Socio-Economic Planning Sciences 44 (2010) 19–24

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

journal homepage: www.elsevier.com/locate/seps

Might randomization in queue discipline be useful when waiting cost is a concave function of waiting time?

Jonathan P. Caulkins*

Carnegie Mellon University, Qatar Campus and H. John Heinz III College, 5000 Forbes Avenue, Pitttsburgh, PA 15213-3890, USA

ARTICLE INFO

Article history:
Available online 18 April 2009

Keywords:
Queueing
Queue discipline
Service in random order
Psychology of queueing

ABSTRACT

This paper suggests that introducing randomization in queue discipline might be welfare enhancing in certain queues for which the cost of waiting is a concave function of waiting time. Concavity can make increased variability in waiting times good not bad for aggregate customer welfare. Such concavity may occur if the costs of waiting asymptotically approach some maximum or if the customer incurs a fixed cost if there is any wait at all. As examples, cost might asymptotically approach a maximum for patients seeking organ transplants who will not live beyond a certain threshold time, and fixed costs could pertain for knowledge workers seeking a piece of information that is required to proceed with their current task, so any delay creates a "set up charge" associated with switching tasks.

1. Introduction

"First Come First Served" (FCFS) is often thought of as the best queue discipline for queues whose customers are people drawn from one homogenous population with no distinctions based on priority or processing time. "Last Come First Served" (LCFS) might be useful for computer scientists programming stacks or when service time increases when customers are not served immediately. "Service In Random Order" (SIRO) might be fine when the customers are inanimate objects, but FCFS is more common for queues comprised of homogeneous people. This predilection is not without basis. Deviating from FCFS leads to "slips" and "skips" that violate a sense of justice, sometimes with severe erosion of customer satisfaction [14].

However, this paper explores the possibility that FCFS may not always be preferred even for homogeneous customers. Two examples explored here are when not everyone will make it to the server and/or if there is unusual value in having a very short wait. In those cases, occasionally pulling someone into the server from further back in the queue might be preferable with respect to certain performance metrics.

The arguments here are entirely distinct from priority schemes or the Shortest Processing Time first (SPT) rule; with homogenous customers, waiting cost functions and processing time distributions do not vary from customer to customer.

To illustrate the idea, consider a contrived example. Suppose there were two, not just one, international space stations. Moments apart, both suffer catastrophic damage in the same meteor shower, disabling their escape modules and oxygen regeneration capacity, leaving the crews just one week to live. There is a single rocket on earth that can recover one crew at a time. The rocket can be launched in a week, but not again for another week, at which point it would only be recovering the second crews' bodies for burial on Earth.

How will outcomes differ if the distress calls are handled with an FCFS vs. SIRO queue discipline? First note what will not differ. There will be no change in the number of deaths or life-years saved or lost since one of the two crews will die in one week regardless of the queue discipline. Nor will there be any difference in waiting time.

One thing that will differ is the perception of time spent in queue. With SIRO, both crews spend a week knowing that they have a 50/50 chance of living or dying at the end of the week. With FCFS, one crew merely waits for a week, knowing with certainty that they will get rescued. The other essentially sits not in queue but on death row, knowing death will come at the appointed hour.

Which is preferable? Arrow [5] assures us we cannot answer that in any objective sense by adding the crews' utilities. We can, however, use a "veil of ignorance" test [19]. Imagine getting to chose between being born into a world with FCFS or with SIRO discipline given that we are equally likely to be on either crew, the one who called in the distress signal first by a few moments or the one who was second. Essentially that boils down to the following choice, sketched as a decision tree with position on the tree's arms indicating passage of time (Fig. 1).

* Tel.: +1 412 268 9590; fax: +1 412 268 5338.
E-mail address: caulkins@andrew.cmu.edu

0038-0121/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.seps.2009.04.001

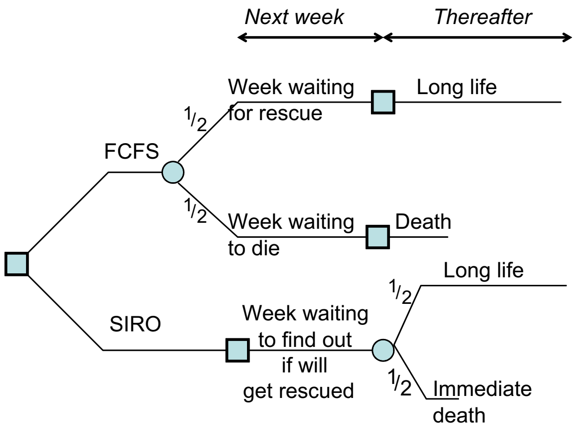

Fig. 1. FCFS and SIRO compared when one crew can be rescued in one week and there are two crews facing certain death in one week if they are not rescued.

To use the jargon of Quality Adjusted Life Years (QALYs), preferences for the FCFS vs. SIRO branches depend on how much life quality is lost during a week of facing certain death vs. spending a week facing a coin toss that will determine whether one lives or dies. If the first is less than twice the second, one prefers the FCFS branch. Otherwise, one prefers SIRO.

There is another way to think about this sort of choice. Suppose you are confronted today with the following terrible coin-toss: heads you live; tails you die one year from today, irrespective of when the coin is tossed. There is no way to avoid this lottery or to alter its odds. All you control is when the coin is tossed. When would you like to toss the coin? Now? In six months? Or not until a moment or a day before the death sentence would be carried out?

I would not argue that most people would or should prefer the SIRO branch. I personally would have a hard time making a choice. However, this example illustrates two points. First, the cost of waiting can be a concave function of waiting time; for the crews, the cost is the same whether the wait is two weeks or three or four. Second, the choice of queue discipline can matter, even if it has no effect on the standard performance metrics, because the queuing discipline can itself affect the perceived cost of waiting a given amount of time. Furthermore, there exist utility functions such that SIRO can be preferable to FCFS.

Given this motivation, the next section reviews the standard case for FCFS to remind ourselves what assumptions underpin that preference. This helps make clear why certain queues might be exceptions to the rule. The following sections examine stylized queuing models pertaining to organ transplantation and generic knowledge work to show how introducing randomness into the queuing discipline may be useful in diverse domains.

2. Reminder of why FCFS queues are appealing

The most obvious drawback to deviating from FCFS is the injustice of creating “slips” and “skips” meaning some people will be served sooner than others who have “seniority” by virtue of having waited longer [13]. The intuitive notion that FCFS is the fairest discipline can be made rigorous. Avi-Itzhak and Levy [7,8] argue from axiomatic principles that for a wide class of service disciplines, the variance of waiting time is a good measure of (departures from) fairness, and Vasicek [23] showed that under fairly general circumstances, variability is minimized by FCFS, maximized by LCFS, and takes on an intermediate value for SIRO. Indeed, this rank order of FCFS beating SIRO with LCFS bringing up the rear holds for any *convex* function of waiting times [23]. The qualification concerning convexity is key. The sound bite message of this note is that convexity may be the norm, but it is not universal.

Under fairly general circumstances, the rules governing which customer next enters service when a server becomes free do not affect *average* waiting times. Fuhrmann and Iliadis [12, p. 250–1] outline three conditions that jointly produce such invariance.

- Service is non-preemptive.
- Selection of customers for service is independent of their subsequent service times.
- If the queue is non-empty, the next customer enters service as soon as the previous customer completes service.

To the extent that queue discipline does not affect average waiting time and FCFS guarantees are most just, one needs a compelling additional or outside consideration to not implement FCFS. The literature describes three categories of such considerations.

The most familiar involve heterogeneous customers, with preference given to customers who are simpler (e.g., Shortest Processing Time first, fast track lanes in emergency rooms, express lanes in grocery stores, etc.) or more important (priority queues such as in 911 emergency dispatch). Maister [15] articulates this point eloquently, and Avi-Itzhak and Levy [7] note that their axiomatic proof concerning variance applies only to fairness with respect to *queue discipline*. It does not necessarily apply to fairness of the overall *queuing system*. Avi-Itzhak et al. [8] make the distinction clear with a simple example. If an emergency room identifies the next patient to enter service by asking “Who is the sickest?” the result might violate the principle of seniority, but it does not violate intuitive notions of fairness because it serves another (some would say higher) sense of justice.

Second, there are times, particularly in telecommunications applications, when keeping track of queue positions poses a computational burden that is significant relative to other cost considerations, so some simpler discipline, such as SIRO, may be preferred. As a related point, in network switching, there are queues of information packets, but the real customers are associated with streams of packets, not individual packets. So discussions of fairness focus on “flow-fairness”, “stream-fairness”, and other considerations beyond FCFS within an individual queue [11].

Finally, sometimes service time increases with waiting time [17]. Two examples would be solving crimes and tracing people who have come in contact with an infected person since in both cases memories and evidence fade over time. In such cases, LCFS might be preferred.

We do not dispute the merits of any of these traditional arguments for deviating from FCFS. Rather, we advance an additional, entirely distinct argument, one that applies to homogenous customers whose service time is independent of time in queue and is the sole determinant of system cost.

3. An organ transplant queue example

Organ transplant queues do already depart from FCFS by giving priority to sicker patients. Indeed, Alagoz et al. [2] describe seniority as being a second- or third-tier criteria in United Network for Organ Sharing (UNOS) rules for allocating cadaveric livers, and even then seniority is not literally time in queue but rather time in queue at a given health state. Also, Su and Zenios [20] observe that not all organs are equally desirable, so since the patient at the front of the queue does not have to accept an organ, FCFS may make the person at the front too selective about which organs to accept, generating excessive organ wastage (cf. also ref. [2]).

The model below abstracts away from such considerations. Realistic analyses of organ transplant queues are quite elaborate [1,2,20,21,25], and this note is meant merely to be a think piece that puts forward a contrarian idea. It highlights, though, three additional reasons why adding randomness might sometimes be preferred to pure FCFS.

(1) Utility might be concave, not convex, function of waiting time at least over some ranges,
(2) Queue discipline might directly affect the perceived cost of waiting, and
(3) Queue discipline might directly affect objective outcomes associated with waiting.

Waiting cost being concave in time spent waiting could stem from death creating a morbid form of reneging that shields customers from progressively increasing costs due to extremely long waits. For someone who will die within a year without a transplant, the disutility of what would have been a ten year wait is no worse than that of a two year wait.

The second factor relates to whether having some reasonable probability of getting an organ makes time spent waiting more tolerable. Hope is rarely factored into queuing analyses, but it does not exist only in Pollyanna's world; hope is empirically measurable in a scientific sense when defined as utility being enhanced by delaying the resolution of uncertainty [10]. Whether delaying uncertainty resolution is a net gain or loss for transplant patients is an empirical question because of the complicated ways people process risk-related outcomes [4,18,22]. Empirical studies of genetic testing reach conflicting conclusions about effects on psychosocial welfare (cf. [6,24]). However, in Chu and Ho's specific sense [10], randomness can create hope, and hope could reduce subjectively experienced waiting costs in some circumstances.

The third argument is that hope might not only make the wait more tolerable, but also improve actual health outcomes. There is a large and contested literature investigating whether hope or optimism improves health outcomes. It is beyond the scope of this note to review that literature. However, notwithstanding the fact that some people disagree, it is clear that (1) at least some people believe having hope can improve health outcomes and (2) there are literature reviews and empirical studies that offer support for that belief [3,16].

A fourth factor merits mention. Organ transplant queues are "blind" in the sense that customers do not observe each other waiting. Avi-Itzhak et al. [7] argue that blindness does not make deviations from FCFS any less unfair. However, if one is considering giving up FCFS, then it is reasonable to ask in a utilitarian sense, how much skip-related outrage will be generated? If a queues' blindness softens some of the outrage because customers merely know in general that the discipline allows skipping, but never knows if and when one has actually suffered a skip, that softening is germane.

Given these observations, it is useful to compare FCFS and SIRO with respect to a specific model. Suppose N people register to obtain a kidney at essentially the same time and in the same health state, namely one that gives them 12 months and only 12 months to live without a transplant. Suppose further that at the end of the year the number of kidneys available is a random variable that is binomially distributed with $n = N$ and some probability θ that might be on the order of 0.25 (meaning in expected value terms, one-quarter of those needing a transplant will get one, and three-quarters will die without getting a transplant).

A SIRO queue offers a 100% chance of spending a year facing probability θ of surviving beyond the end of that year. With FCFS you are equally likely to spend that year knowing your survival probability is $P\{X \geq n\}$, for $n = 1,2,\ldots,N$, where $P\{X \geq n\}$ is the probability a binomial random variable with parameters N and θ takes on a value of n or larger. Fig. 2 shows what this looks like for the case of $N = 4$.

The expected number of lives saved is the same in either case since for the FCFS case,

$$\sum_{n=1}^{N} P\{X \geq n\} = \sum_{n=0}^{N-1} P\{X > n\} = \sum_{n=0}^{N} P\{X > n\} = N\theta. \qquad (1)$$

A key question is whether SIRO or FCFS offers better life quality during the year spent waiting. Let $U(x)$ be the quality of life while in queue as a function of the probability of surviving beyond that first year. A Rawl's test would prefer SIRO if

$$U(\theta) < \sum_{n=1}^{N} \frac{1}{N} U(P\{X \geq n\})$$

which is true for any concave function $U(\ )$. Indeed, SIRO would be optimal since for any concave function $U(\ )$, the maximum over the vector of survival probabilities $\boldsymbol{p} = [p_1, p_2, \ldots, p_N]$

$$\max Z = \sum_{n=1}^{N} \frac{1}{N} U(p_i) \quad \text{subject to} \quad \sum_{n=1}^{N} p_i = N\theta \qquad (2)$$

occurs when $p_i = \theta$ for all i.

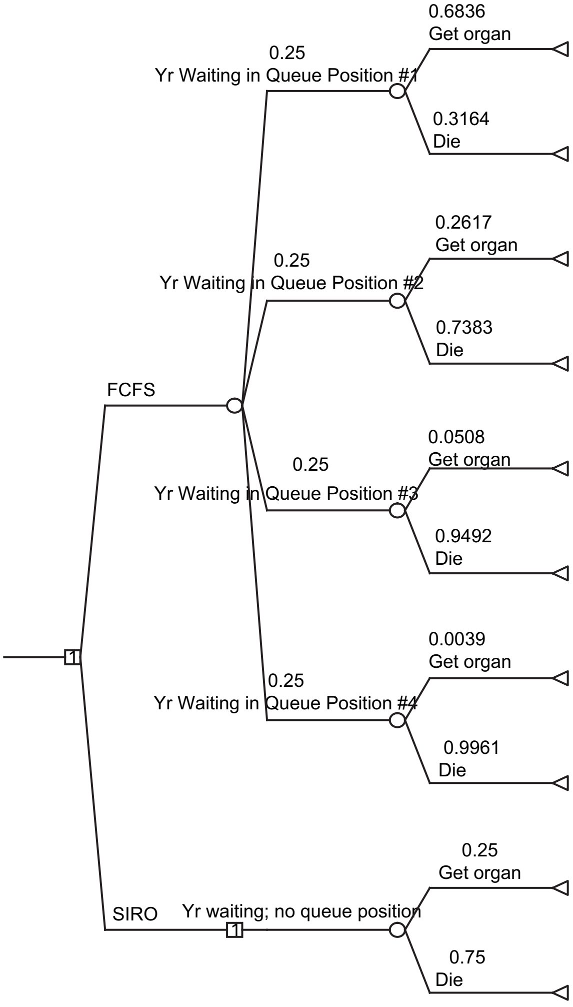
0.6836
Get organ
0.25
Yr Waiting in Queue Position #1
0.3164
Die
0.2617
Get organ
0.25
Yr Waiting in Queue Position #2
0.7383
Die
FCFS
0.0508
Get organ
0.25
Yr Waiting in Queue Position #3
0.9492
Die
1
0.0039
Get organ
0.25
Yr Waiting in Queue Position #4
0.9961
Die
0.25
Get organ
SIRO
Yr waiting; no queue position
1
0.75
Die

Fig. 2. Choosing FCFS or SIRO from behind a veil of ignorance about one's position in an organ transplant queue; 4 patients and $X \sim$ binomial(4,0.25) transplantable kidneys.

Numerical calculations for standard, stylized concave utility functions such as $U(x) = \ln(x)$ or x^{α} suggest that the gap between FCFS and SIRO in summed individual utilities can be nontrivial.[1] Such calculations are not terribly meaningful because they do not use empirically validated functions, and the whole notion of summing individual utility has dubious value. However, Fig. 3 illustrates the reason for this numerical result in a manner that is independent of the specific utility function $U(\ )$. It shows, as a function of the number of patients N, and for three different values of θ, what proportion of people in an FCFS queue have an intermediate probability of receiving a transplant ($0.05 < p_i < 0.5$). For even modestly large N, most people in an FCFS queue have either a very high probability of receiving a kidney (if they are near the front of the queue) or a very low probability (if they are near the back of the queue), whereas for moderate θ values, an SIRO queue gives all individuals a p_i large enough to get fairly far up a concave utility function.

4. Knowledge worker information request example

Many information-age workers juggle multiple tasks that require assembling information from disparate sources. Sometimes batches of requests for information can be submitted in parallel, but often the process is sequential. The knowledge worker discovers a need for information and submits an information request immediately, but it is inefficient to proceed further with the current task until that bit of information is obtained. If the (animate or inanimate) information server has no queue, then the information is obtained without delay so the knowledge worker continues working on the current task. However, if there is a queuing delay of more than a few moments, the knowledge worker does not usually sit in queue. Rather, he or she puts aside the current task and works on a different task until the requested information comes through.

An example familiar to readers is looking up information from the literature while working on a research project. Sometimes the library has a full text on-line subscription to the relevant journal, and the article of interest can be obtained without delay. Other times it must be ordered via interlibrary loan and so will not be available until the next day. In that case, the delay sometimes prompts one to put aside the current task (e.g., reviewing the literature) and switch to a different task, whether it is related (perhaps writing computer code for the same project) or not (grading homework). Other examples of task-critical information sought on a just in time basis might include the result of a database query, a manager's approval to proceed with a particular course of action, or a discrete piece of tacit knowledge resident in a colleague's head.

Shifting from one task to another can impose a productivity cost. The inefficiency can come from having to close down one set of computer files and open another or from cognitive processing limitations – less formally, needing time to change gears and get one's head around the new task. This amounts to a fixed cost paid whenever waiting time is more than negligible.

Variability in service time can still be undesirable, as in the classic analysis that favors FCFS, so cost might be a convex function of waiting time everywhere beyond this initial step up around 0. However, an example illustrates that the initial step function can be enough to make it optimal to introduce randomness into the queuing discipline. Since this is a proof by example that the converse is false, we examine an atypical queue whose unusual properties make the analysis particularly convenient.

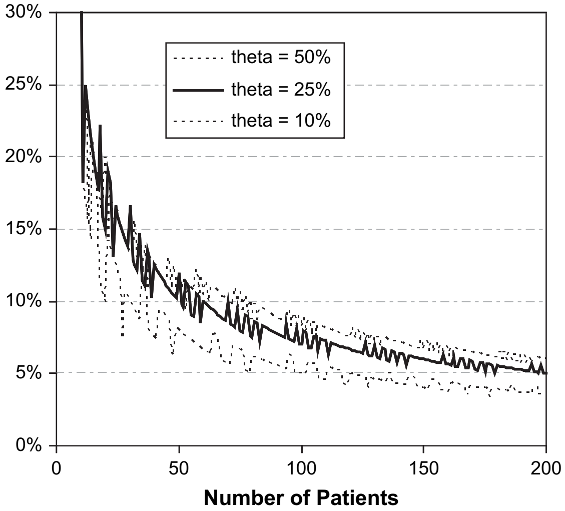

Fig. 3. Proportion of patients whose probability of receiving a kidney under FCFS is between 5% and 50% as a function of the number of patients, for three different p values (ratios of organs to patients).

Consider a D/D/1 queue for which the service and inter-arrival times are the same fixed constant, say 1 minute. For convenience, suppose that the arrival and service completion time epochs are coordinated, so the number of customers in queue never changes, $N(t) = N_0$ for all time t. With an FCFS queuing discipline the waiting time in queue is always exactly N_0 minutes with zero variability in waiting time.

Now consider introducing randomness to the queue discipline in the following way. Flip a coin each time a customer arrives (which also happens to be when the server becomes free). If it comes up heads (with probability p) the customer joins the queue normally; if it comes up tails, the customer enters the server immediately, skipping ahead of everyone waiting in queue.

Now with probability $(1 - p)$ the waiting time in queue is zero and the fixed cost of swapping tasks is averted. With probability p, the new customer tossed a head and went to the back of the queue. In that case, the waiting time in queue takes on its standard value of N_0 if and only if all of the next N_0 arrivals tosses a head, an event which occurs with probability p^{N0}. Otherwise, the randomization increases waiting time in queue. More generally, if one enters the queue, the time spent in queue equals the number of tosses it takes to observe N_0 heads, so it has a negative binomial distribution with parameters N_0 and p. So the overall, unconditional waiting time distribution with randomization is

$$p_W(w) = \begin{cases} 1-p & \text{for} \quad w = 0 \\ p\binom{w-1}{N_0-1}p^{N_0}(1-p)^{w-N_0} & \text{for} \quad w = N_0, N_0+1, \ldots \end{cases} \tag{3}$$

The variance of this distribution is

$$\sigma_W^2 = N_0\left(\frac{N_0+1}{p} - 1\right) - N_0^2 \tag{4}$$

Suppose the overall cost of waiting included a fixed cost c_0 if there is any wait plus a term that is proportional to the variance, with proportionality constant c. Slips and skips are not incorporated explicitly because, as with the transplant example, the queue would usually be "blind" since the knowledge workers would not literally

[1] For $N = 20$, $\theta = 0.25$, and $U(x) = \ln(1 + x)$, $U(x) = x^{0.8}$, or $U(x) = 1 - e^{-x}$, an SIRO queue achieves the same summed utility with a smaller θ of only around 0.2. For $U(x) = x^{\alpha}$ with a smaller α, the required θ for an SIRO queue is even smaller, $\theta = 0.103$ for $\alpha = 0.5$ and $\theta = 0.02$ for $\alpha = 0.2$.

stand in a physical queue where they could observe other people skipping ahead of them. Differentiating this cost function

$$\text{Cost} = c_0 p + c\left(N_0\left(\frac{N_0+1}{p} - 1\right) - N_0^2\right) \tag{5}$$

with respect to p yields first order conditions that are quadratic in p. They imply a minimum cost when the probability of entering the queue normally, p, is

$$p^* = \begin{cases} \sqrt{\frac{N_0(N_0+1)c}{c_0}} & \text{provided } \sqrt{\frac{N_0(N_0+1)c}{c_0}} < 1 \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

So if the penalty on increasing the variance in waiting time (c) and the queue length (N_0) are small relative to the fixed cost of having to wait at all (c_0), then a pure FCFS queue discipline is inferior to one that introduces some randomness.

Reflection on typical office etiquette suggests that we may already depart from FCFS queue discipline in ways that mitigate these fixed costs. It is one justification for answering email in reverse chronological order. Likewise, the general practice of leaving email turned on throughout the work day can be inconsistent with FCFS. Most office workers have a queue of tasks lined up when they arrive in the morning. A pure FCFS strategy would suggest not reading incoming messages until that queue is cleared. Few people do that. Those who occasionally refuse to turn on email in order to complete some overdue task can be seen as making an atypical special effort to abide by FCFS for a task that has already suffered a large number of skips.

Asking people to flip coins to decide whether to put an incoming job at the front or the back of the queue is not realistic, but introducing randomization is entirely possible in contexts such as responding to computer help desk inquiries. Many help desks use software to manage service requests. It would be easy for such software to implement randomization scheme like the one in the example above.

A similar application could be within email systems. Most already allow users to sort the inbox by ascending or descending date. They could add, as an option, sorting oldest to newest but with some random fraction of new incoming emails jumping to the front of the queue.

A mathematically parallel but substantively different application would be processing a request for government service, such as access to a civil court for contract dispute resolution; processing of a passport, zoning variance, or building permit application; or emergency housing placement. Contract disputes can stall construction and other projects. If the average delay with FCFS is long enough that contractors have to lay-off or redeploy workers, move equipment, and/or take actions to secure a worksite so that it does not become an attractive nuisance, then a partially random queue discipline that offers a 50/50 chance of no delay or twice the average delay might be preferable to FCFS. Likewise, if a family suddenly becomes homeless, they might prefer a 50/50 chance of getting emergency housing immediately or in two months over a certainty of having to wait a month if a month is long enough to force the family to find another more or less stable arrangement, to lose their job, and/or to sell their household possessions.

A related example is the US "green card lottery" through which the US annually offers permanent resident cards to 50,000 people from eligible countries. (The program, officially called the Diversity Immigrant Visa Program, excludes applicants from countries from which 50,000 people immigrated into the US in any one of the last five years.) The program attracted 6.4 million applicants in 2008. Under FCFS, someone applying in 2009 would have to wait a century to obtain a visa. The lottery, which is essentially SIRO, gives all applicants some hope that they will not have to wait so long. It also spares prospective applicants the burden of applying before they are sure they want to come just to save a place in line.

5. Discussion and further work

First Come, First Served (FCFS) is generally preferred for queues with homogenous customers, but this paper makes the contrarian suggestion that sometimes introducing randomness into the service discipline can improve customer welfare. Larson [14] observed that utility as a function of waiting time is often nonlinear. When cost is a convex function of time waited, the greater variability induced by randomness aggravates social cost. However, in certain applications the costs of waiting might be concave in waiting time at least over some ranges. This recognition can be seen as a special case of Bitran et al.'s [9] call for Operations Research models to embrace psychological richer and more realistic models of how customers experience temporal aspects of service encounters.

In organ transplantation, once the waiting time is so long that the patient will die before being served, additional waiting time ceases to be relevant. Furthermore, perceived waiting cost may depend not only on time spent in queue, but also on some integral of time weighted by the probability perceived at that time of ever being served. In other contexts there may be fixed costs associated with suffering any non-negligible wait, potentially yielding concave–convex cost functions. An example is knowledge workers seeking information that is necessary in order to proceed with a given task rather than switching to some different task that entails some "set up charge".

The traditional queue discipline rule of thumb might be summarized as, "bits and bytes may suffer injustice, but when the queues involve homogenous people drawn from the same priority and service time distribution, the discipline should be FCFS." This note amends that rule to be, "When people are literally standing on line, FCFS is the default, but for spatially distributed or virtual queues comprised of people and/or their processing requests, strategies that blend randomness and FCFS might have a role in creating the hope of being lucky, whether that is the existential hope of getting an organ transplant or the everyday hope of getting a quick answer to an information request."

The goal of this paper was merely to articulate a contrarian idea. Further work along these lines could take any of at least three forms: (1) empirical evaluation of changes in customer satisfaction after implementing the scheme (e.g., for computer help desk requests), (2) mathematical analysis of how much and what type of randomization is optimal for various waiting cost functions and queuing system structures, and (3) simulations to explore the consequences of introducing randomized queuing disciplines into more realistic models.

Acknowledgements

Ed Kaplan, Cindy Bryce, and two anonymous referees made useful comments on earlier drafts. This work was funded in part by the Qatar Foundation.

References

[1] Alagoz O, Maillart LM, Schaefer AJ, Roberts MS. The optimal timing of living-donor liver transplantation. Manag Sci 2004;50(10):1420–30.

[2] Alagoz O, Maillart LM, Schaefer AJ, Roberts MS. Determining the acceptance of cadaveric livers using an implicit model of the waiting list. 42–36. Oper Res 2007;55(1).

[3] Allison PJ, Guichard C, Fung K, Gilain L. Dispositional optimism predicts survival status 1 year after diagnosis in head and neck cancer patients. J Clin Oncol 2003;21:543–8.

[4] Andorno J. The right not to know: an autonomy based approach. J Med Ethics 2004;30:435–9.

[5] Arrow KJ. Social choice and individual values. New York: Wiley; 1951.

[6] Arver B, Haegermark A, Platten U, Lindblom A, Brandberg Y. Evaluation of psychosocial effects of pre-symptomatic testing for breast/ovarian and colon cancer pre-disposing genes: a 12-month follow-up. Familial Cancer 2004;3:109–16.

[7] Avi-Itzhak B, Levy H. On measuring fairness in queues. Adv Appl Probability 2004;36(3):919–36.

[8] Avi-Itzhak B, Levy H, Raz D. Quantifying fairness in queueing systems: principles and applications. RUTCOR research report RRR 26-2004, July. Piscataway, NJ: Rutgers University; 2004.

[9] Bitran GR, Ferrer JC, e Oliveria PR. Managing customer experiences: perspectives on the temporal aspects of service encounters. MSOM 2008;10(1):61–83.

[10] Chew SH, Ho JL. Hope: an empirical study of attitude toward the timing of uncertainty resolution. J Risk Uncertainty 1994;8(3):267–88.

[11] Floyd S, Fall K. Promoting the use of end-to-end congestion control in the internet. IEEE/ACM Trans Networking 1999;7(4):458–72.

[12] Fuhrmann S, Iliadis I. A comparison of three random disciplines. Queueing Systems 1994;18:249–71.

[13] Gordon E. New problems in queues: social injustice and server production management. PhD dissertation. MIT, Cambridge, MA: Operations Research Center; 1987.

[14] Larson RC. Perspectives on queues: social justice and the psychology of queueing. Oper Res 1987;35(6):895–905.

[15] Maister DH. The Psychology of Waiting Lines. In: Czepiel JA, Solomon MR, Suprenant CF, editors. The service encounter. Lexington, MA: Lexington Books; 1985. p. 113–24.

[16] Mondloch MV, Cole DC, Frank JW. Does how you do depend on how you think you'll do? A systematic review of the evidence for a relation between patients' recovery expectations and health outcomes. CMAJ 2001;165(2):174–9.

[17] Posner M. Single-server queues with service time dependent on waiting time. Oper Res 1973;21(2):610–6.

[18] Prelec D. Compound invariant weighting functions in prospect theory. In: Kahneman D, Tverky A, editors. Choices, values, and frames. New York: Cambridge University Press; 2000.

[19] Rawls J. A theory of justice. Cambridge, MA: Harvard University Press; 1971.

[20] Su X, Zenios S. Patient choice in kidney allocation: the role of the queueing discipline. MSOM 2004;6:280–301.

[21] Su X, Zenios S. Recipient choice can address the efficiency-equity trade-off in kidney transplantation: a mechanism design model. Manag Sci 2006;52: 1647–60.

[22] Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. J Risk Uncertainty 1992;5:297–323.

[23] Vasicek OA. An inequality for the variance of waiting time under a general queuing discipline. Oper Res 1977;25(5):879–84.

[24] Wermer MJH, van der Schaaf IC, Van Nunen P, Bossuyt PMM, Anderson CS, Rinkel GJE. Psychosocial impact of screening for intracranial aneurysms in relatives with familial subarachnoid hemorrhage. Stroke 2005;36:836–40.

[25] Zenios SA. Optimal control of a paired-kidney exchange program. Manag Sci 2002;48:328–42.

Jonathan P. Caulkins, PhD, is Professor of Operations Research and Public Policy at Carnegie Mellon University's Qatar campus, and Heinz College, Pittsburgh, PA. He received a BS and MS in Systems Science from Washington University, St. Louis, MO, an SM in Electrical Engineering and Computer Science, and a PhD in Operations Research, both from MIT. Professor Caulkins specializes in mathematical modeling and systems analysis with particular focus on social policy systems pertaining to drugs, crime, terror, violence, and prevention – work that won the David Kershaw Award from the Association of Public Policy Analysis and Management. Other interests include software quality, optimal control, black markets, airline operations, and personnel performance evaluation. He has taught his quantitative decision making course on four continents to students from 47 countries at every level, from undergraduate through PhD and executive education, and has published seven books and monographs, and more than 85 articles in such journals as *Operations Research*, *Management Science*, *JASA*, *Automatica*, *Decision Support Systems*, *JPAM*, *The American Journal of Public Health*, *Mathematical Biosciences*, *IEEE Security & Privacy, The Journal of Environmental Economics and Management*, *The Journal of Economic Dynamics and Control*, and the *Journal of Optimization Theory and Applications*, among other outlets. At RAND, he has been a consultant, visiting scientist, Co-Director of their Drug Policy Research Center (1994–1996), and Founding Director of its Pittsburgh office (1999–2001).