Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Operational Transparency: Showing When Work Gets Done

Robert L. Bray

To cite this article:
Robert L. Bray (2023) Operational Transparency: Showing When Work Gets Done. Manufacturing & Service Operations Management 25(3):812-826. https://doi.org/10.1287/msom.2020.0899

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 25, No. 3, May–June 2023, pp. 812–826
ISSN 1523-4614 (print), ISSN 1526-5498 (online)
https://pubsonline.informs.org/journal/msom

Winner—2018 MSOM Data Driven Research Challenge

Operational Transparency: Showing When Work Gets Done

Robert L. Bray[a]

[a] Operations Management Department, Kellogg School of Management, Northwestern University, Evanston, Illinois 60208
Contact: r-bray@kellogg.northwestern.edu, https://orcid.org/0000-0003-2773-0663 (RLB)

Received: October 25, 2018
Revised: April 17, 2019; September 5, 2019
Accepted: November 13, 2019
Published Online in Articles in Advance: October 29, 2020

https://doi.org/10.1287/msom.2020.0899

Abstract. ***Problem definition***: Do the benefits of operational transparency depend on when the work is done? ***Academic/practical relevance***: This work connects the operations management literature on operational transparency with the psychology literature on the peak-end effect. ***Methodology***: This study examines how customers respond to operational transparency with parcel delivery data from the Cainiao Network, the logistics arm of Alibaba. The sample comprises 4.68 million deliveries. Each delivery has between 4 and 10 track-package activities, which customers can check in real time, and a delivery service score, which customers leave after receiving the package. Instrumental-variable regressions quantify the causal effect of track-package-activity times on delivery scores. ***Results***: The regressions suggest that customers punish early idleness less than late idleness, leaving higher delivery service scores when track-package activities cluster toward the end of the shipping horizon. For example, if a shipment takes 100 hours, then delaying the time of the average action from hour 20 to hour 80 increases the expected delivery score by approximately the same amount as expediting the arrival time from hour 100 to hour 73. ***Managerial implications***: Memory limitations make customers especially sensitive to how service operations end.

History: This paper has been accepted as part of the 2018 MSOM Data Driven Research Challenge.

Keywords: operational transparency • package delivery • peak-end rule • empirical operations management

1. Introduction

Customers rate service more highly when effort is visible. For example, Buell and Norton (2011, p. 1564) argue that "engaging in operational transparency, by making the work that a website is purportedly doing more salient, leads consumers to value that service more highly." And Buell et al. (2017, p. 1673) note that

> The introduction of [operational] transparency contributed to a 22.2% increase in customer-reported quality and reduced throughput times by 19.2%. Laboratory studies revealed that customers who observed process transparency perceived greater employee effort and thus were more appreciative of the employees and valued the service more.

These authors, however, implicitly assume that customers will perceive worker effort but not worker loafing. Yet true operational transparency will make both industry and idleness visible. For example, consider the following track-package records that the Cainiao Network shared with its customers:

Shipment	Action	Facility	Timestamp	Score
3144672	Order	—	2017-02-05 15:05	—
3144672	Consign	—	2017-02-05 17:37	—
3144672	Receive	105638	2017-02-05 18:40	—
3144672	Depart	105638	2017-02-05 21:52	—
3144672	Arrive	65132	2017-02-06 04:15	—
3144672	Depart	65132	2017-02-06 05:20	—
3144672	Arrive	29048	2017-02-06 08:22	—
3144672	Scan	29048	2017-02-06 08:44	—
3144672	Sign	29048	2017-02-10 21:58	1
15007307	Order	—	2017-03-08 13:15	—
15007307	Consign	—	2017-03-10 17:14	—
15007307	Receive	49199	2017-03-14 19:27	—
15007307	Depart	49199	2017-03-14 19:51	—
15007307	Arrive	162115	2017-03-14 20:48	—
15007307	Depart	162115	2017-03-15 05:12	—
15007307	Arrive	166957	2017-03-15 06:29	—
15007307	Scan	166957	2017-03-15 07:28	—
15007307	Sign	166957	2017-03-15 10:04	5

Shipment 3144672 ended with an idle spell between the sixth and tenth of February, and shipment 15007307 began with an idle spell between the tenth and fourteenth of March. Cainiao's operational transparency exposed this inactivity.

But the delivery service scores suggest that shipment 3144672's late idleness was worse than shipment 15007307's early idleness. This makes sense. First, imagine waiting for shipment 3144672: After seeing the package zip through three facilities in two

days, you anticipate its arrival at any moment, only to suffer four additional days of delay. Moreover, the quick start makes you more conscious of the subsequent silence—the parcel appears to vanish as its track-package signals abruptly end. Thus, when the package finally arrives, you give it the worst possible score (one out of five).

Now imagine waiting for shipment 15007307: You see little progress in the first six days of your order. This is unsettling, but you're not sure whether to attribute the lack of reported actions to a lack of reporting or to a lack of actions—only in the last two days do you learn that this shipper thoroughly records its activities. And by this time, you're reassured by a steady stream of updates. This final hustle is fresh in your mind when you give the delivery the best possible score (five out of five).

These cherry-picked examples are extreme, but they illustrate my thesis: Consumers leave higher parcel delivery ratings when track-package activities occur near the final delivery time. Thus, the goodwill garnered by operational transparency depends on when the work is done—and when it's not done.

I support this claim with the Cainiao Network's track-package records. Each shipment in my sample has a customer delivery score and a sequence of actions with corresponding timestamps. I regress the delivery scores on the action times with five different specifications. Each indicates that later actions yield higher scores. For example, if the shipping time is 100 hours, then the first regression suggests that shifting a *single action* from hour 20 to hour 80 increases the expected delivery score by 0.021 standard deviations; the second, third, fourth, and fifth regressions suggest that shifting the *average action* from hour 20 to hour 80 increases the expected delivery score by 0.075, 0.185, 0.197, and 0.037 standard deviations, respectively. For perspective, decreasing the shipping time from four to three days increases the expected delivery score by 0.064 standard deviations.

The first two regressions use ordinary least squares (OLS). The third regression uses weekend lulls to instrument for action times. For example, because Saturdays and Sundays have the least activity, shipments that start on Fridays tend to have slower starts; hence, later average action times and shipments that end on Mondays tend to have slower finishes and, hence, earlier average action times. The fourth regression generalizes this instrumental variables (IV) specification to factor other temporal shocks, such as national holidays. And the fifth regression uses preshipment delays to instrument for action times. Warehouse-to-shipper consignment is always the first action to follow the customer's order, so delaying this consignment delays all subsequent actions. But this consignment happens *before* the shipment, so the consignment time should not directly affect the shipment quality (conditional on the final delivery time).

2. Theory

Delayed activity can increase scores in several ways. First, psychology's peak-end rule states that "the final moments of an extended episode appear to exert a strong influence on the overall judgment [of its utility]" (Varey and Kahneman 1992, p. 169). For example, in Kahneman et al.'s (1993) peak-end study,

> [s]ubjects were exposed to two aversive experiences: In the short trial, they immersed one hand in water at 14°C for 60 s; in the long trial, they immersed the other hand at 14°C for 60 s, then kept the hand in the water 30 s longer as the temperature of the water was gradually raised to 15°C, still painful but distinctly less so for most subjects. Subjects were later given a choice of which trial to repeat. A significant majority chose to repeat the long trial, apparently preferring more pain over less (Kahneman et al. 1993, p. 401).

According to the peak-end rule, a shipment's ending will be especially memorable, which suggests that it's best to finish on a strong note with a burst of activity at the end.

Second, customer satisfaction depends on service quality *relative to expectations*. Surveying psychology's satisfaction literature, Oliver (1980, p. 460) found that

> [a]lmost without exception, reviewers and early researchers in the areas of job, life, self, and patient satisfaction agree that satisfaction is a function of an initial standard and some perceived discrepancy from the initial reference point. . . . Specifically, expectations are thought to create a frame of reference about which one makes a comparative judgment. Thus, outcomes poorer than expected (a negative disconfirmation) are rated below this reference point, whereas those better than expected (a positive disconfirmation) are evaluated above this base.

In this light, early activity can be counterproductive, as it gives customers unrealistic expectations about the speed of delivery—starting fast raises customer hopes and ending slow dashes them. Moreover, the unfulfilled expectations can make customers apprehensive, as Harvard's Ryan Buell explained to me in an email (Buell 2018):

> Reading through your paper made me think of the work by Osuna (1985), which basically shows how customer uncertainty can increase frustration and anger, undermining people's satisfaction. That's the paper that basically became the reason we see progress bars everywhere—people value the certainty of knowing when a task will be complete or a service will be delivered. A package making fitful progress toward delivery . . . only to be stalled at the last minute could amp up uncertainty.

Third, inactivity vexes customers only after they've learned to expect steady status updates. Most customers don't know how much track-package activity to expect from a given shipper, so they can attribute a silent start to a silent shipper. But a lively start establishes a high benchmark: After a few days of

consistent posting, a day of inactivity seems an ominous halt to momentum. Once trained to expect progress reports, customers will notice their absence.

And fourth, customers will check the track-package logs more frequently near the expected arrival time, so activities reported near this time are more likely to be noticed and appreciated.

3. Data

I use data provided for the 2018 MSOM Data Driven Research Challenge by the Cainiao Network.[1] An affiliate of Alibaba Group, Cainiao runs an online logistics platform for managing the delivery of goods purchased through Alibaba's websites. The company was founded in 2013 with the goal of "realiz[ing] delivery anywhere in China within 24 hours, and across the globe within 72 hours" (Cainiao Network 2020).

The data set comprises (i) a 10.02-GB table that describes customer orders; (ii) a 507-MB table that describes warehouse inventories; (iii) a 2.52-GB table that describes products; (iv) a 77-MB table that describes merchants; and (v) a 74-GB table that describes package delivery logistics. The first table provides my dependent variable—the *customer delivery score*; the last table provides my primary independent variables—the *track-package action timestamps*.

There are several types of track-package action:[2]

- Order: the customer places the order
- Consign: the warehouse dispatches the package
- Receive: the carrier receives the package
- Depart: the package exits a facility
- Arrive: the package enters a facility
- Scan: the shipper scans the package for final delivery
- Sign: the customer receives the package
- Failure: the shipper fails to deliver the package

The Tmall and Cainiao mobile apps—which account for the lion's share of Cainiao's business—disclose these track-package actions in real time (e.g., see Figure 1).

I filter my sample along several dimensions, removing all shipments

- with a failure action (0.74% of observations),
- with an origin warehouse not managed by Cainiao (73% of observations),
- without a shipment score or shipment times (64% of observations),
- with actions reported before the order action (0.024% of observations),
- with actions reported after the sign action (3.2% of observations),
- without exactly one sign action (2.6% of observations),
- without exactly one consign action (1.3% of observations),
- without the slowest shipping speed (15% of observations),
- with multiple shippers (0.0010% of observations),
- with multiple product types (6.7% of observations),
- with shipment times in excess of eight days (1.6% of observations),

Figure 1. Example of What the Customer Sees

Notes. This is a screenshot of the actions the Tmall app reported for a representative package. The package was ordered on December 12, consigned to the shipper on December 13, moved from Handan to Xingtai on December 13, moved from Xingtai to Shijiazhuang and then to Beijing on December 14, and delivered on December 15.

• with more than 10 posted actions (6.0% of observations), or

• with fewer than 4 posted actions (6.3% of observations).

The resulting sample comprises 101 thousand facilities, 4.68 million shipments, and 40.10 million actions from January 1, 2017 to July 31, 2017. It includes the following variables:

• Delivery Score is a delivery logistics quality score left by the customer. The customer uploads this information via a mobile app or website after receiving the package. The variable takes values in $\{1,\cdots,5\}$, where 1 is the worst and 5 the best. It has a mean of 4.82 and a standard deviation of 0.64 (see Table 1).

• Action Time is the time of a particular action, measured as a fraction of the shipping time. The variable takes values in $[0,1]$ and has a mean of 0.49 and a standard deviation of 0.36 (see Table 2). For example, Table 3 reports shipment 3144672's Action Times: The order Action Time is 0.00 because this action starts the shipment; the sign Action Time is 1.00 because this action ends the shipment; and the consign Action Time is 0.02 because this action happens after 2% of the shipping time has elapsed. I henceforth disregard order and sign actions, because their Action Times mechanically equal 0.00 and 1.00, respectively.

• Average Action Time is the shipment's average Action Time (excluding order and sign actions). This variable takes values in $[0,1]$ and has a mean of 0.49 and a standard deviation of 0.13. For example, shipment 3144672's Average Action Time is $(0.020 + 0.028 + 0.054 + 0.104 + 0.112 + 0.136 + 0.139)/7 = 0.085$ (see Table 3).

• Action Count is the number of distinct actions—other than order and sign—reported on the shipment's track-package log. This variable takes values in $\{4,\cdots,10\}$ (until Section 6 removes this constraint) and has a mean of 6.57 and a standard deviation of 1.72. For example, shipment 3144672's Action Count is seven (see Table 3).

• Action Count$[t,s)$ is the number of distinct actions with Action Times in range $[t,s)$. This variable takes values in $\{0,\cdots,10\}$. For example, shipment 3144672 has an Action Count$[0,0.05)$ of two (see Table 3).

• Shipping Time is the time between the shipment's order and sign actions, measured in days. This variable takes values in $[0,8]$ (until Section 6 removes this constraint) and has a mean of 2.74 and a standard deviation of 1.16. For example, shipment 3144672's Shipping Time is 5.29 (see Table 3). A regression of Delivery Score on Shipping Time suggests that increasing the latter by one day decreases the former by 0.0443 points.

• Day Count is Shipping Time rounded up to the nearest day. This variable takes values in $\{1,\cdots,8\}$ (until Section 6 removes this constraint) and has a mean of 3.22 and a standard deviation of 1.19. For example, shipment 3144672's Day Count is six (see Table 3).

• Consign Count, Receive Count, Arrive Count, Depart Count, and Scan Count are the number of distinct consign, receive, arrive, depart, and scan actions reported on the shipment's track-package log. These variables take values in $\{0,\cdots,10\}$ (until Section 6 removes these constraints). For example, shipment 3144672 has a Receive Count of one and a Depart Count of two (see Table 3).

• Facility Count is the number of distinct facilities reported on the shipment's track-package log.[3] This variable takes values in $\{0,\cdots,8\}$ and has a mean of 3.52 and a standard deviation of 1.59. For example, shipment 3144672's Facility Count is three (see Table 3).

• Day is the day of the shipment's order action. This variable takes values in $\{1,\cdots,212\}$, where Day = 1 corresponds to January 1, 2017, the first date in my sample, and Day = 212 corresponds to July 31, 2017, the last date in my sample. For example, shipment 3144672's Day is 36 (see Table 3).

• Week is the week of the shipment's order action. This variable takes values in $\{1,\cdots,31\}$, where Week = 1 corresponds to the week starting on January 1, 2017, the first Sunday in my sample, and Week = 31 corresponds to the week starting on July 30, 2017, the last Sunday in my sample. For example, shipment 3144672's Week is 6 (see Table 3).

Table 1. Average Delivery Scores

	Day Count								
Action Count	1	2	3	4	5	6	7	8	Total
4	4.88	4.86	4.84	4.80	4.74	4.71	4.69	4.65	4.83
5	4.89	4.86	4.83	4.80	4.75	4.71	4.68	4.68	4.82
6	4.89	4.87	4.85	4.80	4.73	4.66	4.57	4.56	4.82
7	4.88	4.86	4.83	4.80	4.75	4.69	4.64	4.59	4.82
8	4.88	4.86	4.84	4.78	4.70	4.62	4.56	4.49	4.82
9	4.88	4.85	4.83	4.78	4.72	4.65	4.62	4.53	4.81
10	4.85	4.85	4.82	4.78	4.73	4.66	4.61	4.53	4.79
Total	4.88	4.86	4.84	4.79	4.74	4.68	4.63	4.59	4.82

Notes. I tabulate the average Delivery Score by Action Count and Day Count. For example, four-action shipments that arrive within a day have an average Delivery Score of 4.88.

Table 2. Action Time Deciles

Action Type	10%	20%	30%	40%	50%	60%	70%	80%	90%
Order	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Consign	0.03	0.04	0.06	0.09	0.13	0.18	0.22	0.27	0.35
Receive	0.07	0.10	0.14	0.19	0.23	0.28	0.33	0.40	0.51
Depart	0.17	0.27	0.35	0.43	0.52	0.59	0.67	0.74	0.83
Arrive	0.23	0.36	0.48	0.57	0.65	0.73	0.81	0.86	0.92
Scan	0.69	0.81	0.85	0.88	0.91	0.93	0.95	0.97	0.99
Sign	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Notes. I tabulate the Action Time deciles by action type. For example, the median consign action happens after 13% of the shipping time has elapsed. Note that the order Action Time is always 0.00 and the sign Action Time is always 1.00 because these actions bookend the shipment.

- Day of Week is the day of the week of the shipment's order action. This variable takes values in $\{1,\cdots,7\}$, where 1 corresponds to Sunday and 7 to Saturday. For example, shipment 3144672's Day of Week is 1 (see Table 3).
- Buyer, Brand, Category, Merchant, and Shipper are the ID numbers of the customer, product brand, product category, merchant, and shipper, respectively. For example, shipment 3144672's Buyer is 61581582.
- Shipping Speed is the shipping speed selected by the customer. This variable takes values in $\{1,2,3,\infty\}$, where the first three options guarantee delivery within one, two, and three days, respectively; the last option provides no delivery date guarantee. Shipping Speeds are restricted to ∞ until Section 6.
- Postmedian Average Action Time is the average Action Time of the actions that occur after the median Action Time. This variable takes values in $[0,1]$ and has a mean of 0.746 and a standard deviation of 0.138. For example, shipment 3144672's median Action Time is 0.1038, and its Postmedian Average Action Time is $(0.1123 + 0.1362 + 0.1391)/3 = 0.1292$ (see Table 3).

4. Empirical Definition

Before proceeding to the analysis, I must explain what operational transparency means in my context. I explained in Section 3 that customers can observe the track package records with the Tmall and Cainiao mobile apps. However, customers must explicitly log in to one of these apps to access these records, as the apps do not push package-update notifications to their phones. Thus, I study a new version of operational transparency. In previous operational transparency studies, customers couldn't avoid the process updates, which elbowed their way into customers' consciousness whether they were wanted or not. For example, as soon as Buell and Norton's (2011) subjects input their flight requests, they were immediately prompted with a "waiting screen [that] displayed a continually changing list of which sites were being searched and showed an animation of the fares being compiled as they were 'found'" (p. 1566); these subjects couldn't help but observe the ticket-finding process, unless they closed their eyes the moment they input their flight requests. But operational transparency usually isn't so invasive. For example, when we say that a government is transparent, we don't mean that it compels every citizen to watch every critical decision in real time; instead, we mean that the government makes this information available upon request. My process exhibits this more subtle flavor of operational transparency.

To be clear, although not every customer in my sample observes the process, every customer in my sample receives the "treatment" of operational transparency. This treatment isn't to be given process information; this treatment is to be given *access to* process information. So I don't estimate the effect of giving customers process information; I estimate the effect of giving customers *access to* process information.

5. Results

Figure 2 demonstrates that shipments with different Delivery Scores have different Action Time distributions. The plots depict the Action Time probability density functions (PDFs) conditional on the Delivery Score minus the Action Time PDFs unconditional on the Delivery Score (I subtract away the unconditional distributions to highlight the across-score differences). Each action type yields the same pattern: To the left, the score-1–2 PDFs are the highest, followed by the score-3 PDFs, then the score-4 PDFs, and then the score-5 PDFs; to the right, this order is reversed. Thus, the score-1–2 actions occur earlier than the score-3 actions, which occur earlier than the score-4 actions, which occur earlier than the score-5 actions.

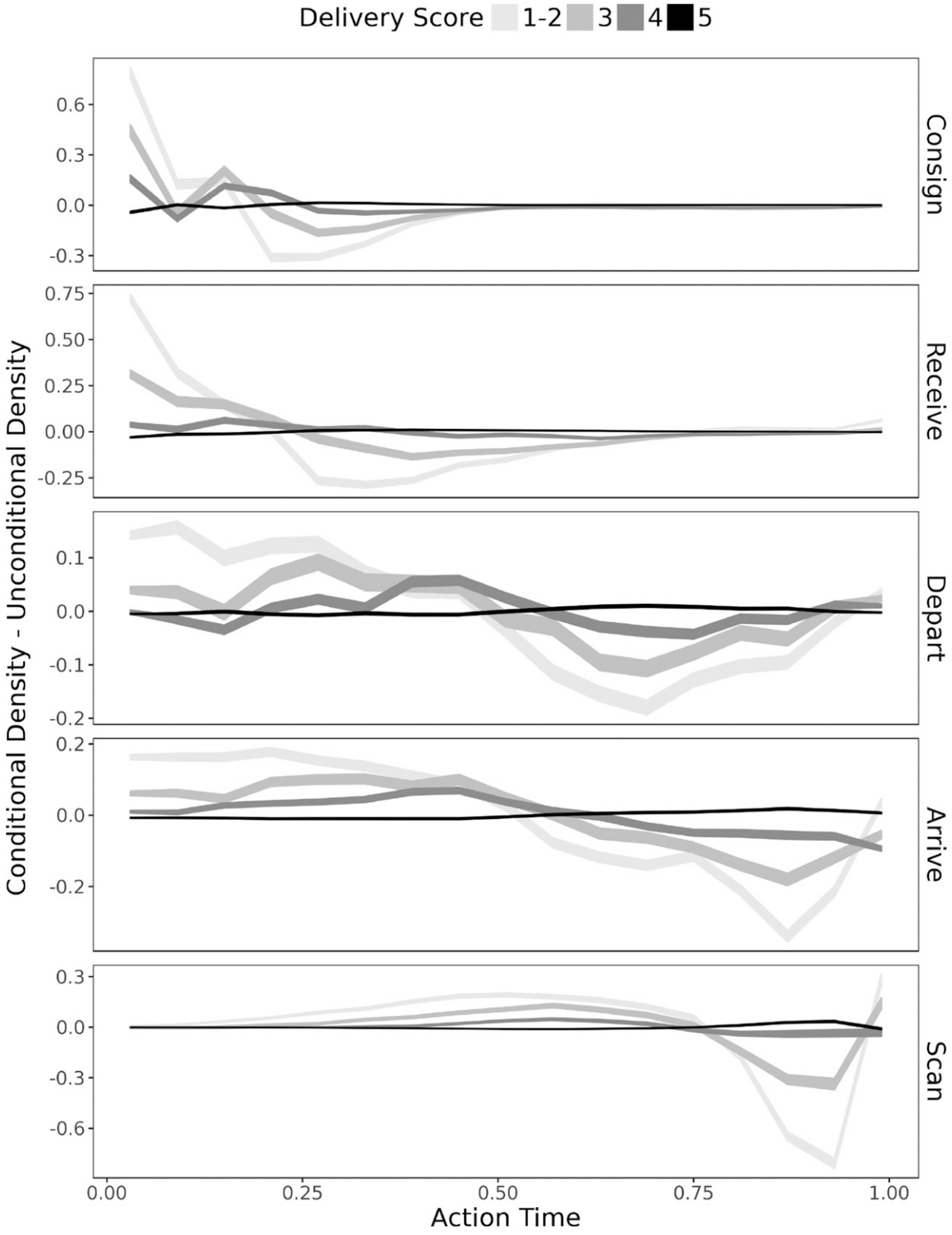
I now formalize the relationship between Action Times and Delivery Scores with a set of regressions. I run 14 OLS regressions across 14 subsamples; these subsamples correspond to the seven distinct Action Count values and the seven distinct Day Count values, respectively. The regressions' dependent variable is

Table 3. Action Time Definition

Shipment	Action	Facility	Timestamp	Action Time
3144672	Order	–	2017-02-05 15:05	0.0000
3144672	Consign	–	2017-02-05 17:37	0.0200
3144672	Receive	105638	2017-02-05 18:40	0.0282
3144672	Depart	105638	2017-02-05 21:52	0.0535
3144672	Arrive	65132	2017-02-06 04:15	0.1038
3144672	Depart	65132	2017-02-06 05:20	0.1123
3144672	Arrive	29048	2017-02-06 08:22	0.1362
3144672	Scan	29048	2017-02-06 08:44	0.1391
3144672	Sign	29048	2017-02-10 21:58	1.0000

Notes. I tabulate the Action Times of shipment 3144672. I derive these values from the action timestamps, setting $\text{Action Time}_n = \frac{\text{Timestamp}_n - \min_m(\text{Timestamp}_m)}{\max_m(\text{Timestamp}_m) - \min_m(\text{Timestamp}_m)}$.

Figure 2. Action Time Distributions



Notes. I estimate each action type's Action Time probability density functions, both conditional and unconditional on the Delivery Score. I then subtract the unconditional density estimates from the conditional density estimates; I plot the differences' 90% confidence intervals with lines of varying thickness, thinner lines indicating more precise estimates. For example, the score-5 lines are the thinnest because score-5 estimates are the most precise, because the score-5 sample is the largest (accounting for 89% of the total sample). Because so many scores are 5, the distributions conditional on the score being 5 resemble the unconditional distributions, which explains why the score-5 lines are so near zero. I combine the score-1 and score-2 observations because only 0.5% of scores are 2.

the Delivery Score. The control variables are a set of dummies that specify (i) the Day Count, Facility Count, Receive Count, Arrive Count, Depart Count, and Scan Count values and (ii) the Brand, Category, Merchant, Shipper, and Week values that have at least 5,000 observations (see Table 4).[4] Finally, the primary independent variables are the Action Count$[t,s)$ values corresponding to the following 19 time ranges:

[0.00, 0.05), [0.05, 0.10), [0.10, 0.15), [0.15, 0.20), [0.20, 0.25),
[0.25, 0.30), [0.30, 0.35), [0.35, 0.40), [0.40, 0.45), [0.45, 0.50),
[0.50, 0.55), [0.55, 0.60), [0.60, 0.65), [0.65, 0.70), [0.70, 0.75),
[0.75, 0.80), [0.80, 0.85), [0.85, 0.90), and [0.90, 0.95).

For example, the Action Times of shipment 3144672 are

$$\underbrace{0.020,\ 0.028,}_{\in[0.00,0.05)}\ \underbrace{0.054,}_{\in[0.05,0.10)}\ \underbrace{0.104,\ 0.112,\ 0.136,\ \text{and}\ 0.139.}_{\in[0.10,0.15)}$$

So, for this observation, Action Count [0.00, 0.05) is two, Action Count [0.05, 0.10) is one, Action Count [0.10, 0.15) is four, and the rest are zero. That said, my first set of regressions have the following form:

$$\begin{aligned}\text{Delivery Score} \sim\ & \text{Action Count}[0.00,\ 0.05) + \cdots \\ & + \text{Action Count}[0.90,\ 0.95) \\ & + \text{Controls},\end{aligned}$$

where Controls are the variables listed in Table 4.

Figure 3 plots the 19 coefficient estimates corresponding to the 19 time ranges.[5] These estimates report the amount an action in the given time range increases the expected score minus the amount an action in the [0.95, 1.00] time range increases the expected score. For example, the far-left estimate of the four-action plot is −0.017, which suggests that shifting one Action Time from [0.95, 1.00] to [0.00, 0.05) decreases the expected Delivery Score by 0.017 points (or 0.027 standard deviations). Overall, the estimates suggest that actions that occur after 95% of the shipping time has elapsed increase scores more than

- actions that occur before 10% of the shipping time has elapsed: Of the 28 Action Count[0.00, 0.05) and Action Count[0.05, 0.10) estimates, 27 are negative and 21 are significantly negative at the $p = 0.01$ level;
- actions that occur before 25% of the shipping time has elapsed: Of the 70 Action Count[0.00, 0.05)–Action Count[0.20, 0.25) estimates, 67 are negative and 51 are significantly negative at the $p = 0.01$ level; and
- actions that occur before 50% of the shipping time has elapsed: Of the 140 Action Count[0.00, 0.05)–Action Count[0.45, 0.50) estimates, 136 are negative and 99 are significantly negative at the $p = 0.01$ level.

However, the most valuable actions appear to be those that occur between Action Times 0.85 and 0.95: Of the 28 Action Count [0.85, 0.90)–Action Count [0.90, 0.95) estimates, 24 are positive and 7 are significantly positive at the $p = 0.01$ level. The average difference between the Action Count[0.80, 0.85) and Action Count[0.15, 0.20) estimates is 0.0137, which suggests that shifting one action from the [0.15, 0.20) time range to the [0.80, 0.85) time range increases the expected Delivery Score by an average of 0.0137 points (which is quite a lot, considering that the average shipment comprises 6.57 actions).

The [0.85, 0.95) time range is the sweet spot because it's late enough to enjoy a peak-end effect but not so late that the package arrives before the action is noticed. Because customers aren't notified when actions are uploaded, many actions posted in the [0.85, 0.95) time range won't be noticed until the [0.95, 1.00) time range and many actions posted in the [0.95, 1.00) time range won't be noticed until they're moot. However, despite this time lag, the correlation between Action Times and Delivery Scores is undeniably positive: 13 out of 14 trend lines fit through the regression estimates are significantly positive at the $p = 0.01$ level.

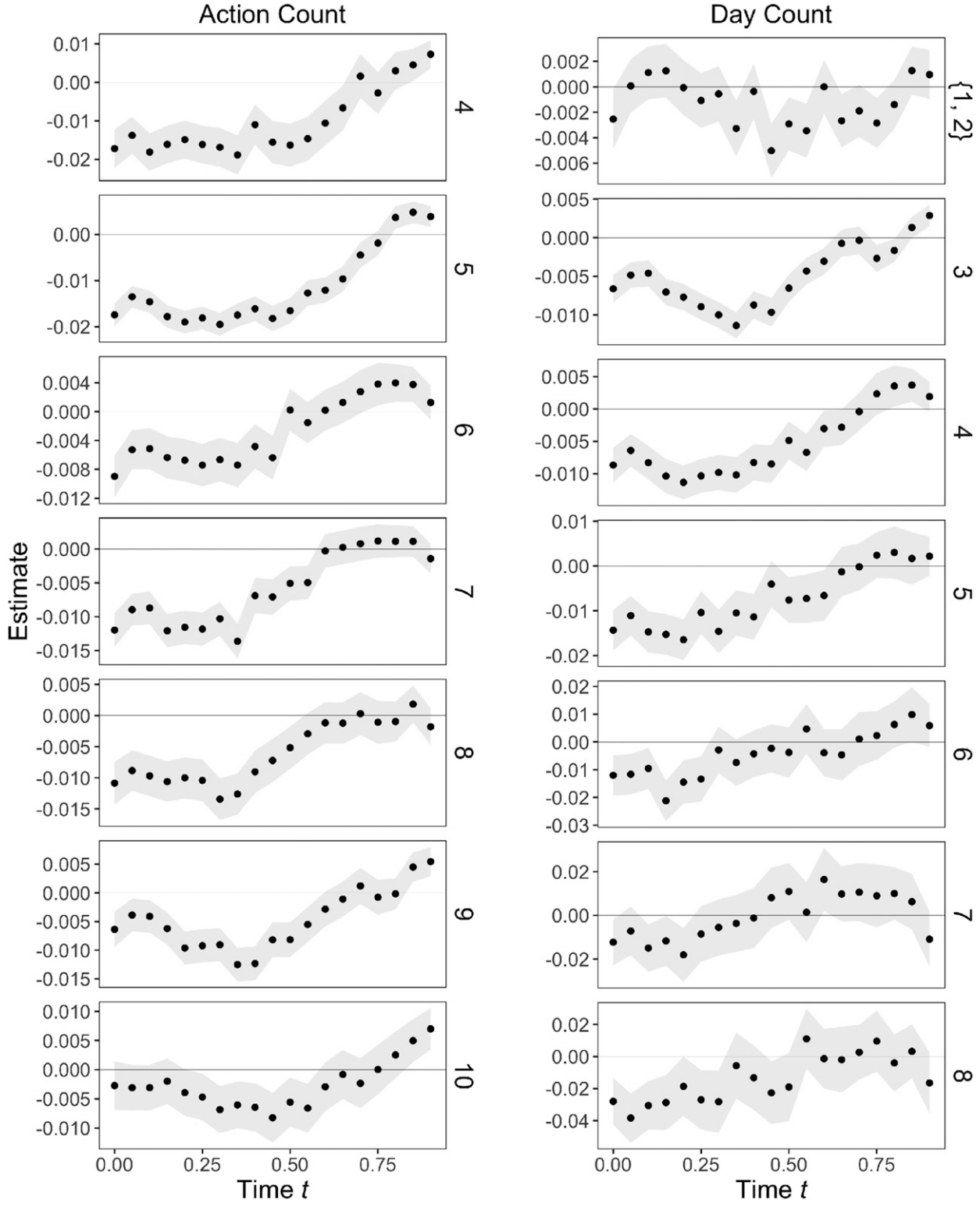
To establish the ubiquity of this effect, I run 100 additional regressions across 50 different subsamples. The subsamples are the observations of the 10 most frequent (i) Brand, (ii) Category, (iii) Merchant, (iv) Shipper, and (v) Week values. For example, the 10 most common Shipper IDs are 247, 565, 674, 724, 532, 431, 149, 132, 270, and 184, each of which has its own subsample. For each subsample, I run two regressions. For both regressions, the dependent variable is the Delivery Score and the primary independent variable is the Average Action Time, but the first regression incorporates Table 4's control variables, whereas the

Table 4. Control Variable Dummies

Merchant	170
Brand	98
Category	15
Shipper	20
Week	30
Day Count	7
Facility Count	8
Arrive Count	6
Depart Count	7
Receive Count	5
Scan Count	8

Notes. This table records the number of dummy variables of each type that I use as controls in my regressions. For Day Count, Facility Count, Receive Count, Arrive Count, Depart Count, and Scan Count, the number of dummy variables equals one less than the number of distinct values (the fully saturated case). For Brand, Category, Merchant, Shipper, and Week, there is one dummy variable for each value with at least 5,000 observations. For example, the sample has eight Day Count values and 186 Brand values, 98 of which appear at least 5,000 times.

Figure 3. OLS Estimates of Delivery Score on Action Count$[t, s)$



Notes. I run 14 regressions across 14 subsamples: the observations with Action Count = n, for $n \in \{4, \cdots, 10\}$, and the observations with Day Count $\in n$, for $n \in \{\{1, 2\}, 3, \cdots, 8\}$. I pool the Day Count = 1 and Day Count = 2 observations because only 2% of Day Counts are 1. The dependent variable is Delivery Score; the control variables are Table 4's dummies; and the primary independent variables are Action Count$[t, t + 0.05)$, for $t \in \{0.00, 0.05, \cdots 0.90\}$. The black dots depict the coefficient estimates of these primary independent variables, with the left-most dot corresponding to the [0.00, 0.05) time range and the right-most dot corresponding to the [0.90, 0.95) time range. The gray bands depict the estimates' 90% confidence intervals.

second does not. That said, these regressions have the following form:

$$\text{Delivery Score} \sim \text{Average Action Time} + \text{Controls}.$$

Table 5 reports the Average Action Time coefficient estimates. Of the 100 estimates, 93 are positive and 84 are significantly positive at the $p = 0.01$ level: The effect is pervasive.[6] And the effect is meaningful: Running the regressions across the entire sample yields an estimate of 0.081 (see Table 6), which suggests that shifting the Average Action Time from 0.2 to 0.8 increases the expected Delivery Score by $0.081 \cdot (.8 - .2) = 0.049$ points (or 0.075 standard deviations). This change would have the same effect on Delivery Scores as a $0.049/0.0443 = 1.11$-day reduction in Shipping Times. (Section 3 establishes that shortening the Shipping Time by one day increases the expected Delivery Score by 0.0443 points.)

To establish the causality of the relationship between Average Action Time and Delivery Score, I run three sets of two-stage least squares (2SLS) regressions with three sets of instrumental variables. These IV regressions are analogous to those reported in Table 5, except they permit the Average Action time to correlate with the error term via unobserved shipping factors. For example, suppose the final deliveryman is either tardy and rude or prompt and courteous; in this case, packages with late delays would tend to arrive in a ruder fashion than those with early delays. To control for such unobserved shipping factors, I use instrumental variables that influence the Average Action Time but not the shipping process. That said, my 2SLS regressions have the following form:

$$\text{Delivery Score} \sim \text{Average Action Time} + \text{Controls}$$
$$\text{and Average Action Time} \sim \text{Instrument} + \text{Controls}.$$

My first 2SLS specification derives instruments from the weekly variation in activity levels. For example, there's more idleness on weekends, so the idleness of Saturday-to-Friday shipments tends to be earlier than the idleness of Monday-to-Sunday shipments; thus, Saturday-to-Friday shipments tend to have larger Average Action Times than Monday-to-Sunday shipments. This logic suggests that I can treat Day of Week and Day Count pairs as exogenous shifters of Average Action Time. I interact these pairings with Table 4's shipper dummies, because different carriers have different weekly trends. My resulting instrumental variables are 1,159 Day of Week × Day Count × Shipper Dummies (plus the control variables, when they are called for). For example, shipment 3144672 starts on the first day of the week, ends on the sixth day, and is handled by shipper 149, so for this observation the {Day of Week = 1} × {Day Count = 6} × {Shipper = 149} dummy variable is one and the other 1,158 dummy variables are zero. These instruments explain 20.9% of the variation in Average Action Times.

Table 5. OLS Estimates of Delivery Score on Average Action Time

Specification	Sample	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
Controls	Brand	0.067	0.039	0.119	0.013	0.137	–0.036	0.114	0.069	0.093	0.084
		(0.006)	(0.010)	(0.013)	(0.013)	(0.015)	(0.014)	(0.011)	(0.015)	(0.018)	(0.017)
	Category	0.078	0.041	0.036	0.119	0.043	0.063	0.086	0.068	0.074	0.017
		(0.006)	(0.005)	(0.008)	(0.006)	(0.008)	(0.013)	(0.017)	(0.016)	(0.020)	(0.025)
	Merchant	0.063	0.119	0.010	0.137	0.031	–0.035	0.032	0.077	–0.059	0.042
		(0.008)	(0.013)	(0.015)	(0.017)	(0.020)	(0.016)	(0.021)	(0.021)	(0.022)	(0.019)
	Shipper	0.079	0.049	0.013	0.119	0.109	0.048	0.179	0.111	0.052	–0.052
		(0.007)	(0.008)	(0.008)	(0.006)	(0.011)	(0.011)	(0.010)	(0.017)	(0.019)	(0.018)
	Week	0.120	0.071	0.064	0.050	0.122	0.046	0.070	0.068	0.047	0.046
		(0.008)	(0.010)	(0.014)	(0.015)	(0.015)	(0.015)	(0.014)	(0.017)	(0.015)	(0.017)
No controls	Brand	0.114	0.111	0.298	0.230	0.167	–0.021	0.099	0.099	0.114	0.169
		(0.005)	(0.008)	(0.011)	(0.010)	(0.012)	(0.012)	(0.010)	(0.012)	(0.015)	(0.013)
	Category	0.174	0.122	0.166	0.110	0.115	0.177	0.219	0.089	0.096	0.158
		(0.005)	(0.005)	(0.007)	(0.006)	(0.007)	(0.010)	(0.013)	(0.013)	(0.017)	(0.021)
	Merchant	0.097	0.298	0.226	0.161	0.130	–0.045	0.114	0.113	–0.019	0.129
		(0.008)	(0.011)	(0.012)	(0.014)	(0.016)	(0.015)	(0.017)	(0.017)	(0.019)	(0.016)
	Shipper	0.187	0.189	0.005	0.108	0.274	0.034	0.290	0.425	0.315	0.028
		(0.006)	(0.006)	(0.007)	(0.006)	(0.010)	(0.009)	(0.010)	(0.014)	(0.017)	(0.016)
	Week	0.100	0.127	0.163	0.190	0.263	0.122	0.118	0.183	0.115	0.108
		(0.007)	(0.009)	(0.013)	(0.012)	(0.013)	(0.012)	(0.012)	(0.014)	(0.013)	(0.014)

Notes. I run two regression specifications across 50 subsamples, for a total of 100 regressions. The subsamples are the observations with the nth most common Brand, Category, Merchant, Shipper, and Week values, for $n \in \{1, \cdots, 10\}$. The first regression specification includes Table 4's control variables, and the second does not. The dependent variable is Delivery Score, and the primary independent variable is Average Action Time. I tabulate the Average Action Time coefficient estimates and corresponding standard errors. For example, the top-left estimate gives the effect the Average Action Time has on the Delivery Score in the sample comprising the most common Brand, and the bottom-left estimate gives the effect the Average Action Time has on the Delivery Score in the sample comprising the tenth most common Week.

Table 6. OLS and 2SLS Estimates of Delivery Score on Average Action Time

Estimator	Controls	No controls
OLS	0.081	0.137
	(0.003)	(0.002)
2SLS: Day of Week	0.200	0.317
	(0.016)	(0.005)
2SLS: Day	0.213	0.289
	(0.007)	(0.004)
2SLS: Consign Action Time	0.040	0.119
	(0.003)	(0.003)

Notes. I run two OLS regressions and six 2SLS regressions. The OLS regressions are the same as Table 5's, except they use the entire sample. The 2SLS regressions are the same as the OLS regressions, except they instrument for the average action time. The first 2SLS specification uses 1,159 Day of Week × Day Count × Shipper dummies as instruments; the second uses 25,617 Day × Day Count × Shipper dummies as instruments; and the third uses 209 consign Action Time decile × Shipper dummies as instruments. Additionally, the control variables serve as exogenous instruments (in the regressions that include them). I tabulate the Average Action time coefficient estimates and corresponding standard errors. 2SLS, two-stage least squares.

My second 2SLS specification is the same as the first except it replaces Day of Week with Day. That is, it uses 25,617 Day × Day Count × Shipper dummies as instruments (plus the control variables, when they are called for). Ryan Buell, at Harvard, and an anonymous reviewer gave me the idea for these instrumental variables; they explained that Day of Week isn't granular enough to capture most temporal shocks, such as national holidays, inclement weather, or site-wide promotions (Buell 2018). Giving each day its own set of instruments enables me to more flexibly exploit calendar events. These instruments explain 32.5% of the variation in Average Action Times.

My final 2SLS specification instruments for the Average Action Time with the consign Action Time. From the customer's perspective, it appears that the shipment starts as soon as the order is placed. But, actually, the shipment doesn't begin until the warehouse consigns the parcel to the shipper. Because the consign Action Time should not directly affect the condition of the package, the consign Action Time should not directly affect the Delivery Score (after controlling for the package's arrival time). I, therefore, treat the warehouse-to-shipper consignment times as exogenous shifters of Average Action Time. Specifically, I create a set of dummy variables that characterize the consign Action Time's decile. For example, the consign Action Time is in the first decile if less than 0.024, the second decile if between 0.024 and 0.039, and the third decile if between 0.039 and 0.057. I then interact these consign Action Time decile dummies with Table 4's shipper dummies, because the relationship between consign Action Time and Average Action Time should vary by shipper. My resulting instrumental variables are 209 consign Action Time decile × Shipper dummies (plus the control variables, when they are called for). For example, shipment 3144672 has a first-decile consign Action Time and a Shipper ID of 149, so for this observation the {consign Action Time decile = 1} × {Shipper = 149} dummy variable is one and the other 208 dummy variables are zero. These instruments explain 62.2% of the variation in Average Action Times.

Table 6 reports the 2SLS estimates.[7] They are all significantly positive at the $p = 0.001$ level and are similar to the corresponding OLS estimates.

6. Replication

I imposed 14 filters on my data in Section 3. I did so to compare like with like and to minimize the effect of unobserved confounding variables: For example, if a shipment required three weeks to arrive, then there's probably something important about that delivery I don't see. But Bill Schmidt at Cornell objected to some of these restrictions.[8] To address his feedback, I reran my regressions with the observations initially left out. Loosely speaking, this analysis provides a replication of my primary results, because I conducted it after distributing Section 5's findings.

To create my replication sample, I begin with the observations I excluded in Section 3 and remove shipments

- with a failure action (0.74% of observations),
- with an origin warehouse not managed by Cainiao (73% of observations),
- without a shipment score or shipment times (64% of observations),
- with actions reported before the order action (0.024% of observations),
- with actions reported after the sign action (3.2% of observations), or
- without exactly one sign action (2.6% of observations).

In other words, my replication sample comprises the observations that satisfy the first six conditions listed in Section 3 but not the last seven. It comprises 7.78 million shipments and 66.4 million actions, none of which appear in my initial sample. And it contains a new variable: Shipping Speed, which I initially restricted to the slowest setting. Table 7 demonstrates that slowing the Shipping Speed from one day to two days, to three days, to ∞ days increases the Shipping Time and Action Count and decreases the Delivery Score.[9]

I rerun Figure 3's regression with my replication sample and plot the estimates in Figure 4. The effect is stronger when the Shipping Speed is slower: From one day to two days to three days to ∞ days, the slope of the trend line through the estimates increases from 0.0025–0.0104 to 0.0137–0.0209. This makes sense: Strengthening the shipping guarantee shortens the

Table 7. Replication Sample Summary Statistics

Variable	One day	Two days	Three days	∞ Days
Delivery Score	4.90	4.88	4.85	4.78
Action Time	0.50	0.48	0.51	0.48
Average Action Time	0.50	0.47	0.51	0.45
Shipping Time	0.53	1.20	1.92	3.72
Action Count	4.68	5.51	5.96	7.23
Facility Count	2.71	2.81	3.03	4.19

Notes. I tabulate the average of six variables in my replication sample by Shipping Speed. The one-day Shipping Speed is the fastest, with an overnight delivery guarantee; the ∞-day Shipping Speed is the slowest, with no guaranteed delivery date.

shipping time, which blurs the distinction between early and late actions.

Next, I rerun Table 6's regressions with my replication sample and tabulate the coefficient estimates in Table 8. As before, the effect is stronger with slower Shipping Speeds: From one day to two days to three days to ∞ days, the primary OLS estimate increases from –0.001–0.043 to 0.068–0.143. All the two-, three-, and ∞-day estimates are significantly positive; and three-quarters of the one-day estimates are significantly positive.

7. Mechanism

The IV regressions of Section 5 should convince most readers that delaying Action Times increases Delivery Scores. Unfortunately, it's difficult to determine the mechanism underlying this relationship. In Section 2, I explained that delaying actions could increase scores by (i) making actions more memorable; (ii) giving customers more conservative arrival time estimates; (iii) making customers believe that actions go unreported; or (iv) making actions occur at more conspicuous times. Of these four potential drivers, I believe the first—the peak-end effect—is most likely, as the other three contradict the data.

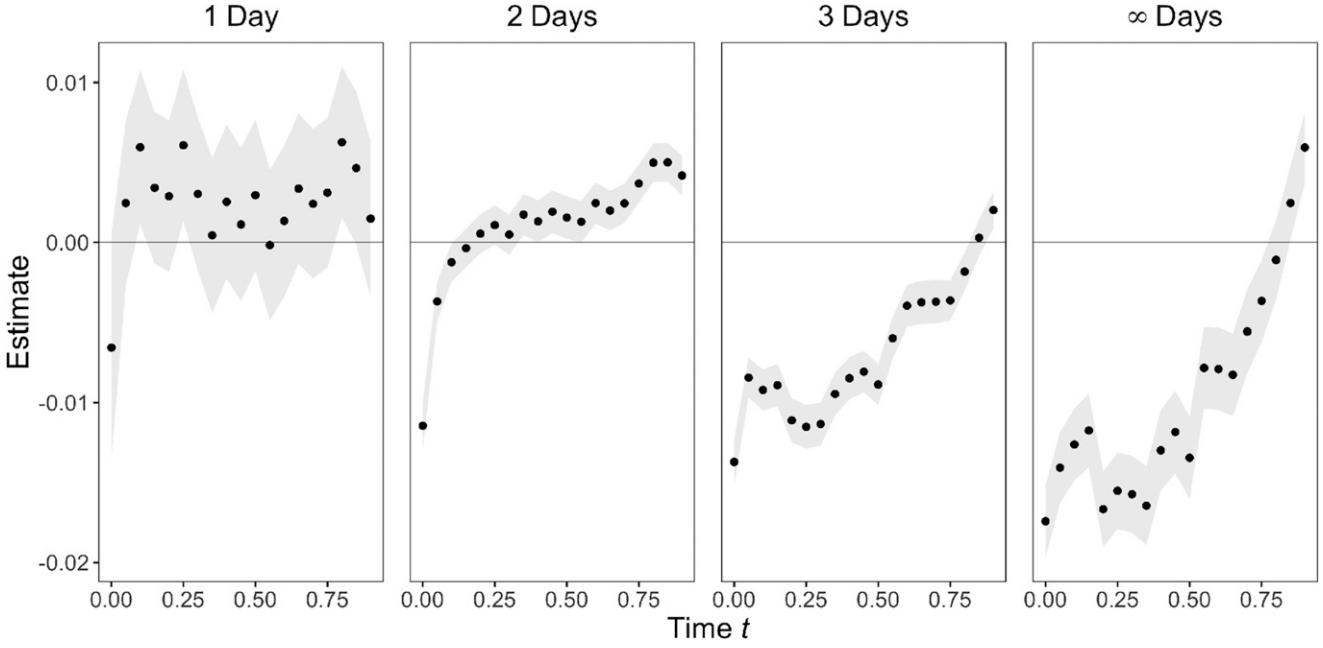
First, only 1.46% of shipments with two- or three-day Shipping Speeds arrive late, so customers should have accurate beliefs about when these shipments will arrive. Thus, the second potential driver—arrival-time expectations management—cannot explain the two- and three-day-Shipping-Speed estimates of Table 8 and Figure 4.

Second, Figure 3's curves increase convexly: Fitting each set of estimates to a second-order polynomial, I find 7 out of the 14 quadratic terms significantly positive at the $p = .01$ level and 0 significantly negative. To establish this convexity more formally, I rerun Table 6's OLS regressions with Postmedian Average Action Time as an additional regressor. Specifically, this new regression has the following form:

$$\begin{aligned}\text{Delivery Score} \sim\ & \text{Average Action Time} \\ & + \text{Postmedian Average Action Time} \\ & + \text{Controls}.\end{aligned}$$

The Postmedian coefficient estimates are significantly positive (see Table 9), which suggests that Delivery Scores

Figure 4. Replication Estimates of Delivery Score on Action Count$[t, s)$



Notes. I rerun Figure 3's regression with my replication sample. However, I perform the regression by Shipping Speed rather than by Action Count or Day Count.

Table 8. Replication Estimates of Delivery Score on Average Action Time

Specification	Estimator	One day	Two days	Three days	∞ Days
Controls	OLS	−0.001	0.043	0.068	0.143
		(0.006)	(0.002)	(0.003)	(0.007)
	2SLS: Day of Week	0.288	0.141	0.195	0.376
		(0.064)	(0.017)	(0.012)	(0.035)
	2SLS: Day	0.090	0.154	0.170	0.193
		(0.016)	(0.005)	(0.007)	(0.009)
	2SLS: Consign Action Time	−0.017	0.031	0.036	0.174
		(0.006)	(0.003)	(0.004)	(0.005)
No controls	OLS	0.057	0.094	0.145	0.137
		(0.008)	(0.002)	(0.003)	(0.010)
	2SLS: Day of Week	0.718	0.414	0.371	0.214
		(0.023)	(0.005)	(0.004)	(0.033)
	2SLS: Day	0.372	0.263	0.288	0.162
		(0.017)	(0.005)	(0.005)	(0.006)
	2SLS: Consign Action Time	0.056	0.091	0.128	0.229
		(0.008)	(0.004)	(0.003)	(0.005)

Notes. I rerun Table 6's regressions with my replication sample. I perform the regressions by Shipping Speed.

are more sensitive to actions that occur after the median action. The third potential driver—status-update-frequency expectations management—cannot explain these findings. Indeed, under this driver, Delivery Scores are especially sensitive to the timing of early actions, which most determine the customers' beliefs about the frequency of status updates. Thus, under the third driver, we would expect Figure 3's curves to be concave and Table 9's Postmedian Average Action Time coefficients to be negative.

Third, Table 10 shows that the results hold in the subset of shipments that arrived more than 24 hours early. To create this table, I reran Table 6's OLS regressions across (i) the subsample of observations with two-day Shipping Speeds and less-than-one-day Shipping Times and (ii) the subsample of observations with three-day Shipping Speeds and less-than-two-day Shipping Times. Because they arrived so far ahead of schedule, these shipments should not have arrived at a particularly conspicuous time. Thus, the fourth potential driver—delayed customer attention—cannot explain these results.

8. Robustness Checks

I now run six robustness checks with my initial sample.

First, an anonymous reviewer asked me to control for the identity of the customer. I control for customer identity with two matching specifications: the first randomly pairs shipments with the same Buyer; the second randomly pairs shipments with the same Buyer, Brand, Category, and Merchant. After pairing the observations, I difference the data across pairs and regress the differenced Delivery Scores on the differenced Average Action Times. For example, I observe the following data for Buyers 84302736 and 84336882:

Shipment	Buyer	Brand	Category	Merchant	Delivery Score	Average Action Time
108254690	84302736	829	54	65	5	0.457
86394137	84302736	253	53	166	5	0.640
90700921	84302736	253	53	166	5	0.483
126280069	84302736	457	11	329	4	0.459
54304079	84302736	457	11	329	5	0.534
29423550	84336882	412	1	134	5	0.509
125927329	84336882	412	1	134	5	0.391
100489141	84336882	412	1	139	5	0.313
69369963	84336882	412	1	139	3	0.306

Matching the sample by Buyer transforms these observations to:

Shipment 1	Shipment 2	ΔDelivery Score	ΔAverage Action Time
108254690	86394137	0	0.183
90700921	126280069	−1	−0.025
29423550	125927329	0	−0.118
100489141	69369963	−2	−0.008

And matching the sample by Buyer, Brand, Category, and Merchant transforms them to:

Shipment 1	Shipment 2	ΔDelivery Score	ΔAverage Action Time
86394137	90700921	0	−0.156
126280069	54304079	1	0.075
29423550	125927329	0	−0.118
100489141	69369963	−2	−0.008

Table 9. OLS Estimates of Delivery Score on Postmedian Average Action Time

Variable	Controls	No controls
Average Action Time	0.036	0.034
	(0.004)	(0.004)
Postmedian Average Action Time	0.054	0.120
	(0.004)	(0.004)

Notes. I rerun Table 6's OLS regressions with an additional covariate: the Postmedian Average Action Time. The Postmedian estimates report the difference between the premedian Action Time effect on Delivery Scores and the postmedian Action Time effect on Delivery Scores. For example, the leftmost column suggests that increasing a premedian Action Time by 0.1 would increase the expected Delivery Score by $0.1 \cdot 0.036 = 0.0036$ points, whereas increasing a postmedian Action Time by 0.1 would increase the expected Delivery Score by $0.1 \cdot (0.036 + 0.054) = 0.0090$ points.

Note that I lose shipment 54304079 in the first case—because buyer 84302736 has an odd number of observations—and I lose shipment 108254690 in the second case—because it's the only Buyer-84302736 observation that corresponds to Category 54. After calculating ΔDelivery Score and ΔAverage Action Time, I run regressions with form

$$\Delta\text{Delivery Score} \sim \Delta\text{Average Action Time}.$$

Table 11 demonstrates that the estimates of these matched regressions are similar to those reported in Table 6. Thus, controlling for the customer identity does not overturn the result.

Second, Dennis Zhang at Washington University and an anonymous reviewer identified a potential problem (Zhang 2018): Both action rates and service quality vary by location. For example, my estimates would be biased upwards if packages moved more expeditiously through the city than through the country and urban customers left systematically higher scores than rural customers. To control for geographic effects, I match my sample by the final facility reported on the track-package log (with 104,000 distinct locations, this final facility variable is granular). I consider two matching specifications: The first randomly pairs shipments with the same final facility and the second randomly pairs shipments with the same final facility, Brand, Category, and Merchant. As before, I difference each pair's Delivery Scores and Average Action Times and regress the differenced Delivery Scores on the differenced Average Action Times. Table 11 demonstrates that controlling for the final facility location does not overturn the result.

Table 10. OLS Estimates from Shipments that Arrive Ahead of Schedule

Sample	Controls	No controls
Two days	0.0060	0.0172
	(0.0034)	(0.0029)
Three days	0.0269	0.0228
	(0.0035)	(0.0031)

Notes. I apply Table 6's OLS regressions to the subset of shipments that arrived at least one day early. The top row corresponds to observations with Shipping Speed = 2 and Day Count = 1, and the bottom row corresponds to observations with Shipping Speed = 3 and Day Count ≤ 2.

Third, Ruomeng Cui at Emory University identified a second potential problem (Cui 2018): Customers can cancel a shipment at no cost before the consign action. This can introduce a selection bias, as the time until the first action influences whether the transaction is represented in my sample (which does not include canceled shipments).[10] To avoid this potential bias, I control for the time until the consign action (measured in hours, not as a fraction of the Shipping Time). I group the consign times into 100 percentile buckets and match the sample by bucket. I consider two matching specifications: The first randomly pairs shipments with the same consign time bucket, and the second randomly pairs shipments with the same consign time bucket, Brand, Category, and Merchant. As before, I difference the data across pairs and regress the differenced Delivery Scores on the differenced Average Action Times. Table 11 demonstrates that controlling for the consignment time does not overturn the result.

Fourth, an anonymous reviewer wondered whether I could reproduce my result without the Delivery Score ≤ two observations. Thus, I rerun Table 6's OLS regressions without these extreme observations. The specification with control variables yields an Average Action Time coefficient estimate of 0.051, with a corresponding t-statistic of 40.5; the specification without control variables yields an Average Action Time coefficient estimate of 0.021, with a corresponding

Table 11. Robustness Check Estimates

Matching scheme	Controls	No controls
Buyer	0.037	0.047
	(0.011)	(0.006)
Final Facility	0.103	0.092
	(0.004)	(0.004)
Consign Time	0.220	0.190
	(0.004)	(0.004)

Notes. I run six regressions, each with its own matching scheme. For a given scheme, I randomly match pairs of observations and difference the Delivery Score and Average Action Time variables by pair. I then regress the differenced Delivery Score variable on an intercept and the differenced Average Action Time variable, reporting the coefficient estimates of the latter. The "no controls" specifications match the shipments by either the Buyer, the final facility reported on the track-package logs, or the consign time percentile. The "controls" specifications match the shipments by these variables in addition to Brand, Category, and Merchant.

t-statistic of 12.2. So removing extreme Delivery Scores does not overturn the result.

Fifth, an anonymous reviewer asked me to control for customer learning, because "if the same buyer has multiple transactions, s/he may gain knowledge on the delivery process over time." I do so in two ways. First, I rerun Table 6's OLS regressions with additional dummy variables that specify the number of shipments a given Buyer has received from Cainiao up until that point.[11] Second, I rerun Table 6's OLS regressions with the sample limited to each customer's first purchase. With control variables, I get an Average Action Time coefficient estimate of 0.138 in my first specification and 0.145 in my second specification, with a corresponding t-statistics of 47.4 and 26.2; without control variables, I get an Average Action Time coefficient estimate of 0.083 in my first specification and 0.086 in my second specification, with corresponding t-statistics of 29.5 and 25.8. So controlling for customer learning does not overturn the result.

Sixth, an anonymous reviewer requested cluster-robust standard errors, clustered by Merchants. I accommodate this request with the panel bootstrap, which resamples the data by Merchant (see Cameron and Trivedi, 2005, p. 377). This approach permits two observations to have correlated errors if and only if they share the same Merchant. The method is valid because the data are dispersed among clusters: The Herfindahl index of Merchant shares is 0.023. Rerunning Table 6's OLS regressions with the panel bootstrap yields the same coefficient estimates—0.081 with controls and 0.137 without—and with slightly more conservative standard errors—0.0062 with controls and 0.0212 without. Nevertheless, the new t-statistics—13.0 with controls and 6.45 without—are still strong. So adopting cluster-robust standard errors does not overturn the result.

9. Conclusion

Cainiao's customers leave higher delivery scores when the track-package activities they see gravitate toward the end of the shipping horizon. Figure 3's estimates suggest that increasing *one* action time from [0.10, 0.15) to [0.80, 0.85) increases the expected delivery score by an average of 0.0141 points, and Table 6's primary OLS estimates suggest that increasing the *average* action time from 0.15 to 0.85 increases the expected delivery score by 0.0565 points. On average, an extra day of shipping decreases the expected delivery score by 0.0443 points, so these interventions are roughly analogous to decreasing the shipping time by 0.0141/0.0443 = 0.318 and 0.0565/0.0443 = 1.28 days.

These results are consistent with the peak-end rule, which states that customers remember endings more vividly than beginnings. Accordingly, Cainiao should emphasize last-mile logistics, as the last mile is the most memorable mile. Or Cainiao could craft their messages to highlight later actions, for example, not reporting the initial consign actions would increase the average action time from 0.494 to 0.553.

The peak-end rule should apply to most service operations. For example, Redelmeier et al. (2003) showed that adding a needless resting period to the end of a colonoscopy improved patient impressions of the procedure:

> By random assignment, half the patients had a short interval added to the end of their procedure during which the tip of the colonoscope remained in the rectum. . . . As theorized, patients who underwent the extended procedure experienced the final moments as less painful (1.7 vs. 2.5 on a ten point intensity scale, $P < 0.001$), rated the entire experience as less unpleasant (4.4 vs. 4.9 on a 10 cm visual analogue scale, $P = 0.006$), and ranked the procedure as less aversive compared with seven other unpleasant experiences (4.1 vs. 4.6 with eight as the worst, $P = 0.002$).

Although it was published in a medical journal, the work is classic operations management: The researchers modify a repeated process (medical procedure) to reduce its perceived cost (recalled pain). The intervention challenges our operational insights—lengthening the procedure increases the flow time and decreases the throughput rate—and our basic intuitions—lengthening the procedure increases the total experienced pain. But in this case, the relevant quality measure is not the experienced pain but the recollected pain, as "Patients' memories of unpleasant medical procedures influence their decisions about future treatment choices" (Redelmeier et al. 2003, p. 187). And a patient's retrospective evaluation is not a naïve integral of instantaneous utilities, because the final utilities receive extra weight. Accordingly, the researchers

modify the process to end on a (relatively) painless note. This all's-well-that-ends-well logic applies more broadly: Judgment usually comes after the process, not during the process. This means we should pay special attention to how our service operations end.

That said, here's my attempt at an agreeable ending: What's more frustrating, waiting six weeks for the referees and one week for the editor or waiting one week for the referees and six weeks for the editor?

Endnotes

[1] Pendem and Deshpande (2018), Cui et al. (2019a), Li et al. (2019a), and Li et al. (2019b) also analyzed this data set as participants of the MSOM Data Driven Research Challenge. My work is closely related to Pendem and Deshpande's (2018) paper, which studies the sales effect of the customer delivery scores, and is somewhat related to Cui et al.'s (2019a) paper, which studies the sales effect of the temporary ban of a prominent shipper.

[2] There is also a confirmation action that I disregard, because it occurs after the customers file their delivery scores.

[3] Actually, Facility Count is the lesser of the number of reported facilities and eight. However, fewer than 0.02% of shipments have more than eight facilities.

[4] I do not include Consign Count dummies as controls because Consign Count always equals one (for now). And I do not include Action Count dummies as controls, because the Action Count is redundant given the Receive Count, Arrive Count, Depart Count, and Scan Count.

[5] Unless otherwise specified, my standard errors are calculated with the paired bootstrap, which is robust to general heteroskedasticity (Cameron and Trivedi 2005, p. 376).

[6] Suppose I adopted the null hypothesis that the Average Action Time has no effect on the Delivery Score. Under this null, the number of estimates, out of 100, that are significantly positive at the $p = .01$ level has a binomial (100, .01) distribution (ignoring dependencies across regressions). In this case, the probability of deriving at least 84 significantly positive estimates is less than 10^{-10}. Thus, I strongly reject the null hypothesis (Hedges and Olkin 2014).

[7] Following Stock et al. (2002), I confirm the strength of my three sets of instruments with three F tests. The first F test compares the R^2 of a regression of Average Action Time on Day of Week × Day Count × Shipper dummies to the R^2 of a regression of Average Action Time on Shipper dummies; the second F test compares the R^2 of a regression of Average Action Time on Day × Day Count × Shipper dummies to the R^2 of a regression of Average Action Time on Shipper dummies; and the third F test compares the R^2 of a regression of Average Action Time on consign Action Time decile × Shipper dummies to the R^2 of a regression of Average Action Time on Shipper dummies. With respective F-statistics of 298, 47.7, and 5,108, these three tests overwhelmingly reject the weak instruments hypothesis.

[8] By the way, Bill Schmidt has recently written an operational transparency article with Ananth Raman. Schmidt and Raman (2019) show that operational transparency can decrease the information asymmetry between a company and its investors, which in turn can make the company's stock price less sensitive to operational disruptions.

[9] In addition to multiple Shipping Speeds, there are now shipments with multiple shippers, product types, and consign actions. To accommodate these changes, I (i) run my regressions by Shipping Speed, (ii) define Shipper as the first shipper to handle the package, (iii) define Brand and Category as the brand and category of the first listed product type, (iv) include Consign Count dummies as control variables, and (v) derive the consign Action Time decile × Shipper instrumental variables from the time of the first consign action.

[10] Incidentally, Dennis Zhang and Ruomeng Cui have recently written an operational transparency paper with Achal Bassamboo. Cui et al. (2019b) exogenously shifted the inventory levels posted in Amazon Lightning Deals by randomly adding products to 10 fictitious Amazon carts. Cui et al. (2019b, p. 16) showed that reducing the available inventory levels increased demand rates, concluding that "real-time inventory information could serve as an effective lever for signaling popularity and attracting future customers."

[11] More specifically, I include 30 such dummy variables, with the last corresponding to the observations with 30 or more prior shipments.

References

Buell RW, Norton MI (2011) The labor illusion: How operational transparency increases perceived value. *Management Sci.* 57(9): 1564–1579.

Buell RW, Kim T, Tsay C-J (2017) Creating reciprocal value through operational transparency. *Management Sci.* 63(6):1673–1695.

Buell RW (2018) Email regarding operational transparency, August 28th.

Cainiao Network (2020) Cainiao tracking. Accessed July 21, 2020, https://www.ship24.com/couriers/cainiao-tracking.

Cameron AC, Trivedi PK (2005) *Microeconometrics: Methods and Applications* (Cambridge University Press, Cambridge, UK).

Cui R (2018) Personal communication with the author regarding feedback on robustness checks, August.

Cui R, Li M, Li Q (2019a) Value of high-quality logistics: Evidence from a clash between SF express and Alibaba. *Management Sci.*, ePub ahead of print December 5, https://doi.org/10.1287/mnsc.2019.3411.

Cui R, Zhang DJ, Bassamboo A (2019b) Learning from inventory availability information: Evidence from field experiments on Amazon. *Management Sci.* 65(3):1216–1235.

Hedges LV, Olkin I (2014) *Statistical Methods for Meta-analysis* (Academic press, Cambridge, MA).

Kahneman D, Fredrickson BL, Schreiber CA, Redelmeier DA (1993) When more pain is preferred to less. *Psych. Sci.* 4(36):401–405.

Li M, Liu X, Huang Y, Shi C (2019a) Integrating empirical estimation and assortment personalization for E-commerce: A consider-then-choose model. Working paper, Stephen M. Ross School of Business, University of Michigan.

Li X, Zheng Y, Zhou Z, Zheng Z (2019b) Demand prediction, predictive shipping, and product allocation for large-scale E-commerce. Working paper, Department of Management Sciences and Engineering, Stanford University.

Oliver RL (1980) A cognitive model of the antecedents and consequences of satisfaction decisions. *J. Marketing Res.* 17(4):460–469.

Osuna EE (1985) The psychological cost of waiting. *J. Math. Psych.* 29(1):82–105.

Pendem P, Deshpande V (2018) Logistics performance, ratings, and its impact on customer purchasing behavior and sales in E-commerce platforms. Working paper, Charles H. Lundquist College of Business, University of Oregon.

Redelmeier DA, Katz J, Kahneman D (2003) Memories of colonoscopy: A randomized trial. *Pain* 104(1-2):187–194.

Schmidt W, Raman A (2019) Operational disruptions, firm risk, and control systems. Working paper, Samuel Curtis Johnson Graduate School of Management, Cornell University.

Stock JH, Wright JH, Yogo M (2002) A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econom. Statist* 20(4):518–529.

Varey C, Kahneman D (1992) Experiences extended across time: Evaluation of moments and episodes. *J. Behavior Decision Making* 5(3):169–185.

Zhang D (2018) Personal communication with author regarding potential problems with robustness checks, August.