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,

Problem Definition: People often make service quality judgments based on information about the quality of each server even though they care primarily about the quality each customer experiences. When and how do server-level quality metrics differ from customer-experienced ones? Can people properly account for these differences, or do they drive human judgment and decision biases?

Relevance: Biased judgments about service quality can cause governments to fund programs suboptimally, organizations to promote the wrong employees, and customers to make disappointing purchases. We further our understanding of the role that cognitive biases play in services and how to manage quality information in light of them.

Methodology: We use a mathematical model to define the gap between server-level and customer-experienced quality metrics. We use secondary data in the context of the higher education industry to quantify the customer-server quality gap in practice. We construct a behavioral model to derive hypotheses about how environmental factors impact the direction and magnitude of judgment biases. Controlled laboratory experiments test the hypothesized biases and mitigation techniques.

Results: Our empirical study reveals that the two measures differ enough to drive significant differences in the rank order of school majors, teachers, and airports. Our experiments support our main conjecture that judgments and decisions about customer-experienced metrics are biased towards server-level metrics. Consequently, (1) judgments about customer-experienced quality are biased high/low when quality and server load are negatively/positively correlated; (2) judgments about a server’s absolute impact on customer experience are biased high/low when a server has smaller/larger load than average; and (3) providing customer-experienced quality metrics mitigate these biases.

Managerial Implications: Our results help identify when and why service quality metrics are likely to mislead judgments and bias decisions, as well as who is likely to benefit from such biases. The results also guide system designers on how to report metrics when seeking to help support effective decision-making.

Key words: Behavioral Operations, Service Operations, Experiments, Empirical Research, Cognitive Judgment Bias, Inspection Paradox

History: February 2019
1. Introduction

Services represent about 80% ($10 trillion) of the US private-sector gross domestic product (Central Intelligence Agency 2017). Biased judgment of the quality of such services can have important and widespread consequences. It can cause governments to fund programs suboptimally, organizations to promote the wrong employees, and customers to make undesirable purchases. This paper argues that a seemingly innocuous data aggregation technique commonly used in services can potentially lead to systematic and meaningful biases in how people judge service quality.

To illustrate this data aggregation issue, consider the following two examples in education:

**Example 1.** Ben is a college freshman picking a major. Being resourceful, he decides to use course ratings information available from the school website to inform his decision. Each major posts the average course rating over all its classes. He chooses a major for which the average course rating is 8/10. Three years later, Ben is disappointed with his major experience. He looks back at the ratings he gave his classes and realizes that they average to only 6/10. He asks others within his major and finds that they also gave average ratings significantly less than 8/10. He is convinced that the major falsely advertised: How can the average class rating be 8/10 when he and his fellow classmates have all experienced a lower average rating?

**Example 2.** Sarah is a college dean heading the faculty teaching awards committee. The purpose of the committee is to recognize the top five instructors who have the greatest impact on the school based on course evaluation scores. Someone suggests they simply allocate awards based on the average score received by each instructor over the past year. Sarah feels that something may not be quite right: she has a lot of interactions with students and her impression is that the instructors who have the largest impact at the school don’t make the list. However, she wants the recommendations to be backed by data, so she convinces herself that her opinions must be biased: How can her impressions of the most impactful instructors be correct if they don’t have the highest average evaluations?

Both Ben and Sarah feel their experience doesn’t match the data, and conclude that either the data was wrong or their experience wasn’t normal. While their conclusions may be true, we argue in this paper that there is a completely different plausible explanation for this mismatch: the data presented to Ben and Sarah reflect the quality at the course or instructor level, but Ben and Sarah are implicitly concerned with the quality experienced by students. These two perspectives yield different metrics, but Ben and Sarah fail to fully recognize this difference.

For Ben, it is mathematically possible that the average course rating across the major is in fact 8/10 while simultaneously the average student-experienced rating in the major is much lower – even 6/10. This is possible because small classes (which only a few students experience) may have high ratings while large classes (which many students experience) have low ratings - creating a high rating
Table 1 A Small Ratings Example

<table>
<thead>
<tr>
<th>Class/Teacher</th>
<th>Avg. Rating</th>
<th># Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9.7</td>
<td>9</td>
</tr>
<tr>
<td>B</td>
<td>9.6</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>9.6</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>9.5</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>9.4</td>
<td>15</td>
</tr>
<tr>
<td>F</td>
<td>8.9</td>
<td>79</td>
</tr>
<tr>
<td>G</td>
<td>8.9</td>
<td>18</td>
</tr>
<tr>
<td>H</td>
<td>7.1</td>
<td>47</td>
</tr>
<tr>
<td>I</td>
<td>5.7</td>
<td>247</td>
</tr>
<tr>
<td>J</td>
<td>5.3</td>
<td>327</td>
</tr>
<tr>
<td>K</td>
<td>4.2</td>
<td>98</td>
</tr>
</tbody>
</table>

Class Mean: 8.0 78
Student Mean: 6.0 218

when averaging over the classes, but a low rating when averaging over students’ experiences. To illustrate, Table 1 presents an example of a major where the average over the classes is indeed 8.0, but the average over the students’ experiences is only 6.0 (assuming students take the same number of classes). Therefore, Ben is disappointed not because the majors necessarily lied about their ratings, but because they presented ratings that Ben could not correctly interpret. The underlying reason for this difference also drives the phenomenon that the average class size over the classes is only 78 students, but the average class size over student experiences is a much larger 218 - only a few students experience small classes, but many students experience large classes.

Similarly for Sarah, it is mathematically possible that she has perfect impressions about which instructors have the highest impact on the average student-experienced ratings while at the same time those instructors do not have the highest average course evaluations - the metric used by the committee. This is possible because an instructor’s impact on the average quality that students experience takes into account the fact that more students are impacted by instructors who teach larger classes. In contrast, the committee’s ranking implicitly considers the impact an instructor has on the the average quality across instructors, which does not depend on the number of students per instructor. For example, in Table 1, teacher A has the highest average rating (9.7), but teacher F (8.9) raises the overall student-experienced average significantly more because she teaches more students. In fact, even teacher H (7.1) has a greater positive impact on the average student experience than teacher A. Therefore, the mismatch between Sarah’s impressions and the committee’s rankings isn’t necessarily because her impressions are wrong, but because the committee didn’t calculate an impact statistic consistent with her intuitive student-focused definition of impact.

Both Ben’s and Sarah’s examples point to a larger and more general issue that can impact any service organization that aggregates data by servers when it actually cares about average customer-experienced quality. To illustrate the issue, we introduce the following organization structure and
terminology (Figure 1). We consider service organizations with multiple servers each serving multiple customers. Groups of servers are organized into units, and multiple units may be grouped into an organization. We refer to customers as Level 1, servers as Level 2, units as Level 3, and organizations as Level 4.

In Example 1 above, the college is the organization, each major is a unit, each class within a major is a server in the unit, and Ben is a customer trying to select the best unit within the organization. Ben was misled because the majors reported the quality of the average class (Level 2), but he cared about the quality experienced by the average student (Level 1). Many students refer to reports from sources like the US News and World Report college rankings, which uses quality metrics from the instructor-perspective, such as class size, in their rankings (Morse et al. 2017), but these may not be reflective of the class sizes the average student experiences at these schools.

In Example 2, Sarah is the dean of a school (unit) who is trying to determine which teacher (server) has the largest impact on the average quality experienced by the students (customers). Sarah was misled because she cared about the impact a teacher had on the average student (Level 1), but the committee reported rankings based on the impact to the average teacher (Level 2). Similarly, many schools may promote professor A but not professor B because A has better average course ratings, even though B may actually be better for the overall student-experienced ratings.

We can see similar service metric confusion in other industries as well. In the airline industry, federal law requires airlines to provide monthly on-time performance (OTP) metrics by flight number on their websites. These statistics influence industry rankings and customer behavior. For example, we can think of the passengers as customers, a single flight instance as a server, and a flight number (in a given month) as a unit. Customers buying a flight from Delta may observe that 95% of the flight
instances of a given flight number arrived on-time (Level 2). However, the percent of passengers who experienced on-time arrivals on that flight number (Level 1) could be significantly different than 95% if the individual instances of the flight number carry different numbers of passengers.

Another example occurs in the restaurant industry. As one employee at a large fast food chain told us, workers and managers in that industry are often incentivized based on the percent of orders they complete correctly and within a target time. If we consider the diners as customers, the order to be a server, and the store to be the unit, we see that because orders serve varying numbers of customers an order-level metric (Level 2) does not necessarily reflect the percent of customers that receive on-time service (Level 1), which is a metric that is more closely aligned with what services care about.

In this paper, we mathematically formalize the definitions of server- and customer-experienced quality metrics via a stylized model setting, and we use secondary data from the context of higher education to examine whether there are meaningful differences between the two types of metrics in practice. We also develop a behavioral model which we use to derive four hypotheses which we test with a series of controlled laboratory experiments. Our main conjecture is that when presented server-level data, average judgments about customer-experienced quality are biased towards server-level metrics. The results provide support for our theorized mechanism, and they also demonstrate the potential consequences for decision-making and the value of implementing mitigation strategies by providing decision-makers with data and metrics that correspond to their objectives.

2. Related Literature

This paper contributes to a body of work sometimes referred to in the psychology literature as naïve sampling, the naïve intuitive statistician, or a cognitive-ecological approach to studying judgment biases (Fiedler 2000, Fiedler and Juslin 2006, Juslin et al. 2007). A common theme in this literature is an emphasis on studying the structure of the information given to the decision-maker and the biases inherent in that information, rather than focusing on ways that humans introduce bias when processing or interpreting given data. Thus, the researcher often explains observed biases by assuming that people are quite proficient at basic descriptive statistics (i.e., calculating the mean and variance of a given sample), but fail to make necessary sophisticated statistical corrections to correct for non-representative samples they are given. For example, people generally fail to correct for the fact that the variance of a small sample is a biased estimator for the underlying variance (Kareev 2000) or that sample success percentages are biased if the user does not sample data in a representative manner (Fiedler et al. 2000). Such lack of metacognition to correct for the mis-representativeness of given samples is not limited to students or managers, but can also persist in researchers trained in statistics, too (e.g., Tversky and Kahneman 1971, Miller and Sanjurjo 2016).
In this paper, the decision-maker is provided a sample of server qualities that are not a representative sample of customer-experienced qualities. We focus on how this structure, combined with the lack of metacognition to correct for it, explains quality judgment biases. In this way, our paper can be viewed as contributing to this body of work. We further discuss this body of work in §5 when we motivate our hypotheses.

In operations management, our paper contributes to a burgeoning area that examines the implications of cognitive judgment biases in operations settings. Much of this work concentrates on inventory and supply chain management settings. Ren and Croson (2013) provide evidence that an overprecision bias can drive the pull-to-center effect in newsvendor order decisions (Bolton and Katok 2008, Schweitzer and Cachon 2000). Individuals also tend to underweight the supply line in serial supply chains, which is a contributor to the behavioral bullwhip effect (Croson and Donohue 2006, Narayanan and Moritz 2015). In time-series demand forecasting, Kremer et al. (2011) show that people may over- or under-react to fluctuations in demand due to neglecting the underlying system dynamics (see also Massey and Wu 2005). By examining human behavior in hierarchical forecasting processes, Kremer et al. (2016) also studies cognitive biases and aggregation issues. Feiler et al. (2013) provide evidence that people make biased judgments about customer demand when basing those judgments on sales data which do not reflect lost sales. Cui and Zhang (2018) provide evidence that cognitive limitations in strategic-reasoning capabilities as proposed in the cognitive hierarchy model (Camerer et al. 2004) well model behavior in supply chain capacity allocation games. Stangl and Thonemann (2017) is perhaps most closely related to our work in that it also emphasizes how, given mathematically equivalent information, the metric used can significantly affect judgments because of human cognitive limitations. Specifically, they show that managers tend to judge an “inventory turn rate” metric differently than an equivalent “days of supply” metric because of a tendency to use linear thinking (Larrick and Soll 2008).

In service settings, some recent papers present laboratory results on how customers perceive service quality. Gans et al. (2007) study how customers behave in multi-arm bandit problems, which mimic how services compete on quality over time, and examines which heuristics better capture human behavior. Kremer and Debo (2016) study human customer purchasing behavior in a setting where customers try to infer service quality by wait times, and customers may make random decision errors (see also Su 2008). Buell et al. (2017) show that operational transparency, allowing customers and servers to see each other during service, increases both perceived and actual service quality. Our paper complements this literature by focusing on a different cognitive bias that often arises in the context of service quality judgment.

Other related work in service operations includes papers that incorporate behavioral deviations from rationality with regards to preferences (as opposed to judgments). For example, even if total...
average quality over time is the same, customers may care about the sequence of the service quality they receive – which impacts how firms sequence the services they provide (Das Gupta et al. 2015). Similarly, customers may prefer not to undergo an immediate unpleasant experience even if it generates significant future benefit – which impacts how services design their pricing contracts with customers (Plambeck and Wang 2013). In access services, customers may also incur a psychological cost associated with sunk costs as well as a psychological transaction cost for payment at the time of consumption (Leider and Şahin 2014). Service performance metrics also play a central role in the empirical research in Song et al. (2017), which studies their role in learning best-practices and achieving operational improvement.

Finally, the mathematical characterization of the customer-server quality gap in this paper is closely related to other “paradoxes” that have been discussed in mathematics, operations research, accounting, and economics. Perhaps most closely related is the so-called “inspection paradox,” (Ross 2014) which has been notably applied to renewal theory (Ross 2003, Angus 1997). A fundamentally similar mathematical phenomenon drives the so-called “class size paradox” (Feld and Grofman 1977) and “friendship paradox” (Feld 1991, Feiler and Kleinbaum 2015). In each of these paradoxes, some attribute (e.g., class size, number of friends, lifetime) of a randomly chosen entity (e.g., student, friend, circuit board) is larger than the average over all the entities because the attribute value is actually a driver of the probability of being selected at random. We discuss the relationship between our settings and these types of paradoxes in §5. In short, our paper can be viewed as an extension of this type of mathematical paradox formalized to address the issue of service quality metrics, with empirical evidence of the magnitude of the potential bias in service settings and experimental evidence on how people fail to appreciate the paradox.

3. Model Setting and Quality Metrics

To formalize our main arguments, we consider the following simple model setting using the terminology introduced in §1. A unit consists of \( m \) servers. Let \( s_j \) denote the number of customers served by server \( j \) (a server’s “load”), where the subscript \( j = 1, ..., m \) refers to the server index. Assume that not all servers serve the exact same number of customers. Also, assume each customer is assigned to exactly one server. Then, \( n = \sum_{j=1}^{m} s_j \) is the total number of customers. Finally, define \( q_{ij}, i = 1, ..., s_j, j = 1, ..., m \) the quality delivered by server \( j \) to its \( i \)th customer, where higher values are associated with better quality.

**Average Quality Metrics** We now define two average quality metrics, one across the servers and the other across the customers. At the server level (Level 2), the average quality is

\[
\bar{q} = \frac{1}{m} \sum_{j=1}^{m} \bar{q}_j
\]
where $\bar{q}_j = \frac{1}{s_j} \sum_{i=1}^{s_j} q_{ij}$ is the average quality server $j$ delivers its customers. We refer to this statistic as the server-level quality because it is an average across servers.

From the customers’ perspective (Level 1) the average quality experienced is

$$\bar{c}q = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{s_j} q_{ij} = \frac{1}{n} \sum_{j=1}^{m} s_j \bar{q}_j.$$  \hfill (2)

We refer to this statistic as the customer-experienced quality because it is an average across customers.

Now, $\bar{s}q$ does not necessarily equal $\bar{c}q$. We define the customer-server quality gap as $\bar{c}q - \bar{s}q$.

**Server Impact Metrics** What is a server’s impact on the average quality metrics considered above? Consider a server of interest $j'$, and denote $\bar{q}'$ the average quality delivered to its $s'$ customers. Let $\bar{q} = \{\bar{q}_j\}_{j=1,...,m}$ be a vector of all the servers’ average qualities, and denote $\bar{q} \setminus \bar{q}'$ the vector of all the servers’ average qualities minus that of the server under consideration (size $m - 1$).

To assess a server’s impact on the server-level quality, we evaluate how much $\bar{s}q$ changes when one removes a server with average quality $\bar{q}'$. Then, we can define a server’s $\bar{s}q$ impact as:

$$isq(\bar{q}') = \bar{s}q(\bar{q}) - \bar{s}q(\bar{q} \setminus \bar{q}')$$

$$= \frac{1}{m} \sum_{j=1}^{m} \bar{q}_j - \frac{1}{m-1} \sum_{j=1,j\neq j'}^{m} \bar{q}_j. \hfill (3)$$

Similarly, we can assess a server’s impact on the customer-experienced quality by calculating how much $\bar{c}q$ changes when one removes a server with quality $\bar{q}'$. Define a server’s $\bar{c}q$ impact as

$$icq(\bar{q}') = \bar{c}q(\bar{q}) - \bar{c}q(\bar{q} \setminus \bar{q}')$$

$$= \frac{1}{n} \sum_{j=1}^{m} \bar{q}_j s_j - \frac{1}{n - s_j'} \sum_{j=1,j\neq j'}^{m} \bar{q}_j s_j. \hfill (4)$$

Finally, we define the server impact gap as $icq(\bar{q}') - isq(\bar{q}')$.


Having defined sever-level quality and customer-experienced quality metrics, a natural question is whether there exist meaningful differences between these two types of metrics in practice. To address this question, we study field data from the higher education industry.

As seen in the motivating hypothetical examples in §1, higher education is a setting in which the data commonly made available to customers (students) may not represent well the information that customers need to make informed decisions. For example, as previously mentioned, the highly influential US News and World Report college rankings takes into account class level metrics such
as the percent of classes with 20 or fewer students (Morse et al. 2017), but this does not necessarily represent the average proportion of classes that students experience with under 20 students. Yet, prior research has shown that the US News and World Report ranking has a large effect on the choices prospective students make and the resulting profile of a school’s student body (Monks and Ehrenberg 1999, Meredith 2004). Administrators can also be influenced by the rankings. For example, Richard Freeland, former president of Northeastern University in Boston, made it his explicit goal to help the school climb from 162 in the US News ranking into the top 100 when he became university president in 1996. Knowing that class size was a factor in the ranking, Freeland capped many classes at 19 students (Kutner 2014). Prior work has documented many such effects rankings have had on administrative decisions (see Dill and Soo (2005) for a review of this literature).

To get a sense of magnitude of the customer-server quality gap and the server impact gap in a higher education setting, we focus on average quality metrics used by the students and administration at a large business school. We collect student feedback data from 2,320 undergraduate business classes from 2011 through 2016 at a major American university. This data is collected in class at the end of each semester via an anonymous paper survey instrument. The survey presents 25 statements about various aspects of quality (e.g., “The course was well organized.”, “The instructor was well prepared for class.”). Students respond on a five-point Likert scale ranging from “Strongly Agree” to “Strongly Disagree.” These responses are coded to values 1 through 5 with “Strongly Agree” being 5, and high scores are preferred on all questions. Once the information is collected and tabulated, summary data (mean and standard deviation for each question) for each class is made available to students. The summary information also contains the number of students in each class. In this setting, students are the customers, classes are the servers, majors are the units, and the business school is the organization.

### Table 2 Enrollment statistics

<table>
<thead>
<tr>
<th>Dept.</th>
<th>Classes</th>
<th>Total Students</th>
<th>Mean ($sq$)</th>
<th>Standard Deviation</th>
<th>Mean ($cq$)</th>
<th>Rank (by $sq$)</th>
<th>Rank (by $cq$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>142</td>
<td>5,142</td>
<td>36.2 (1.2)</td>
<td>14.5</td>
<td>42.0 (0.2)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>534</td>
<td>23,575</td>
<td>44.1 (1.1)</td>
<td>25</td>
<td>58.2 (0.3)</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>276</td>
<td>12,736</td>
<td>46.1 (2.2)</td>
<td>37.3</td>
<td>76.2 (0.5)</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>A</td>
<td>572</td>
<td>26,528</td>
<td>46.4 (2.5)</td>
<td>39.6</td>
<td>122.9 (0.8)</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>G</td>
<td>126</td>
<td>6,225</td>
<td>49.4 (5.6)</td>
<td>63.1</td>
<td>129.3 (1.3)</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>D</td>
<td>230</td>
<td>13,909</td>
<td>55.6 (3.5)</td>
<td>56.1</td>
<td>112.0 (0.7)</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>I</td>
<td>77</td>
<td>4,804</td>
<td>62.4 (7.9)</td>
<td>69.3</td>
<td>138.4 (1.2)</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>E</td>
<td>240</td>
<td>15,410</td>
<td>64.2 (4.2)</td>
<td>64.7</td>
<td>129.2 (0.8)</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>H</td>
<td>103</td>
<td>6,879</td>
<td>66.8 (4.5)</td>
<td>45.4</td>
<td>97.3 (0.5)</td>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

Consider a student that is trying to select a major from amongst the nine majors offered in the business school (e.g., marketing, accounting, operations). One measure of quality is simply the class
size itself (Feld and Grofman 1977). Table 2 presents information about class enrollment across the majors. For some majors the difference is quite large (e.g., Major G, 49.4 vs. 129.3). Perhaps, a student picks Major A under the belief that the mean class size will be about 46, which isn’t much larger than Major F with the smallest mean enrollment of 36.2. She is likely to be greatly disappointed at the end of college because her true (customer-level) expected class size is closer to 123 students, 2.7 times larger than she expected. In this data, \( \bar{s}q \) and \( \bar{c}q \) are never identical, and their differences are enough to change the rankings of majors by mean enrollment for six of the nine majors. Thus, for example, Major H might be overlooked by students due to its last place ranking based on \( \bar{s}q \), when by \( \bar{c}q \) it ranks 4th smallest.

<table>
<thead>
<tr>
<th>Major</th>
<th>Mean (( \bar{s}q ))</th>
<th>Mean (( \bar{c}q ))</th>
<th>Quality Gap ( (\bar{c}q - \bar{s}q) )</th>
<th>Standardized Difference ((\bar{c}q - \bar{s}q) / \text{Standard Deviation of } \bar{c}q )</th>
<th>Rank (by ( \bar{s}q ))</th>
<th>Rank (by ( \bar{c}q ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>4.65</td>
<td>4.38</td>
<td>-0.07</td>
<td>-0.9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>4.54</td>
<td>4.50</td>
<td>-0.04</td>
<td>-0.5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>F</td>
<td>4.47</td>
<td>4.48</td>
<td>0.01</td>
<td>0.1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>H</td>
<td>4.46</td>
<td>4.51</td>
<td>0.05</td>
<td>0.6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>4.46</td>
<td>4.34</td>
<td>-0.12</td>
<td>-1.5</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>A</td>
<td>4.44</td>
<td>4.35</td>
<td>-0.09</td>
<td>-1.1</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>D</td>
<td>4.41</td>
<td>4.46</td>
<td>0.05</td>
<td>0.6</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>4.41</td>
<td>4.39</td>
<td>-0.02</td>
<td>-0.3</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>G</td>
<td>4.37</td>
<td>4.43</td>
<td>0.06</td>
<td>0.8</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

Similar errors can occur with quality metrics other than class size. To demonstrate the potential for such error, we consider one summary metric on which the school administration focuses: Instructor Quality Metric (IQM). IQM is constructed for each survey by taking the average score over 14 of the survey questions which are focused on the instructor (e.g., organization, preparedness, enthusiasm, empathy).\(^1\) Table 3 shows \( \bar{s}q \) and \( \bar{c}q \) for IQM. We see that the customer-server quality gap ranges from 0.06 to -0.12. The standardized difference column helps put this in perspective by dividing by the standard deviation of the \( \bar{c}q \) column. Observe that the difference between \( \bar{c}q \) and \( \bar{s}q \) can be up to 1.5 standard deviations in magnitude. We note that (perhaps surprisingly) some majors exhibit a positive customer-server quality gap, while others exhibit a negative quality gap. Thus, if students form (biased) expectations of quality when they select a major, majors with a negative quality gap will likely create disappointment and majors with a positive quality gap will likely exceed expectations. We return to the issue of the sign of the quality gap in §5.

The differences between the \( \bar{s}q \) and \( \bar{c}q \) quality metrics in Table 3 also have possible implications from an administrator’s perspective, for example, when the teaching quality of majors is a factor

\(^1\) Similar results can be shown for each of the survey questions individually, as well.
in allocating resources or in allowing departments to hire new faculty. We see that the ranking of the majors differs depending on whether $\bar{s}q$ or $\bar{c}q$ is used, with differences in ranking by as much as four positions (Major C, 5 vs. 9).

**Figure 2** Comparison of Customer and Server Impact

(a) $\bar{s}q$ and $\bar{c}q$ impact

(b) $\bar{s}q$ and $\bar{c}q$ impact rank

Another way that administrators can use class quality metrics is to identify the instructors that have the greatest impact on the average student experience for recognition or reward. For example, at the school that provided the data for this analysis, the Instructor Quality Metric $isq(\bar{q})$ rank is a major factor in determining teaching award recipients each year. We use (3) and (4) to calculate IQM $isq(\bar{q})$ and $icq(\bar{q})$ for the 2015-2016 academic year (448 classes), and we plot these two impact metrics in Figure 2a.\(^2\) We see that there is very little agreement between $isq(\bar{q})$ and $icq(\bar{q})$ (few points on the 45 degree line), indicating that a server impact gap exists. The vertical distance between a marker point and the 45 degree line represents the server impact gap. We note that the server-impact gap is positive for some servers, while negative for others. Interestingly, Figure 2a shows that the range of the magnitude of impact on $\bar{s}q$ is quite small (horizontal spread), whereas the range of the magnitude of impact on $\bar{c}q$ (vertical spread) is much larger. Further, the size of the circles indicates the relative class size, and we see clearly that it is the largest classes that have by far the largest impact on $\bar{c}q$, both positive and negative, which is as expected given Equation 4. We return to the issue of the sign and magnitude of the server impact gap in §5.

\(^2\) For ease of exposition, assume each instructor teaches one class. With minor modifications, we could allow for instructors who teach multiple classes.
Figure 2b shows a similar plot for the ranks of $isq(\bar{q}')$ and $icq(\bar{q}')$ rather than the values themselves. Again, we see many points off the 45 degree line indicating disagreement between the server-level and customer-level rankings. In fact, there is no overlap in the rankings of the top 20 instructors according to the two metrics, and of the 10 instructors with the largest impact on $icq(\bar{q}')$, the best any of them do in the $isq(\bar{q}')$ rankings is 27th. As mentioned previously, at the data-source school, $isq(\bar{q}')$ rank is a major factor in annual teaching awards. If administrators incorrectly use $isq(\bar{q}')$ rankings as a measure of $icq(\bar{q}')$, then teachers that fall above the 45 degree line are under-ranked by administrators and teachers below the line are over-ranked. The marker size again indicates relative class size and we see that it is the large classes that tend to be most underappreciated. For example, a large class with 269 students and an IQM $\bar{q}$ of 4.51 is ranked 37th according to the $icq(\bar{q}')$ ranking, but is ranked 257th according to the $isq(\bar{q}')$ ranking, a mismatch of 220 positions.

5. Quality Gap Drivers & Behavioral Hypotheses Development

What drives the differences empirically observed above? Are humans likely to be able to account for them? In this section, we show that certain factors have systematically predictable implications on how server-level (Level 2) metrics differ from customer-experience (Level 1) metrics. We then combine these implications with a psychology-informed behavioral model to derive several testable hypotheses.

5.1. Quality Gap Drivers

Given that customer-experienced metrics ($cq$ and $icq(\bar{q}')$) differ from server-level metrics ($sq$ and $isq(\bar{q}')$), we now mathematically examine the relationship between these two metric types to illuminate factors that contribute to their differences.

**Average Quality Metrics** Because more customers experience the quality of larger-load servers, one may conjecture that correlation between server load and server quality may impact the customer-server quality gap. Indeed, we can establish such a result formally. To do so, we assume the following stylized linear model of quality provided to the $i$th customer by server $j$:

$$Q_{ij} = \alpha_j + \beta_j s_j + Y_j + Z_{ij}. \quad (5)$$

Here, $\alpha_j$, $\beta_j$, and $Y_j$ are random variables capturing the variation unique to each server such that, for all $j$, $E[\alpha_j] = \alpha$, $E[\beta_j] = \beta$, $E[Y_j] = 0$. Thus, each server may have a unique relationship between quality and the number of customers. The random variables $Z_{ij}$ capture the variation unique to each customer, such that for all $i$ and $j$, $E[Z_{ij}] = 0$. With this model and the definitions of $cq$ and $sq$, we can derive the following result, which characterizes the expected customer-server quality gap. (All proofs are in the online supplement).
Proposition 1. \( E[CQ - SQ] = \beta \sigma^2 s_m \), where \( \sigma^2_s \) is the variance of the number of customers per server.

Thus, the expected customer-experienced quality is larger than the expected server-level quality if and only if there is a positive relationship between server quality and server load (i.e. \( \beta > 0 \)). The magnitude of the difference depends on the strength of this relationship, the variance in the number of customers per server, and the ratio between the number of servers and the number of customers.

To gain some intuition behind Proposition 1, assume \( \alpha = 0, \beta = 1 \) and \( \alpha_j, \beta_j, Y_j \) and \( Z_{ij} \) have zero variance. These assumptions make quality deterministic and equal to the number of customers per server, \( Q_{ij} = s_j \), just as in the class size paradox (Feld and Grofman 1977). In this case, Proposition 1 simplifies to \( \bar{cq} - \bar{sq} = \sigma^2 s_m > 0 \). That is, the average customer always experiences more people in its service than the average server because customers are more likely to be served by high-load servers (Note that we saw this effect in Table 1 and Table 2). The result that the server load that a randomly-chosen customer experiences tends to be greater than that of a randomly chosen server is an example of the so-called inspection paradox (Ross 2003). Thus, Proposition 1 can also be viewed as a generalization of the class size paradox and as the application of the inspection paradox to quality measures in services: Customers and servers differ in their experiences of server load, quality depends on server load, so customers and servers differ in their experiences of quality.

Server Impact Metrics

The next result characterizes the server impact gap - the difference between the impact a server has on the customer-experienced average quality versus the server-level average quality.

Proposition 2. Let \( q' \) and \( s' \) be sufficiently different from \( \bar{sq} \) and \( n/m \), respectively, such that \( |(q' - \bar{sq})(s' - \bar{s/m})| > |(\bar{sq} - \bar{cq})(\bar{s/m} - \bar{m})| \). Then,

- If \( q' > \max\{\bar{sq}, \bar{cq}\}, s' > n/m \), then, \( 0 < isq(q') < icq(q') \).
- If \( q' > \max\{\bar{sq}, \bar{cq}\}, s' < n/m \), then, \( 0 < icq(q') < isq(q') \).
- If \( q' < \min\{\bar{sq}, \bar{cq}\}, s' > n/m \), then, \( icq(q') < isq(q') < 0 \).
- If \( q' < \min\{\bar{sq}, \bar{cq}\}, s' < n/m \), then, \( isq(q') < icq(q') < 0 \).

Recall that \( isq(q') \) and \( icq(q') \) measure the changes in quality metrics when server \( j' \) is added to the system. When adding a server with better-than-average service quality (\( q' > \max sq, \bar{cq} \)), having this server serve a larger population of customers (\( s' > n/m \)) will have a greater positive impact on customer-experienced quality (Level 1, \( \bar{cq} \)) than server level quality (Level 2, \( \bar{sq} \)). On the other hand, when a better-than-average server serves a smaller population (\( s' < n/m \)), she will have a lower impact on customer level quality. Similarly, adding into the system a worse-than-average server would have a greater negative impact on customer-experienced quality if she serves a larger
population of customers, but a greater negative impact on server-level quality if her service load is small.

Another implication of the above results (and one that we leverage in our experimental designs) is as follows. Given that two servers have the same impact on $\bar{cq}$ (i.e., the same value of $icq$), a smaller-than-average server will have a larger absolute impact on $\bar{sq}$ than a larger-than-average server. Thus, the quality of a server with a small load must be more extreme (i.e., further from $\bar{cq}$ and $\bar{sq}$) in order to generate the same impact on $\bar{cq}$ as a server with a large load.

The interpretation of the condition $\left|(q' - sq)(s' - \bar{s})\right| > \left|\bar{sq} - \bar{cq}\right|$ is that these results will hold as long as the server in question is not too close to the average quality and load, or if $\bar{cq}$ is not substantially different than $\bar{sq}$ (e.g., recall from Proposition 1 that the difference between $\bar{cq}$ and $\bar{sq}$ has expectation zero when $\beta = 0$). Finally, we point out that these results do not depend on the underlying quality generation model, e.g., we do not need to assume $Q_{ij}$ follows the model in (5).

5.2. Hypotheses Development

As discussed in the introduction and as demonstrated in §4, people in many settings are naturally exposed to data that is aggregated at a level different from the level most directly relevant to the decision they must make. In particular, in this paper we focus on service settings in which the decision-maker is exposed to server-level data (Level 2: $\bar{q}_j$, $s_j$, and $\bar{sq}$) and is tasked with making judgments about overall customer-experience (Level 1: $\bar{cq}$ and $icq(q')$). Our main conjecture is that people in such settings make judgments that are systematically biased away from the correct customer-experienced metrics towards the server-level ones. There is significant empirical evidence and psychological theory to motivate this conjecture.

As discussed in §2, the psychological perspective of humans as naive intuitive statisticians (Fiedler and Juslin 2006, Juslin et al. 2007) generally assumes that while people are quite proficient at basic descriptive statistics, they fail to make necessary sophisticated statistical corrections to correct for non-representative samples they are given. In this setting, the given server qualities $\{\bar{q}_j\}$ are not a representative sample of customer-experienced qualities $\{q_{ij}\}$. Thus, this perspective suggests that while people are generally proficient at estimating $\bar{sq}$ because it is simply the descriptive mean of the given $\{\bar{q}_j\}$, they will fail to make the necessary statistical correction for how it may be a biased estimator for $\bar{cq}$. In short, people are biased towards the server-level metrics because they are easier to calculate and (at least a portion of people) mistakenly fail to recognize that they may systematically differ from the customer-experienced metrics.

Similarly, from the probabilistic (as opposed to statistical) perspective, observe that server-level metrics are consistent with naively assuming that if one were to randomly sample a quality of service, there is an equal probability of selecting a quality from each server. A natural tendency to
assume an equal probability among alternatives has been documented in judgments of the likelihood of events in a variety of problems (e.g., Fox and Clemen 2005, Falk and Lann 2008). Thus, this perspective also suggests that people are biased toward the server-level metrics because a uniform probability distribution is easier to conceptualize, and they fail to recognize that the probability of arriving to each server is not uniformly distributed for a randomly-arriving customer.

We refer to the above shortcomings in statistical/probabilistic thinking as "naive statistical thinking." Of course, not necessarily all individuals are fully naive. However, even if only a portion of the population is statistically naive, a population-level bias towards server-level metrics emerges. Furthermore, even if all individuals correctly recognize that server-level metrics are biased, some may still use them as easily-calculable anchors from which to make an adjustment. In such a case, research on anchoring and adjustment suggests that such adjustments are likely to be insufficient because they require cognitive effort (Epley and Gilovich 2006) so that people may stop as soon as they reach a plausible region (Quattrone 1982). Thus, such an insufficient adjustment also leads to a population-level bias towards the server-level metrics.

To formalize our main conjecture, we define a simple behavioral model. Given the decision-maker is presented server-level data, we define a participant \(k\)'s behavioral judgment of \(\bar{cq}\) as

\[
\bar{cq}^b_k = \bar{s}q + \gamma_{cq}^k (\bar{cq} - \bar{s}q) + \varepsilon_{cq}^k \tag{6}
\]

where \(\gamma_{cq}^k\) denotes the degree of adjustment from \(\bar{cq}\) towards \(\bar{s}q\), with \(0 < E[\gamma_{cq}^k] = \gamma_{cq} < 1\), and \(\varepsilon_{cq}^k\) is a mean zero random variable. Thus, the behavioral judgment is a weighted average between the server-level and customer level metrics. Similarly, we define participant \(k\)'s behavioral judgment of \(icq(q')\) as

\[
\bar{icq}^b_k(q') = \bar{isq}(q') + \gamma_{icq}^k (\bar{icq}(q') - \bar{isq}(q')) + \varepsilon_{icq}^k
\]

where \(\gamma_{icq}^k\) denotes the degree of adjustment from \(\bar{ICQ}\) towards \(\bar{ISQ}\), with \(0 < E[\gamma_{icq}^k] = \gamma_{icq} < 1\).

We refer to the special cases when \(\gamma_{cq}^k = 0\) and \(\gamma_{icq}^k = 0\) as the fully statistically naive benchmarks, or simply the "naive" benchmarks. Conversely, note that when \(\gamma_{cq}^k = 1\) and \(\gamma_{icq}^k = 1\), the above behavioral models reduce to the rational solutions, \(\bar{cq}\) and \(icq(q')\).

The following hypothesis summarizes our main conjecture that judgements will be biased toward server-level metrics:

**Hypothesis 1.** When presented server-level data, average judgments about \(\bar{cq}\) and \(icq(q')\) are biased towards \(\bar{s}q\) and \(\bar{isq}(q')\), respectively.

Further, if individuals’ judgments are biased towards server-level metrics according to Hypothesis 1, we expect their judgments to be systematically biased in the same ways that the server-level metrics deviate from the customer-experienced ones. The following two hypotheses, which follow from the behavioral model and Propositions 1 and 2, present directional predictions:
Hypothesis 2. When presented server-level data, judgments about customer-experienced quality are biased high \((E[\bar{cq}_k] > \bar{cq})\) when server quality and load are negatively correlated, but biased low \((E[\bar{cq}_k] < \bar{cq})\) when server quality and load are positively correlated.

Hypothesis 3. When presented server-level data, (i) for a server with larger-than-average load \((s' > n/m)\), judgments underestimate its absolute impact on customer-experienced quality \((E[|icq'_k(\bar{q}')|] < |icq(\bar{q}')|)\), (ii) for a server with smaller-than-average load \((s' < n/m)\), judgments overestimate its absolute impact on customer-experienced quality \((E[|icq'_k(\bar{q}')|] > |icq(\bar{q}')|)\).

Finally, if the above systematic biases are driven by the decision-makers having cognitive challenges interpreting server-level data, then simply showing customer-experienced data ought to improve behavior and reduce systematic bias. Therefore, we have:

Hypothesis 4. Judgments about \(\bar{cq}\) and \(icq(\bar{q}')\) are less biased when decision makers are presented customer-experienced data than when presented server-level data.

We test these hypotheses in the following sections.

6. Experimental Evidence: Quality Judgment Bias

This section presents experimental results from two studies testing whether there exists human quality judgment biases consistent with the predictions derived in §5. The first study focused on whether customers who are presented with data aggregated at the server level make biased quality judgments about average customer-experienced quality \(\bar{cq}\) (Hypotheses 1 and 2). The second study focused on whether administrators presented with data aggregated at the server level make biased quality inferences about the impact of an individual server on customer-experienced quality \(icq(\bar{q}')\) (Hypotheses 1 and 3).

We set our studies in the context of education. In Study 1, students are customers, classes are servers, and schools are units. The subjects were asked to evaluate the quality of multiple schools. In Study 2, students are customers, instructors are servers, and the school is the unit. Subjects were asked to evaluate the impact of instructors on the school. We chose the education context because it allowed us to make the problem concrete to aid comprehension, because our subjects were university students, and because education is an industry in which our findings have important implications (see Alekseev et al. 2017).

6.1. Recruitment and Subjects for 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. We targeted a sample size of 100 subjects by posting 12 laboratory sessions with 10 slots per session. This process resulted in a total of 92 subjects participating in the
study, which was conducted with pen/pencil and paper in the behavioral lab. Subjects completed both Study 1 and Study 2, in sequential order, and subjects took between 15 and 30 minutes to complete both studies. Subjects received $5 compensation for completing the study and an additional bonus of $0 to $2 based on the accuracy of their answers.

6.2. Study 1: Evaluating Schools

6.2.1. Task Subjects were asked to assess students’ experiences at schools by estimating the course ratings students experience on average (i.e., customer-experienced quality) at each of five different schools. Subsequently, they were asked to rank these five schools from best to worst. Subjects were told that at the end of the task they would receive an additional $1 bonus if their answer was within 0.5 points of the correct answer for one randomly selected school.

Figure 3 Study 1 experimental stimulus example

<table>
<thead>
<tr>
<th>School ID#</th>
<th>Average</th>
<th>Standard Dev.</th>
<th># Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.3</td>
<td>1.0</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td>5.9</td>
<td>1.5</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td>9.7</td>
<td>0.5</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>2.5</td>
<td>1.4</td>
<td>84</td>
</tr>
<tr>
<td>5</td>
<td>9.3</td>
<td>0.9</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>6.2</td>
<td>1.6</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>7.5</td>
<td>1.3</td>
<td>71</td>
</tr>
<tr>
<td>8</td>
<td>6.1</td>
<td>1.6</td>
<td>130</td>
</tr>
<tr>
<td>9</td>
<td>3.8</td>
<td>1.4</td>
<td>104</td>
</tr>
<tr>
<td>10</td>
<td>7.6</td>
<td>1.5</td>
<td>109</td>
</tr>
<tr>
<td>11</td>
<td>3.0</td>
<td>1.6</td>
<td>118</td>
</tr>
<tr>
<td>12</td>
<td>7.7</td>
<td>1.2</td>
<td>35</td>
</tr>
<tr>
<td>13</td>
<td>8.7</td>
<td>1.3</td>
<td>31</td>
</tr>
<tr>
<td>14</td>
<td>9.4</td>
<td>0.7</td>
<td>35</td>
</tr>
<tr>
<td>15</td>
<td>2.9</td>
<td>1.4</td>
<td>113</td>
</tr>
</tbody>
</table>

Imagine you randomly surveyed 1,000 students from this school and collected the rating they gave their class. What should we expect the average of these 1,000 responses to be? (Assume each student takes only 1 class.)

Expected average of 1,000 student ratings: ______________

Note: Subjects see five random schools. We only show one as an example.

On each of five pages, subjects saw information about course ratings and enrollments for 15 courses at a school (Figure 3). Courses were rated on a 0-10 scale (0=worst, 10=best). Specifically, for each course at a school they saw the following columns: course identification number, average course rating, standard deviation of course rating, and number of students. They were then asked

Full experimental stimuli materials are available in the “Experimental Stimuli” supplementary document available from the authors.
the following question: “Imagine you randomly selected 1,000 students from this school and collected the rating they gave their class. What should we expect the average of these 1,000 responses to be? (Assume each student takes only 1 class.)” Finally, on the sixth page, subjects were asked to re-record their responses for each school and rank the five schools from best to worst.

The study was a 2 × 1 within-subject design, where we manipulated whether there was positive or negative correlation between course size and average course rating at the school level. For each school, we first randomly determined whether it was a positive or negative correlation school, with equal probability. We then generated 15 courses each of which had either large enrollment (N(100, 20)) or small enrollment (N(30, 10)) also with equal probability, and constructed the positive or negative correlation schools by assigning high (N(9, 2)) mean course ratings to large enrollment courses and low (N(5, 2)) mean course ratings to small enrollment courses in the positive correlation condition, and vice versa for the negative correlation condition. This process resulted in ex post estimated model parameters of $\alpha = 4.48, \beta = .033$ for the positive correlation condition and $\alpha = 8.88, \beta = -.036$ for the negative correlation condition (see the Online Supplement for full details of the data generation process).

6.2.2. Results The primary dependent variable of interest was the error of the subjects’ estimate of the customer-experienced average quality at each school (subject’s response – $\bar{cq}$), where positive error corresponds to overestimation and negative error corresponds to underestimation. All 92 subjects completed 5 estimations of the customer-experienced average rating (one estimation per school), yielding 460 observations. All data are included in the analyses.

<table>
<thead>
<tr>
<th>Subject Mean Error</th>
<th>Positive Correlation</th>
<th>Negative Correlation</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.358***</td>
<td>0.372***</td>
<td>0.723***</td>
</tr>
<tr>
<td></td>
<td>(0.078)</td>
<td>(0.095)</td>
<td>(0.153)</td>
</tr>
<tr>
<td>Naive Mean Error</td>
<td>-0.885***</td>
<td>0.872***</td>
<td>1.757***</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.019)</td>
<td>(0.029)</td>
</tr>
</tbody>
</table>

Cluster robust standard errors shown in parentheses.

*** $p < 0.001$

We estimate mean subject errors and cluster robust standard errors clustered by subject for each experimental condition (Table 4). For schools with positive correlation between course enrollment and rating, subjects’ estimations were too low with negative average estimation error, $M = -0.358$, $SE = 0.078$. For schools with negative correlation, subjects’ estimations were too high with positive average estimation error, $M = 0.372, SE = 0.095$. Thus, we find support for Hypotheses 1 and 2 in that subjects exhibit bias directionally consistent with such a naive estimate ($\bar{sq}$). We also compute the naive error, that is the error that would result if subjects computed the correct server-level
average quality (i.e., completely ignoring class size information). We see that the naive mean error is $-0.885$ in the positive correlation condition and $0.872$ in the negative correlation condition.

<table>
<thead>
<tr>
<th>Table 5 Study 1: School quality rank error (N=460)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Subject Rank Error</td>
</tr>
<tr>
<td>Positive Correlation</td>
</tr>
<tr>
<td>Negative Correlation</td>
</tr>
<tr>
<td>Naive Rank Error</td>
</tr>
<tr>
<td>Positive Correlation</td>
</tr>
<tr>
<td>Negative Correlation</td>
</tr>
</tbody>
</table>

Values shown are proportions within each row

A secondary measure of interest is corresponding rank errors for each school, where “overranking” corresponds to ranking higher than the true customer-experienced rank and “underranking” corresponds to a ranking lower than the true rank (Table 5). Unsurprisingly, the distribution of the rank errors corresponds to the pattern of over and underestimation found in the estimations for customer-experienced average quality. Subjects tended to rank schools with positive correlation too low and schools with negative correlation too high (A Pearson’s Chi-squared test confirms that the distributions of rank errors is significantly different between the two correlation conditions ($p < 0.001$)). Again, we calculate a naive error based on ranking schools on the server-level quality. Similar to the error in rating, the naive error in ranking is directionally the same as the subjects’ errors.

6.3. Study 2: Evaluating Teacher Impact

6.3.1. Task In this study, subjects were asked to play the role of an administrator. Their job was to assess teachers’ value to schools according to the impact they have on the average ratings that students experience (i.e., $\bar{q}$ impact). They were asked to rank the top five and bottom five teachers at a given school. Subjects were told that at the end of the task they would receive an additional $1 bonus if a randomly selected teacher from the actual top five or bottom five was correctly in the subject’s top five or bottom five.

In this study, subjects considered information about teacher quality ratings and enrollments at one school. Teacher ratings were again on a 0-10 scale (0=worst, 10=best). Specifically, for each teacher they saw the following columns: teacher name, average teacher rating, standard deviation of teacher rating, and number of students (Figure 4). They were then asked the following question: “Imagine you are an administrator at this school. Your job is to identify and rank the top five most valuable teachers and bottom five least valuable teachers. When assessing a teacher’s ‘value,’ only consider the impact he/she has on the average rating that students experience.” They then wrote in the names of the teachers they believed to be ranked 1-5 (top five) and 26-30 (bottom five).
Imagine you are an administrator at this school. Your job is to identify and rank the top 5 most valuable teachers and bottom 5 least valuable teachers.

When assessing a teacher’s “value,” only consider the impact he/she has on the average rating that students experience.

Fill in the blanks below with the teacher associated with rank requested (1=most valuable, 30=least valuable):

Top 5 (rank 1-5) teachers:
1. (best) ______________
2. ______________
3. ______________
4. ______________
5. ______________

Bottom 5 (rank 26-30) teachers:
26. ______________
27. ______________
28. ______________
29. ______________
30. (worst) ______________

Note: Each subject observes a different randomly-generated school.

This study was a $2 \times 2$ within-subject design, where we manipulated whether teachers taught a large or small number of students and whether subjects were ranking the top five or bottom five teachers. The goal of the data generation process for this study was to construct pairs of teachers that had identical $icq$ values. The benefit of constructing pairs of teachers with identical $icq$ values is that it allows us to test whether people systematically tend to favor more extreme $\bar{q}$ and smaller $s$ or less extreme $\bar{q}$ and larger $s$, controlling for the true $icq$.

For each subject in the study, we randomly generated course evaluation data for the 30 teachers shown as follows. First, we randomly generated evaluation scores for 15 “small enrollment” teachers with enrollments $N(30,10)$ and mean scores $N(6,2)$. Then, for each small enrollment teacher, we generated a corresponding “large enrollment” teacher with three times the enrollment but adjusting the course ratings to keep the impact on the student experienced average score approximately equal. (Although paired teachers’ impact on $\bar{q}$ are extremely close, a large class has a slightly greater absolute impact than its associated small class due to the simulation process and the number of decimal places we showed subjects. See the Online Supplement for full details.) Thus, by construction, for each small/large teacher pair, the large enrollment teacher had a rating closer to the mean
(6) relative to the small enrollment teacher. Also note that by construction, for each small/large teacher pair, in general, if one teacher in the pair is in the top (bottom) five, the other teacher in the pair is as well, except for the pair in 5th (26th) place, for which only the large enrollment teacher is in the top (bottom) five. There is also no correlation between teacher enrollment and mean scores. Finally, we assigned each of the 30 scores a randomly-generated teacher last name and then sorted the information table by the last name (see the Online Supplement for full details of this data generation process).

6.3.2. Results 91 subjects listed 10 names each, yielding 455 listed names for the top five and 455 listed names for the bottom five. (One of the 92 total subjects did not provide answers for Study 2.) We find that subjects are more likely to list small enrollment teachers in the top five, with 71% of the teachers they listed being small classes. Similarly, subjects were also more likely to list small enrollment teachers in the bottom five, with 78% of teachers they listed being small classes. These percentages are significantly different from the 40% small classes that would be expected if subjects accurately assessed the precise teacher impact on student experience. Further, these percentages are significantly different from the 50% small classes that would be expected if subjects assessed approximate teacher impact on student experience and then randomly chose between the two teachers with almost identical impact. If one were to evaluate all teachers using their listed mean ratings but assume they all taught an equal number of students, 93% of the top five would be small enrollment teachers, and 88% of the bottom five would be small enrollment teachers. Thus the subjects’ estimations were biased in the direction predicted in Hypothesis 3.

Another way to test our theorized mechanism for Hypothesis 3 is to use regression to determine whether actually being in the top (bottom) five by server-level impact or customer-level impact is more predictive of being selected by a subject as being in the top (bottom) five. Because the probability of any given teacher being selected as a top (bottom) teacher is a function of the attributes of all the teachers (i.e., rating and class size), we use conditional logistic regression grouped by school (Greene 2012, Sec. 18.2.3). In this model, the dependent variable is a binary variable indicating if teacher \( j \) was selected as a top (bottom) five teacher by subject \( i \). For explanatory variables, we use binary variables indicating if teacher \( j \) is actually a top (bottom) five teacher based on server-level quality impact \( isq(q') \) and customer-experienced quality impact \( icq(q') \).

Table 6 shows the estimated coefficients. Comparing Models 1 and 2, we see that being in the top five by \( isq(q') \) is more predictive of being selected as a top five than is being in the top five by \( icq(q') \) (the metric that the subjects were asked to identify). The results show this by the larger coefficient

\footnote{There are a few exceptions to this due to the randomness of the data generation process. In these cases, the unmatched teacher is in the 4th (27th) position rather than the 5th (26th).}
Table 6  Study 2: Predictors of teacher being selected top/bottom five by subject (conditional logistic model)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(isq(q')) Top 5 (Y/N)</td>
<td>3.62***</td>
<td>3.43***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.26)</td>
<td>(0.30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(icq(q')) Top 5 (Y/N)</td>
<td></td>
<td></td>
<td>2.31***</td>
<td>1.98***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.11)</td>
<td>(0.20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(isq(q')) Bottom 5 (Y/N)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.51***</td>
<td>4.30***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.20)</td>
<td>(0.23)</td>
</tr>
<tr>
<td>(icq(q')) Bottom 5 (Y/N)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.47***</td>
<td>2.06***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.10)</td>
<td>(0.25)</td>
</tr>
<tr>
<td>N</td>
<td>2,730</td>
<td>2,730</td>
<td>2,730</td>
<td>2,730</td>
<td>2,730</td>
<td>2,730</td>
</tr>
<tr>
<td>Pseudo-(R^2)</td>
<td>0.44</td>
<td>0.19</td>
<td>0.52</td>
<td>0.61</td>
<td>0.22</td>
<td>0.68</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>-607.53</td>
<td>-874.42</td>
<td>-528.37</td>
<td>-417.41</td>
<td>-843.35</td>
<td>-362.19</td>
</tr>
</tbody>
</table>

Robust standard errors shown in parentheses
*** \(p < 0.001\)

in Model 1, as well as the larger Pseudo-\(R^2\) and log-likelihood (see Alcacer and Delgado (2016) for a similar use of conditional logistic models). Models 4 and 5 show a similar pattern for selection as a bottom five teacher; \(isq(q')\) is more predictive of selection. Because some teachers are top (bottom) five by both impact metrics, in Models 3 and 6 we include both explanatory binary variables. We see that both variables continue to have explanatory power, but the server-level quality impact variable still has a larger effect (\(p < 0.001\) for test of coefficient equality in both Models 3 and 6). Thus, subjects’ behavior is more consistent with selecting teachers based on their impact on \(\bar{s}q\) (i.e., ignoring class size) than on the teachers’ impact on \(\bar{c}q\). This pattern of biased rankings of teachers suggest that high quality, low-enrollment teachers are more likely to be overvalued while moderate (but above average) quality, large-enrollment teachers are more likely to be undervalued. Similarly, it suggests that administrators will tend to pay too much attention to low quality, low-enrollment teachers but not enough attention to low quality, high-enrollment teachers.

7. Experimental Evidence: Mitigation of Bias and Implication for Decision Making

Having established in Studies 1 and 2 that service quality judgment biases exist, we are now interested in two main questions. First, are these judgment biases sufficiently large to affect decision-making? Second, can we mitigate these judgment biases by providing customer-experienced quality statistics (Hypothesis 4)?

To address these questions, Studies 3 and 4 contain very similar tasks as Studies 1 and 2 with two important exceptions: (1) We compare behavior across three conditions in which we either show server-level data (as in Study 1), customer-experienced data, or both server-level and customer-experience data simultaneously, and (2) We ask subjects to make hypothetical decisions about school choice (Study 3) and which teachers to award and discipline (Study 4) to see whether these quality judgment biases can generate differences in decision behavior.
7.1. Recruitment and Subjects for Studies 3 and 4
We used the same online recruiting and scheduling system as in Studies 1 and 2. We targeted a sample size of 150 subjects by posting 25 laboratory sessions with 7 slots per session. This process resulted in a total of 166 subjects participating in the study, which was conducted with pen/pencil and paper in the behavioral lab. Subjects completed both Study 3 and Study 4, in sequential order, and subjects took between 10 and 20 minutes to complete both studies. Subjects received $7 compensation for completing the study and an additional bonus of $0 to $4 based on the accuracy of their answers.

7.2. Study 3: Choosing Schools

7.2.1. Task and Conditions
The task in Study 3 was the same as in Study 1 except for the following main differences:

- The structure of the quality data and statistics differed from Study 1 by condition (outlined in the conditions below).

- The task framing was for the subject to play the role of a student making a hypothetical school choice decision (not merely quality judgments). Specifically, after reviewing the information, the subjects were told to “Imagine that you are a student selecting a new school and that you are trying to select the school that will provide you with the highest expected course quality.” Subjects were then asked to rank order the five schools and to provide an estimate of the quality they expect from each school.

- The incentives for subjects was to receive an additional $1 bonus if they correctly picked the school with the highest expected student experienced quality school, and another $1 bonus if their quality estimate for a randomly selected school was within 0.5 of the correct answer.

- The underlying data generation process differed slightly from Study 1. In this study, the quality generation followed the model in (5) with $\alpha = 1, \beta = 0.066$ in the positive correlation condition but $\alpha = 9, \beta = -0.066$ in the negative correlation condition. Full details on this data generation process are located in the Online Supplement.

- Calculators were provided to subjects.

Subjects were assigned to one of three conditions, where we manipulated the type of information provided to the subject (Figure 5).

Server-Level Quality (SQ). As in Study 1, this condition provided subjects with the ratings by class. Each row contained a class’ average rating and standard deviation, along with the number of students in that class. Finally, the average across classes ($\bar{sq}$) was shown at the bottom of the table.

Customer-Experience Quality (CQ). This condition provided subjects with the ratings by student. Each row contained a student’s rating and the class ID he/she was rating. While some rows were omitted for space, at the bottom of the table, it showed the average across students ($\bar{cq}$).
### Figure 5 Study 3 experimental stimulus example

<table>
<thead>
<tr>
<th>School ID#</th>
<th>818</th>
</tr>
</thead>
</table>

This table shows the rating each student gave the class he or she took.

<table>
<thead>
<tr>
<th>Student ID</th>
<th>Class ID</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>6297</td>
<td>1</td>
<td>2.1</td>
</tr>
<tr>
<td>6005</td>
<td>1</td>
<td>4.6</td>
</tr>
<tr>
<td>9762</td>
<td>1</td>
<td>5.0</td>
</tr>
<tr>
<td>8687</td>
<td>2</td>
<td>3.4</td>
</tr>
<tr>
<td>7092</td>
<td>2</td>
<td>3.9</td>
</tr>
<tr>
<td>2597</td>
<td>2</td>
<td>3.8</td>
</tr>
</tbody>
</table>

The rest of the 1032 student ratings are not shown in the interest of space.

<table>
<thead>
<tr>
<th>Class ID</th>
<th>Ratings</th>
<th># Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.4</td>
<td>84</td>
</tr>
<tr>
<td>2</td>
<td>3.4</td>
<td>83</td>
</tr>
<tr>
<td>3</td>
<td>2.6</td>
<td>97</td>
</tr>
<tr>
<td>4</td>
<td>7.7</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>3.6</td>
<td>80</td>
</tr>
<tr>
<td>6</td>
<td>6.3</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>1.8</td>
<td>106</td>
</tr>
<tr>
<td>8</td>
<td>2.7</td>
<td>95</td>
</tr>
<tr>
<td>9</td>
<td>4.2</td>
<td>73</td>
</tr>
<tr>
<td>10</td>
<td>2.5</td>
<td>103</td>
</tr>
<tr>
<td>11</td>
<td>3.6</td>
<td>82</td>
</tr>
<tr>
<td>12</td>
<td>6.1</td>
<td>35</td>
</tr>
<tr>
<td>13</td>
<td>4.4</td>
<td>72</td>
</tr>
<tr>
<td>14</td>
<td>7.5</td>
<td>25</td>
</tr>
<tr>
<td>15</td>
<td>6.8</td>
<td>39</td>
</tr>
</tbody>
</table>

This table shows the summary statistics of ratings each class received.

<table>
<thead>
<tr>
<th>Class Average</th>
<th>Standard Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Note: Only the “Both” condition shown. Subjects see five random schools. We only show one as an example.

**Both Server-Level and Customer-Experience (Both).** This condition provided subjects with both of the above tables side-by-side.

#### 7.2.2. Results

The primary dependent variable of interest was the error of the subject’s estimate of the customer-experienced average quality at each school. We define the sign of the estimation error relative to the server-level average quality ($\bar{sq}$). That is, a positive error indicates an error in the direction of the $sq$, while a negative error is away from $\bar{sq}$. 162 subjects completed five estimations of the customer-experienced average quality, yielding 810 observations (53 subjects in condition SQ, 55 subjects in condition Both, and 54 subjects in condition CQ).\(^5\) For comparison, in the following results we also create an artificial “Naive” condition based on the schools shown to the subjects in the SQ condition and assuming that all responses are equal to the server-level metrics (i.e., the artificial subject believes $\bar{sq}$ to be the customer-experienced average quality).

We calculate the mean estimation error for each condition along with cluster-robust standard errors of the mean clustered by subject (Figure 6a). We see that in experimental condition SQ, subject quality estimation has a mean error of 0.84 ($SE = 0.09$). However, this error is significantly

\(^5\) Four subjects were dropped due to providing answers as textual descriptions (e.g., “excellent”) or for values outside the 0-10 range of possible quality.
smaller than the Naive condition ($p < 0.01$ for test of equality of means), suggesting that subjects are potentially anchoring on and adjusting from the provided server-level information to try to estimate the customer quality. Experimental conditions Both and CQ have mean errors significantly lower than the SQ condition ($p < 0.01$ and $p < 0.001$ for the Both and CQ conditions, respectively), thus providing support for Hypothesis 4 that judgment bias will be reduced when customer-level information is provided.

It is interesting to note that mean error is significantly larger in the Both condition than in the CQ condition ($p < 0.01$) despite the fact that in the two conditions subjects are provided with the customer-experienced information. To explore this result further, we create a normalized estimation error metric by dividing the estimation error for each school by the absolute value of the customer-server quality gap ($|\bar{cq} - \bar{sq}|$). Thus, the normalized estimation error is an estimate of $\gamma_k^{cq}$: a subject response which is equal to $\bar{cq}$ will have a normalized estimation error of 0, and a subject response which is equal to $\bar{sq}$ will have a normalized estimation error of 1.

Figure 6b displays the histograms of the normalized estimation error for each of the experimental conditions. The figure shows that in conditions in which only one type of information is provided (SQ and CQ), the provided information provides a strong anchor point for subject responses and many subjects respond with precisely the information provided. Perhaps more interesting is the pattern observed for the Both condition. Recall that in this condition, subjects are provided with both server-level and customer-level information. We see small spikes at both 0 and 1, indicating clusters of subjects responding with either the $\bar{sq}$ or $\bar{cq}$ information. However, we also observe a (larger) spike around 0.5, indicating subjects that are providing answers that “split the difference”
between the two sets of information given. This helps explain why the mean error of the Both condition is larger than the CQ condition and it highlights the danger in providing decision makers with superfluous information. There appears to be a risk that they will incorporate the unnecessary information ($\bar{s}q$), leading to poorer judgments.

We also observe that in the CQ condition, there is a small but statistically significant bias toward $\bar{s}q$. One possible explanation for this small effect is the asymmetry in the plausible magnitude of positive versus negative errors (see Su (2008) and Kremer et al. (2010) for a similar argument in newsvendor decision making). By construction, school quality is bounded at 0 and 10, the mean $\bar{cq}$ for the schools is either 4 or 6, and $\bar{s}q$ tends to lie toward the middle of the scale, which is the side to which there is room for larger errors. Together, these facts allow for a larger error tail in the direction of $\bar{s}q$, likely causing the small positive mean error in the CQ condition. We note that the CQ condition median error is 0.01 with a confidence interval that includes zero, which further suggests that the mean error is driven by asymmetric estimation errors.

Having shown that quantitative judgments of customer quality can be biased, we also want to examine if these biases are sufficiently large to cause errors in decisions. Recall that in Study 3 subjects were asked to select the school to attend that would provide them with the highest expected quality, and they were incentivized based on selecting the correct school (i.e., the one with the highest $\bar{cq}$). Figure 7a shows the proportion of subjects that selected the correct top school, and Figure 7b shows the average reduction in expected quality due to incorrectly selecting which school to attend.

![Figure 7](image_url)

**Figure 7** Study 3: Selecting the best school

(a) Proportion of subjects selecting the best school

(b) Quality loss due to selection error

Subjects in the SQ condition selected the correct school only 42% of the time, which is no better than the Naive condition in which the highest $\bar{s}q$ school is selected. In contrast, subjects in the Both
and CQ conditions selected the correct school 78% and 80% of the time, respectively. These errors in selection led to a reduction of expected student quality of 0.72 for the SQ condition and only 0.19 and 0.14 for the Both and CQ conditions respectively, relative to always selecting the correct school.

Altogether, the results of Study 3 suggest that biased quality judgments are sufficient to impact school choice decisions, and that providing the correct customer-experienced quality information is an effective mitigation technique.

7.3. Study 4: Identifying High and Low Performing Teachers

7.3.1. Task and Conditions The task in Study 4 was the same as in Study 2 except for the following differences:

- The structure of the quality data and statistics differed from Study 2 by condition (outlined in the conditions below).
- The framing of the task was to play the role of a school administrator to make a hypothetical teacher identification decision. Specifically, after reviewing the information, the subjects were told to “Imagine that you are serving on the faculty teaching performance review committee. Your job is to identify the 4 most positive impact teachers and the 4 most negative impact teachers. The 4 most positive impact teachers you identify will be given teaching awards. The 4 most negative impact teachers you identify will be required to participate in a teaching-improvement program. To make these assessments as objective as possible, your committee has been instructed to assess a teacher’s “impact” according only to the impact he/she has on the average rating that students experience.”

- We had subjects only list the top 4 and bottom 4 (as opposed to top five and bottom five in Study 2). We incentivized subjects by randomly selecting one teacher in the top 4 or bottom 4 and awarding $1 bonus if they correctly listed that teacher in the appropriate group.

- We had subjects perform the task twice (i.e., for 2 schools instead of 1).\(^6\)

- Calculators were provided to subjects.

Subjects were assigned to one of three conditions, where we manipulated the impact statistics provided to the subjects (Figure 8). We describe the differences in information provided across conditions below:

*Server-Level Quality (SQ).* This condition provided the impact on teacher average rating \((isq)\) for each teacher as a percent of the average rating across teachers \((\bar{sq})\).

*Customer-Experience Quality (CQ).* This condition provided the impact on student average rating \((icq)\) for each teacher as a percent of the average rating across students \((\bar{cq})\).

\(^6\) We have subjects perform the task twice to increase the sample size, which is necessary to have sufficient power across the three experimental conditions. No feedback is received between repetitions. We test for a change in performance between tasks and do not find evidence of any change (i.e., learning).
### Figure 8: Study 4 experimental stimulus example

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Ratings</th>
<th>Impact on Teacher</th>
<th>Impact on Student</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Dev.</td>
<td>Average Rating</td>
</tr>
<tr>
<td>Adams</td>
<td>5.8</td>
<td>1.6</td>
<td>111</td>
</tr>
<tr>
<td>Allen</td>
<td>6.9</td>
<td>1.6</td>
<td>30</td>
</tr>
<tr>
<td>Bailey</td>
<td>5.7</td>
<td>1.4</td>
<td>40</td>
</tr>
<tr>
<td>Brown</td>
<td>6.8</td>
<td>1.8</td>
<td>19</td>
</tr>
<tr>
<td>Butler</td>
<td>5.7</td>
<td>1.4</td>
<td>84</td>
</tr>
<tr>
<td>Campbell</td>
<td>8.7</td>
<td>1.2</td>
<td>34</td>
</tr>
<tr>
<td>Coleman</td>
<td>6.6</td>
<td>1.4</td>
<td>15</td>
</tr>
<tr>
<td>Collins</td>
<td>6.9</td>
<td>1.4</td>
<td>102</td>
</tr>
<tr>
<td>Cooper</td>
<td>6.9</td>
<td>1.5</td>
<td>105</td>
</tr>
<tr>
<td>Diaz</td>
<td>8.8</td>
<td>1.1</td>
<td>10</td>
</tr>
<tr>
<td>Edwards</td>
<td>6.1</td>
<td>1.6</td>
<td>87</td>
</tr>
<tr>
<td>Gray</td>
<td>3.8</td>
<td>1.3</td>
<td>28</td>
</tr>
<tr>
<td>Howard</td>
<td>8.4</td>
<td>1.4</td>
<td>20</td>
</tr>
<tr>
<td>Hughes</td>
<td>6.2</td>
<td>1.4</td>
<td>45</td>
</tr>
<tr>
<td>James</td>
<td>6.2</td>
<td>1.6</td>
<td>29</td>
</tr>
<tr>
<td>Kelly</td>
<td>5.7</td>
<td>1.6</td>
<td>57</td>
</tr>
<tr>
<td>Lee</td>
<td>5.2</td>
<td>1.6</td>
<td>28</td>
</tr>
<tr>
<td>Lewis</td>
<td>4.7</td>
<td>1.5</td>
<td>75</td>
</tr>
<tr>
<td>Martin</td>
<td>5.3</td>
<td>1.7</td>
<td>84</td>
</tr>
<tr>
<td>Morgan</td>
<td>6.3</td>
<td>1.6</td>
<td>57</td>
</tr>
<tr>
<td>Murphy</td>
<td>3.5</td>
<td>1.4</td>
<td>49</td>
</tr>
<tr>
<td>Nelson</td>
<td>5.2</td>
<td>1.5</td>
<td>147</td>
</tr>
<tr>
<td>Parker</td>
<td>6.6</td>
<td>1.5</td>
<td>132</td>
</tr>
<tr>
<td>Powell</td>
<td>8.6</td>
<td>1.5</td>
<td>35</td>
</tr>
<tr>
<td>Scott</td>
<td>6.8</td>
<td>1.7</td>
<td>60</td>
</tr>
<tr>
<td>Smith</td>
<td>2.0</td>
<td>1.5</td>
<td>25</td>
</tr>
<tr>
<td>Turner</td>
<td>5.2</td>
<td>1.3</td>
<td>19</td>
</tr>
<tr>
<td>Washington</td>
<td>5.9</td>
<td>1.5</td>
<td>120</td>
</tr>
<tr>
<td>Weight</td>
<td>5.4</td>
<td>1.5</td>
<td>37</td>
</tr>
<tr>
<td>Young</td>
<td>7.7</td>
<td>1.4</td>
<td>44</td>
</tr>
</tbody>
</table>

**Note:** Only the “Both” condition shown. Only one school shown as an example.

**Both Server-Level and Customer-Experience (Both).** This condition provided subjects with both of the above information side-by-side.

#### 7.3.2. Results

165 subjects listed 8 names for each of two schools (an additional subject only provided answers for one school), yielding 1,324 listed names for the top four and 1,324 listed names for the bottom four. Just as in Study 2, teachers are paired such that for every small enrollment teacher there is a large enrollment teacher with equivalent impact on the customer-experienced average quality, however the small enrollment teacher always has a more extreme quality rating and thus a larger impact on server-level quality. As shown in Study 2, this leads to a disproportionate number of teachers with small classes being selected as top and bottom performers. We now test to see if providing customer-experienced quality information mitigates this bias.

Figure 9 displays the proportion of subject-selected top(bottom) teachers that had a small class size. We see that 90% of the teachers selected by subjects in the SQ condition had small classes, compared to the 46% of actual top(bottom) teachers that had small classes. Subjects in the Both and CQ conditions were less biased ($p < 0.001$) in their selection of small-class teachers with roughly 62% and 55% small, respectively (these conditions are statistically indistinguishable from each...
other). Thus, providing customer-experienced information helps mitigate the bias toward server-level quality, in support of Hypothesis 4. Note that both the Both and CQ condition results are significantly higher than the actual level of small-classes teachers, likely due to the fact that in this study subjects were provided class size and average quality ($\bar{q}$) in all experimental conditions, perhaps leading some subjects to (incorrectly) adjust toward server-level metrics.

As with Study 3, we examine whether the overall accuracy of decision making is improved by providing customer-experienced data. Figure 10 presents the proportion of true top(bottom) teachers according to $icq(\bar{q})$ that were selected by the study subjects. The results show a similar pattern to Study 3 (Figure 7a). Subjects in the SQ condition selected true top(bottom) performers approximately 50% of the time, which is accuracy no different from selecting purely on server-level impact. In contrast, subjects in the Both and CQ conditions were approximately 80% accurate with their
selections. These two conditions were statistically indistinguishable from each other, but are both significantly better than the SQ condition ($p < 0.001$). This supports Hypothesis 4.

8. General Discussion

Service quality judgments are a key input to customers’ and organizations’ decision making. These judgments are often formed based on server-level data (Level 2), even though the decision-maker cares more about customer-experience metrics (Level 1). We empirically documented that the differences between the two types of metrics are substantial and sufficient to yield different rankings of servers. We theorized and provided experimental evidence that people also tend to make judgments about customer-experienced quality that are systematically biased away from the correct customer-experienced metrics towards the server-level ones. We also demonstrated how these judgment biases can lead to systematic decision-making biases. Lastly, we demonstrate a way to help mitigate these biases.

8.1. Implications

Our findings show that people tend not to be able to interpret server-level data to infer customer-experienced metrics, even if the information provided is sufficient to do so. In all conditions across all four experiments, subjects have sufficient information (i.e., quality and load information) to correctly judge and make decisions based on customer-experienced quality, yet their judgments and decisions are biased towards the server-level metrics. Therefore, when reporting service quality data, it is important to consider who will use the data and what kinds of decisions they will make. If the decision-maker cares about customer experience, then report the appropriate customer-experienced metrics, and do not rely on the decision-maker to correctly infer them from server-level data.

Our paper also helps to identify settings when the customer-server quality gap is likely to drive the largest judgment biases, and who is likely to benefit/suffer the most from such a judgment bias. For example, we show that the gap will be largest when there is a strong correlation between load and quality (positive or negative) and a large variance in the server load. Thus, settings that possess these features are strong candidates for intervention. Without such interventions, organizations with negative (positive) correlation between server load and quality are likely to benefit (suffer) the most, and small high (low) quality servers are likely to benefit (suffer) the most.

Studies 1 and 2 show that merely showing data aggregated at the server level is sufficient to cause decision makers who are trying to judge customer-experienced quality to be biased towards the server-level metrics. Some decision-makers are likely statistically naive and mistakenly try to calculate the server-level metrics, while others may realize they should calculate the customer-experienced metrics but first calculate the server-level metrics to use as an anchor from which to make an insufficient adjustment. Subsequently, Studies 3 and 4 show that decision-making accuracy improves
significantly if one provides decision-makers with the correctly calculated customer-experienced metrics. It also shows that providing only these customer-level metrics yields more robust mitigation than merely adding them to server-level ones because a portion of the population will either continue to mistakenly use server-level metrics or take the average between the two.

From a different perspective, mitigation of these judgment biases may not be the goal. When interacting with other decision-makers, one can take their judgment biases as given, and then revise one’s own actions strategically to anticipate them. For example, from the administrator’s perspective in Studies 1 and 3, schools can take advantage of the judgment biases of students (and ranking organizations) by offering a few very large (and poorly rated) classes along with many small (and well-received) classes, so that the average class size is much smaller (and quality higher), even if the average student-experience may not be. Similarly, from the teacher’s perspective in Studies 2 and 4, a teacher may be able to strategically take advantage of the judgment biases of administrators by teaching smaller classes in which she can score a higher average rating, although it comes at a cost to the organization of a lower average student-experienced quality.

8.2. Other Applications

We illustrated our findings with experiments and secondary data from the specific context of education, but we believe our findings have broader applicability. For example, as mentioned in §1, airlines are required by law to report monthly on-time performance metrics. However, we suspect that because OTP is reported at the flight number level as opposed to the customer level, there are meaningful judgment biases that occur. In particular, it would be interesting to investigate whether there is a positive correlation between the occupancy of flights and the probability of late departure. Our results would suggest that if such a correlation exists, then the on-time statistics reported to customers overestimate the customer-experienced on-time departure. In the Online Supplement, we use on-time performance data from May 2017 to show that, much like in the Higher Education setting of §4, airport OTP can differ significantly depending on whether the metric is calculated at the flight (server) level or the passenger (customer) level.

We also note that one can take various perspectives as to what is a “server.” For instance, a time interval can be a server. If one considers a server as a day, then we can apply our results to any setting in which quality is measured at the day level. For example, hospitals in Singapore report the average admission wait time of each day (Singapore Ministry of Health 2018). Just as subjects in Study 1 generated estimates of customer-experienced quality that were biased toward the server-level mean, we conjecture that the same bias would exist for potential patients in Singapore. In fact, the bias is likely stronger because hospitals do not report the number of patients admitted each day, making it impossible to calculate customer-experience metrics, and because is generally a strong
correlation between the number of patients in an emergency department and wait time (Batt and Terwiesch 2016).

Finally, we note that the customer versus server perspective can also be applied to inventory management. In this setting, the item in-stock rate (the server-perspective) may differ significantly from the order in-stock rate (the customer perspective) (Song 1998), and failure to report the more relevant metric may lead to biased judgments.

8.3. Limitations and Future Work
We conclude with a couple of comments about limitations of our findings and future work. First, the empirical evidence presented in this paper is a combination of experimental evidence that provides support for directional bias according to the behavior of the customer-server quality gap, and empirical evidence that suggests that the gap is meaningfully large. However, we do not have field evidence of the magnitude of judgment biases stemming from to our proposed mechanism. Our results suggest that the true magnitude of judgment biases in practice will depend not only on the magnitude of the customer-server quality gap, but also on the framing of the information and the characteristics of the individual making the judgment and the judgment environment. For example, in our experiments we clearly stated the server load information (i.e., class size), but this information may not even be available in many cases (e.g., passenger count is not available in airline OTP reports) - in which case we suspect the biases to be larger in magnitude than in our experiments.

Last, we have studied quality judgments in a static and non-strategic setting. Because quality judgments also serve as an input to decision-making in dynamic and strategic settings, future work may examine the implications of the judgment biases documented here on system outcomes. For example, a service provider trying to manage the dynamics of customer reviews being posted about its quality must consider how later customers will interpret the earlier customers’ posted reviews (e.g., see Papanastasiou et al. 2017). By incorporating customers’ quality judgment biases into such models, researchers may be able to yield better insight for organizations.

References


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:

$$\bar{c} - \bar{s} = \frac{n}{n} \sum_{j=1}^{m} s_j^2 - \frac{n}{m} \sum_{j=1}^{m} s_j - n \frac{m}{mn} \left( \sum_{j=1}^{m} s_j^2 \right) - n^2 \frac{m}{mn} \left( \sum_{j=1}^{m} s_j \right)^2$$

$$= \frac{n}{m} \left[ \left( \frac{\sum_{j=1}^{m} s_j^2}{m} \right) - \frac{1}{m} \left( \sum_{j=1}^{m} s_j \right)^2 \right] = \frac{m}{n} \left[ \left( \frac{1}{m} \sum_{j=1}^{m} s_j^2 \right) - \left( \frac{1}{m} \sum_{j=1}^{m} s_j \right)^2 \right] = \frac{\sigma^2}{m}. $$

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

$$E[SQ] = E \left[ \frac{1}{m} \sum_{j=1}^{m} \frac{1}{s_j} \sum_{i=1}^{s_j} Q_{ij} \right] = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{s_j} \sum_{i=1}^{s_j} E[\alpha_j + \beta_j s_j + Y_j + Z_{ij}]$$

$$= \frac{1}{m} \sum_{j=1}^{m} (\alpha + \beta s_j) + \frac{1}{m} \sum_{j=1}^{m} \frac{1}{s_j} \sum_{i=1}^{s_j} E[Y_j + Z_{ij}]$$

$$= \alpha + \beta \frac{n}{m} = \alpha + \beta \bar{s}$$

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

$$= \beta \left( \frac{1}{n} \sum_{j=1}^{m} s_j^2 \right) + \frac{1}{n} \sum_{j=1}^{m} (s_j \alpha) = \beta \bar{c} + \alpha$$

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


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

$$\text{isq}(\bar{q}') = \frac{1}{m} \sum_{j=1}^{m} \bar{q}_j - \frac{1}{m-1} \sum_{j=1; j \neq j'}^{m} \bar{q}_j = \frac{1}{m} \sum_{j=1}^{m} \bar{q}_j - \left[ \frac{1}{m-1} \sum_{j=1}^{m} \bar{q}_j - \frac{1}{m-1} \bar{q}' \right]$$

$$= \frac{m-1-m}{m(m-1)} \sum_{j=1}^{m} \bar{q}_j + \frac{1}{m-1} \bar{q}' = \frac{1}{m} \sum_{j=1}^{m} \bar{q}_j - \frac{1}{m-1} \left[ \bar{q}' - \frac{1}{m} \sum_{j=1}^{m} \bar{q}_j \right] = \frac{1}{m-1} [\bar{q}' - \bar{s}q] .$$

$$\text{icq}(\bar{q}') = \frac{1}{n} \sum_{j=1}^{n} \bar{s}q_j - \frac{1}{n-s'} \sum_{j=1; j \neq j'}^{n} \bar{s}q_j = \frac{1}{n} \sum_{j=1}^{n} \bar{s}q_j - \left[ \frac{1}{n-s'} \sum_{j=1}^{n} \bar{s}q_j - \frac{1}{n-s'} \bar{s}'q' \right]$$

$$= \left[ \frac{1}{n} - \frac{1}{n-s'} \right] \sum_{j=1}^{m} \bar{s}q_j + \frac{1}{n-s'} \bar{s}'q' = \frac{-s'}{n(n-s')} \sum_{j=1}^{n} \bar{s}q_j + \frac{1}{n-s'} \bar{s}'q'$$

$$= \frac{s'}{n-s'} \left[ \bar{q}' - \frac{1}{n} \sum_{j=1}^{n} \bar{s}q_j \right] = \frac{s'}{n-s'} [\bar{q}' - \bar{c}q] .$$

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

$$\Delta(\bar{q}') = \frac{s'}{n-s'} (\bar{q}' - \bar{c}q) - \frac{1}{m-1} (\bar{q}' - \bar{s}q)$$

$$= \frac{s'm-n}{(n-s')(m-1)} - \bar{c}q \frac{s'}{n-s'} + \bar{s}q \frac{1}{m-1}$$

$$= \frac{s'm-n}{(n-s')(m-1)} - \bar{c}q \frac{s'}{n-s'} + \bar{s}q \frac{1}{m-1} - (\bar{s}q - \bar{c}q) \frac{s'}{n-s'} + (\bar{s}q - \bar{c}q) \frac{s'}{n-s'}$$

$$= (\bar{s}'q - \bar{s}q) \frac{s'm-n}{(n-s')(m-1)} + (\bar{s}q - \bar{c}q) \frac{s'}{n-s'}$$

$$= \frac{s'm}{(n-s')(m-1)} \left[ (\bar{s}'q - \bar{s}q)(\frac{s'-m}{s'}) + (\bar{s}q - \bar{c}q) \frac{m-1}{m} \right]$$

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

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

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

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

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

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

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


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 $s' \left[ q' - CQ \right]$. Thus, a teacher who teaches $s'' = 3s'$ students who has the same impact must teach at the quality $q''$ which solves $s' \left[ q' - CQ \right] = 3s' \left[ q'' - CQ \right]$. Thus, $q'' = \frac{3}{4} \left( 1 - \frac{2s}{N - s'} \right) \left[ q' - CQ \right] + CQ$. We chose to implement $\left[ q' - CQ \right]/3 + CQ$ 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 $\alpha_j$ and $\beta_j$ 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. Empirical 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 Arikan (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) $\bar{s}q$ and $\bar{c}q$ impact  
(b) $\bar{s}q$ and $\bar{c}q$ impact rank

References


