Quality - Speed Conundrum

Tradeoffs in Customer Intensive Services

Senthil K. Veeraraghavan

joint work with:
Krishnan Anand, David Eccles School of Business, University of Utah
M. Fazıl Paç, Wharton School, University of Pennsylvania

University of Utah
Utah Winter Operations Conference, January 2010
Generally service **speed** and service **value** are considered as independent performance metrics

- Focal point of service quality in queueing literature: service delivery lead times/waiting times
- Fast service/low waiting times often considered better.

Customer intensive services: Primary health care, hospitality and education.

- Value provided increases with time spent with customer.
- Longer service times are preferred and provide more value...
- ... but no one likes to wait.
For customer-intensive services, market coverage is often partial. (i.e. many customers remain unserved).

In equilibrium, the servers “slow” down.

Equilibrium demand may increase with market price.

Service Competition may *increase* prices!
Psychology of Waiting vs. Service Duration

In many scenarios, customers prefer to be served well, but dislike waiting for the same service (Schelling 1975).
Customer-Intensive Services

Some services are surely more customer-intensive!

“It's not that complicated. Just give the man a haircut. It ain't brain surgery.”
Examples

- Beauty-care services (hairdressers, spas ... )
- Consulting services (business, legal, financial ...)
- Informative call centers (nurse-lines)
Empirical evidence: Productivity in customer-intensive (labor-intensive) service industries is low, although it is high for services in general (Varian 2004, Triplett and Bosworth 2003, Yarnall et. al. 2003).

Cost Disease and Customer Intensity:

"You can control drug costs and limit expensive new procedures, but when it comes to hospital care and doctor visits, the only way to improve productivity is to shrink the size of the staff and have doctors spend less time with patients. To lower prices you have to lower quality." (Surowiecki 2003)

Cutting Corners \iff Increase in Productivity.
Service Value in Primary Care

"Regular" Doctors

- Rushing between patients.  "I was seeing 30 people a day and always rushing. Patients were dissatisfied.... I was dissatisfied." (Dr. Kaminetsky, NYU, Testimony to Joint Economic Committee U.S. Congress)
  - Not enough time for preventive care.
- Service delays are likely, due to demand related congestion.
- Charge "relatively" lower prices.

"Concierge" ("Boutique") Doctors

(MDVIP, $MD^2$, Current Health, Qliance Primary Care,...)

- More time with each patient (highly customized treatment).
- Offer immediate response (no service delay),
- Charge high prices.
Service Value in Primary Care

"Regular" Doctors

- Rushing between patients. "I was seeing 30 people a day and always rushing. Patients were dissatisfied.... I was dissatisfied." (Dr. Kaminetsky, NYU, Testimony to Joint Economic Committee U.S. Congress)
  - Not enough time for preventive care.
- Service delays are likely, due to demand related congestion.
- Charge "relatively" lower prices.

"Concierge" ("Boutique") Doctors

(MDVIP, MD2, Current Health, Qliance Primary Care,...)

- More time with each patient (highly customized treatment).
- Offer immediate response (no service delay),
- Charge high prices.
"Regular" Doctors

- Rushing between patients. "I was seeing 30 people a day and always rushing. Patients were dissatisfied.... I was dissatisfied." (Dr. Kaminetsky, NYU, Testimony to Joint Economic Committee U.S. Congress)
  - Not enough time for preventive care.
- Service delays are likely, due to demand related congestion.
- Charge "relatively" lower prices.

"Concierge" ("Boutique") Doctors

(MDVIP, MD², Current Health, Qliance Primary Care,...)

- More time with each patient (highly customized treatment).
- Offer immediate response (no service delay),
- Charge high prices.
Speed vs. Quality in Primary Care

<table>
<thead>
<tr>
<th></th>
<th>Regular Doctor</th>
<th>&quot;Concierge Doctors&quot; (MDVIP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price (Primary Care)</td>
<td>$100/visit</td>
<td>$1,500-$10,000 annual</td>
</tr>
<tr>
<td>Average Visits / Patient</td>
<td>6.8</td>
<td>8 - 9</td>
</tr>
<tr>
<td>Patient Load</td>
<td>2,500-4,000</td>
<td>Limited to 600</td>
</tr>
<tr>
<td>Average Time Spent with Patient</td>
<td>10.6 minutes</td>
<td>30 minutes</td>
</tr>
<tr>
<td>Patients per Week</td>
<td>112</td>
<td>30</td>
</tr>
</tbody>
</table>

Sources: Medical Economics Magazine; American Academy of Family Physicians; New England Journal of Medicine; FierceHealthcare

- Does Good care take longer? [Chen et al 2009, Health Care Reform].
- Making Sense of Primary Health Care. [NY times, January 17, 2010]
### Speed vs. Quality in Primary Care

<table>
<thead>
<tr>
<th></th>
<th>Regular Doctor</th>
<th>&quot;Concierge Doctors&quot; (MDVIP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price (Primary Care)</td>
<td>$100/visit</td>
<td>$1,500-$10,000 annual</td>
</tr>
<tr>
<td>Average Visits / Patient</td>
<td>6.8</td>
<td>8 - 9</td>
</tr>
<tr>
<td>Patient Load</td>
<td>2,500-4,000</td>
<td>Limited to 600</td>
</tr>
<tr>
<td>Average Time Spent with Patient</td>
<td>10.6 minutes</td>
<td>30 minutes</td>
</tr>
<tr>
<td>Patients per Week</td>
<td>112</td>
<td>30</td>
</tr>
</tbody>
</table>

*Sources: Medical Economics Magazine; American Academy of Family Physicians; New England Journal of Medicine; FierceHealthcare*

- Does Good care take longer? [Chen et al 2009, Health Care Reform].
- Making Sense of Primary Health Care. [NY times, January 17, 2010]
Majority of the literature considers service **speed** and service **value** as independent performance metrics.

Quality-speed tradeoff in services: Interaction between service speed and service quality.

Model of Customer-Intensive Service

- Potential demand for the service, $\Lambda$
- Service rate, $\mu$
- Joining decision based on:
  - Price, $p$
  - Expected waiting cost, $WC(\lambda, \mu)$
  - Value of the service, $V(\mu)$
  - Service value, $V(\mu)$, decreasing in service rate.

Queueing models of advance access in primary care settings (Bowers and Mould 2002, Green and Savin 2008)
Value of Customer-Intensive Service

- Service value increasing in the time spent on the service, \( t \)
  - \( V(t) = V_b + \frac{\alpha}{t_b} - \frac{\alpha}{t} \)
  - \( \alpha \): quality sensitivity of the service
    - \( V_b \): typical service value
    - \( t_b \): typical service time
Value of Customer-Intensive Service

- Service value increasing in the time spent on the service, $t$
  - $V(t) = V_b + \alpha/t_b - \alpha/t$
  - $\alpha$: quality sensitivity of the service
    - $V_b$: typical service value
    - $t_b$: typical service time

- Service value decreasing in the service rate, $\mu$
  - $V(\mu) = V_b + \alpha \mu_b - \alpha \mu$
  - $\mu_b$: typical service rate
Modeling Value of a Customer-Intensive Service

\( V = V_b + \alpha g(\mu_b) - \alpha g(\mu) \)  

(a) \( V = V_b + \alpha f(t_b) - \alpha f(t) \)  

Figure: Service Value for labor-intensive services varies with service time

- Valuation as a function consistent with diagnostic services with sequential hypothesis testing (Wang et al 2008)
Service Provider’s Operating Region

- Service has to be fast enough: $\mu > A_1(\alpha)$
  - No one will wait forever even if the service value is high.
- Service cannot be "too fast": $\mu < A_2(\alpha)$
  - Not possible to provide valuable service at really high speeds.
Queueing Game

Customers Know
- Service rate, $\mu$ → Service value, $V(\mu)$
- Potential demand, $\Lambda$
- Price, $p$

Queue Joining Decision

Join IF: $V(\mu) - p \geq \text{Expected Waiting Cost}$

Leave IF: $V(\mu) - p < \text{Expected Waiting Cost}$

- No Market Coverage: No customer joins the queue if $V(\mu) - p \leq c/\mu$.
- Full Coverage: All customers join the queue IF $V(\mu) - p \geq c/(\mu - \Lambda)$.
- Partial Coverage: $V(\mu) - p < c/(\mu - \Lambda)$ and $V(\mu) - p > c/\mu$, there exists some mixed equilibrium such that some customers join, and others balk.
Service provider maximizes revenues with respect to price and service rate:

$$\max \{p \geq 0, \mu \in \mathcal{F}(\alpha)\} \ \{ R(\mu, p) = p\lambda_e(\mu, p) \}$$

Equilibrium demand:

$$\lambda_e(\mu, p) = \begin{cases} \Lambda & \text{if } 0 \leq p \leq V(\mu) - WC(\mu, \Lambda) \\ \mu - \frac{c}{V(\mu) - p} & \text{if } V(\mu) - WC(\mu, \Lambda) < p \leq V(\mu) - WC(\mu, 0) \\ 0 & \text{if } V(\mu) - WC(\mu, 0) < p. \end{cases}$$
Service Provider’s Price Decision

- For a fixed service rate $\mu$ in the operating region $\mathcal{F}(\alpha)$, service provider maximizes revenues with respect to price.
  - Equilibrium outcome depends on the threshold $\hat{\lambda}(\mu) = \mu - \sqrt{\frac{c\mu}{V(\mu)}}$.
  - Optimal price:
    \[
    p^*(\mu) = \begin{cases} 
    V(\mu) - WC(\mu, \Lambda) & \text{if } 0 \leq \Lambda \leq \hat{\lambda}(\mu) \\
    V(\mu) - \sqrt{cV(\mu)}/\mu & \text{if } \hat{\lambda}(\mu) < \Lambda.
    \end{cases}
    \tag{1}
    \]
  - Resulting equilibrium arrival rate:
    \[
    \lambda_e(\mu, p^*(\mu)) = \begin{cases} 
    \Lambda & \text{if } 0 \leq \Lambda \leq \hat{\lambda}(\mu) \\
    \hat{\lambda}(\mu) & \text{if } \hat{\lambda}(\mu) < \Lambda.
    \end{cases}
    \tag{2}
    \]
  - Corresponding equilibrium revenue:
    \[
    R(\mu, p^*(\mu)) = \begin{cases} 
    (V(\mu) - WC(\mu, \Lambda))\Lambda & \text{if } 0 \leq \Lambda \leq \hat{\lambda}(\mu) \\
    \mu V(\mu) - 2\sqrt{c\mu V(\mu)} + c & \text{if } \hat{\lambda}(\mu) < \Lambda.
    \end{cases}
    \tag{3}
    \]
As services become customer-intensive, monopoly is less likely to cover the market.

⇒ Higher customer-intensity leads to partial coverage.
Partial Market Coverage: Equilibrium Price, Throughput and Revenue

- **Region 1 (Low $\mu$):** Increase price and admit more customers
- **Region 2 (Intermediate $\mu$):** Decrease price and admit more customers
- **Region 3 (High $\mu$):** Decrease price and admit less customers

![Graph showing price and throughput in different regions]
Partial Market Coverage: Equilibrium Price, Throughput and Revenue

- Region 1 (Low $\mu$): Increase price and admit more customers
- Region 2 (Intermediate $\mu$): Decrease price and admit more customers
- Region 3 (High $\mu$): Decrease price and admit less customers

$V_b = 10, \mu_b = 5, c = 2, \alpha = 2$
Partial Market Coverage: Price-Throughput Tradeoffs - An Example

- Increase service rate from $\mu = 4$ to $\mu = 6$.
- Examining equilibrium price-demand curve ...

![Graph showing the price-demand curve with a service rate increase from \( \mu = 4 \) to \( \mu = 6 \).]
Partial Market Coverage: Price-Throughput Tradeoffs - An Example

- Increase service rate from $\mu = 4$ to $\mu = 6$.
- Examining equilibrium price-demand curve ...
Partial Market Coverage: Price-Throughput Tradeoffs - An Example

- Increase service rate from $\mu = 4$ to $\mu = 6$.
- Examining equilibrium price-demand curve ...
Partial Market Coverage: Price-Throughput Tradeoffs - An Example

- Increase service rate from $\mu = 4$ to $\mu = 6$.
- Examining equilibrium price-demand curve ...
Partial Market Coverage ($\Lambda > \lambda^*_\alpha$)

- **Optimal service rate:** $\mu^* = \frac{V_b + \alpha \mu_b}{2\alpha}$.
  - As the service becomes more labor intensive ($\alpha \uparrow$) the service provider spends more time on each customer ($\mu^* \downarrow$).

- **Optimal price:** $p^*(\mu^*) = \frac{V_b + \alpha \mu_b - 2\sqrt{c\alpha}}{2}$
  - Optimal price falls (increases) as labor intensity increases for $\alpha < (>) c/\mu_b^2$.

- **Equilibrium Demand:** $\lambda_e(\mu^*, p^*(\mu^*)) = \frac{V_b + \alpha \mu_b - 2\sqrt{c\alpha}}{2\alpha}$.
  - Equilibrium demand falls (increases) as labor intensity increases for $\alpha < (>) \frac{V_b^2}{c}$.
  - Higher waiting costs ($c \uparrow$) lead to lower prices and lower equilibrium demand.
Full Market Coverage: \((\Lambda < \lambda^*_\alpha)\)

- **Service provider serves all potential customers:**
  \[
  \lambda_e(\mu^*, p^*(\mu^*)) = \Lambda
  \]

- **Optimal service rate:**
  \[
  \mu^* = \Lambda + \sqrt{c/\alpha}.
  \]
  As the service becomes more labor intensive \((\alpha \uparrow)\) the service provider spends more time on each customer \((\mu^* \downarrow)\).

- **Optimal price:**
  \[
  p^*(\mu^*) = V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{\alpha c}.
  \]
  If \(\Lambda\) is low, then optimal price falls (increases) as labor intensity increases for \(\alpha < (>) \frac{c}{(\mu_b - \Lambda)^2}\).

- Higher waiting costs lead to faster service at lower price.
Multiple \((n)\) servers owned by a single service provider providing service of intensity \(\alpha\).

- Service provider sets price to maximize total revenues.
- Servers set their own service speed to maximize their revenues.
- Surgeons who belong to the same health network (or hospital).
- Tax consultants at a single tax firm.
Large Market ($\Lambda \geq n\lambda^*_\alpha$)

- Servers act as monopolists.
- Competition has no effect on the optimal operating setting.
- Optimal service rate of server $i$ is $\mu^*_i = \frac{V_b + \alpha \mu_b}{2\alpha}$.
- Optimal price is $p^* = \frac{V_b + \alpha \mu_b - 2\sqrt{\alpha c}}{2}$.
Service Provider with Multiple Servers: Full Market Coverage

- When $\Lambda < n\lambda^*_\alpha$, servers share the market demand in equilibrium, serving at rate $\mu_i^e = \frac{\Lambda}{n} + \sqrt{c/\alpha}$.
  - Server spends more time on each customer than in the single server case.
  - Equilibrium expected waiting cost of a customer is identical to the equilibrium waiting cost of the single server case.
  - The optimal admission price is increasing in the number of servers.
Concluding Remarks

- A Queueing Model based explanation for Baumol’s cost disease.
- Normative Model: Service quality vs. time in customer-intensive services.
  - The quality speed trade-off analyzed from both customers’ and service providers’ perspectives.
  - Slow service is necessary to meet the service-level (quality) requirements.
    - Necessitating higher capacity investments $\Rightarrow$ *Baumol’s Cost Disease*.
  - As customer-intensity of the service increases, server must spend more time with each customer.
    - Faster service preferred over more valuable service when customer-intensity is low.
Price and demand for the service move in **tandem** as the customer-intensity changes.

In contrast with the downward-sloping demand curve because:
- Price ⇔ service quality ⇔ service rate.
- Revenues are determined by ⇒ the quality-speed trade-off.

Revenues are decreasing in the customer-intensity if the customer-intensity is low.

Revenues are increasing in the customer-intensity if the customer-intensity is high.

**Competition** in customer-intensive services may:
- Increase the prices,
- Enhance the service value,
- Without increasing the congestion.
Acknowledgements

Thank You

Fishman-Davidson Center for Service Research.

Philipp Afeche, Baris Ata, Gérard Cachon, Francis de Vericourt, Jack Hershey, Ananth Iyer, Raj Rajagopalan, Rob Shumsky, Alan Scheller-Wolf, and anonymous reviewers.

Participants at MSOM Service SIG 2009, Carnegie Mellon University, PennState, Northwestern University, University of Maryland, University of Pennsylvania, University of Rochester, INFORMS JFIG Competition Committee.
Research Questions...

Strategic Service and Expert Diagnosis (Pac and Veeraraghavan 2010).

- Heterogeneity in Customer-intensity $\alpha$.
- Credence good services and information-asymmetry
- Multi-stage Diagnostic Services.