Fuel Load	Eff Flight	Eff Taxi	Fuel Load	Eff Flight	Eff Taxi
Post	0.043**	0.239***	-0.009	0.013	0.007	-0.019
	(0.021)	(0.022)	(0.030)	(0.018)	(0.019)	(0.025)
Post · Information	-0.004	-0.038*	0.034	-0.016	-0.046**	-0.035
	(0.020)	(0.023)	(0.032)	(0.020)	(0.022)	(0.029)
Post · Targets	0.010	-0.030	0.078**	-0.007	-0.063***	-0.032
	(0.020)	(0.025)	(0.030)	(0.019)	(0.023)	(0.027)
Post · Prosocial	-0.030	-0.030	0.062**	-0.047**	-0.052**	-0.021
	(0.021)	(0.023)	(0.027)	(0.021)	(0.022)	(0.028)
Observations	29,087	29,087	18,363	21,115	21,115	12,890
# of Captains	335	335	335	335	335	335
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table shows the results of two difference-in-difference regression specifications with captain fixed effects comparing pre-experimental behavior (January 2013-January 2014) to post-experimental behavior ("Post": October 2014-March 2015). The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior is performed, and since predicted values are not constrained between 0 and 1, we do not report a constant and instead focus on treatment effects. As such, the coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest is successfully performed. We provide conventional robust standard errors clustered at the captain level. Total flight observations are provided. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed an annual training. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

Table 6
Data-Supported Estimates of Overall Fuel Savings (in tons)

	(1) Flight Plan Savings	(2) Fuel Load	(3) Efficient Flight	(4) Efficient Taxi
Monitoring	1,647.79***	586.70***	850.43***	56.52***
	(207.54)	(97.71)	(140.27)	(16.87)
Information	2,152.39***	494.45***	932.59***	69.11***
	(215.51)	(95.56)	(121.46)	(15.15)
Targets	1,994.37***	574.66***	1089.04***	66.05***
	(194.10)	(85.24)	(110.89)	(16.15)
Prosocial	1,973.99***	680.33***	1074.74***	38.29**
	(206.15)	(96.29)	(119.55)	(15.03)
Fuel Savings from Behavior Change	-	2,336.14	3,946.80	229.96
Total Fuel Savings	7,768.54		6,512.90	

Notes: The table presents estimates of total fuel savings by treatment group. Savings are based on regression coefficients from a difference-in-difference specification with captain fixed effects and Newey-West standard errors (lag=1) comparing pre-experimental behavior (January 2013-January 2014) to behavior during the experiment (February 2014-September 2014). The dependent variable in column (1) is the deviation between actual fuel used and predicted fuel use in the flight plan. The dependent variable in columns (2)-(4) is the deviation from ideal fuel usage in each of the three flight periods as described in the text. We calculate fuel savings with an intent-to-treat approach where the regression coefficient of each group (i.e., the group's average treatment effect) and the average monitoring effect (i.e., the coefficient of the experimental-period indicator) are multiplied by the number of flights in each group (3,321; 3,330; 3,016; and 3,258, respectively). In other words, we assume that the monitoring effect is proportional to the number of flights. The per-flight fuel savings estimates corresponding to column (1) for the control, information, targets, and prosocial incentives groups—controlling for a linear time trend—are (respectively): 496.17 ($p<0.01$), 150.19 ($p<0.05$), 165.09 ($p<0.05$), and 109.72 ($p=0.10$). Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed an annual training. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

ONLINE APPENDIX

The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains

Greer Gosnell, John List, Robert Metcalfe

December 17, 2018

Contents

A Theoretical Model

A.1 Model Setup

We consider a static choice problem that determines a captain's chosen effort on the job in a certain period. In our model, we assume that captains, who have vast flying experience, are at an equilibrium fuel usage with respect to their wealth, experience, effort, and concerns for safety, the environment, and company profitability.[41]

A captain faces the following additively separable utility function:

$$U(w, s, e, f, g) = u(w, e, g) + a \cdot v(d(e) \cdot g, g_0, G_{-i}) + y(s, e, f) - c(e) - s(e) \tag{1}$$

where $u(\cdot)$ is utility from monetary wealth, $v(\cdot)$ is utility from giving to charity (pro-social behavior), $y(\cdot)$ is utility from job performance, $c(\cdot)$ is disutility from exerting effort, and $s(\cdot)$ describes disutility from social pressure. Effort is chosen for all three flight tasks, j, i.e., Fuel Load, Efficient Flight, and Efficient Taxi. Captains observe their effort perfectly. They also receive a noisy signal of fuel usage $f_{it} + \epsilon_{it} = \bar{f}_{it}$. f_{it} describes the estimated fuel usage by captain i for flight t which depends on the chosen effort for the fuel-efficient activities. $\bar{f}_{it}$ is actual fuel use, observed *ex post* by the airline, which also includes a random component.[42] Furthermore, each captain has an ideal fuel usage f_I, which is based on her own experience and environmental and firm profit preferences. By revealed preference the equilibrium pre-study fuel usage is $f_I = \bar{f}$.

Experimental treatments in this study alter three model parameters. First, receiving information on fuel use, $i = 1$(information), removes the noisiness of the fuel signal, i.e., $f_{it} + (1 - i)\epsilon_{it} = \bar{f}_{it}$. Second, provision of a target, $r = 1$(target), changes the captain's ideal fuel usage, f_I, because the employer exogenously imposes a target level. Then, $f_I = f_T$ if $r = 1$ where f_T reflects the signaled optimal usage from the point of view of the airline. Third, in the pro-social behavior treatment a donation, g, is made by the airline in the name of the captain. This donation is conditional on meeting the target which has a probability of $d(e)$ in this treatment.[43] In all other treatments, reaching particular fuel use levels does not lead to donations, i.e., $d(e) = 0$. Parameters and elements of the utility function are explained in more detail below.

(Dis)utility from social pressure. In the spirit of DellaVigna et al. (2012, "DLM" hereafter) and Bénabou and Tirole (2006), we assume that captains are either affected by social pressure due to their actions being observed or exhibit some sort of social signaling in which

[41] In a MIT survey, commercial airline captains expressed a concern over fuel usage and fuel cost, both for environmental reasons and company profitability. To become an airline captain requires many years of training and experience within an airline; if a captain loses her job with one airline and seeks employment in another, she loses her prior seniority and must work for many years to reinstate it. Thus, for the sake of their own job security, captains care about minimizing fuel costs.

[42] Due to the vast experience of captains, we assume $E(\epsilon_{it}) = 0$, i.e., captains predict fuel usage correctly, on average.

[43] Captains can directly influence the probability with their effort. That is, captains can be certain that they do not meet a target if they put in little effort, and they can be certain that they have achieved the target if they put in sufficient effort.

they want to appear to be good employees. In this framework, captains are aware of an optimal, social effort level, e^{social}. Because exerting effort is costly to the captain and because her actions are imperfectly observed with probability $\pi^{\text{observed}} < 1$, generally $e > e^{\text{social}}$.[44] In this study, captains in both the control group and treatment group are made aware that their actions are monitored and data on their effort are used for an internal and academic study. Consequently, we expect the probability of detection of deviations from the social effort level to increase for all participants in the study relative to the pre-study period, i.e., $\pi^{\text{observed}}_{\text{study}} > \pi^{\text{observed}}_{\text{pre}}$. We parameterize social pressure as follows:

$$s(e) = [\pi^{\text{observed}} \cdot (e^{\text{social}} - e) + (1 - \pi^{\text{observed}}) \cdot 0] \cdot 1(e < e^{\text{social}})$$

Social pressure decreases utility if the chosen effort level is below the socially optimal effort level of the captain. This disutility is increasing in the distance from the optimal effort level and in the probability of these actions being observed by the airline. The second term is an indicator function implying that unobserved deviations do not lead to disutility.. For agents that exert more effort than e^{social}, $s(e)$ simply drops out of their utility function. Consequently, captains can directly impact the level of disutility by exerting more (costly) effort.

Note that $s(e)$ enters the utility of every captain below the social effort level, regardless of treatment assignment. If social pressure is important, even control captains should respond to this increased cost of low effort.[45] Because $s(e)$ is orthogonal to treatment, we omit it in the following discussion and in the derivation of comparative statics.[46]

Utility from wealth. Similar to DLM, for wealth w, charitable giving from the airline g for meeting the target (if applicable), and other charitable giving g_0, u is defined as follows:

$$u(w, e, g) = u(w - g_0(d(e) \cdot g) + \widetilde{a} \cdot d(e) \cdot g)$$

$$\text{where } \widetilde{a} = \begin{cases} 0 & \text{if } a < 0 \\ a & \text{if } 0 \leq a \leq 1 \\ 1 & \text{if } a > 1 \end{cases}$$

Private consumption is an individual's wealth minus the amount given to charity from that person's wealth (i.e., not from this study). However, to ensure that u is continuously differentiable, we need to account for the effect of charitable donations resulting from our treatments on utility from private consumption. To capture this effect, we multiply the individual's expected donation, $d(e) \cdot g$, by a function of a—a parameter capturing preferences for giving—which we call $\widetilde{a}$. As in DLM, the parameter a is non-negative in the case of pure

[44] It is plausible to argue that effort is perfectly observed in the aviation industry with modern technology. However, captains might not expect these data to be analyzed on a regular basis.

[45] Alternatively, we could interpret e^{social} as a level of effort induced by the researcher, leading to experimenter-demand effects. Put differently, captains in the study could think they are expected to increase effort and not doing so imposes utility costs on them.

[46] Social pressure is additively separable from other utility elements in a linear model. Consequently, it does not affect the sign of comparative statics derived below and, if interactions are present, only attenuates treatment effect estimates.

or impure altruism and negative in the case of spite[47], and $\widetilde{a}$ is simply a truncated at 0 and 1.

The reasons for creating such boundaries on the term capturing preferences for giving are twofold. First, an individual with spiteful preferences ($a < 0$) does not get less utility from private consumption when she donates to charity than when she does not donate to charity; therefore, $\widetilde{a}$ is censored from below at 0. Second, an individual with pure or impure altruistic preferences will get additional utility from her private consumption by giving to charity through our treatment because it corresponds to an outward shift in the budget constraint in the dimension of giving to the chosen charity. However, $\widetilde{a}$ is censored from above at 1 because an individual will experience weakly more utility from increases in w than from giving to the chosen charity (i.e., $\frac{\partial u}{\partial w} \geq \frac{\partial u}{\partial g}$). This relation holds since increases in w shift the budget constraint outward in all dimensions—including the charitable giving dimension—so these must be weakly preferred to shifts in only one dimension. This stipulation is important to assume differentiability in u in a standard expected utility framework, as in DLM.

Please note that the amount an individual gives to other charities will be related to the amount that she gives to charity in the context of this study. Captains will smooth their consumption for giving. If a captain normally gives \$100 to charity each year and this year she gives \$10 through the context of the study, we would expect her total giving to be between \$100 and \$110, or $g_0 + g \in [100, 110]$. The realization of the sum depends on the value of a and whether a stems from pure altruism, impure altruism, or spite. We should expect that an individual who has a negative a value does not donate to charity outside of the context of this study since donating to charity decreases that individual's utility.

Utility from charitable giving. The v term is also adapted from DLM and follows the same properties for each type of individual (pure or impure altruistic and spiteful). The main difference between the v term in this study and that in DLM is that in this study, not everyone has the opportunity to donate to charity (i.e., $d(e) > 0$ for only one treatment group). We also assume that v is separable in its parameters, as follows:

$$v(d(e) \cdot g, g_0, G_{-i}) = v_1(d(e) \cdot g, G_{-i}) + v_2(\theta g_0, G_{-i})$$

where θ is the cost of giving through channels other than the study and G_{-i} is total giving by other individuals. In this specification, v_1 represents utility from giving in the study context and v_2 represents utility from giving from one's personal wealth. Note that $v_1(0, G_{-i}) = 0$ since if $d(e) = 0$, then a captain is not able to donate to the charity through the context of the study, so v_1 should not affect the utility function (similar to the spite case). Based on the arguments made above with respect to consumption smoothing, $v|_{d=0} \leq v|_{d=p}$, $0 = v_1|_{d=0} \leq v_1|_{d=p}$, $v_2|_{d=0} \geq v_2|_{d=p}$. That is, utility from giving is at least as high for those captains for whom $d(e) = p$ as it is for those captains for whom $d(e) = 0$, which follows from our assumption that giving in the study context can only decrease giving from one's own wealth or not affect it at all. Finally, since $\frac{\partial p}{\partial e} > 0$, we have $\frac{\partial v}{\partial e} \geq 0$.

[47] As defined in Andreoni (1989, 1990), pure and impure altruism capture two possible motivations for giving. The first stems from a preference solely for provision of the public good, so that an individual's donations are entirely crowded out by donations from other sources. Impure altruism, on the other hand, refers to the phenomenon whereby individuals receive direct utility from the act of giving itself, i.e., through "warm glow". Spite, as defined in DLM, exists when an individual gets disutility from donating to the charity.

In the case of pure altruism, an individual should get the same utility from giving to charity from her personal wealth as from giving to charity through the context of the study, since the benefit to the charity is identical. In this sense, v can be thought to represent the charity's production function. In the case of impure altruism, an individual should also get the same utility from donating to charity through her personal wealth as she does from donating through the context of the study because the amount donated on her behalf is the same. Lastly, in the case of spite, $g_0 = 0$ since giving to charity decreases utility and so those individuals will not give to charity independently of the study. Note, $v(0) = 0$ because if a person does not give to charity in person then her utility from giving to charity in person is 0.

Utility from job performance. Since captains care about fuel efficiency, and since imposing exogenous targets on performance affects a captain's perception of how well she is doing her job, we include a parameter y capturing job performance.[48] We assume y is separable in safety (s) and fuel (f) because changes in fuel as a result of the study do not affect safety levels, as argued in our assumption above. A captain whose performance exceeds her target will achieve higher utility under this parameter than a captain who does not achieve her target. Similarly, a captain will experience less (more) utility the further below (above) the target is her performance. We therefore incorporate job performance into the model as follows:

$$y(s, e, f) = y_1(s) + y_2(e, f) = y_1(s) + y_2(-\bar{f}| - f_I)$$

where

$$y_2(-\bar{f}| - f_I) = y_{2m}(-\bar{f}) + y_{2n}(-\bar{f}| - f_I)$$

and

$$y_{2n}(-\bar{f}| - f_I) = r \cdot \mu(y_{2m}(-\bar{f}) - y_{2m}(-f_I))$$

Here, y_2 is defined as in Köszegi and Rabin (2006, KR hereafter). We denote the components of y_2 "m" and "n" to mirror the notation in KR. As in KR, m represents the "consumption utility" and n represents the "gain-loss utility." These terms are separable across dimensions. Finally, μ is the "universal gain-loss function" and has the associate properties outlined in KR. To be clear, we assume that captains who receive exogenous targets perceive these targets as reference points for their own attainment.

Note that captains get utility from using less fuel $\frac{\partial y_2}{\partial f} \leq 0$ and, conditional on receiving a reference point, get utility (disutility) from performing above (below) the target, which increases with distance from the target according to μ. We assume μ is linear and $\mu(x) = \eta x$ if $x > 0$ and $\mu = \eta\lambda x$ if $x \leq 0$ for $\eta > 0$, $\lambda > 1$, in accordance with theories of loss aversion. Moreover, following naturally from our definition of μ, we assume $y(x) = x$. If a captain

[48] Evidence indicates that influencing job performance positively influences job satisfaction (or utility), whether through increased self-esteem or perceived managerial support for autonomous decision-making (Christen et al., 2006; Pugno and Depedri, 2009).

does not receive a reference point, her utility does not comprise gain-loss utility, so for these individuals $y_2 = y_{2m}$. That is, if $r = 0$, captains do not receive information regarding ideal performance with respect to fuel efficiency, so their job performance parameter depends solely on fuel consumption.[49]

Additionally, based on industry standards and emphasis on safety—as well as the design of the treatments—we assume that captains' job performance utility from flying safely is constant across treatments, therefore:

$$\frac{\partial y}{\partial s} = S \geq 0$$

(Dis)utility from effort. Finally, c represents the cost of effort. Importantly, the individual cost functions for each fuel-efficient task are allowed to differ to convey that various tasks have different costs associated with them. The cost structure is a function of the difficulty of the task itself (e.g., it may be easier to turn off one engine after landing than to have an efficient flight for several hours) and resistance due to previous habit formation (e.g., captains who for many years have not properly performed the Zero Weight Fuel calculation may find it difficult or bothersome to begin doing so). Additionally, the costs for each task are separable since the tasks are done independently. Therefore,

$$c(e) = \sum_j c_j(e_j)$$

For a captain to decrease her fuel use, she must also increase her effort, i.e., $\frac{\partial f}{\partial e} < 0$. Note that $c(e)$ is subtracted in the utility equation, so $\frac{\partial U}{\partial c} < 0$, $\frac{\partial c}{\partial e} > 0$. Based on interviews with captains, the cost of effort increases at an increasing rate. Defining the cost of effort as a quadratic function of effort implies that the cost of effort increases with the amount of effort exerted (i.e., $\frac{\partial^2 c}{\partial e^2} > 0$).

A.2 Model Predictions

Captains will choose how much effort to exert based on the treatments (information, targets, prosocial incentives) as in the moral hazard model (see Holmström, 1979). The model is simplified because agents are current employees whose base salaries are not affected by the study. The treatments do affect job satisfaction and charitable giving, however. Different treatments represent different contracts.

We now define $V(-f)$ to be the utility of the firm (the principal) from the perspective of the employee (the agent) as a function of firm costs, i.e., fuel costs. V is highly related to y since an employee's job satisfaction is linked to the well-being of the firm itself. We assume V is independent of treatment status, τ, because the marginal benefit and marginal cost to the firm do not depend directly on treatment, but rather on the amount of fuel used (i.e., for the same level of fuel but two different treatments, V is the same). Additionally, salaries are fixed and donations to charity are paid by an outside donor.

[49] To be clear, given that our reference point is exogenously imposed, one cannot clearly assess whether the individual captain is better off in the targets group than in another group.

We further define $U(e, \tau)$ to be the utility function under treatment τ with effort e and $\bar{U}$ as a captain's outside option.[50] Let $\dot{e}$ be the pre-study amount of effort and $\ddot{e}$ be the chosen effort under τ. Note that the profit-maximizing principal (VAA) wants to design contracts (treatments) that induce the optimal level of effort from the point of view of the principal. In this case, the principal observes both the outcome (fuel usage) and the effort by the agent, but is restricted from making contractual changes that introduce monetary compensation based on effort levels.

Therefore, the problem becomes:

$$\begin{aligned} \max_{e, g_0} \quad & E[V(-f)] \\ \text{s.t.} \quad & E[U(w, s, \ddot{e}, f_I, g, \tau)] \geq \bar{U} \\ \text{and} \quad & \ddot{e} \in \operatorname*{argmax}_{\ddot{e}'} E[U(w, s, \ddot{e}', f_I, g, \tau)] \end{aligned}$$

The first-order condition is $\frac{V'(-f)}{U'(w,s,\ddot{e},f_I,g,\tau)} = \lambda$ and so $U'(w, s, \ddot{e}, f_I, g, \tau) = \lambda \cdot V'(-f)$. Captains choose the effort level that satisfies the marginal conditions.

Proposition 1. *Captains in the control group will change their behavior if they are influenced by social pressure. That is, they will generally increase effort if their effort level is below the social effort level.*

Proof: We argued above that scrutiny due to the intervention is likely to (weakly) increase the probability of detection of a sub-optimal effort level ($\pi^{observed}$) or the perceived socially optimal level of effort (e^{social}), or both. Both effects increase the social cost component of the utility function for captains in all treatment cells, including the control group. Put differently, for a given level of effort $\bar{e} < e^{social}$, the intervention increases the marginal social cost of exerting low effort $\frac{\partial U}{\partial s}|\bar{e}$. Consequently, captains respond to these new marginal conditions and increase their effort if they are below the (perceived) socially optimal level.[51]

Proposition 2. *Information will cause captains to increase or decrease their effort and thereby increase or decrease fuel usage respectively or choose the outside option, depending on the realization of the difference between estimated (f_{it}) and actual ($\bar{f}_{it}$) fuel usage (i.e., the value of the parameter ϵ_{it}).*

Proof: Let the pre-study period be $t = 0$ and the study period be $t = 1$.

Assume in period $t = 0$, $\epsilon_{i0} < 0$, then $f_{i0} > \bar{f}_{i0}$, so that when captains receive information in $t = 1$, they learn that $y_{2m}(-\bar{f}) > E[y_{2m}(-\bar{f})]$. In other words, they were more fuel-efficient in $t = 0$ than they expected to be. Therefore, if they provide the same level of effort in period $t = 1$, they will experience a level of utility greater than their pre-study equilibrium. They pay the same cost of effort but receive more utility from job performance. They will then

[50] Our notation differs slightly from Holmström (1979) since the cost of the action is embedded in the utility function of the agent.

[51] Because of orthogonality to treatment, the condition of being observed simply increases baseline effort. Furthermore, because utility is additively separable, qualitative findings from the subsequent comparative statics analysis are unchanged. If there are interactions between social pressure and the treatments, these interactions just attenuate point estimates because all treatments are designed to increase effort against a now greater baseline.

weakly decrease their chosen level of effort. How much depends on the functional form of the y and c functions and their pre-study effort level. Captains in the information or targets treatments—where wealth and the charities' production functions are independent of effort—will not decrease their effort if y is steeper than c around their chosen values. This scenario is possible since there is a random shock of ϵ_{i0} to their location of $-\bar{f}$ and we are agnostic about the functional form of y. Without the shock, they would not be in equilibrium if y were steeper with respect to effort than c at the chosen level of effort because they could increase effort and pay a slightly higher cost but get much more utility from job performance. They will not choose their outside option since if

$$E[U(w, s, \dot{e}, f_I, g, \tau = \text{“pre-study, no treatment”})] \geq \bar{U},$$

then

$$E[U(w, s, \ddot{e}, f_I, g, \tau = \text{“information”})] \geq \bar{U}.$$

In other words, they can hold y constant and decrease effort and thereby increase U, while $\bar{U}$ is held fixed.

Now assume $\epsilon_{i0} > 0$, then $f_{i0} < \bar{f}_{i0}$ and so when captains receive information, they learn that $y_{2m}(-\bar{f}) < E[y_{2m}(-\bar{f})]$, i.e., they were less fuel-efficient than expected. Therefore, if they provide the same level of effort in period $t = 1$, they will receive below their pre-study equilibrium amount of utility. They pay the same cost of effort but receive less utility from job performance. They will weakly increase their effort if the change in y is more than the change in c, which depends on the functional form of these functions and their pre-study effort level. They will not increase their effort if c is steeper than y for similar reasons described in the previous case. They will choose their outside option if the change in y leads to $E[U(w, s, \ddot{e}, f_I, g, \tau = \text{“information”})] < \bar{U}$, which could occur if increases in effort lead to larger increases in c than in y. Whether or not it occurs also depends on the captains' outside option.

Finally, assume $\epsilon_{i0} = 0$. Then captains are at their equilibrium with $y_{2m}(-\bar{f}) = y_{2m}(-f_I)$ and do not change their effort.

Proposition 3. *Targets set above pre-study use will cause captains to weakly increase their effort or choose their outside option.*[52]

Proof: Since the target is set above pre-study use (i.e., captains are meeting the targets fewer times than is optimal from the perspective of the firm), upon receiving a target, the captains learn $f > f_T$ and get reference-dependent loss utility equal to $y_{2n} < 0$. Therefore, captains are strictly below their equilibrium in effort and strictly above in fuel usage since in the pre-study period $y_{2n} = 0$ from the assumption that $f_I = \bar{f}$.

Captains will not increase their effort if the increased cost of effort is larger than the gain from the associated decrease in fuel usage in the job performance function. Captains will increase their effort if the gain from the associated decrease in fuel usage is more than the cost of effort. This depends on the functional form of these functions, the value of μ, and

[52] All targets were set above the pre-study attainment level, so this is the only case we consider.

the captains' initial values during the pre-study period. Their chosen level of effort comes from the first-order condition with $\tau =$ "receive targets".

Since captains experience a negative utility shock from receiving a target, they will choose the outside option if $E[U(w, s, \ddot{e}, f_I, g, \tau =$ "receive targets"$)] \leq \bar{U}$.

Proposition 4. *Donations made to charity for meeting targets will weakly increase effort if captains' altruism is strictly positive and the donations do not affect their effort otherwise.*

Proof: Let $V_c(d(e), g)$ be the production function of the charity. Note that in the case of pure altruism $V_c = v_1$, as defined in the previous section. $\forall d(e) \cdot g \geq 0$, we have $V_c > 0$ and $V_c = 0$ if and only if $d(e) \cdot g = 0$. Then, captains solve the following optimization problem:

$$\begin{aligned} \max_{e, g_0} \quad & E[V(-f) + \tilde{a} \cdot V_c] \\ \text{s.t.} \quad & E[U(w, s, \ddot{e}, f_I, g, \tau)] \geq \bar{U} \\ \text{and} \quad & \ddot{e} \in \operatorname*{argmax}_{\ddot{e}'} E[w, s, \ddot{e}', f_I, g, \tau)] \end{aligned}$$

with first-order condition $\frac{V'(-f)+\tilde{a}\cdot V_c'}{U'(w,s,\ddot{e},f,g,\tau)} = \lambda$. If a captain has zero altruism, i.e., $\tilde{a} = 0$, then this equation reduces to the original and effort does not increase above the effect described in Proposition 1. If $\tilde{a} > 0$, then the numerator of the first-order condition is weakly larger than the control case. It is strictly larger if $d > 0$. Captains with strictly positive altruism may choose an effort level corresponding to $d = 0$ if the additional cost of increased effort required for meeting the target is more than the gain in utility from donating to charity. The probability of this outcome occurring is decreasing in the level of altruism.

Since λ is a constant, increases in the sum of the production functions of the firm and charity cause increases in effort, $\dot{e} < \ddot{e}$.

Proposition 5. *Captains in the targets and prosocial conditions will choose to increase their effort the most in tasks for which the targets are easiest to meet.*

Proof: Since the firm sets the targets and donations exogenously[53], the utility for meeting a target is constant across tasks. The donation to charity is the same across tasks as exogenously determined, and since the targets are also exogenously determined, the captains believe that the firm values them all equally by revealed preference. If the firm did not value them equally, then it would not offer the same reward. However, the cost function is not constant across tasks for reasons described earlier, which implies that the captains will choose to increase their effort on tasks for which targets are easiest to meet.[54] Within our context, the least effortful behavior to attain is Efficient Taxi, followed by Fuel Load, then Efficient Flight. The determination of this ordering is based on discussions with many airline captains and the trusted pilot group.

[53]Note that the "firm" here refers to both VAA and the academic researchers, who jointly made most decisions with respect to experimental design.

[54]Our theory and interventions are rooted in Holmström's (1979) "Informativeness Principle", which states that any accessible information about an agent's effort should be used in the design and enforcement of optimal contracts. Our interventions are not aimed at the efficient allocation of effort across these tasks—as proposed in Holmström and Milgrom (1991) and Baker (1992)—since we assume our three behaviors are not substitutable (i.e., since they occur during different phases of flight). We acknowledge the possibility that additional fuel-efficient behaviors exist that we do not measure that may be fully or partially neglected due to our treatments.

B Map

Figure 5
Global destinations of VAA

C Examples of Treatment Groups

Figure A1
Treatment Group 1: Information

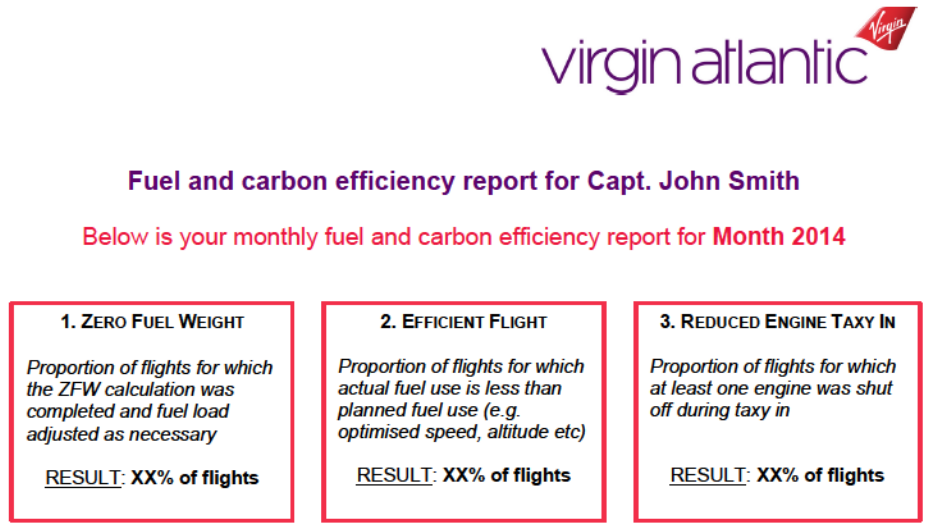


Figure A2
Treatment Group 2: Targets

Fuel and carbon efficiency report for Capt. John Smith

Below is your monthly fuel and carbon efficiency report for **Month 2014**

1. Zero Fuel Weight

Proportion of flights for which the ZFW calculation was completed and fuel load adjusted as necessary

TARGET: **XX% of flights**

RESULT: **XX% of flights**

You ACHIEVED/MISSED your target.

2. Efficient Flight

Proportion of flights for which actual fuel use is less than planned fuel use (e.g. optimised speed, altitude etc)

TARGET: **XX% of flights**

RESULT: **XX% of flights**

You ACHIEVED/MISSED your target.

3. Reduced Engine Taxy In

Proportion of flights for which at least one engine was shut off during taxy in

TARGET: **XX% of flights**

RESULT: **XX% of flights**

You ACHIEVED/MISSED your target.

What Was Your Overall Outcome?

You achieved X of your 3 targets last month.

WELL DONE! We will continue to keep you updated on your monthly performance for the next **X months**, John.

Please continue to fly efficiently next month to achieve your targets.

Please see reverse side for further details of the three behaviours.

Questions? We are here to help! Please email us at project.uoc@fly.virgin.com.

All data gathered during this study will remain anonymous and confidential. Safety remains the absolute and overriding priority. This study will be carried out within Virgin's existing and highly robust safety standards, using our existing fuel procedures and policies. Captains retain full authority, as they always have done in VAA, to make decisions based on their professional judgment and experience.

Figure A3
Treatment Group 3: Prosocial Incentives

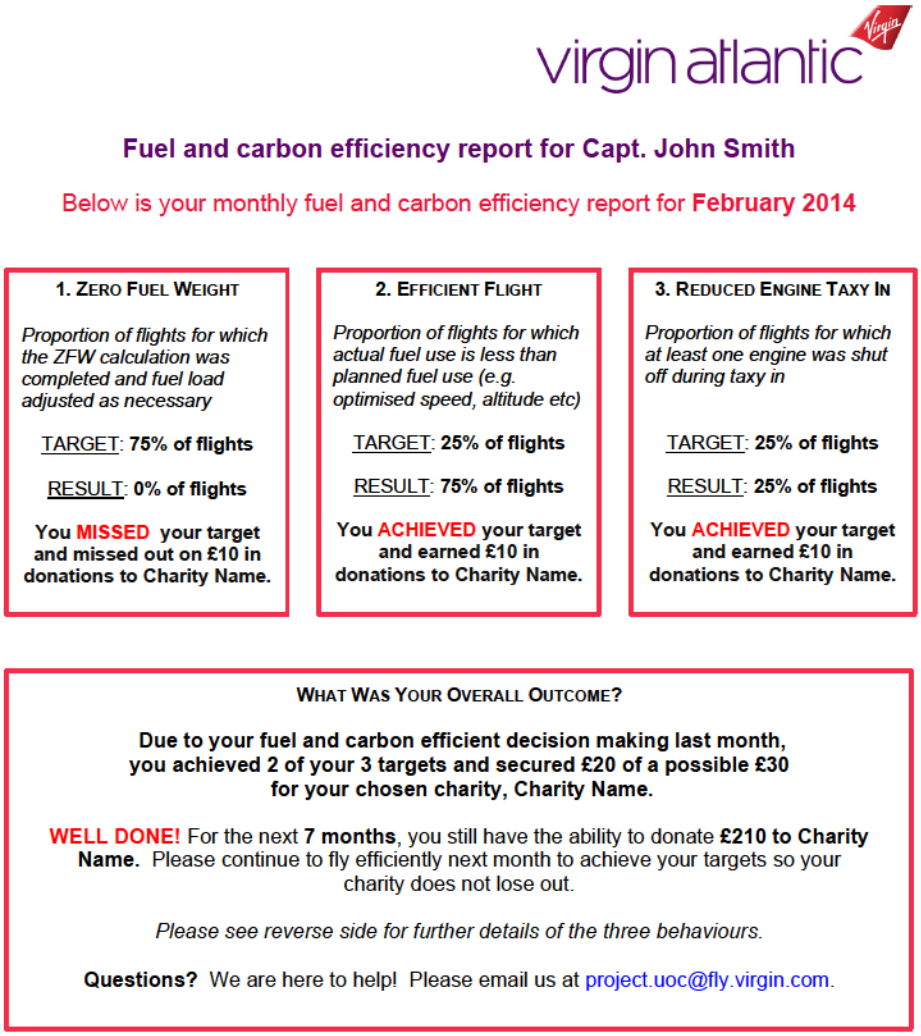


Figure A4
All Treatment Groups: Reverse Side of Report

THE THREE BEHAVIOURS WE ARE MEASURING

Behaviour 1: Zero Fuel Weight Adjustment (ZFW) - Pre Flight

This measure compares Actual Ramp against Plan Ramp adjusted for changes in ZFW. It captures whether a double iteration adjustment has been implemented for ZFW in line with Plan Burn Adjustment and any further amendments to flight plan fuel that have been entered into ACARS. This behaviour has a tolerance of 200kg, which ensures that rounding in the fuel request / loading procedure will not adversely affect the result.

Behaviour 2: Efficient Flight (EF) - During Flight

This measure examines the actual fuel burn per minute compared against the expected fuel burn per minute from OFF to ON (expected fuel burn is Plan Trip adjusted for ZFW). It highlights pilot technique (e.g. optimum settings are realised, optimum levels are sought, speed is optimised, etc.).

Behaviour 3: Reduced Engine Taxy In (RETI) - Post Flight

This measure observes if an engine has been shut down during taxy in. RETI is considered to have taken place if one engine burns less than 70% of the average of other engines during taxy in. If taxy in is shorter than the cool down period required, the flight is omitted, as RETI was not possible.

We hope the above information is beneficial to you. If you require more information about any of the behaviours, please email us at project.uoc@fly.virgin.com.

D Randomization Strategy

Procedure: We conducted the randomization in four steps using panel flight-level data for September through November of 2013.

- **Step 1: Identify eligible population.** All captains who were anticipated to be active during the study period were eligible for the randomization (N=340).
  - Note: Of these captains, two had been previously on long-term sick leave and were anticipated to possibly come back during the study period, two went on long-term sick leave between November 2013 and early January 2014 (prior to sending out study communications), and one retired in December 2013, resulting in a total of 335 captains ultimately taking part in the study.
  - We control for this 'attrition' in our analysis, and results are robust to exclusion of the quads to which these five captains had been allocated.
- **Step 2: Generate quadruplets.** Our data contained information for each of the match variables (i.e. number of engines on the aircraft, number of flights flown by the captain each month, and the three targeted fuel-relevant behaviors). These match variables were selected as they were either determinants of fuel use in the data used for randomization (September-November 2013) or were outcome variables of interest. The 340 captains identified in Step 1 were allocated across 85 quadruplets within which these variables took on as close to identical values as possible.
- **Step 3: Randomize within quadruplets.** Within each quadruplet, one captain was randomly selected for each of the four conditions: Control, Information, Targets, and Prosocial Incentives.
- **Step 4: Check for balance.** Once each captain was assigned to an experimental condition, we checked to make sure that the means (numerical variables) and frequencies (categorical variables) were not statistically significantly different between conditions. For every pairing of study groups, we ran a t-test for each numerical balance variable and a χ^2 test for each of the categorical balance variables.

Match variables:

1. Average flights flown per month (categorical: 0-2, 2-4, 4-6, 6-8, 8-10)
2. Number of engines on aircraft flown (binary: all four-engine or not; 51% of captains had flown only four-engine aircraft)
3. Proportion of flights for which zero fuel weight was conducted and fuel load was adjusted within 200kg of the correct fuel uptake

4. Proportion of flights for which actual fuel burned in flight was less than predicted fuel burn in the flight plan
5. Proportion of flights for which at least one engine was turned off during taxi-in

Balance Check Variables:

- Gender
- Seniority
- Age
- Trainer
- Trusted pilot
- Planned ramp fuel (proxy for flight distance)
- Number of flights
- Above average planned ramp fuel (binary)
- Above average number of engines
- Above average flights per month
- Fuel Load
- Efficient Flight
- Efficient Taxi

Note that we focused on behavior change as opposed to fuel use in the randomization and in our analysis because we targeted captains' behavior in the interventions. We did not directly target captains' fuel use since it is anecdotally very difficult for captains to map their behaviors to their fuel use on a given flight (as experimentally demonstrated in the context of residential energy use; e.g., see Jessoe and Rapson, 2014), and behavioral inputs are more readily accessible than outputs. We therefore performed all power calculations and randomization based on the three behavioral outcome variables, and we provide fuel savings estimates as exploratory analysis to glean insights on resulting fuel savings.

E Additional Tables and Figures

Table A1
T-tests (p-values) of Difference in Pre-Experimental Behavioral Trends

	Fuel Load			**Efficient Flight**			**Efficient Taxi**		
	Monitoring	Info	Targets	Monitoring	Info	Targets	Monitoring	Info	Targets
Info	0.405			0.351			0.417		
Targets	0.089	0.447		0.346	0.913		0.485	0.138	
Prosocial	0.250	0.777	0.634	0.324	0.082	0.066	0.522	0.147	0.938

Notes: For each fuel-related behavior and for each group, we regress the behavior on the time trend and controls in our main regression. The table presents three t-test matrices that provide the p-value for each comparison of trend coefficients derived from the aforementioned regressions.

Table A2
T-tests (p-values) of Difference in Pre-Experimental Fuel Use Trends

	Monitoring	Info	Targets
Info	0.386		
Targets	0.899	0.471	
Prosocial	0.281	0.739	0.346

Notes: For each experimental group, we regress flight-level fuel use on the time trend and controls in our main regression. The table presents a t-test matrix that provides the p-value for each comparison of trend coefficients derived from the aforementioned regressions.

Table A3
Treatment Effect Identification using Difference-in-Difference Regression (Controlling for Attrition)

	(1) Fuel Load	(2) Eff Flight	(3) Eff Taxi
Expt	0.034*	0.132***	0.037*
	(0.014)	(0.013)	(0.017)
Expt · Information	-0.001	0.016	0.079***
	(0.015)	(0.015)	(0.018)
Expt · Targets	0.021	0.037*	0.094***
	(0.015)	(0.015)	(0.019)
Expt · Prosocial	0.029	0.051**	0.093***
	(0.015)	(0.015)	(0.018)
Observations	32,310	32,310	20,747
# of Captains	335	335	335
Controls	Yes	Yes	Yes

Notes: The table shows the results of a panel difference-in-difference regression specification with captain fixed effects and Newey-West standard errors (lag=1), controlling for linear trends in the data and excluding the quadruplets of captains who attrited. The regressions compare pre-experiment behavior (January 2013-January 2014) to behavior during the experiment ("Expt": February 2014-September 2014). The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior is performed, and since predicted values are not constrained between 0 and 1, we do not report a constant and instead focus on treatment effects. As such, the coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest is successfully performed. Robust errors are clustered at the captain level. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed an annual training. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

Table A4
Specification Building: Treatment Effect Identification using Difference-in-Difference Regression

	(1) Fuel Load	(2) Eff Flight	(3) Eff Taxi	(4) Fuel Load	(5) Eff Flight	(6) Eff Taxi	(7) Fuel Load	(8) Eff Flight	(9) Eff Taxi
Expt	0.030***	0.171***	0.153***	0.028**	0.144***	0.038**	0.033**	0.132***	0.038**
	(0.010)	(0.011)	(0.013)	(0.013)	(0.013)	(0.017)	(0.013)	(0.013)	(0.016)
Expt · Information	0.011	0.014	0.069***	0.011	0.013	0.066***	0.007	0.017	0.079***
	(0.015)	(0.015)	(0.018)	(0.015)	(0.015)	(0.018)	(0.015)	(0.014)	(0.017)
Expt · Targets	0.016	0.029*	0.097***	0.016	0.028*	0.095***	0.022	0.037**	0.096***
	(0.015)	(0.015)	(0.019)	(0.015)	(0.016)	(0.019)	(0.015)	(0.015)	(0.018)
Expt · Prosocial	0.018	0.027*	0.080***	0.018	0.026*	0.077***	0.025*	0.047***	0.088***
	(0.015)	(0.015)	(0.018)	(0.015)	(0.015)	(0.018)	(0.015)	(0.014)	(0.018)
Observations	33,822	33,822	21,681	33,822	33,822	21,681	33,822	33,822	21,681
# of Captains	335	335	335	335	335	335	335	335	335
Controls	No	No	No	No	No	No	Yes	Yes	Yes
Time trends	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table shows the results of a panel difference-in-difference regression specification with captain fixed effects and Newey-West standard errors (lag=1). The regressions compare pre-experiment behavior (January 2013-January 2014) to behavior during the experiment ("Expt": February 2014-September 2014). The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior is performed, and since predicted values are not constrained between 0 and 1, we do not report a constant and instead focus on treatment effects. As such, the coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest is successfully performed. Robust errors are clustered at the captain level. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed an annual training. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

Table A5
Treatment Effects During the Experimental Period

	(1) Fuel Load	(2) Eff Flight	(3) Eff Taxi	(4) Fuel Load	(5) Eff Flight	(6) Eff Taxi
Information	0.018	0.027	0.081**	0.021	0.023	0.078***
	(0.030)	(0.022)	(0.034)	(0.025)	(0.018)	(0.026)
Targets	0.032	0.051**	0.115***	0.040	0.045**	0.110***
	(0.030)	(0.024)	(0.035)	(0.025)	(0.018)	(0.026)
Prosocial	0.015	0.033	0.083**	0.020	0.041**	0.093***
	(0.029)	(0.024)	(0.037)	(0.026)	(0.018)	(0.027)
Observations	12,925	12,925	8,104	12,925	12,925	8,104
# of Captains	335	335	335	335	335	335
Controls	No	No	No	Yes	Yes	Yes

Notes: The table shows the results of an OLS regression considering captains' fuel-relevant behavior during the experimental period. The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior is performed, and since predicted values are not constrained between 0 and 1, we do not report a constant and instead focus on treatment effects. As such, the coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest is successfully performed during the experimental period. Robust errors are clustered at the captain level. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed an annual training. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

Table A6
Dependence of Treatment Effect on Salience of Monthly Feedback Report

	(1) Fuel Load	(2) Eff Flight	(3) Eff Taxi
Expt	0.033**	0.128***	0.037**
	(0.014)	(0.013)	(0.017)
Expt · Info	0.000	0.018	0.083***
	(0.016)	(0.015)	(0.019)
Expt · Targets	0.020	0.042***	0.096***
	(0.016)	(0.016)	(0.020)
Expt · Prosocial	0.015	0.045***	0.085***
	(0.016)	(0.015)	(0.019)
Expt · Info · Salient	0.034	-0.008	-0.020
	(0.028)	(0.028)	(0.033)
Expt · Targets · Salient	0.006	-0.027	-0.004
	(0.029)	(0.029)	(0.034)
Expt · Prosocial · Salient	0.049*	0.007	0.014
	(0.029)	(0.028)	(0.033)
Observations	33,822	33,822	21,681
# of Captains	335	335	335
Controls	Yes	Yes	Yes

Notes: The table shows the results of a panel difference-in-difference regression specification with captain fixed effects and Newey-West standard errors (lag=1), controlling for linear trends in the data. The regressions compare pre-experiment behavior (January 2013-January 2014) to behavior during the experiment ("Expt":February 2014-September 2014). The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior is performed, and since predicted values are not constrained between 0 and 1, we do not report a constant and instead focus on treatment effects. As such, the single interaction coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest is successfully performed, and the double interaction coefficients indicate this treatment effect after the first seven days of receiving the feedback report. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed an annual training. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

Table A7
Impact on Delays

	(1) Delays	(2) Short Delays	(3) Long Delays
Expt	-0.043***	-0.034***	-0.020*
	(0.014)	(0.013)	(0.103)
Information	-0.022	-0.017	-0.007
	(0.016)	(0.014)	(0.013)
Targets	0.001	-0.001	0.004
	(0.017)	(0.015)	(0.014)
Prosocial	-0.026*	-0.023*	-0.007
	(0.016)	(0.013)	(0.013)

Notes: The table presents estimates of delays leaving the departure gate by experimental condition. Delays are based on regression coefficients from a difference-in-difference specification with captain fixed effects comparing pre-experiment delays (January 2013-January 2014) to delays during the experiment (February 2014-September 2014), controlling for a linear trend. The dependent variable is whether a delay occurs at all (column 1), whether a short delay occurs (1-15 minutes; column 2), or whether a long delay occurs (greater than 15 minutes; column 3). Standard errors are clustered at the captain level. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed an annual training. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

Table A8
Job Satisfaction and Treatment Group

	Job Satisfaction
Information	0.212
	(0.221)
Targets	0.0242
	(0.230)
Prosocial	0.365*
	(0.220)
Constant	5.58***
	(0.158)
# of Captains	202

Notes: The dependent variable in this regression is a 7-point scale of job satisfaction, where self-reported job satisfaction increases in the scale. Standard errors are reported below estimates in parentheses. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

Table A9
Job Satisfaction and Job Performance

Groups:	Control and Information		Targets and Prosocial	
	Job Satisfaction		Job Satisfaction	
Fuel Load Targets Met	0.093		0.065	
	(0.062)		(0.060)	
Eff Flight Targets Met	-0.074		-0.017	
	(0.056)		(0.054)	
Eff Taxi Targets Met	0.025		0.120**	
	(0.043)		(0.054)	
Overall Targets Met		0.006		0.058*
		(0.028)		(0.031)
Constant	5.691***	5.632***	5.341***	5.326***
	(0.291)	(0.300)	(0.358)	(0.358)
Observations	103		99	
# of Captains	103		99	
Controls	No		No	

Notes: The dependent variable in these regressions is a 7-point scale of job satisfaction, where self-reported job satisfaction increases in the scale. Robust standard errors are reported below estimates in parentheses. The independent variables indicate the number of targets met per behavior and overall over the course of the study. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

F Additional Analysis on the Correlations Across Behaviors

The objective of this section is to understand whether all captains changed their performance across all three behaviors, or whether some captains changed one or two dimensions while others improved on other dimensions (i.e., the change in the behaviors are not highly correlated across and within captains). First we observe the raw correlations in all pairwise combinations of the behaviors for the 13 months prior to the experiment to get an idea of how the behaviors move together prior to our intervention (Figure A5). We then consider correlations (again, for each pairwise combination of the behaviors) in the average within-captain difference in implementation before versus during the experiment to discern changes in the slopes of the first set of graphs (Figure A6).

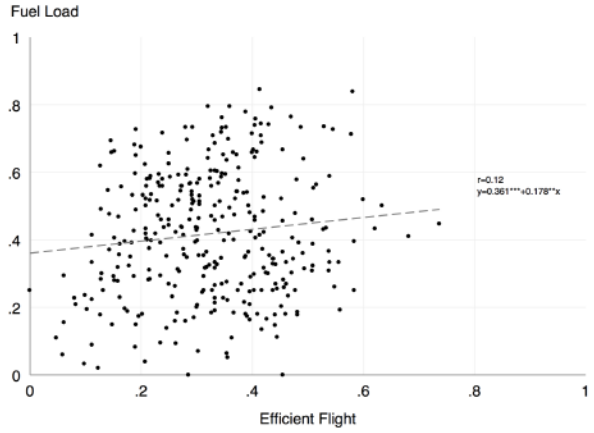
For the first approach, we provide scatter plots of the raw data for each captain for all combinations of the behaviors averaged for the 13 months of pre-experimental data. In Figure A5a, we see that Fuel Load and Efficient Flight have a correlation coefficient of 0.12 (line of best fit: $FL = 0.361^{***}+0.178^{**}EF$), suggesting that Fuel Load and Efficient Flight are positively correlated, with a one-unit increase in the former associated with a 0.18 unit increase in the latter. In Figure A5b, we see that Fuel Load and Efficient Taxi have a correlation coefficient of -0.13 (line of best fit: $FL = 0.455^{***}-0.105^{**}ET$), suggesting that these two behaviors are negatively correlated. In Figure A5c, we see that Efficient Flight and Efficient Taxi have a correlation coefficient of 0.35 (line of best fit: $EF = 0.257^{***}+0.189^{***}ET$), suggesting that Efficient Flight and Efficient Taxi are positively correlated. These plots suggest that the behaviors are related to each other before the experiment.

In Figure A6, we average each captain's implementation before and during the experiment and subtract the former average from the latter. This exercise provides us with the pre- versus during-experiment difference for each of the three behaviors for each captain. Figure A6a shows that the experiment increased the correlation between Fuel Load and Efficient Flight (r=0.18, line of best fit: $FL = 0.012 + 0.169^{***}EF$), suggesting that the experiment increased the correlation between these two behaviors. Figures A6b and A6c indicate that the experiment did not increase the correlation in implementation between Efficient Taxi and either of the other two behaviors. Overall, it appears that on average, (some) captains were more likely to implement Fuel Load and Efficient Flight, but these captains did not necessarily also improve on Efficient Taxi, and similarly captains who improved on Efficient Taxi may not have been more likely to improve on Fuel Load or Efficient Flight.

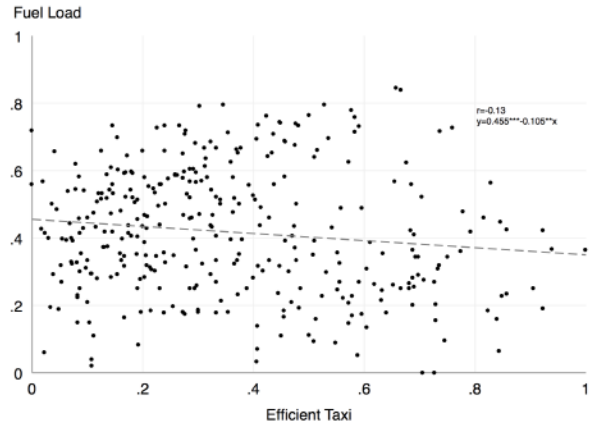
In Table A10, we break down the correlation coefficient for within-captain differences in

Figure A5
Pairwise Correlations in Behaviors Before the Experiment

(a) Fuel Load and Efficient Flight



(b) Fuel Load and Efficient Taxi



(c) Efficient Flight and Efficient Taxi

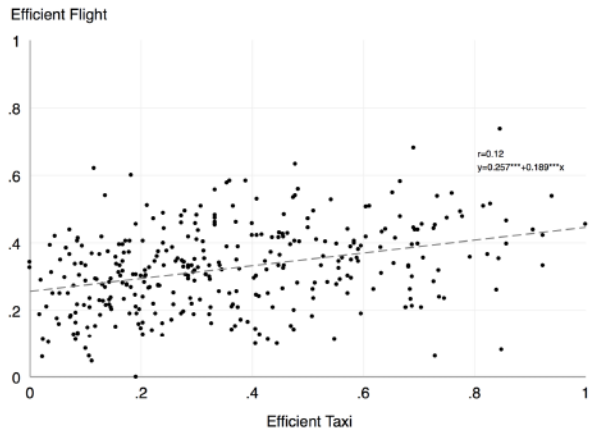
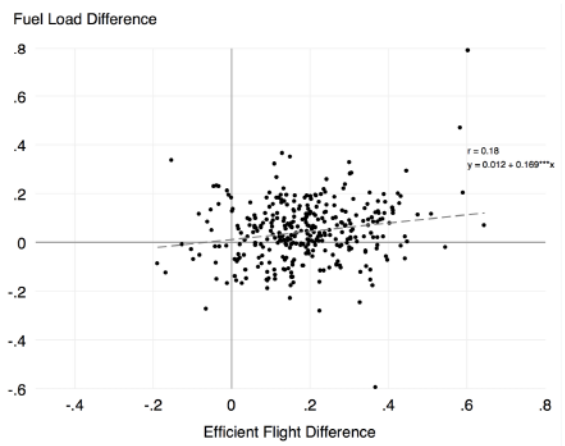
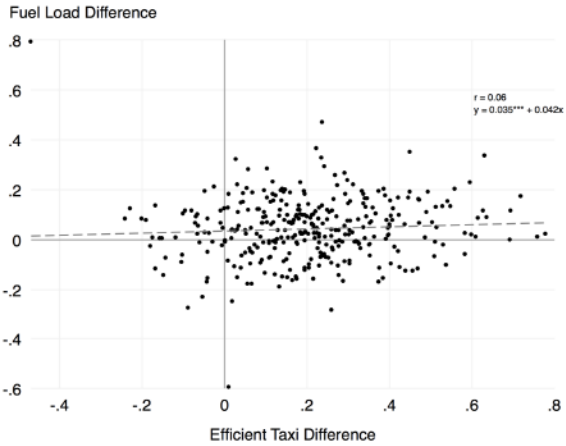


Figure A6
Pairwise Correlations in Within-Captain Behavior Change

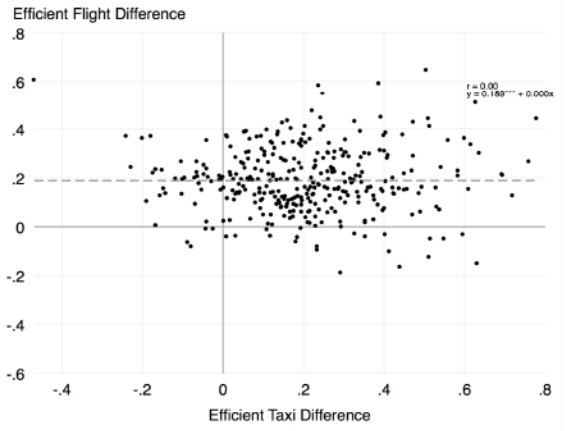
(a) Fuel Load and Efficient Flight



(b) Fuel Load and Efficient Taxi



(c) Efficient Flight and Efficient Taxi



implementation (pre- versus during) by study group. When examining the correlation coefficient for the changes in Fuel Load and Efficient Flight, we see consistent coefficients between 0.15 and 0.21. For the other two combinations of the behaviors, the coefficients are less consistent although not drastically different from each other. For instance, for Efficient Flight and Efficient Taxi, we get correlation coefficients of -0.10, 0.07, -0.02, and -0.03 for control, information, targets, and prosocial incentives, respectively. For Fuel Load and Efficient Taxi, we get correlation coefficients of -0.06, 0.17, -0.01, and 0.13, respectively. Overall, there is no consistent pattern when we separately assess correlations within each experimental group.

Table A10
Change in Correlation Coefficients in Behavior Implementation

	Monitoring		**Information**		**Targets**		**Prosocial**	
	Eff Flight	Eff Taxi	Eff Flight	Eff Taxi	Eff Flight	Eff Taxi	Eff Flight	Eff Taxi
Fuel Load	0.18	-0.06	0.15	0.17	0.21	-0.01	0.16	0.13
Eff Flight		-0.10		0.07		-0.02		-0.03

Notes: The table shows the correlation coefficients of within-captain differences in behavioral implementation (pre- versus during experiment) for each pairwise combination of behaviors within each experimental group.

G Additional Analysis on Heterogeneous Treatment Response

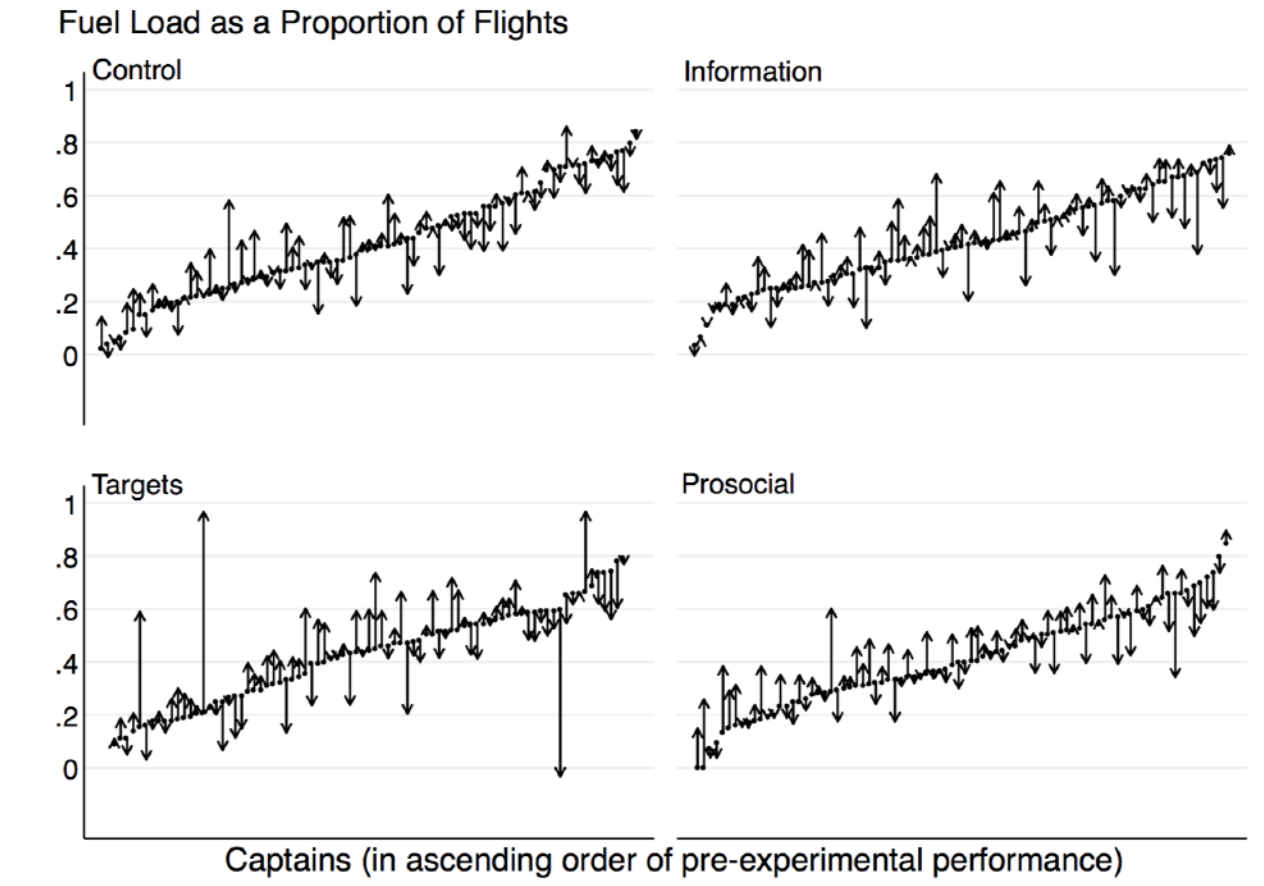
From Figure A7 we glean several qualitative insights that provide motivation for further nuanced research. The figure provides evidence of variance in the response to the interventions, perhaps suggesting that airlines (and possibly other high-skilled labor organizations) may benefit from tailoring management practices to individuals depending on their own or their "type's" response to various management practices. For instance, for Fuel Load (Figure A7a), a handful of captains are extremely motivated by targets (whereas one captain is entirely put off by them). Furthermore, in the targets group, high-achieving captains appear to perform worse than they had pre-experiment once they start receiving feedback reports, while low-achieving captains seem to improve on average. For Fuel Load, the lowest-achieving captains appear to be unaffected by the treatments with the exception of the prosocial incentive treatment group, where the lowest-achieving captains seem to be motivated by the intervention to implement this particularly sticky behavior.

As for Efficient Flight, Figure A7b indicates that information and targets motivate the worst performers, while prosocial incentives have no impact on (or even de-motivate) them. Consistent with our main difference-in-difference findings in Table 4, we see more upward arrows for captains in the targets and prosocial groups, indicating that it is not just a few captains driving up our average treatment effects on Efficient Flight implementation.

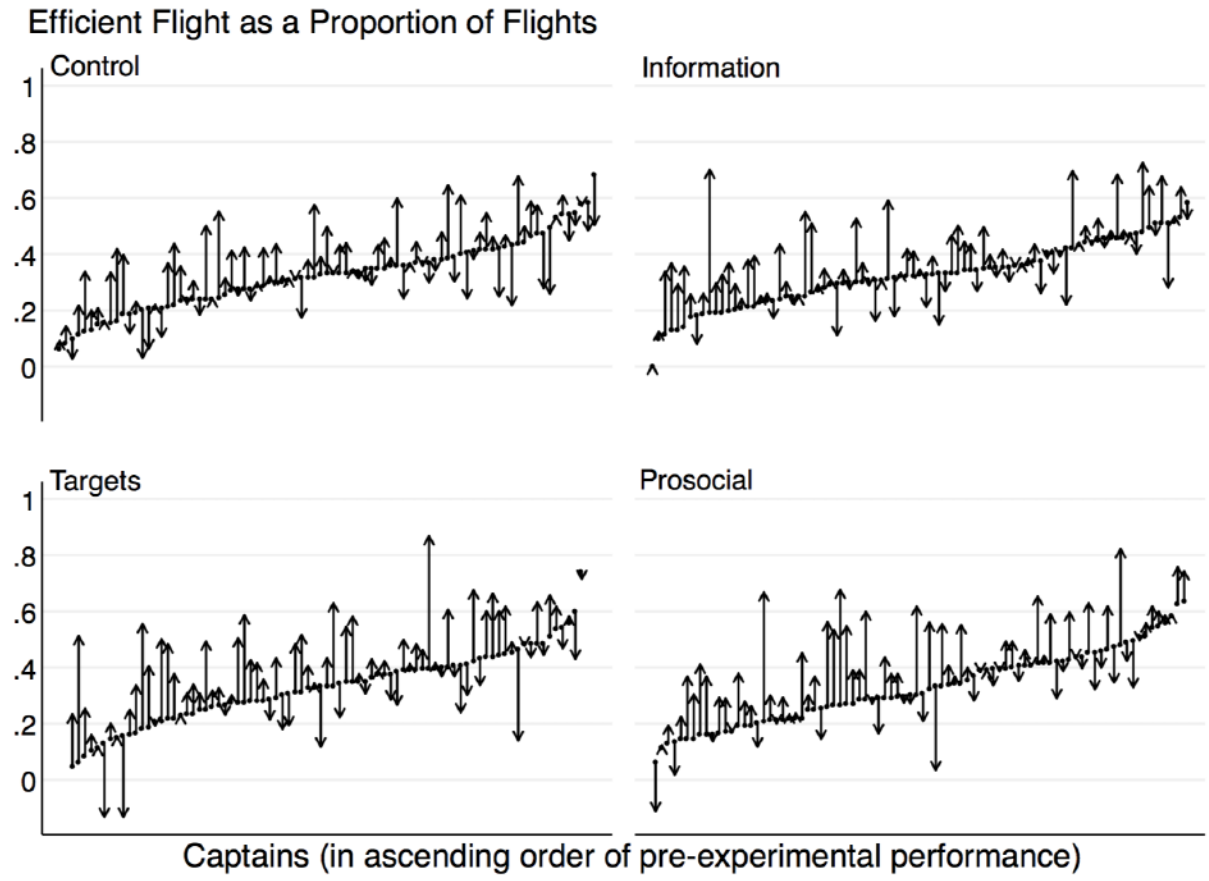
A vast majority of captains in the treatment groups improve on Efficient Taxi (as opposed to the control group, which we would expect given that the graphs show individual changes in behavior net of the monitoring effect; Figure A7c). Interestingly, most improvements come from captains who generally implement Efficient Taxi on 50% or fewer flights prior to the experiment (with very few captains reducing their implementation during the experiment in the information and targets groups), indicating that the management practices deployed in our treatments may best target lower performers on this behavioral dimension.

Figure A7
Within-subject Changes in All Groups due to Experiment (Net of Monitoring Effects)

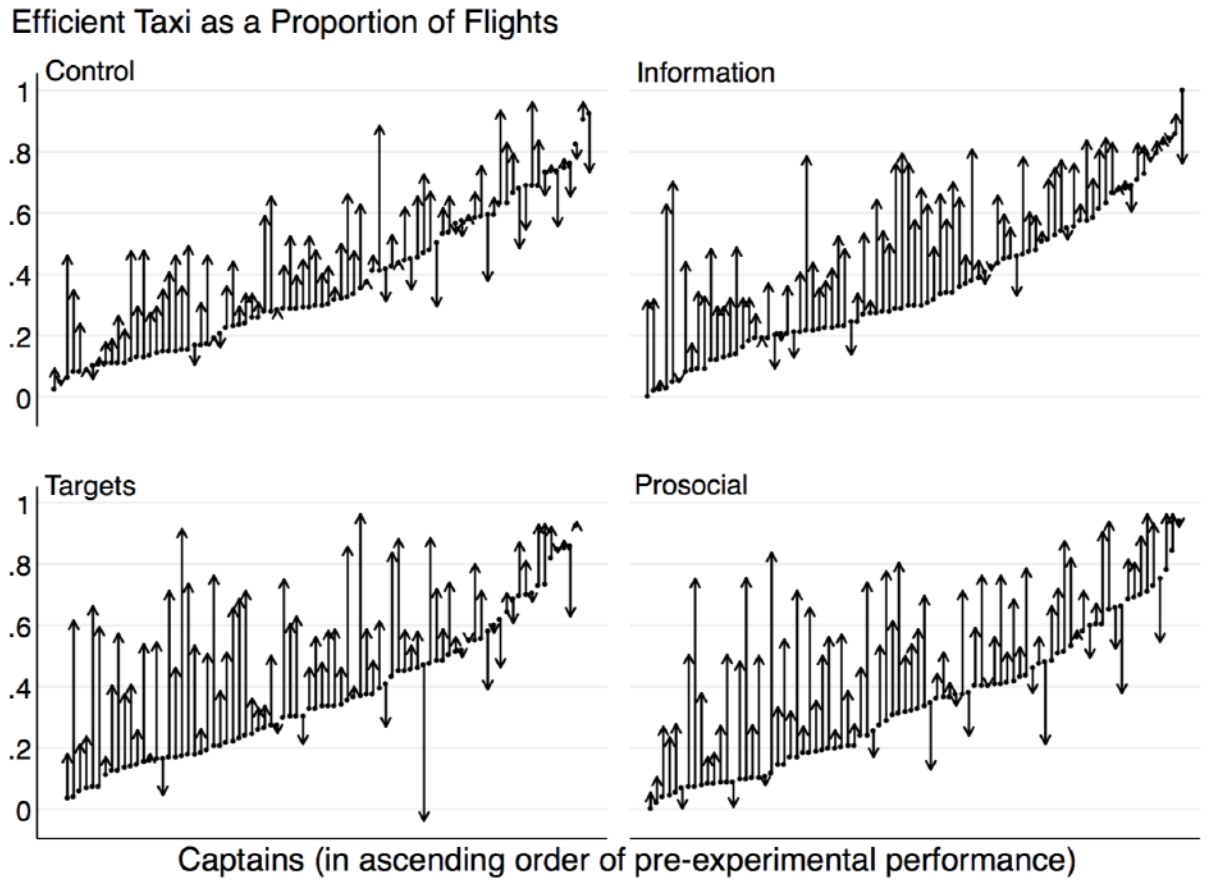
(a) Fuel Load



(b) Efficient Flight



(c) Efficient Taxi



Notes: The data points in the graph represent the proportion of flights for which each captain in the experimental groups implemented the fuel-related behaviors before the experiment (January 2013 - January 2014), in ascending order of pre-experimental performance. The vertical arrows represent the same proportion during the experimental period (February 2014 - September 2014), net of the average monitoring effect identified in our main difference-in-difference specification. An upward (downward) arrow indicates an improvement (decline) in implementation of the behavior net of the monitoring effect.

H Price Elasticity of Fuel Demand

To gain an understanding of the magnitude of the estimated fuel savings in a broader industry context, we estimate the requisite change in price to induce the fuel savings reported in 3.21 of the manuscript using estimates of the price elasticity of jet fuel demand in the aviation industry. To do this, we need to first provide the overall percentage change in fuel use resulting from our study, and then use a credible price elasticity of jet fuel demand to estimate the comparative effect. For the latter, there are few credible estimates in the literature due to endogeneity concerns, as airlines can react to changing prices by, for example, altering the number of seats on their aircraft, thereby influencing prices. However, a review of the literature suggests that jet fuel demand is quite price inelastic, with correlational estimates ranging between 0.04 and 0.30 (Mazraati and Alyousif, 2009).

According to our estimates, the experiment led to 7,768 tons of fuel savings, which is equivalent to reducing 0.7% of overall fuel use by Virgin Atlantic during the time period under investigation. Using the price elasticity ranges above, we can estimate the jet fuel price increase that would bring about a reduction in market jet fuel use by 0.7%. The exercise reveals that the experimental fuel savings are equivalent to those that would result from an increase in jet fuel prices between 2.3% and 17.5%.

I The Effect of Simulator Trainings on Fuel Use

Airline captains regularly participate in flight simulations both to learn and practice new flying techniques and to allow the airline to monitor and provide feedback on their performance in a simulated environment. Simulator trainings regularly cover the ZFW calculation (i.e. efficient Fuel Load procedures) and Efficient Taxi procedures.[55] Importantly for the purpose of our analysis, simulator trainings are scheduled every year and at different times for each captain. As a result, we can treat the date of simulator training as random.

We therefore use a regression discontinuity design (RDD) to examine the performance of fuel-related behaviors before and after captains receive simulator trainings. We estimate the impact of the trainings on efficiency using the following fixed effects panel specification:

$$\text{EfficientBehavior}_{it} = \alpha + \text{PostWindow}_{it}\beta + \text{Window}_{it}\gamma + X_{it}\zeta + \omega_i + e_{it}$$

where the Window indicator is equal to one if the flight is within the relevant time period on either side of the simulator training for captain i (and equals zero otherwise), and the PostWindow indicator is equal to one if the flight took place during the relevant time period after the simulator training for captain i. For robustness, we consider two different window lengths: seven days and thirty days.

Table A11 presents the results of these regressions for each behavior during the pre-experimental and experimental time periods. Columns (1) to (3) present the regressions using seven-day windows and columns (4) to (6) present the regressions using thirty-day windows. From the coefficients presented, there is no consistent evidence that simulator trainings improve fuel efficiency. In fact, we find that after seven days, the simulator training has no impact on any of the fuel efficiency measures, though for the thirty-day window Efficient Taxi reduces by 3.8 percentage points ($p < 0.01$); in other words, the simulator training may actually reduce fuel efficiency on this dimension.

Finally, we explore whether the experiment had a significant impact on the effects of simulator trainings on fuel efficiency. We interact our main specification with the RDD interaction terms defined above in a triple differenced specification, and we find that there is no difference in the effect of simulator training on fuel-efficient behaviors in the experimental period compared to the pre-experimental period. The triple interaction does not produce any strong or consistent results, which suggests that neither awareness of the experiment nor its treatments had any effect on the fuel efficiency impact of simulator trainings.

[55] Efficient Flight is achieved via a mixture of various in-flight techniques that would be encouraged and analyzed during training flights rather than in the simulator.

Table A11
Effects of Simulator Trainings using Regression Discontinuity

	(1) Fuel Load	(2) Eff Flight	(3) Eff Taxi	(4) Fuel Load	(5) Eff Flight	(6) Eff Taxi
Week Before or After	0.016	0.058***	0.024	-	-	-
	(0.018)	(0.018)	(0.023)	-	-	-
Week After	0.003	-0.029	-0.027	-	-	-
	(0.024)	(0.022)	(0.027)	-	-	-
Month Before or After				0.019*	0.006	0.048***
	-	-	-	(0.010)	(0.009)	(0.013)
Month After				-0.014	0.011	-0.038***
	-	-	-	(0.012)	(0.011)	(0.014)
Observations	33,822	33,822	21,681	33,822	33,822	21,681
# of Captains	335	335	335	335	335	335
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table shows the results of a regression discontinuity specification with captain fixed effects and clustered standard errors, controlling for linear trends in the data. 'Week (Month) Before or After' is a dummy variable equal to one if the flight took place within one week (month) of a flight simulation, and equal to zero otherwise. Similarly, 'Week (Month) After' is a dummy variable equal to one if the flight took place within the seven- (thirty-)day period following a simulation. The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior was performed, and since predicted values are not constrained between 0 and 1, we do not report a constant. Robust errors are clustered at the captain level. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed training. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

J Alternative Method for Data-Supported Fuel Savings Estimates

To provide robustness to the data-supported estimates discussed in section 3.3, we provide an alternative measure of fuel savings derived from the data where savings are attributed directly to each of the three behaviors targeted in the study. For Fuel Load, we measure the deviation of the actual fuel load from the "ideal" fuel load—the latter stemming from the double iteration calculation. We identify the average group-level deviation, which is positive if the captain over-fuels relative to the ideal. We then estimate average fuel savings per flight for each treatment group, which entails summing the treatment's effect on per-flight fuel savings from Fuel Load with the control group's per-flight fuel savings from Fuel Load (the monitoring effect; see Table A12). In doing so, we assume that the monitoring effect is constant across groups. On average, captains in the control group decreased fuel load relative to the ideal by 177.7 kg, those in the information group by 148.5 kg per flight, those in the targets group by 190.5 kg per flight, and those in the prosocial group by 208.8 kg per flight.

Similarly, for Efficient Flight, we examine changes in captains' fuel use relative to the "ideal" fuel use, or the anticipated fuel use according to the flight plan (adjusted for updates made during Fuel Load). We find that captains in the control, information, targets, and prosocial groups reduced in-flight fuel use by 256.1, 280.1, 361.1, and 329.9 kg per flight, respectively. Finally, for Efficient Taxi, we examine changes to fuel use during taxi-in. Fuel savings per flight amounted to 17.0 kg, 20.8 kg, 21.9 kg, and 11.8 kg for the control, information, targets, and prosocial incentives groups, respectively.

As a next step, we take these group-level effects and scale them up by the number of flights per treatment group. Put differently, total savings for a given treatment cell are the sum of the per-flight fuel savings for each behavior (i.e., the sum of the average treatment effect and average monitoring effect) multiplied by the number of unique flights during the experimental period flown by captains in that group. Using the data-supported estimates, our interventions led to roughly 6.51 million kg in fuel savings in aggregate. Using the same conversions as in the manuscript, total savings correspond to cost savings of $5.12 million (equivalent to a reduction of 0.53% of overall fuel costs) and CO_2 savings of 20.5 million kg.

Table A12
Data-Supported Estimates of Average Fuel Savings per Flight (in kilograms)

	(1) Fuel Load	(2) Efficient Flight	(3) Efficient Taxi
Control	177.66***	256.08***	17.02***
	(29.42)	(42.24)	(5.08)
Information	148.48***	280.06***	20.75***
	(28.70)	(36.07)	(4.55)
Targets	190.54***	361.08***	21.90***
	(28.26)	(36.77)	(5.36)
Prosocial	208.82***	329.87***	11.75**
	(29.55)	(36.69)	(4.61)

Notes: The table presents estimates of average fuel savings by treatment group. Savings are based on regression coefficients from a difference-in-difference specification with captain fixed effects comparing pre-experiment behavior (January 2013-January 2014) to behavior during the experiment (February 2014-September 2014), controlling for a linear trend. The dependent variable is the deviation from ideal fuel usage in each of the three flight periods as described in the text. We calculate fuel savings with an intent-to-treat approach where we sum the regression coefficient of each group (i.e., the group's average treatment effect) and the average monitoring effect (i.e., the coefficient of the experimental-period indicator). In other words, we assume that the monitoring effect is constant across groups. Standard error calculations are based on Newey-West standard errors (lag=1). Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed an annual training. $^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

K Engineering Estimates of Fuel Savings

We apply engineering estimates to assess fuel savings without requiring data on actual fuel usage or statistical power to detect differences in fuel use pre- and post-intervention. However, the engineering estimates do not account for *actual* changes to fuel usage as a result of behavior change.[56] Since the data-supported estimates incorporate actual changes to fuel use as a result of the study, we have the unique ability in our study to compare engineering estimates with various data-driven estimates.

VAA projects an average fuel savings of 250 kg per flight as a result of proper execution of Fuel Load. The 0.7%, 2.2% and 2.5% treatment effects for the information, targets, and prosocial incentives groups (respectively) correspond to an increase in the implementation of Fuel Load by 169 flights (saving 250 kg each flight), equivalent to a savings of 42,250 kg of fuel over an eight-month period. Moreover, VAA estimates that an Efficient Flight uses (at least) 500 kg less fuel than the alternative, on average. The effect sizes for the three groups were 1.7%, 3.7%, and 4.7% (respectively), which translates to 323 additional "efficient" flights over the eight-month period, or 161,500 kg in fuel savings. Finally, VAA estimates an average fuel wastage of 9 kg per minute if no engines are shut down while taxiing, and the average treatment effects for the three groups were 7.9%, 9.6%, and 8.8%, respectively. Given an average taxi-in time of 8 minutes in the dataset and allowing for a three-minute cooling-off period before engine shutdown, we approximate fuel savings per flight to be 45 kg. An additional 840 extra flights having met Efficient Taxi corresponds to a fuel savings of 37,800 kg over the eight-month study period.

Summing these savings, the engineering estimates indicate that the interventions led to more than 242,000 kg of fuel saved over the course of the study (rounded to the nearest thousand). Combining the industry's standard conversion of 3.1497 kg of CO_2 per kg of fuel burned with the February 2014 global jet fuel price of \$786 per 1000 kg, we estimate a cost savings of \$190,000 and a CO_2 savings of 763,000 kg (i.e., \$28,000 environmental savings using \$37/ton of CO_2 at 3% discount rate in 2015; Interagency Working Group on Social Cost of Carbon, 2013). These calculations constitute fuel and cost savings stemming directly from the treatments and do not incorporate the sizable monitoring effects, which increase the overall CO_2 savings to 3,335,000 kg. The savings associated with the monitoring effects come from captains having performed Fuel Load on 427 more flights, Efficient Flight on 1,706 more flights, and Efficient Taxi on 491 more flights (a savings of 982,000kg of fuel).

[56] There is increasing evidence that engineering estimates diverge from estimates coming from observed changes in behavior derived from clear identification strategies (see Fowlie et al., 2015).

Interestingly, there are substantial differences between the engineering and data-supported estimates from our study. The disparity may be attributable to underestimates of average savings from the three behaviors—especially for the Efficient Flight metric—as well as differences in the nature of the estimations. That is, unlike the engineering estimates, the data-supported estimates do not account for differences in percentages of flights for which a behavior was met. Rather, they estimate overall average fuel use changes in the study itself and apply these changes to all flights. Even if we apply the most conservative fuel savings estimates to the changes in behavior, we find that the study interventions, especially the provision of targets, led to remarkable cost-savings and return on investment for the airline.

L Survey Materials

L.1 Job Satisfaction

All things considered, how **satisfied** are you with...

	Not at all satisfied 1	2	3	4	5	6	Extremely Satisfied 7
...your present job overall?	○	○	○	○	○	○	○

References

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of Political Economy*, 1447–1458.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal 100*(401), 464–477.

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy*, 598–614.

Bénabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652–1678.

Christen, M., G. Iyer, and D. Soberman (2006). Job satisfaction, job performance, and effort: A reexamination using agency theory. *Journal of Marketing 70*(1), 137–150.

DellaVigna, S., J. A. List, and U. Malmendier (2012). Testing for Altruism and Social Pressure in Charitable Giving. *Quarterly Journal of Economics 127*(1), 1–56.

Fowlie, M., M. Greenstone, and C. Wolfram (2015). Do energy efficiency investments deliver? evidence from the weatherization assistance program. Technical report, National Bureau of Economic Research.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, 74–91.

Holmström, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 24–52.

Interagency Working Group on Social Cost of Carbon (2013). Technical support document: Technical update of the social cost of carbon for regulatory impact analysis—under executive order 12866.

https://www.whitehouse.gov/sites/default/files/omb/assets/inforeg/technical-update-social-cost-of-carbon-for-regulatory-impact-analysis.pdf. Accessed: November 13, 2015.

Jessoe, K. and D. Rapson (2014). Knowledge is (less) power: Experimental evidence from residential energy use. *American Economic Review 104*(4), 1417–38.

Köszegi, B. and M. Rabin (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics 121*(4), 1133–1165.

Mazraati, M. and O. M. Alyousif (2009). Aviation fuel demand modelling in oecd and developing countries: impacts of fuel efficiency. *OPEC energy review 33*(1), 23–46.

Pugno, M. and S. Depedri (2009). Job performance and job satisfaction: An integrated survey. Department of Economics Working Papers 0904, Department of Economics, University of Trento, Italia.