A Simulation Study of the Utilization of a Flexible Appointment Scheduling System

John T. Simon
Governors State University

Many services offer reservations for appointments. Often, the available appointment slots are preset, of equal length, and sequentially arranged throughout the workday. Customers choose among those slots for the most convenient time slot. Instead, if they can be offered a slot precisely to their convenience (if available), it may lead to gaps in the workday that cannot be filled, thus reducing the utilization of the service provider. Here, we estimate the amount of under-utilization introduced in such a flexible appointment system using simulation, assuming uniformly distributed demands. This is useful in balancing the service provider's utilization and the customer's waiting time.

Keywords: appointment scheduling system, flexible appointments, utilization, simulation

INTRODUCTION

The service sector is one of the most significant components of our economy, and therefore, the study of service systems utilization is of high importance. Some of the common methods used in studying utilization make use of queueing and scheduling models.

A basic queuing model (see for example, Shortle, Thompson, Gross, and Harris, 2018) assumes that customers arrive at random, wait in line, and are served on a first-come first-served basis. Given the arrival rate and service rate, one can compute measures such as the server utilization, average waiting time, and average length of the waiting line. The literature on queueing theory is vast, spanning over more than a century, with some of the earliest work done by Erlang in the telephone industry (Erlang, 1909).

In contrast, scheduling models typically allow the service provider to choose the sequence of service, with the objective of minimizing makespan or tardiness (see, for example, Pinedo, 2022). Studies in this area are probably older than the Egyptian pyramids.

In this paper, we examine a specialized scheduling model known as an appointment scheduling system. Here, customers make a request for service in a future time slot, and if available, this slot is reserved for them. As more requests come in, the time slots are filled up. Typically, the goal is to maximize server utilization, and for this reason, the slots are preset, often in sequential and equal intervals. New demand requests are given the option of choosing from the remaining available slots. Given sufficient demand, the entire workday of the server can be utilized.

This offers a rich area of discovery for those interested in the scheduling and management of service systems. Given their complexity, analytical solutions are difficult to derive for many of these models. There is a wealth of literature (see for example, Pritchard, Taylor, and Belford, 2025) that promotes the use of

Monte Carlo simulation in the business curriculum, and these appointment scheduling models are excellent examples of problems that can be investigated through simulation.

Many of the prior research (such as Cayirli and Veral 2003, Kaandorp and Koole 2007) have considered behavioral aspects, probability of no-shows, variability of service times, scheduling rules, etc., but in this paper we focus on a basic question: assuming a clean process (for example, constant service times, no no-shows or tardiness), how much of the server utilization is lost if we permit customers to choose any (i.e., not from a preset group) appointment interval if available? We model the customer request as having a length of one and being uniformly distributed throughout the workday. If a requested appointment interval is not available, the appointment may be denied (Model 1), or the nearest available interval may be offered (Model 2). There are analytical solutions for Model 1 assuming infinite demand (mentioned below). In this paper we provide simulation results for both models under finite demand, and indicate a variety of ways this can be extended to answer interesting questions regarding the trade-off between the cost of waiting time for customers and the cost to the service provider due to reduced utilization. If flexible time slots are permitted, Model 2 seems to be more natural to follow than Model 1. The author is unaware of any prior literature that has considered Model 2.

More explicitly, the goal is to consider models that allow flexibility in allotting the time slots. For example, say that originally the workday consisted of ten hours, and each service request was for one hour. The typical preset time slots would be the intervals (0, 1), (1, 2), …, (9, 10). The first customer to make a reservation may choose the slot (5, 6), the next customer (1, 2), the next (6, 7), and so on – the resulting utilization is shown in Figure 1. Given enough demand, and assuming a discrete uniform distribution of customer choice of slots during the workday, it is possible to fill the entire workday without any gap. On the other hand, without preset time slots, the first customer may ask for the interval (4.6, 5.6), the second customer may request (0.8, 1.8), the next (5.9, 6.9), etc., and if those are available, they will be reserved for them – see Figure 2. Here we will model a customer request as an interval (X, X+1) where X has a continuous uniform distribution in the interval (0, 9). Obviously, in this case, we may end up with gaps of less than one hour (like the interval (5.6, 5.9) in Figure 2), which cannot be filled, thus leading to a reduction in server utilization. The question here is: how much on average is this reduction? An understanding of this would be helpful in determining whether it is worthwhile to offer this flexible appointment schedule. For example, this flexible system would make sense if the cost of the server is near zero (so server utilization is less important), and the cost of the customers' time is very high (so the closer they can get to their desired time slot, the less the total system cost). To understand the two models mentioned earlier, again see Figure 2: if a new appointment for (4.4, 5.4) is requested, in Model 1, we reject the appointment; but in Model 2, we offer the slot (3.6, 4.6) instead (the slot (6.9, 7.9) is also available but is farther away).

FIGURE 1
PRESET TIME SLOTS

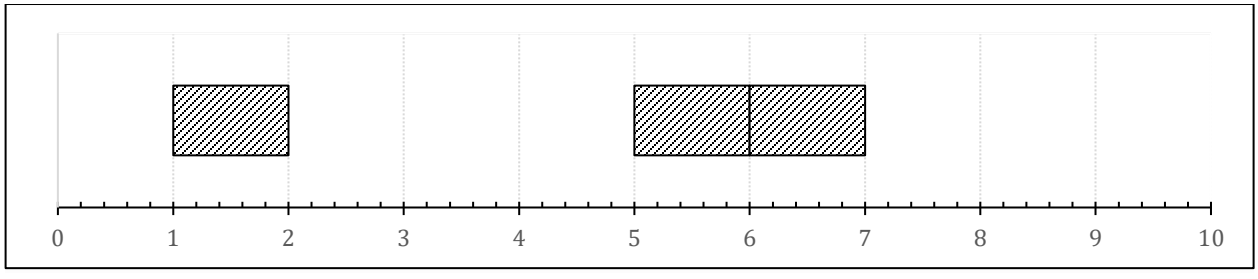
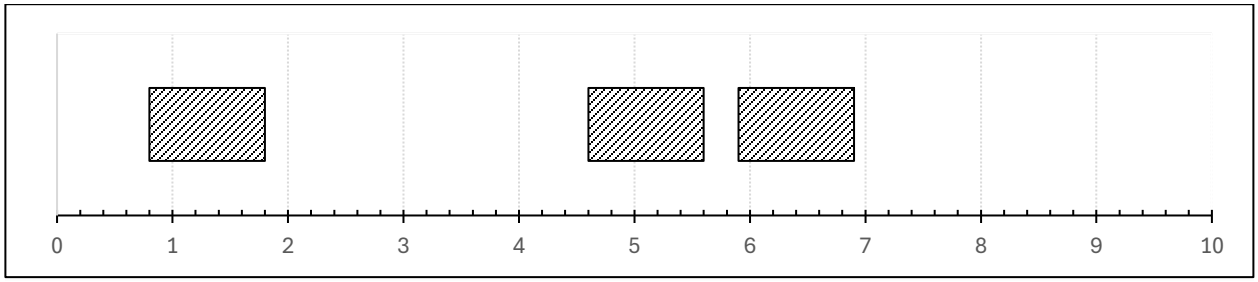


FIGURE 2
FLEXIBLE TIME SLOTS



APPOINTMENT SCHEDULING SYSTEMS: A BRIEF LITERATURE REVIEW

Appointment Scheduling Systems with preset service slots, as outlined in the previous section, are very common in services such as healthcare clinics, hair salons, hotels, and dental offices. A less common example arose when in order to reduce diesel emissions, the State of California imposed a $250 fine for trucks idling more than 30 minutes in line at the ports. A suggested solution was for trucks to make an appointment at the port (Giuliano et. al. 2008).

One of the earliest references in this area is Welch and Bailey (1952), which examined empirical data from a healthcare clinic and proposed an 'individual appointment rule' to minimize patient and facility idle time. Essentially it calls for scheduling a set number of patients at the start of the workday, and to schedule the remaining at equal intervals. Ho and Lau (1992), Ho, Lau, and Li (1995), Kuiper, Mandjes, de Mast, and Brokkelkamp (2021), and Niu, Lei, Guo, Fang, Li, Gao, Yang, and Gao (2024) provide good reviews of the existing research. In these studies, allowance is given for variability in service times, customer arrival times, and server availability times, as well as the possibility of no-shows.

Interestingly there are analogs of the Appointment Scheduling Systems in the studies of street-side parking as well as in physics. Consider parking cars (assuming that all cars are of equal length) along a street with no lines demarcating parking spaces. The first car to arrive chooses a random spot (say, chosen uniformly distributed along the length of the street), and let us assume that subsequent cars also choose similar random spots and park if the spot is available without any overlap with already parked cars (but leave if the spot is unavailable). Clearly this is identical to the appointment scheduling system (Model 1) as described above with flexible time slots (preset time slots would be akin to a street with parking spots marked at equal intervals along the street). Weisstein (2003) provides a dynamic visual model for this. In physics, a related problem is that of random sequential adsorption, as reviewed in Ramsden (1993). Here particles are placed on a solid surface without overlap – this can be modeled in one or more dimensions (the one-dimensional model would be similar to our model).

For the parking problem with an infinite source of cars (with the assumption that each new car has a randomly chosen desired parking spot, and leaves the street if that spot is unavailable – again, this is the Model 1 mentioned in the previous section), an analytical solution is given by Rényi (1958): the ratio of the length of the street utilized to that of the total length of the street, as the length of the street goes to infinity, is approximately 0.7476 (more precisely, it is equal to $\int_0^\infty \exp\left(-2\int_0^t \frac{1-e^{-u}}{u}\,du\right)\,dt$) (Weisstein, 2003). This number is known as Rényi's parking constant. In other words, about 25 percent of the length of the street will be lost due to small gaps that cannot fit a car.

On the other hand, if the desired parking spot is unavailable, the usual practice is to find the nearest available spot. This would lead to Model 2 mentioned in the previous section. As mentioned earlier, prior literature does not seem to cover this model. Our result is that in this case, the loss is about 16 percent for a considerably long street (see below).

SIMULATION MODELS AND RESULTS

Let us set the service slot length to be one, and the length of the workday to be the variable 'WorkdayLength' (in general, the service slot length is chosen by the service provider based on the service, and the length of the workday is fixed – but for our analysis only their ratio matters, and hence our choice is sufficient). Let us also set the variable 'Demand' to indicate the number of possible service requests for that workday. The utilization of the service facility would be the ratio of the number of services scheduled (recall that they are all of length one, and hence the count and the length are the same) to the WorkdayLength.

In our first model, we will use flexible slots with the request being rejected if the slot is not available. A pseudocode for the simulation is as follows:

FIGURE 3
PSEUDOCODE FOR MODEL 1

```
For a given WorkdayLength and Demand:
Schedule = an empty list
For i = 1 to Demand
        RandomValue = a Uniform (continuous) random number between 0 and (WorkdayLength – 1)
        NewAppointment = (RandomValue, RandomValue+1)
        If Schedule is free during NewAppointment, add NewAppointment to Schedule
                Else ignore the NewAppointment
Next i

Count = number of appointments in Schedule
Utilization = Count / WorkdayLength
```

The average utilization will change with WorkdayLength – for example, if the WorkdayLength is 1.5, it is easy to see that the average utilization would only be 0.67 (= 1/1.5) since only one service can (and will) be scheduled. Similarly if the WorkdayLength is 1.8, the average utilization would be 0.56 (= 1/1.8), and for a value of 2, the utilization would be 0.5 (since the probability of precisely timed requests for (0, 1) and (1, 2) is zero). When WorkdayLength increases to 3, the average utilization would rise to 0.67 (= 2/3). These fluctuations in average utilization will diminish as the value of the WorkdayLength becomes large, and will converge to Rényi's constant assuming infinite Demand. Clearly, the average utilization will also change with Demand – if Demand is low, the probability of WorkdayLength being filled will also be low, leading to lower utilization.

The results of the simulation are given in Table 1. Each result is obtained from 1,000 iterations of the simulation, and the average value and a 90% confidence interval for the percentage utilization of the server are provided for each result. The simulations were run in Python.

TABLE 1
SERVER UTILIZATION (AS A PERCENTAGE) FOR MODEL 1

	Demand					
	5	10	50	100	1000	10000
WorkdayLength						
2	50.00 +/- 0.0	50.00 +/- 0.0	50.00 +/- 0.0	50.00 +/- 0.0	50.00 +/- 0.0	50.00 +/- 0.0
5	47.32 +/- 0.64	56.26 +/- 0.58	67.3 +/- 0.52	67.38 +/- 0.51	69.66 +/- 0.52	69.46 +/- 0.52
10	33.84 +/- 0.41	47.39 +/- 0.44	66.52 +/- 0.37	69.03 +/- 0.35	72.30 +/- 0.33	72.23 +/- 0.35
50	9.22 +/- 0.06	16.89 +/- 0.10	47.05 +/- 0.20	59.1 +/- 0.18	72.60 +/- 0.14	73.90 +/- 0.15
100	4.82 +/- 0.02	9.18 +/- 0.04	32.74 +/- 0.12	47.14 +/- 0.14	71.35 +/- 0.10	74.23 +/- 0.10
1000	0.50 +/- 0.001	0.99 +/- 0.002	4.76 +/- 0.008	9.09 +/- 0.01	47.13 +/- 0.04	71.57 +/- 0.03

Many insights can be drawn from the results. It is intuitive that the utilization increases with Demand for a given value of WorkdayLength. We see that in each row of Table 1, the utilization is increasing (within experimental error) from left to right (except for the row with WorkdayLength of 2 – this is discussed below). As demand becomes very large (as best seen in the row of WorkdayLength = 100), the utilization does climb up close to the limit of 74.76 percent derived by Rényi (1958). For the row with a WorkdayLength of 1000, the demand must be considerably more than 10,000 for the utilization to approach Rényi's limit. Interestingly, if the Demand is equal to the Workday Length, the utilization is approximately 43 percent for the values given in the table. The dependence of utilization on Demand does not seem to have been explored in earlier literature; however, it is relevant since services do not have infinite demand, and often demand and Workday Length are comparable values.

WorkdayLength can be thought of as the ratio of the available work duration to that of the individual service interval (or in the parking analogy, the ratio of the length of the street to that of a car). For small values of this ratio, utilization would be poor. As mentioned earlier, it is easy to see that if this ratio is 2, the utilization would be 50%, since only the first request would be accommodated – the only way to have 100% utilization is to have the first service request to be either (0,1) or (1, 2), and another request to be the reverse of that, and this has zero probability. This explains the row with WorkdayLength of 2. Larger values of WorkdayLength permit more flexibility in accommodating service requests – but now Demand plays an important role, since only a large Demand can fill the possible gaps during the WorkdayLength. Thus, we see in the column for a Demand of 1000 that, as WorkdayLength increases, the utilization initially rises, but after WorkdayLength reaches approximately 50, the utilization begins to decrease again.

For our second model, we again utilize flexible slots; however, if a new requested service interval is not available, we will offer the nearest available interval. Here the following pseudocode can be used for its simulation:

FIGURE 4
PSEUDOCODE FOR MODEL 2

```
For a given WorkdayLength and Demand:
Schedule = an empty list
For i = 1 to Demand
        RandomValue = a Uniform (continuous) random number between 0 and (WorkdayLength – 1)
        NewAppointment = (RandomValue, RandomValue+1)
                If Schedule is free during NewAppointment, add NewAppointment to Schedule
                Else check if an open interval of length one is available to its left or right
                        If yes, set NewAppointment to the nearest open interval
                        Add NewAppointment to the Schedule
                              If no interval is available, exit the For loop
Next i

Count = number of appointments in Schedule
                                        Utilization = Count / WorkdayLength
```

The results of the simulation are given in Table 2. Each result is obtained from a thousand iterations of the simulation, and the average value and a ninety percent confidence interval for the percentage utilization of the server are provided for each result. These simulations were also run in Python.

TABLE 2
SERVER UTILIZATION (AS A PERCENTAGE) FOR MODEL 2

	Demand					
	5	10	50	100	1000	10000
WorkdayLength						
2	50.00 +/- 0.0	50.00 +/- 0.0	50.00 +/- 0.0	50.00 +/- 0.0	50.00 +/- 0.0	50.00 +/- 0.0
5	74.80 +/- 0.46	74.52 +/- 0.47	75.26 +/- 0.44	74.54 +/- 0.46	74.48 +/- 0.46	75.06 +/- 0.45
10	50.00 +/- 0.0	79.51 +/- 0.34	79.47 +/- 0.35	79.54 +/- 0.34	79.26 +/- 0.34	79.87 +/- 0.34
50	10.00 +/- 0.0	20.0 +/- 0.0	83.12 +/- 0.15	83.22 +/- 0.15	83.25 +/- 0.14	83.27 +/- 0.14
100	5.0 +/- 0.0	10.0 +/- 0.0	50.0 +/- 0.0	83.74 +/- 0.10	83.68 +/- 0.10	83.70 +/- 0.10
1000	0.5 +/- 0.0	1.0 +/- 0.0	5.0 +/- 0.0	10.0 +/- 0.0	84.14 +/- 0.04	84.10 +/- 0.03

As before, we see that within experimental errors the utilization increases with WorkdayLength (unless restricted by Demand). Utilization also increases with Demand. However, we observe that utilization increases to a level higher than that in Model 1, and for a given WorkdayLength, utilization quickly reaches its maximum value when Demand is close to the WorkdayLength. This is intuitive since when the Demand equals WorkdayLength, no further gaps would be available to accommodate a new appointment. To find a comparable constant to that of Rényi's parking constant of 74.76 percent, we used a longer run (10000 iterations) for the case of WorkdayLength of 1000 and Demand of 1000, and a ninety percent confidence interval for the mean utilization is 84.12 +/- 0.01 percent (or a loss of utilization of about 16 percent).

EXTENSIONS

These simulation models are accessible to undergraduate business students and can be used to explore the outcomes of various service policies. One possibility would be to introduce variability in service times

sought (but once scheduled, there would be no further variability). For specificity, suppose that each service request is for either 0.5 or 1.0 units of time, with equal probability. In the pseudocode for Model 2, we can modify the NewAppointment interval accordingly, and also change the utilization computation to be the ratio of the length of workday utilized to that of the workday. As an example, a simulation result using a thousand iterations for WorkdayLength of 100.0 and Demand of 1000 shows that a ninety percent confidence interval for utilization is 88.7 +/- 0.06. Instead, if we had preset slots of length one, about half would be filled by services of length 0.5, and about half by services of length 1.0, with an overall utilization of 75 percent. So in this special case, offering flexible slots is more efficient than offering preset slots of length one.

Other possible situations we can assess using minor modifications of the described simulation models are listed below:

1. Consider a cost (to the customer) based on the distance of the obtained appointment interval from the requested appointment interval. This cost can be based on the absolute distance, or the square of the distance, or may even be asymmetric in that an earlier appointment is less costly than a later appointment (or vice versa). Now we can compare the costs of preset slots versus flexible slots. We can include a cost for the lost utilization to the server as well.
2. Consider situations where Demand is less than the WorkdayLength. Here we can compare the utilization (and costs) of preset slots versus flexible slots.
3. Consider situations where if a requested service interval is unavailable, then only a future interval may be offered. In the case of parking, this would be similar to the case that when a required parking spot is unavailable, one must move forward only. In this situation, we can analyze the utilization of the system for a given Demand and WorkdayLength.
4. Consider the case where the demand requests that come early have a higher cost of moving away from the requested interval (assuming that those who make their request early have a higher concern regarding the precise service interval). Here costs may be compared between preset and flexible schedule slots.
5. Consider a general distribution for service times (which are set at the time of request). We could have preset slots of length given by the maximum value of the service times, or we could have some of the preset slots to be at the maximum value and others at a lower value. These can be compared to the flexible slot schedule (without preset slots) in terms of their utilization.
6. Consider customer request distribution that is different from a uniform distribution. For example, instead of requests being uniformly distributed during the WorkdayLength, we may have a higher likelihood of demand during the mid-value of the WorkdayLength. Again, costs and utilizations can be considered in a variety of situations.

CONCLUSION

Flexible time slots have generally not been considered in the literature for Appointment Scheduling Systems. Hence, we have explored it here using simulation to answer the basic question of the loss of utilization in such systems compared to fixed time slots. Given a large workday compared to the length of a single service and infinite demand, if service requests are rejected when they overlap with existing service appointments, it has been analytically demonstrated in the literature (Renyi, 1958) that the average server utilization will be approximately 74.76 percent. In this paper, we have used simulation to extend the analysis to include the computation of utilization in the case of limited demand. In the range of values tested, we found that when demand is approximately equal to the workday length, only about 47 percent of the workday is utilized. For a large value of workday and an even larger value of demand (such as WorkdayLength = 100, Demand = 10000, in Table 1), we find the utilization approaching the theoretical value of 74.76 percent. We have also introduced the possibility of offering the nearest available interval if a service request overlaps with existing service appointments. In this scenario, we find that utilization can be as high as 84.12 percent when the workday is large compared to the length of a single service and demand is sufficient to fill the workday.

This simulation approach seems to be a promising way to explore these models, and several avenues for further exploration have been suggested. It is also interesting to note the analogies between this model and street parking and random adsorption models in physics.

REFERENCES

Cayirli, T., &Veral, E. (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, *12*(4), 519–549.

Erlang, A.K. (1909). The theory of probabilities and telephone conversations. *Nyt Tidsskrift fur Matematik B*, *20*, 33–39.

Giuliano, G., Hayden, S., Dell'aquila, P., & O'Brien, T. (2008). Evaluation of the terminal gate appointment system at the Los Angeles/Long Beach ports. *METRANS Project*, pp. 04–06.

Ho, C.J., & Lau, H.S. (1992). Minimizing total cost in scheduling outpatient appointments. *Management Science*, *38*(12), 1750–1764.

Ho, C.J., Lau, H.S., & Li, J. (1995). Introducing variable-interval appointment scheduling rules in service systems. *International Journal of Operations & Production Management*, *15*(6), 59–68.

Kaandorp, G.C. & Koole, G. (2007). Optimal outpatient appointment scheduling. *Health Care Management Science*, *10*, 217–229.

Kuiper, A., Mandjes, M., de Mast, J., & Brokkelkamp, R. (2021). *Decision Sciences*, *54*(1), 85–100.

Niu, T., Lei, B., Guo, L., Fang, S., Li, Q., Gao, B., . . . Gao, K. (2023). A Review of Optimization Studies for System Appointment Scheduling. *Axioms*, *13*(1), 16. https://doi.org/10.3390/axioms13010016

Pinedo, M.L. (2022). *Scheduling: Theory, algorithms, and systems*. Springer International Publishing.

Pritchard, A., Taylor, D., & Belford, M. (2025). Teaching data-driven decision making for inventory analysis with Monte Carlo simulation. *Decision Sciences Journal of Innovative Education*, *23*, e12328. https://doi.org/10.1111/dsji.12328

Ramsden, J.J. (1993). Review of new experimental techniques for investigating random sequential adsorption. *Journal of Statistical Physics*, *73*, 853–877.

Rényi, A. (1958). On a one-dimensional problem concerning random space-filling. *Publ. Math. Inst. Hungar. Acad. Sci.*, *3*, 109–127.

Shortle, J.F., Thompson, J.M., Gross, D., & Harris, C.M. (2018). *Fundamentals of queueing theory* (5th ed.). John Wiley & Sons, Inc.

Weisstein, E.W. (2003). *Rényi's Parking Constants. From MathWorld--A Wolfram Web Resource. Retrieved* from https://mathworld.wolfram.com/RenyisParkingConstants.html

Welch, J.D., & Bailey, N.J. (1952). Appointment systems in hospital outpatient departments. *The Lancet*, *259*(6718), 1105–1108.