ABSTRACT

In this paper we discuss fairness in queues, view it in the perspective of social justice at large and survey the recently published research work and publications dealing with the issue of measuring fairness of queues. The emphasis is placed on the underlying principles of the different measuring approaches, on reviewing their methodology and on examining their applicability and intuitive appeal. Some quantitative results are also presented.

The paper has three major parts (sections) and a short concluding discussion. In the first part, fairness in queues and its importance are discussed in the broader context of the prevailing conception of social justice at large. A special effort, including illustrative examples, is made to differentiate between fairness of the queue and fairness at large, which derives from favoring the more needy. The second part is dedicated to explaining and discussing the three main properties expected of a fairness measure: conformity to the general concept of social justice, granularity, and intuitive appeal and rationality. The third part reviews the fairness of the queue evaluation and measuring approaches proposed and studied in recent years. We describe the underlying principles of the different approaches, present some of their results and review them in context of the three main properties expected from a measure.

The short discussion that follows centers on future research issues.

1. Introduction

1.1. Preface

Why are we using ordered queues? Why do they serve in many real life applications, such as banks, supermarkets, airports, computer systems, communications systems, Web services, call centers and numerous other systems?

While the major reason for the formation of queues is economic, i.e. scarcity of resources, the dominant reason for using ordered (disciplined) queues is often the strive to maintain some level of social justice, or in other words fairness in treatment of everyone involved.

In this sense, a system serving a queue of people is a microcosm social construct. Emotions and resentment may flare if unfairness is practiced, or is perceived as being practiced, while courtesy, and even camaraderie due to same experience-sharing, may result when fairness in treatment is perceived (see Rafaeli et al (2002)). Notwithstanding its fundamental role, the fairness factor was virtually neglected, or even disregarded, in the published queueing literature until quite
recently. Aspects of fairness in queues were recognized and discussed, or mentioned in passing, quite early by a considerable number of authors: Palm (1953) deals with judging the annoyance caused by congestion, Mann (1969) discusses the queue as a social system and Whitt (1984) addresses overtaking in queues, to mention just three.

While almost every child, if asked, can tell you what is fair and what isn’t, it is not an undemanding undertaking to have a group of people agree on a common definition of fairness, much more so when it comes to defining a quantitative measure of the level of fairness, and when the group is vast. It is not surprising, then, that extensive research aimed at developing fairness measures for queues, in contrast to the traditional “efficiency” measures of sojourn and waiting times, has been slow in coming.

Traditionally, a first-come-first-served (FCFS), or a first-in-first-out (FIFO), queue discipline is considered most fair. This probably derives from experience in queues where the total amount of service the system is able, or willing, to dispose is limited by a maximal number served or by a length of time the system is open for service, i.e. exhaustible-servers systems. In such systems (e.g. a line at a gas pump at a time of energy crisis, a line for basic foods in a refugee camp, or less dramatic, a line for tickets for a show or a sport event), which were very prevalent in the human experience, if you are not early enough in the queue, chances are you will never get the service, or product, or you may have to come again at a future time. Namely, the early bird gets the worm. Placing ahead of you a person, who arrived after you, will be regarded as grossly unfair, particularly if that person is not needier than you. Many present day queueing systems, however, are not of this type, rather, all birds get their worms, not only the early ones, and thus FIFO may not be as crucial in these systems. Fairness of exhaustible-servers queues is an important issue, deserving attention on its own, and is outside the scope of this paper, that focuses on non-exhaustive servers.

Larson (1988) in his discussion paper on the psychology of waiting recognizes the central role played by ‘Social Justice’, (which is another name for fairness). In the first part of his paper, dedicated to social justice in queues, he brings several anecdotal actual situations, experienced by him and others (Martin (1983), Kettelle (1986) and Lewin (1986)), that strongly support the traditional claim of FIFO being the most socially just queue discipline. In fact he practically defines social injustice as violation of FIFO when stating “…customers may become infuriated if they experience social injustice, defined as violation of FIFO.”

What would be a fair service order in a supermarket queue or in an airport waiting line? Many people would instinctively embrace Larson’s view, responding that FIFO is the fairest order, that is, serving the most senior customer first, where seniority is measured in the time the customer has already spent in the line. Already Kingman (1962) pronounces this same view by calling FIFO “the fairest queue discipline”. The underlying principle, or rationale, of this view can be expressed in one sentence, the one who has been waiting longest earned the right to be served first. But, recalling that the server is non-exhaustible, is FIFO undeniably the most fair discipline?

To answer this question, consider a common situation at a supermarket counter, which some readers may associate with their own personal experience: Mr. Short arrives at the supermarket counter holding only one item. In the line ahead of him he finds Mrs. Long carrying a fully loaded cart of items. Long says to Short “Excuse me, I only have one item. Would you mind if I go ahead of you?” Would it be fair to have Mr. Long served ahead of Short and Short waiting for the full
processing of Mrs. Long’s loaded cart? Or, would it be more fair to advance Short in the queue and serve him ahead of Long? This dilemma may cause some to “relax” their strong belief in the absolute fairness of FIFO. In fact, the dilemma brings to the discussion a new factor, that of service requirement. The basic intuition thus suggests that prioritizing short jobs over long jobs may also be fair, based on the underlying principle, the one who demands the least of the server’s time should be served first. It is the trade-off between these two factors, seniority (prioritize Mrs. Long) and service requirement (prioritize Mr. Short), that creates the dilemma in this case. To demonstrate the conflict we continue our scenario in two directions: (i) Long looks at Short, smiles and says “Why don’t you go ahead of me. I have arrived only a few seconds ago and it is not fair that you will wait that long while your short service will delay me very little”. This is one possibility. Alternately, Long may be negative, saying (ii) “Look, I have been waiting in this line forever. If not for this lengthy wait I would have been out of here long before your arrival. You can patiently wait too”. Clearly, Long weighs their seniority difference against their service requirement difference in deciding what is the fair thing to do. This tradeoff, illustrated by the “Long vs. Short” scenario, will accompany us in this paper in attempting to understand fairness in queues. (It should be noted that many supermarkets handle this conflict by allocating some of the counters exclusively to Shorts who also retain the option to select a “regular” counter.)

Figure 1:

1.2. What is “Fairness of the Queue”?

Evidently, there is a need to agree upon the definition of fairness, or at least the underlying principles, or rationale, that forms its foundation. As mentioned earlier, a queueing system is a microcosm social construct and its fairness should conform to the general cultural perception of social justice in the particular society. Social justice has always been, and still is, a cardinal issue in all cultures. It is the cement holding the society together. As such it has been subject to debate by philosophers, prophets and spiritual leaders since the beginning of recorded history. In modern time, many economists and social scientists joined the ongoing debate. Since social justice perception is culture and time dependent we are interested in the modern western societies perspective. As is to be expected, there is a vast ocean of modern research and publications on this issue, mostly by philosophers, economists and social and behavioral scientists. Reviewing and interpreting this literature is much beyond the scope of this paper, and probably also beyond our ability. To readers who would like to dive into this ocean, or just wet their feet at its shores, we recommend to start with
visiting the Stanford Encyclopedia of Philosophy (2005). A most, some would say the most, prominent and comprehensive publication on this issue is Rawls’ book “A Theory of Justice” (1971, 1999). The book does not make for an easy reading, but in essence, Rawls’ general conception of social justice, as summarized in a nutshell by Piccard (2005), is:

> All social primary goods – liberty and opportunity, income and wealth, and the bases for self-respect – are to be distributed equally unless an unequal distribution of any or all of these goods is to the advantage of the least favored.

We are back to the traditional economists’ approach of achieving social justice by appropriately dividing the “pie”, except that the pie here is made of a mix of tangibles and non-tangibles, while the traditional economists’ pie is wholly tangible. By Rawls conception, if all persons involved are equally non-favored (equally needy) the pie should be equally divided. Obviously, Rawls’ conception, though widely recognized, has its dissenters, as is true for practically any social issue. We will use it here as a guideline.

### 1.2.1. Fairness of the Queue vs. Fairness in a Queueing System

Social justice in a queueing environment, i.e. a queueing system, does not differ from social justice at large and if we accept the above conception it must also apply to queueing environments as well. We therefore need to differentiate between fairness in a queueing system and fairness of a queue, which is, roughly put, the fairness component that is attributable to the queue discipline or structure. For illustration, imagine a waiting room packed with patients. The door to the doctor’s office opens and a nurse appears and asks: “Who is the sickest?” This order of service is near to longest–job–first, LJF. Still, the many, if not most, will say it is fair by the principle that those most at-risk, or those suffering the most, should be attended to first. The fairness issue is cast here in a queueing situation. Alas, it has little to do with fairness of the queue. Very few will categorize a LJF ordering as fair, given that all customers are equally needy.

![Figure 2](image)

---

1 Nussbaum (2001) describes Rawls as “the most distinguished moral and political philosopher of our age”.

---


In this discussion paper we define the *fairness/unfairness of the queue* as the fairness/unfairness that can be related to the discipline or configuration of the queue when all customers are equally needy. Customers will be assumed to be equally needy if they are only discernable by their arrival time and service requirement, and are identical in all other respects. The doctor’s waiting room scenario, which was artificially constructed to make a point, is very realistic when looking at hospitals’ ERs. Arriving patients are categorized into several classes of neediness (urgency, or critical level, of condition) and the classes are prioritized in accordance to their level of neediness\(^2\). The fairness related to the class prioritizing is determined by the nature of the required service and neediness of the customers. These may vary widely from one queueing system to another and a universal measure to quantify the related fairness is not likely to be found\(^3\). The “fairness of the queue” is related to the order of service within each class, under the assumption that same class patients are practically equally needy.

Note that most customers would not differentiate between fairness of the queue and the fairness of the system, unless specifically guided to do so. Rafaeli et al (2005, Study III) conducted an experiment comparing perceived fairness by customers in a multi-server/multi-queue system (each server has its own queue, served in a FIFO order) to that of customers served in the same system that has, in addition to the regular queues, VIP queues (e.g. business class check-in counters in an airport). Only responses of those served in non-VIP queues were considered. Average fairness in the VIP structure was found to be significantly lower than in the structure without VIP queues, unless people knew that those in the VIP queue had paid a special fee in order to join it. That is, the queue was perceived as unfair by participants who thought others are getting a preferential treatment with no justification. However, once they learned that the preference was bought for a special fee they perceived the same system as being fair. In the first situation we are dealing with the perceived fairness of the queue. In the second situation we are dealing with the perception of the system’s fairness at large. Participants perceive buying preferential treatment as fair.

In what follows, fairness is meant to stand for fairness of the queue, unless otherwise specified. The distinction between customers based on their neediness, “value” or other economic factors\(^4\) is not considered.

---

\(^1\) I learned about ER prioritizing both the hard way and the easy way. The hard way was when, about a dozen years ago, I had to “visit” a local ER due to a relative minor injury requiring several stitches. It didn’t take me long to realize that FIFO order was not followed. During a couple of hours wait I observed quite a number of severely injured new patients, mostly accidents victims, being treated before me. I didn’t perceive this as being unfair, in fact, I remember the guilt I felt at taking even the twenty minutes of a doctor’s time, which I felt was taken away from much needier patients. The easy way was when, a couple of years ago, one of our UG seniors approached me, requesting my supervision of an independent project. I found that he was doing an internship in a local hospital, and after some inquiry, he told me that the hospital is considering expanding its ER. He ended up putting together a queueing model to help determine how many beds will be needed. I ended up learning in detail quite a lot about ER operation. Ben Avi-Itzhak.

\(^2\) A possible “universal” way to address this is to model neediness levels by customer weights and study fairness in this framework; in this case the burden of dealing with various problems is on selecting the weights (based on the problem) while the queue fairness measure is uniform. This is a subject for future research.

\(^3\) There is a large body of literature on the aspects of queue pricing and its relation to scheduling and customer behavior in the queue (see, e.g., a recent book by Hassin and Haviv (2002)).
1.2.2. What is the Pie?

Assume all customers are equally needy, what is then the “pie” and how can it be equally divided? Clearly the scarce resource, or the pie, is the service rendered by the servers. Consider an M/D/1 system where service requirement is the same constant for all customers. In such a system all customers seemingly receive an equal share of the server’s attention, hence FIFO and LIFO are equally fair! Not so, the pie’s division is not a one-time act. In an on-going process the pie must be continually divided, i.e. timely divided. Therein lies the key to the just division.

One can take either a circuitous approach or a direct one to attaining a just timely division of the resource. In the circuitous approach the customer is assumed to get the utility of the service plus the disutility of the wait. Therefore in the M/D/1 case, where equal service time is given to all customers, the waiting times must also be identical, to attain absolute fairness. Unfairness in this case is produced by deviations from equal wait, and a “natural” measure for it is the waiting time variance. The fairest discipline must produce the smallest waiting time variance. For the M/D/1 class of disciplines that are non-preemptive (Processor Sharing is considered to be preemptive; see discussion of PS in Section 3.2) and work conserving, the smallest variance is produced by FIFO. This is also true for M/G/1, (Kingman (1962), Avi-Itzhak and Levy (2004)). In extending this approach to the M/G/1 system we note that customers receive unequal shares of the server’s attention, giving rise to the Long-versus-Short dilemma. One way to solve this conflict is to assume that absolute fairness is achieved if the waiting disutility of each customer is proportional to his service utility (assuming, for simplicity, linearity of both utility functions). The unfairness measure can be derived from the variability, or the normalized variability, of the deviations from this proportionality. In both cases of the circuitous approach, the particular “equal and timely pie division” results from a pragmatic perceived fairness of the queue, instead of vice versa. The conformity to the conception of general social justice is an after-the-fact rationalizing of the two pragmatic fairness-of-the-queue principles used.

The direct approach to dividing the pie, assumes that the community entitled to a slice of it at any time point is made of the customers present in the system at that time. If there are \( N \) customers present, each is entitled to receive \((1/N)\)-th of the servers’ attention (service rate), to achieve absolute fairness. Thus the pie is equally divided at all points in time. Deviations from this division of the server’s rate are unfair, and a summary measure of their variability can serve as an unfairness measure. In this approach the absolute fairness results from the just division of the pie, in contrast to the circuitous approach. Still, it remains to agree upon the definition of a deviation from the defined just division of the servers’ rate.

1.3. Importance and applicability of fairness of the queue

As already mentioned, the fairness factor has long been recognized in queueing literature. Larson (1987) brings several actual situations where fairness considerations play a role in deciding the structure and discipline of service systems. Nevertheless, queueing theory has been mostly occupied with the performance metrics of waiting time. This has been the main quantity (perhaps almost the sole

The relation between pricing and fairness is beyond the scope of this article and requires further study. See further discussion at the end of Section 4.
quantity) used in queueing theory to evaluate queueing systems (see, e.g., text books on the subject, Kleinrock (1975, 1976), Hall (1991), Cooper (1981), Daigle (1992)) and is frequently being looked at via the expected delay or its variance in steady state. Under this quantity, customer satisfaction decreases with the delay experienced by the job and thus customer’s objective is to minimize delay. The use of this quantity seems to be appropriate when the major performance issue associated with job queueing is indeed the delay experienced in the system. The fairness factor, though playing an important role in the design and operation of actual waiting systems, has recently become a topic of interest also to queueing theorists. Rothkopf and Rech (1987), in their paper discussing perceptions in queues, bring an impressive list of quantifiable considerations showing that combining queues may not be economically advantageous, contra to the “common” belief. At the end they concede however, that all these considerations may not have sufficient weight to overcome the unfairness perceived by customers (as suggested by Larson (1987), based on a private communication by A. Lewin) served in a separate queues structure.

Experimental evidence of the importance of fairness in queues was recently provided in Rafaeli et. al. (2002), who studied, using an experimental psychology approach, the reaction of humans to waiting in queues and to various queueing and scheduling policies. The studies revealed that for humans waiting in queues the issue of fairness is highly important, perhaps some times more important than the duration of the wait. For the case of common queue versus a separate one at each server, they found that the common queue was perceived as more fair. Probably for this reason we find separate queues mostly in systems where a common queue is physically not practical, e.g. traffic toll boots and supermarkets.

Fairness and efficiency are the major reasons for the need for disciplined queues. In queues, like in most situations of limited resources, there is a need to utilize, or share, the resources in an efficient and fair way. Thus an ordered queue is a fairness and efficiency management facility and is perceived as such by most service systems operators, particularly those subject to competition. Supermarkets, where common queues are not always practical, try to increase both fairness and efficiency by assigning some of the counters to Shorts only. The same practice is common to toll booths as well. An alternate solution is to make a common queue feasible by allocating the necessary additional resources. For example, if you arrive to Newark airport on an international flight you find that the passport control queue is common and an extra attendant is assigned, to orderly direct people to the next available server as to reduce overtaking.

1.3.1. How Applicable is Fairness of the Queue in a ‘Blind Queue’

In the course of our study of fairness in queues we were asked more than once “is fairness relevant at all in a blind queue?” There are many situations where customers cannot see each other and are not informed of the state of the system and the discipline used. Call centers know from experience that some customers are impatient and are likely to renege after a relatively short wait. More patient customers will hang on for quite a while before hanging up (pun not intended). Therefore, a waiting customer is more likely to be a patient one, as compared to a new arrival. Using LIFO waiting line discipline will result in retaining more customers and increased profit. However, most customers would consider LIFO as unfair, even if informed of it ahead of time, and outrageous if it is concealed and then revealed to them somehow. In fact, in today’s information age it is hard to expect such practice to remain concealed for long time. Suppose, nonetheless, that such LIFO practice can
indeed be hidden. Does it make the practice fair? No. Is fairness in this case relevant? This is a question of ethics. Is cheating right if it never gets disclosed and the cheater can get away with it unscathed? The answer to this question, like the answer to the question of whether fairness is relevant in a blind queue, depends on your ethical values.

In fact, making the queue less blind might be quite important to customers. Many call centers will inform you of your place in the line and sometimes provide you with an estimate of the wait involved. This allows you to be aware that the order of service is FIFO and enables you to renege now, instead of wasting so much of your time before reneging anyway. Both are fairness considerations. Along these lines, surveys of 911 callers who were classified by the police as “low priority”, and kept waiting a long time for police arrival, found that callers were not dissatisfied with the service, provided they were told that the police are busy with higher priority calls and tasks, and were also told to expect a long delay (see Larson (1987), Chan and Tien (1981) and McEwen, Connors and Cohen (1984)). In this case, though we are dealing with fairness based on need rather than fairness of the queue, the knowledge that the system is fair strongly influences the callers’ degree of satisfaction and prevents repeated calls and complaints.

1.4. Flow Related Fairness

Queueing model applications can be classified into 1) Job-based systems, and 2) Flow-based systems. In the former, the i-th customer, say \( C_i \), is associated with a single job \( J_i \) arriving at epoch \( a_i \). Of interest is therefore the performance experienced by that individual job, which is synonym to customer in this paper. In the latter, customer \( C_i \) is associated with a stream (or flow) of jobs \( J^1_i, J^2_i, \ldots \) arriving at epochs \( a^1_i, a^2_i, \ldots \) respectively. Of interest is the performance experienced by the whole flow. The applications associated with this latter model are communications networks applications where a customer (sometimes called source) is associated with a stream of packets that are sent through a communications device, e.g., a router.

Much work has been done and published in the context of communications networks where the concern is with flows traversing a communications node and in allocating the bandwidth fairly among the flows. This is in contrast to the present work that focuses on fairness to jobs. One of the earliest attempts to define flow-fairness is Wang and Morris (1985) where the Q-factor is defined. Later on, the research on flow fairness has flourished with the introduction of Weighted Fair Queueing (WFQ), which deals with the fair scheduling of packet flows. Some early papers on the subject are Demers, Keshav and Shenker (1990), Greenberg and Madras (1992), Parekh (1992), Parekh and Gallager (1993), (1994), Golestani (1994), Rexford, Greenberg and Bonomi (1996), Bennett and Zhang (1997). Many other papers have been published on this subject. A popular measure of fairness within that context is the relative fairness bound (Used by Golestani (1994) and others) which captures the maximum possible difference between the (normalized) service received by any two streams. As such it measures "fairness of throughput of streams"\(^5\).

\(^5\) Within the context of a network, the literature deals with fair allocation of bandwidth (e.g. the Max-Min fairness allocation (Jaffe (1981)), Proportional fairness (Kelly (1997)), and Balanced Fairness (Bonald and Proutiere (2004)), which is orthogonal to our work.
Our focus in this work is on job-based systems. In what follows, customer \( C_i \) and job \( J_i \) are synonymous and will be used interchangeably. Applications that are associated with this model are:

1. **Banks, supermarkets, public offices and the like**, in which customers physically enter queues where they wait for service and then get served.
2. **Some computer systems**, in which a customer (or a customer’s computer application) submits a job to the system and the customer gets satisfied when the service of the job is completed.
3. **Call Centers**, in which customers call into a call center to receive service, possibly wait in a virtual queue (while listening to some music) until being answered by "the next available agent". Call center queueing systems are conceptually identical to physical queueing facilities, such as banks or airlines counters, except that the queue can be blind unless the operator decides otherwise.

## 2. Properties expected of a fairness measure

When dealing with the introduction of a new queueing performance measure for an entity that is somewhat abstract and not very tangible, several questions should be brought up and discussed. What is the underlying principle or conception that is in the foundation of the measure? Does this principle conform to the wider, non-queueing related, approach to dealing with this entity? What is the physical quantity, or performance objective that should be dealt with? What are the physical properties that affect the measure? At what level of detail should the system be measured? How intuitive and appealing is the measure? These questions relate to three major properties characterizing the measure: (i) conformity, (ii) granularity and (iii) intuitive appeal and rationality. In this section we discuss these properties, to be used later in examining the fairness measures proposed recently in the literature.

### 2.1. Conformity to the general concept of social justice

For many people, fairness perception is very intuitive, almost instinctive. Why do people consider FIFO to be most fair in many situations? The answer is that it is "naturally" fair. Thus, approaches towards fairness of the queue are mostly based on pragmatic principles, e.g. seniority must be respected, or, customers requiring little should get priority (Short versus Long), or, waiting time should be in proportion to the service required. These pragmatic approaches are not necessarily directly based on an abstract general conception offered by “deep thinkers”. Nevertheless, also the general conception worked out by the deep thinkers (mostly philosophers) emerges from the same “natural” pragmatic cultural attitudes of the society, and is a product of much discussion and lengthy discourse by these thinkers who, frequently, represent some of the finest minds of the society. The underlying principle of a fairness measure should conform to the general cultural perception of social justice prevailing in the particular society, either directly or indirectly. If it doesn’t, its acceptance and usefulness may be deterred by inconsistencies and “surprises” in the form of counter-intuitive and unaccepted results.

### 2.2. Granularity

At what granularity level should the fairness performance metric conform to the underlying fairness principle? Our conclusion is that conformity is desirable on all three granularity levels of the system: (i) the individual customer level, (ii) the
scenario level (scenario is defined as a sample path of the stochastic process), and (iii) the system level. The measure should be useful in assigning a consistent and meaningful fairness (or unfairness) value to each individual customer, to each possible scenario and to the system as a whole. The following is a more precise description of the three levels:

(i) **Individual Customer (Job) Unfairness (Discrimination):** This is a quantity attributed to the individual job (customer). It represents the deviation of the treatment given to the customer, in a particular scenario, from the absolutely fair treatment as defined by the underlying fairness principle of the measure.

(ii) **Scenario (Sample path) fairness:** A summary-statistic that summarizes the discrimination as experienced by a (finite or infinite) set of jobs in a particular scenario (a sample path).

(iii) **System fairness:** A summary statistic of a probabilistic measure (e.g. expected value or variance) of the performance as experienced by an arbitrary job, when the system is in steady state. This can be extended to a similar measure for transient behavior of the system.

Addressing fairness at all three levels is similar to the addressing of the waiting time measure, which can also be evaluated at these three levels. It should be noted that queueing theory has dealt explicitly mainly with the third type of quantity (expected delay or its variance), as the other quantities are somewhat trivial in the context of customer delay. In the context of fairness, it is nonetheless essential to make explicit use of the individual and scenario quantities as well, since humans can feel them better and associate with them better than with the third quantity. This is important to building confidence in the fairness measure, which is somewhat abstract, non-tangible and difficult to feel.

In deriving the system fairness one may take two different approaches for dealing with the stochastic nature of the system:

1. **Fairness of actual measures:** This approach first computes the fairness for all situations in the system, and then uses some summary statistics function (e.g. the max operation or some type of expectation) to yield the system fairness measure. Thus, the approach compares the actual performance measure (as opposed to comparing the expected performance measure) observed by the individuals.

2. **Fairness of the mean:** This approach classifies the customers into classes and computes the expected performance (e.g. expected delay) of each of the classes; then all these expected values are compared to each other (by some summary statistics function, e.g. the max operation) to yield a measure of system fairness. Thus, in the comparison stage the entities that are compared to each other are the expected performance measures of the individuals and not the actual performance of the individuals.

To illustrate the difference between these approaches and the importance of granularity, consider any “pie division” problem, for example, a bonus $b$ divided by an employer among $n$ equally deserving employees. One approach is to consider the actual bonuses, $\{b_1, b_2, \ldots, b_n; b_1+b_2+ \ldots +b_n=b\}$, given to the employees, compare them to each other and then use a summary statistic to summarize them. Since all employees are equally deserving the absolutely fair slicing of the pie is into equal shares, namely, $b_1=b_2= \ldots =b_n = b/n$, then the discrimination (positive or negative) of employee $i$ is expressible as $(b_i-b/n)$, namely, the deviation from absolute fairness. We note that the sum of discriminations is always zero, since this is a zero-sum situation; if one employee gets more it is taken away from other employees. The unfairness of
the scenario can then be given by the averaged absolute values of the individual discriminations, or more convenient for analysis, by the averaged squared discrimination, $\Sigma(b_i-b/n)^2/n$.

Suppose now that the employer decides to use a probabilistic mechanism for slicing the pie. As a result, the bonuses are random variables $B_1, B_2, \ldots, B_n$ summing to $b$. The unfairness of the system, resulting from the unfairness of the scenario (synonym to realization) when using the first approach, is given by $\Sigma E[(B_i-b/n)^2]/n$. In this approach the unfairness is defined and computable for all three granularity levels. An alternate approach is to use the concept of fairness of the mean. In this, alternate approach the distribution of the pie is fair if $E(B_i)=b/n$ for $i=1,2,\ldots,n$. This yields a criterion for classifying systems into fair ones and unfair ones. What it really classifies is the probabilistic mechanism, or the lottery. The criterion is not applicable at the individual level or the scenario level. If one tries to apply it to a scenario, it will, in most cases, classify all possible scenarios (realizations) of a “fair” lottery as being unfair. To further illustrate, suppose the employer takes an “all or none” approach, by which it will be decided by lottery to grant one of the employees all the bonus money $b$, and all others get nothing. The fairness-of-the-mean principle will classify this system as fair, provided that the lottery gives even odds to all employees. Nevertheless, all possible scenarios (realizations) will be classified as unfair by the same criterion. In a gambling environment, where participants are psychologically prepared and willing to gamble, this lottery will be considered to be fair. It is very doubtful that employees will view it as gamblers do. Employees (and likewise customers in a queue) are individuals and experience unfairness individually and personally. Most of them are likely to view the implementation of this lottery as highly unfair, when compared to a deterministic even division of the bonus money. Putting it differently, the employees, unlike the fabled statistician, are wary of the possibility of drowning in the lake, notwithstanding its six inches mean depth. The first approach recognizes the difference in unfairness between the fair deterministic division of the bonus money and the “fair” division by use of the lottery. It assigns system unfairness value of zero to the former and $b^2/(n-1)/n^2$ to the later, $(b-\bar{b})^2/n$ if normalized by setting $\bar{b}=1$.

2.3. Intuitive appeal and rationality

Producing intuitively acceptable results is a highly important, maybe the most important, property expected of a fairness measure. Surprising results, whose disagreement with intuition cannot be rationally and convincingly explained, are most likely to be rejected. A measure producing such “surprises” is not likely to achieve wide acceptance and might be viewed, instead, as an interesting curiosity. A good measure is not supposed to invent “new” fairness; it is supposed to quantify the prevailing widely accepted conceptions of fairness. Reactions like “What’s so interesting about this measure? I knew intuitively that discipline $\phi$ is the fairest, without the help of the measure”, are perhaps the strongest proof of the validity of the measure. The question often is how much more fair is discipline $\phi$ as compared to other disciplines, under various operating conditions.

6 The approach can also yield a measure of unfairness in the mean. Suppose the lottery is such that $E(B_i) \neq b/n$ for some values of $i$. Then $\Sigma [E(B_i)-b/n]^2/n$ can be used as a measure of unfairness in the mean.
In this section we propose four, intuitively based, simple tests for the validity of a measure. These do not suffice to label a measure as valid, rather, not passing them is a red light that the measure is questionable.

Two fundamental quantities determine the queueing process and the job scheduling decisions. These are the arrival epochs and service times, $a_i, s_i$, of the customer $C_i$, $i=1,2,...$. As our goal is to focus on the fairness of the queue and neutralize other external parameters, we will deal with these variables only. (And, as discussed earlier, we will not account for external parameters, such as neediness of customers, payments made by customers to obtain preferential service, or a gold/silver/bronze classification of customers.) Since these quantities are the only remaining ones determining the queueing and scheduling process, they also serve as the fundamental variables for determining scheduling fairness. For convenience of presentation, we use the terms seniority, and service requirement. The seniority of $J_i$ at epoch $t$ is given by $t - a_i$. The service requirement of $J_i$ is $s_i$. One may recall that seniority and service-requirement were in the heart of the dilemma in the Short vs. Long scenario.

It is natural to expect that a “fair” scheduling discipline will give preferential service to highly senior jobs, and to low service-requirement jobs. This can be stated formally in the following two tests:

1. **(Weak) Service-requirement Preference Test:** If all jobs in the system have the same arrival time, then for jobs $J_i$ and $J_j$, arriving at the same time and residing concurrently in the system, if $s_i < s_j$ then it will be more fair to complete service of $J_i$ ahead of $J_j$ than vice versa.

2. **(Weak) Seniority Preference Test:** If all jobs in the system have the same service times, then for jobs $J_i$ and $J_j$, residing concurrently in the system, if $a_i < a_j$ then it will be more fair to complete service of $J_i$ ahead of $J_j$ than vice versa.

A stronger form of the preference tests is as follows:

3. **Strong Service-requirement Preference Test:** For jobs $J_i$ and $J_j$, arriving at the same time and residing concurrently in the system, if $s_i < s_j$ then it will be more fair to complete service of $J_i$ ahead of $J_j$ than vice versa.

4. **Strong Seniority Preference Test:** For jobs $J_i$ and $J_j$, residing concurrently in the system and requiring equal service times, if $a_i < a_j$ then it will be more fair to complete service of $J_i$ ahead of $J_j$ than vice versa.

The seniority preference test is rooted in the common belief that jobs arriving at the system earlier “deserve” to leave it earlier. The service-requirement preference test is rooted in the belief that it is “less fair” to have short jobs wait for long ones. It should be noted that when $a_i < a_j$ and $s_i > s_j$ (the Short vs. Long case) the two

---

7 An alternative view to “leave it earlier” is “enter service earlier”. These two alternative concepts are equivalent when service times are identical and service is uninterruptible. The latter concept might lead to difficulties in the case of service interrupting scheduling (e.g. any preemptive regime).
principles conflict with each other, and thus the relative fairness of the possible scheduling of \( J_i \) and \( J_j \) is likely to depend on the relative values of the parameters.

One may view these two preference tests as two axioms expressing one’s basic belief in queue fairness. As such, one may expect that a fairness measure will satisfy these tests. A fairness measure is said to satisfy a preference test if it associates higher fairness values with schedules that are more fair. A formal definition is given next:

**Definition:** Consider jobs \( J_i \) and \( J_j \), requiring equal service times and obeying \( a_i < a_j \). Let \( \pi \) be a scheduling policy where the service of \( J_i \) is completed before that of \( J_j \) and \( \pi' \) be identical to \( \pi \), except for exchanging the service schedule of \( J_i \) and \( J_j \). A fairness measure is said to satisfy the strong seniority preference test if the fairness value it associates with \( \pi \) is higher than that it associates with \( \pi' \).

Similar definitions can be given to the service-time preference test and to the weak-versions of the preference tests.

It is easy to see that if a fairness measure satisfies the strong preference test (either Service-requirement or Seniority) then it must also satisfy the corresponding weak preference test.

To illustrate the preference tests in the context of scheduling policies we review several common policies and examine whether they follow the preference tests. A formal definition is:

**Definition:** A scheduling policy \( \pi \) is said to satisfy the strong seniority preference test if for every two jobs \( J_i \) and \( J_j \), requiring equal service times and obeying \( a_i < a_j \), \( \pi \) completes the service of \( J_i \) ahead of that of \( J_j \).

A similar definition can be given for the strong service-time preference test and for the two weak preference tests.

Using these definitions, one can classify common scheduling policies as follows:

**a. FIFO:** The First-In-First-Out scheduling satisfies the strong seniority preference test. On the other hand, since it gives no special consideration to shorter jobs, it does not satisfy the service-time preference tests (weak or strong).

**b. LIFO and ROS:** The Last-In-First-Out and Random Order of Service policies do not satisfy the seniority preference test (either strong or weak). Furthermore, neither do they satisfy the service-time preference test.

**c. SJF and LJF:** The Shortest Job First (SJF) satisfies the strong service-time preference test. Nonetheless – it does not satisfy the seniority preference test (both strong and weak). The longest Job First (LJF) satisfies none of the tests.

**d. PS:** The Processor Sharing policy satisfies both the strong seniority preference and the strong service-time preference tests.

**e. FQ:** Fair Queueing, which is the non-weighted version of Weighted Fair Queueing (Parekh (1992) and Parekh and Gallager (1993)), serves the jobs in the order they complete service under Processor Sharing (unless some of the jobs are not present at the server at the time that the service decisions must be taken). This property and the fact that PS satisfies both of the strong preference tests, imply that FQ satisfies both the strong seniority preference and the strong service-time preference tests.