ONLINE SUPPLEMENT:
Mean Service Metrics: Biased Quality Judgment and the Customer-Server Quality Gap

Robert J. Batt, Jordan D. Tong
Wisconsin School of Business, University of Wisconsin-Madison, Madison, WI 53706,bob.batt@wisc.edu, jordan.tong@wisc.edu,

---

OS1. Proofs

Proof of Proposition 1. First, denote $\bar{s} = \frac{1}{m}\sum_{j=1}^{m} s_j$ and $\bar{c} = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{s_j} s_j$. We have:

$$\begin{aligned}\bar{c} - \bar{s} &= \frac{\sum_{j=1}^{m} s_j^2}{n} - \frac{n}{m} = \frac{m\left(\sum_{j=1}^{m} s_j^2\right) - n^2}{mn} = \frac{m\left(\sum_{j=1}^{m} s_j^2\right) - \left(\sum_{j=1}^{m} s_j\right)^2}{mn} \\ &= \frac{m}{n}\left[\frac{\left(\sum_{j=1}^{m} s_j^2\right) - \frac{1}{m}\left(\sum_{j=1}^{m} s_j\right)^2}{m}\right] = \frac{m}{n}\left[\left(\frac{1}{m}\sum_{j=1}^{m} s_j^2\right) - \left(\frac{1}{m}\sum_{j=1}^{m} s_j\right)^2\right] = \frac{\sigma_s^2 m}{n}.\end{aligned}$$

We can rewrite the expected server-level average quality and the expected customer-experienced quality as follows

$$\begin{aligned}E[\bar{SQ}] &= E\left[\frac{1}{m}\sum_{j=1}^{m}\frac{1}{s_j}\sum_{i=1}^{s_j} Q_{ij}\right] = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{s_j}\sum_{i=1}^{s_j} E[\alpha_j + \beta_j s_j + Y_j + Z_{ij}] \\ &= \frac{1}{m}\sum_{j=1}^{m}(\alpha + \beta s_j) + \frac{1}{m}\sum_{j=1}^{m}\frac{1}{s_j}\sum_{i=1}^{s_j} E[Y_j + Z_{ij}] \\ &= \alpha + \beta\frac{n}{m} = \alpha + \beta\bar{s} \\ E[\bar{CQ}] &= E\left[\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{s_j} Q_{ij}\right] = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{s_j} E[\alpha_j + \beta_j s_j + Y_j + Z_{ij}] = \frac{1}{n}\sum_{j=1}^{m}\left(s_j\alpha + \beta s_j^2\right) \\ &= \beta\left(\frac{1}{n}\sum_{j=1}^{m} s_j^2\right) + \frac{1}{n}\sum_{j=1}^{m}(s_j\alpha) = \beta\bar{c} + \alpha\end{aligned}$$

Taking the difference yields $E\left[\bar{CQ} - \bar{SQ}\right] = \beta(\bar{c} - \bar{s}) = \beta\frac{\sigma_s^2 m}{n}$.

Proof of Proposition 2. We can rewrite the server's impact on the server-level quality and customer-experience quality as follows:

$$\begin{aligned}
isq(\bar{q}') &= \frac{1}{m}\sum_{j=1}^{m}\bar{q}_j - \frac{1}{m-1}\sum_{j=1;j\neq j'}^{m}\bar{q}_j = \frac{1}{m}\sum_{j=1}^{m}\bar{q}_j - \left[\frac{1}{m-1}\sum_{j=1}^{m}\bar{q}_j - \frac{1}{m-1}\bar{q}'\right] \\
&= \frac{m-1-m}{m(m-1)}\sum_{j=1}^{m}\bar{q}_j + \frac{1}{m-1}\bar{q}' = \frac{1}{m-1}\left[\bar{q}' - \frac{1}{m}\sum_{j=1}^{m}\bar{q}_j\right] = \frac{1}{m-1}\left[\bar{q}' - \bar{sq}\right]. \\
icq(\bar{q}') &= \frac{1}{n}\sum_{j=1}^{m}\bar{q}_j s_j - \frac{1}{n-s'}\sum_{j=1;j\neq j'}^{m}\bar{q}_j s_j = \frac{1}{n}\sum_{j=1}^{m}\bar{q}_j s_j - \left[\frac{1}{n-s'}\sum_{j=1}^{m}\bar{q}_j s_j - \frac{1}{n-s'}\bar{q}'s'\right] \\
&= \left[\frac{1}{n} - \frac{1}{n-s'}\right]\sum_{j=1}^{m}\bar{q}_j s_j + \frac{1}{n-s'}\bar{q}'s' = \frac{-s'}{n(n-s')}\sum_{j=1}^{m}\bar{q}_j s_j + \frac{1}{n-s'}\bar{q}'s' \\
&= \frac{s'}{n-s'}\left[\bar{q}' - \frac{1}{n}\sum_{j=1}^{m}\bar{q}_j s_j\right] = \frac{s'}{n-s'}\left[\bar{q}' - \bar{cq}\right].
\end{aligned}$$

Define $\Delta(\bar{q}') = icq(\bar{q}') - isq(\bar{q}')$. We have:

$$\begin{aligned}
\Delta(\bar{q}') &= \frac{s'}{n-s'}\left(\bar{q}' - \bar{cq}\right) - \frac{1}{m-1}\left(\bar{q}' - \bar{sq}\right) \\
&= \bar{q}'\frac{s'm-n}{(n-s')(m-1)} - \bar{cq}\frac{s'}{n-s'} + \bar{sq}\frac{1}{m-1} \\
&= \bar{q}'\frac{s'm-n}{(n-s')(m-1)} - \bar{cq}\frac{s'}{n-s'} + \bar{sq}\frac{1}{m-1} - (\bar{sq}-\bar{cq})\frac{s'}{n-s'} + (\bar{sq}-\bar{cq})\frac{s'}{n-s'} \\
&= (\bar{q}'-\bar{sq})\frac{s'm-n}{(n-s')(m-1)} + (\bar{sq}-\bar{cq})\frac{s'}{n-s'} \\
&= \frac{s'm}{(n-s')(m-1)}\left[(\bar{q}'-\bar{sq})(\frac{s'-\frac{n}{m}}{s'}) + (\bar{sq}-\bar{cq})\frac{m-1}{m}\right]
\end{aligned}$$

The first term is always positive. Therefore, $\Delta(\bar{q}') > 0$ if and only if the following inequality holds:

$$(\bar{q}'-\bar{sq})(\frac{s'-\frac{n}{m}}{s'}) > (\bar{sq}-\bar{cq})\frac{m-1}{m} \tag{1}$$

Now, under the proposition's condition that $\left|(\bar{q}'-\bar{sq})(\frac{s'-\frac{n}{m}}{s'})\right| > \left|(\bar{sq}-\bar{cq})\frac{m-1}{m}\right|$, we can prove the 4 results:

- Let $\bar{q}' > \max\{\bar{sq}, \bar{cq}\}, s' > n/m$. Then, $icq(\bar{q}') > isq(\bar{q}')$ by (1), and $isq(\bar{q}') > 0$ since $\bar{q}' > \bar{sq}$. Thus, $0 < isq(\bar{q}') < icq(\bar{q}')$.
- Let $\bar{q}' > \max\{\bar{sq}, \bar{cq}\}, s' < n/m$. Then, $isq(\bar{q}') > icq(\bar{q}')$ by (1), and $isq(\bar{q}') > 0$ since $\bar{q}' > \bar{sq}$. Thus, $0 < icq(\bar{q}') < isq(\bar{q}')$.
- Let $\bar{q}' < \min\{\bar{sq}, \bar{cq}\}, s' > n/m$. Then, $isq(\bar{q}') > icq(\bar{q}')$ by (1), and $icq(\bar{q}') < 0$ since $\bar{q}' < \bar{cq}$. Thus, $icq(\bar{q}') < isq(\bar{q}') < 0$.
- Let $\bar{q}' < \min\{\bar{sq}, \bar{cq}\}, s' < n/m$. Then, $icq(\bar{q}') > isq(\bar{q}')$ by (1), and $icq(\bar{q}') < 0$ since $\bar{q}' < \bar{cq}$. Thus, $isq(\bar{q}') < icq(\bar{q}') < 0$.

OS2. Experiment Data Generation Details

Study 1 First, randomly decide whether the school is a positive or negative correlation school (with equal probability). Then, generate 15 courses in a school and their ratings using the following procedure for each course:

1. Generate a large course with probability 0.5, and a small course otherwise. The size of large courses are drawn from $N(100, 20)$, while the size of small courses are drawn from $N(30, 10)$.

2. For each course, generate underlying mean ratings from a "high" or "low" distribution depending on whether the school is designated as a positive or negative size/quality correlation school (i.e., draw from the high distribution if the school is large and the condition is positive correlation or if the school is small and the condition is negative correlation, and correspondingly for the low distribution). The "high" distribution is $N(9, 2)$, while the "low" is $N(5, 2)$, truncated to be between 0 and 10.

3. For each course, generate as many "student ratings" as indicated by the size of the course determined in Step 1. Each student rating is a random draw from a normal distribution with the underlying mean determined in Step 2 and a standard deviation of 1.5, again truncated between 0 and 10. From these student ratings, calculate the realized class mean and standard deviation. The realized size, mean, and standard deviation for each class are shown on the experiment document.

Study 2 We generated evaluation scores for the 30 teachers in a school using the following procedure:

1. We randomly generated evaluation scores for 15 "small enrollment" teachers with random enrollment sizes drawn from $N(30, 10)$. For each teacher, we randomly generated a single underlying target mean from $N(6, 2)$, truncated to be between 0 and 10. Then, the realized mean and standard deviations of the teacher ratings were simulated using random draws from a normal distribution with the mean equal to the underlying target mean and standard deviation of 1.5, again truncated to be between 0 and 10. We generate as many random draws as the enrollment size drawn above.

2. For each of the 15 "small enrollment" teachers, we generated a corresponding "large enrollment" teacher with similar impact to the student experienced average score. To do this, we set the enrollment to be three times the enrollment of the corresponding small enrollment teacher. Then, we set this teacher's average score such that its impact on students is approximately equal but slightly higher than its corresponding small enrollment teacher. To accomplish this matching, recall that a teacher who teaches s' students with average quality q' impacts the student experienced average quality by $\frac{s'}{N-s'}\left[q' - \bar{CQ}\right]$. Thus, a teacher who teaches $s'' = 3s'$ students who has the same impact must teach at the quality q'' which solves $\frac{s'}{N-s'}\left[q' - \bar{CQ}\right] = \frac{3s'}{N-3s'}\left[q'' - \bar{CQ}\right]$. Thus, $q'' = \frac{1}{3}\left(1 - \frac{2s'}{N-s'}\right)\left[q' - \bar{CQ}\right] + \bar{CQ}$. We chose to implement $\left[q' - \bar{CQ}\right]/3 + \bar{CQ}$ because its absolute

value is larger than $|q''|$, thereby making the large enrollment classes slightly more "deserving" to be assigned to the top five or bottom five and setting up a stronger test of our hypothesis. In practice, this difference was very small due to the large value of N in our experiment. It served only to break ties between matched teacher-pairs, resulting in exactly 3/5 of the true top five designated teachers being large enrollment teachers, and exactly 3/5 of the true bottom five being large enrollment teachers.

3. Finally, we generated 30 random teacher last names (http://random-name-generator.info), assigned them to the 30 scores, and sorted the list by the generated last name.

Study 3 First, randomly decide whether the school is a positive or negative correlation school (with equal probability). Then, generate 15 courses in a school and their ratings using the following procedure for each course:

1. Generate a large course with probability 0.5, and a small course otherwise. The size of large courses are drawn from $N(90, 10)$, while the size of small courses are drawn from $N(30, 10)$.

2. For each course, generate as many "student ratings" as indicated by the size of the course determined in Step 1. Each student rating is a random draw from from the distribution of $Q_{ij} = \alpha_j + \beta_j s_j + Y_j + Z_{ij}$, where $Y_j = 0$, $Z_{ij} \sim N(0, 1.5)$, but α_j and β_j differ by condition. In the positive correlation condition, $\alpha_j = 1, \beta_j = 2/30$. In the negative correlation condition, $\alpha_j = 9, \beta_j = -2/30$.

3. Finally, from these student ratings, calculate the realized class mean and standard deviation. The realized size, mean, and standard deviation for each class are shown on the experiment document.

Study 4 The random generation process for Study 4 was identical to Study 2.

OS3. Experimental Study Subject Characteristics

Studies 1 and 2 We used an online recruiting and scheduling system at a behavioral lab at a public university in the United States to recruit subjects. A total of 92 subjects participated in the study. 97% of participating subjects were full-time students, of which 82% were undergraduate and 18% were graduate students. 74% were female; 55% self-identified as white, 37% as Asian, and 3% as black. 95% lived in the United States for at least one year, while 74% lived in the United States for at least five years. 63% had taken at least one semester of statistics. Subjects completed both Study 1 and Study 2, in sequential order, and subjects took between 15 and 30 minutes to complete both studies.

Studies 3 and 4 We used the same online recruiting and scheduling system as in Studies 1 and 2. A total of 166 subjects participated in the study. 66% of participating subjects were full-time students, of which 78% were undergraduate and 22% were graduate students. 74% were female; 62%

self-identified as white, 27% as Asian, and 4% as black. 96% lived in the United States for at least one year, while 81% lived in the United States for at least five years. Subjects completed both Study 3 and Study 4, in sequential order, and subjects took between 10 and 20 minutes to complete both studies.

OS4. Emprical Evidence from Airlines

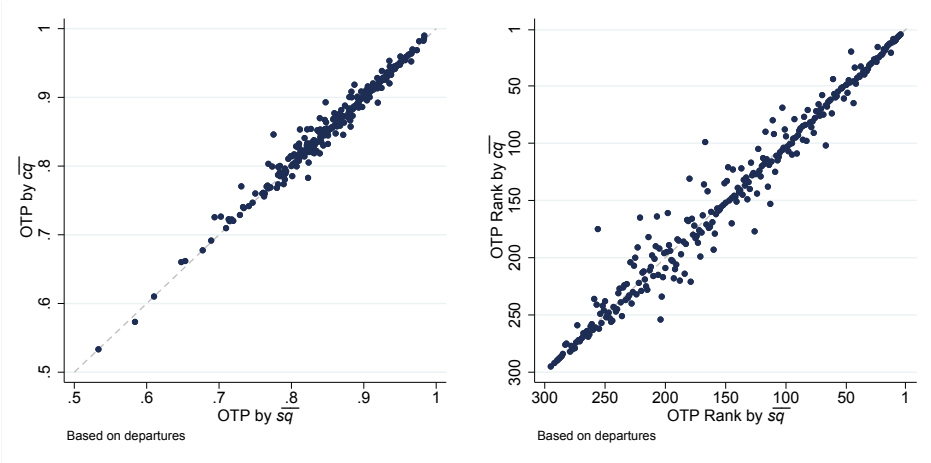
Another industry with plentiful quality data that is reported at the server level is the commercial airline industry. The quality metric on which we focus is on-time performance (OTP) which is defined by the federal government as the percent of departures or arrivals that occur prior to 15 minutes past the scheduled time. By law, airlines which carry at least 1% of total domestic scheduled-service passenger revenues must monthly report detailed flight and OTP information for domestic flights to the DOT (Code of Federal Regulations 1987). This data is publicly available and is used by various third-party organizations to create airline and airport performance rankings. For example, airline industry information aggregator OAG Aviation Worldwide Limited publishes a monthly report ranking airports by OTP (OAG Aviation Worldwide Limited 2017). Airline and airport managers tout these results as they seek to woo both consumers and industry partners (e.g., Delta Air Lines 2018, Southwest Airlines Co. 2018). However, the industry standard OTP metric is a server-level quality metric and does not account for the average customer experience. Thus, just as with the Higher Education setting, the metrics may differ significantly, leading to differences in ranking.

We first focus on the OTP of airports, thus we consider passengers (Level 1), flights (Level 2), and airports (Level 3). This metric may be used by airport managers trying to attract a new airline, or by airline managers comparing performances across airports in their network. We collect flight departure data from the DOT for the month of May 2017. These data contain departure timing for all domestic flights flown by the 12 largest domestic airlines from 296 domestic airports. Ideally, we would use the actual number of passengers on each flight to construct the customer experience OTP metric, however this information is not publicly available and airlines view it as highly confidential. Thus, we proxy for the number of passengers with the number of seats on the plane (Deshpande and Arıkan (2012) similarly use seats as a proxy for passengers).

For each airport we calculate $\bar{sq}$ and $\bar{cq}$ OTP and the corresponding $\bar{sq}$ and $\bar{cq}$ ranks. Figure 1a plots the airports by the two means. Points above the 45 degree line (dashed line) have $\bar{cq}$ greater than $\bar{sq}$ and thus a positive customer-server quality gap, and points below the line have $\bar{cq}$ less than $\bar{sq}$, and a negative customer-server quality gap. The mean absolute difference between $\bar{cq}$ and $\bar{sq}$ OTP is 0.6 percentage points, and for 25% of the airports the mean absolute difference is greater than 1 percentage point. While these differences are small in absolute terms, they are sufficiently

Figure 1 **Airport on-time performance (all airlines)**

(a) OTP mean (b) OTP rank



large to lead to many rank differences. Figure 1b plots the airports by the OTP ranks. Only 10% of the markers fall on the 45 degree line. Thus, 90% have a different $\bar{cq}$ rank than $\bar{sq}$ rank, with the median absolute rank error of 4, mean of 8.7, and maximum of 80 (the Blountville, TN airport (TRI) is ranked 250 by $\bar{sq}$ and 170 by $\bar{cq}$).

Within a single airline, managers may want to evaluate the performance of the airports already in their network to determine the impact each airport has on the airline's overall OTP metric. We calculate the impact that each airport in American Airlines' network has on the overall airline $\bar{cq}$ and $\bar{sq}$ (We find similar results for other airlines.). In this case, the airport is the server and the passenger is the customer. Thus the server-level metric simply averages the mean OTP by flight of each airport regardless of the number of flights and passengers at the airport. The customer-experienced metric is created by taking the weighted average across the airports, weighting by the the number of passengers on each flight.

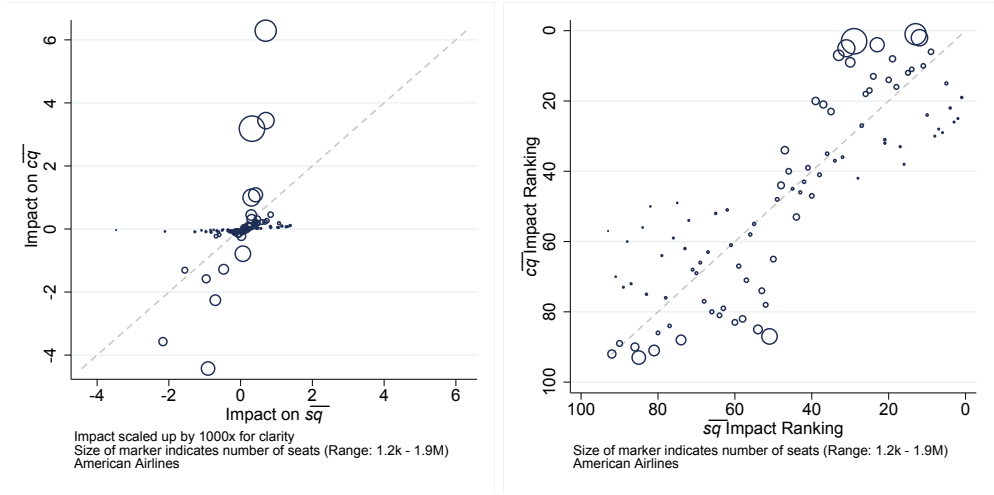
For American Airlines, the May 2017 $\bar{sq}$ is 0.819 and $\bar{cq}$ is 0.832. Figure 2 compares the 93 airports used by American Airlines by these metrics and related rankings. Figure 2a plots each airport by its impact on $\bar{sq}$ and $\bar{cq}$. We see that there is very little agreement between the metrics (few points on the 45 degree line), and there is a striking difference in the horizontal and vertical range of the points. This shows that most airports have relatively little impact on $\bar{cq}$ (points close to zero on the y-axis), and it is the airports with the largest number of flights, as indicated by the size of the marker, that have by far the largest impact on $\bar{cq}$ (both positive and negative). In contrast, the distribution of impacts on $\bar{sq}$ appears more even, and busy airports do not have particularly large impacts.

Figure 2b compares the impact ranks based on the two metrics. Again, there is little agreement between the rankings. Points above the 45 degree line are under-ranked by the server-level metric and

points below the line are over-ranked by the server-level metric. The vertical distance between the point and the line indicates the degree of under/over ranking. Again, the size of the marker indicates the number of flights from the airport, and we note that the server-level ranking systematically under-ranks the impact of large airports that positively impact OTP and systematically under-ranks the impact of large airports that negatively impact OTP. Together, these graphs show that if managers use server-level metrics to allocate resources, direct improvement efforts, or reward or punish performance, they are unlikely to focus on the airports that will yield the most benefit to the average customer experience.

Figure 2 Airport OTP impact (American Airlines)

(a) $\bar{sq}$ and $\bar{cq}$ impact (b) $\bar{sq}$ and $\bar{cq}$ impact rank



References

Code of Federal Regulations. 1987. Title 14, chapter 2, part 234.

Delta Air Lines. 2018. Delta tops U.S. global carriers in 2017 on-time performance. online. URL `https://pro.delta.com/content/agency/us/en/news/news-archive/2018/january-2018/delta-tops-u-s--global-carriers-in-2017-on-time-performance.html`.

Deshpande, Vinayak, Mazhar Arıkan. 2012. The impact of airline flight schedules on flight delays. *Manufacturing & Service Operations Management* **14**(3) 423–440.

OAG Aviation Worldwide Limited. 2017. OAG flightview: Airports monthly OTP URL `https://www.oag.com/may-2017-airports-on-time-performance`.

Southwest Airlines Co. 2018. *2017 Annual Report to Shareholders*. URL `http://investors.southwest.com/financials/company-reports/annual-reports`.