JOHN J. REGAZZI and RODNEY M. HERSBERGER

Queues and Reference Service: Some Implications for Staffing

In most organizations activities occur that produce waiting lines or queues. This study, undertaken at a medium-sized academic library, attempts through the use of a simulation technique (1) to analyze the extent to which queues develop at a reference desk during peak periods, (2) to propose alternative staffing models to reduce queues, and (3) through the use of a cost-effectiveness formula to examine the merits of the proposed alternatives.

THIS PAPER IS a case history of some staffing and service patterns at the reference desk of a medium-sized academic library. The study, however, is also an effort to apply standard quantitative techniques, i.e., queuing models and simulation, to a library management problem in the area of public services. These techniques, developed in business, industry, and science, have been widely used in such library areas as circulation,[1] library administration,[2] and technical services.[3] This particular case attempts to evaluate or "measure" reference services beyond the compilation of "reference statistics" that emphasize the evaluation of past performance or defend the value of a reference service.[4]

The emphasis here is on present services and staffing requirements for optimizing that service within the budgetary constraints of a particular library. In 1974 in a "Symposium on Measurement of Reference," it was stated that "the number one need for statistical information centered on information for staffing patterns including peak and idle periods, subject specialization and non-desk time."[5] Using certain quantitative techniques, this study illustrates how various alternative staffing patterns for peak period services can be compared and evaluated.

John J. Regazzi is director of documentation services, The Foundation Center, New York, and a doctoral student at Rutgers University. Rodney M. Hersberger is assistant to the library director, Northern Illinois University, DeKalb.

DEFINITIONS

As in many service organizations, there exist processes that produce waiting lines or queues. In the operation of an information desk, a queue is defined as occurring when a patron or employee must wait because the desk, operating at capacity, is temporarily unable to provide service.

A second term used in this study is "model." Although there are various types of models, for the purpose of this study the term model refers to the symbolic representation in tabular form of the reference service. Although four models were developed in the study, only one is shown here as an example of the management technique known as a "simulation." The data necessary to develop a simulation model were obtained by sampling patron arrival time and service time during peak periods for several weeks.

PURPOSE

This study, conducted in the library at Northern Illinois University, DeKalb, was initially undertaken because of the authors' observations of long waiting times for reference service during peak periods and because of staff observations relative to increased service demands during these same periods. At the time of the study, there were approximately 18,000 students (FTE) at the university, and reference service in the university library was provided through a general reference desk on the library's main floor.

Although the reference department had kept certain statistics, these data were not helpful in dealing with waiting lines. The study, then, had two purposes: (1) to evaluate the extent to which queues develop at an information desk and (2) through the use of simulation models and cost analyses to develop alternative staffing patterns which might increase the effectiveness of the service.

METHODOLOGY

For approximately one year prior to this study, the reference department kept statistics on an hourly basis for the following categories: (1) the number of persons using the reference room, (2) the number of directional questions answered by the department, and (3) the number of "reference" or research questions answered by the department. In order to develop a queuing model, the arrival time intervals of users at the reference desk and the service times of staff for the various types of questions answered were needed as well.

The library was utilizing a single-channel/single-stage reference station, and thus the reference desk was always staffed by one professional librarian, but very rarely by more than one. It was decided to confine the study to the library's most heavily used periods only. Obviously, the queues would be most severe during these times with any acceptable alternatives applicable to other times of the day if so desired. Through a review of the statistics already gathered by the reference department, the peak periods were established. The arrival times of patrons and the service times of staff during these periods were randomly sampled over a period of six months.

Because of the demand on the reference department during peak periods, many librarians felt that service was necessarily "shortened," i.e., less time was spent with each patron because of the heavy number of requests during these periods. It was also decided to obtain service times for nonpeak periods. These observations were also randomly sampled over a period of six months. As one might expect, the increased service time does not occur in the directional questions, most of which can be answered in .5–1.0 minutes, but rather in the longer research-oriented questions.

PRESENT MODEL

A representation of the present operating reference service was then developed (see table 1). This model used the arrival times and service times for peak periods. Since neither arrival times nor service times are constant or uniform, a random number sequence was used in selecting each. Thus RN represents the random number sequence for the arrival times, and RN1 the sequence for service times. In each case the two-digit number is matched with a ranking sequence for arrival intervals and service times which were obtained in the sampling period. For example, for the first arrival the random number selected was 52, which corresponds to an arrival interval value of 1.5 minutes. Thus our first patron arrived at 1 minute and 30 seconds after the start of the simulation. Similarly, a random number was selected for this patron's service time; the RN1 of 26 corresponds to a service time of .5 minutes; therefore, the first patron arriving :01:30 after the hour experienced no delay, and the service was completed in 30 seconds at :02:00. The librarian, however, in this instance experienced an idle time of 1.5 minutes in waiting for the patron to arrive.

The model was thus developed for 30 minutes of arrivals, with some large queues occurring in the process. It should be noted that some twenty-five patrons in approximately a half-hour period would experience a total delay time of 146.5 minutes during peak periods, while the librarian experiences a total of 3 minutes idle time. The average delay time per patron is 5.86 minutes, with individual delays of up to 12.25 minutes. During the sampling, it was our observation that queues of this magnitude did occur occasionally; however, in certain instances individuals would leave the queue. One of the most serious weaknesses of this one-channel system is for the individual who has a relatively simple direc-

TABLE 1

SIMULATION OF PRESENT SERVICE DURING PEAK HOURS

Arrival	RN	Arrival Interval	CUM Arrival Time	Delay Time	RN1	Service Time	Start	Finish	Idle Time
1	52	1.5	01:30	—	26	0.5	01:03	02:00	1.5
2	75	2.0	03:30	—	63	1.0	03:30	04:30	1.5
3	49	1.0	04:30	—	51	1.0	04:30	05:30	—
4	05	0.0	04:30	1.0	99	9.0	05:30	14:30	—
5	21	0.25	04:45	9.25	64	1.0	14:30	15:30	—
6	18	0.25	05:00	10.5	75	2.0	15:30	17:30	—
7	55	1.5	06:30	11.0	35	0.5	17:30	18:00	—
8	13	0.25	06:45	12.25	07	0.5	18:00	18:30	—
9	51	1.5	08:15	10.25	42	1.0	18:30	19:30	—
10	99	5.5	13:45	5.75	10	0.5	19:30	20:00	—
11	02	0.0	13:45	6.25	04	0.5	20:00	20:30	—
12	43	1.0	14:45	5.75	35	0.5	20:30	21:30	—
13	42	1.0	15:45	5.25	13	0.5	21:30	22:00	—
14	47	1.0	16:45	3.25	63	1.0	22:00	23:00	—
15	77	2.5	19:15	3.75	87	3.0	23:00	26:00	—
16	03	0.0	19:15	6.75	81	2.5	26:00	28:30	—
17	13	0.25	19:30	7.0	31	0.5	28:30	29:00	—
18	63	1.5	21:00	8.0	05	0.5	29:00	29:30	—
19	44	1.0	22:00	7.5	48	1.0	29:30	30:30	—
20	60	1.5	23:30	7.5	49	1.0	30:30	31:30	—
21	05	0.0	23:30	8.0	04	0.5	31:30	32:00	—
22	83	3.0	26:30	5.5	28	0.5	32:00	32:30	—
23	38	1.0	27:30	5.0	48	1.0	32:30	33:30	—
24	67	2.0	29:30	4.0	22	0.5	33:30	34:00	—
25	39	1.0	30:30	3.5	67	1.5	34:00	35:30	—

tional question—which might take only .5 to 1.0 minutes to answer—but who must wait as long as 5 to 12.25 minutes.

Using the same single-channel/single-stage service model and the same arrival patterns, a simulation was done substituting nonpeak period service times for the peak period times. As expected, the queues were substantially longer for the same time period, and the same number of patrons would experience a projected total delay time of 202.5 minutes, with 8.1 minutes of delay time per patron. Thus, according to this simulation the present reference service represents a "shortened" service by 27.6 percent during peak periods compared to non–peak period services.

These delay times represent the worst possible cases, where no patrons leave the queue. Another simulation would include an estimate of the number of patrons who temporarily drop out of the queue and return later; however, such data were not collected and are not available here. Regardless of whether the patron remained in the queue or not, the reference desk was operating at capacity and was, therefore, temporarily unable to provide service to that patron.

ALTERNATIVE MODELS

Given the amount of reduction in the peak-period reference service that had already occurred, it was obvious that procedural changes attempting to further shorten the service seemed to be futile for two reasons. First, the demand for reference service during these periods is so great that even if the reference process could be further shortened, it would not substantially reduce the queue. Second, any further reduction of service might seriously compromise the quality of the service during this time. Therefore, some expansion of staff was necessary, and various alternatives were analyzed.

The alternative staffing patterns attempt to utilize both professional and nonprofessional personnel. There is substantial evidence for using nonprofessional personnel to staff a reference desk. Dawson has remarked "in reference work there is much that can be done by nonprofessionals under supervision. They can answer many directional and informational questions, freeing the professional for the more difficult and extensive 'research' question."[6] Rogers and Weber agree that nonprofessionals can handle many of the directional and interpretive questions asked at reference desks in college libraries.[7] Bloomberg holds a similar view that allowing a nonprofessional to be the patron's first contact will free a librarian from the time-consuming task of answering directional and ready reference questions.[8]

In their survey of colleges and smaller universities with enrollments up to 6,000 students, Boyer and Theimer found that in "69 percent of the reporting libraries, nonprofessionals are used at the reference desk."[9] Of the three alternative models discussed below, two models add nonprofessional staff only, and the third uses a full-time professional librarian.

One proposed alternative, a two-stage, single-service facility, would consist of a student and a librarian working together during busy periods. The student would be trained to conduct the initial reference interview and determine if the question would need to be answered by the librarian. If so, the patron would be referred to the librarian. In order to simulate this procedure, an estimation was needed of what portion of the service could be handled by the student.

Again, by using the statistics compiled by the reference department prior to this study, it was determined that most directional questions required .5 to 1.0 minutes. It was estimated that the student would be able to answer 80 percent of the .5-minute questions and 50 percent of the 1.0-minute questions. These estimations were arrived at through discussions with the reference staff and an analysis of the various subcategories of directional questions.

Under this multistage, single-channel approach, the user experiences a substantially reduced amount of delay time—in fact, only a total of 20.75 minutes or an average of .83 minutes per patron. Moreover, only one patron experienced a delay of over 3 minutes. The librarian has an increased idle time of 7.75 minutes, and the student's idle time is 21.75 minutes. Although a more complete cost analysis will be discussed in the next section, it may be noted that since student workers are hired at minimum wages and on a per diem basis, this model will only increase the cost of reference service about $.05 per patron during peak periods.

Next, a simulation was developed using the same multistage service channel for peak periods but substituting nonpeak service times. The patron under these conditions experienced a total of 49.25 minutes delay time or an average of 1.97 minutes per patron. Two patrons, however, would experience delays of over 10 minutes, but the remainder were substantially lower. Idle time for the librarian is somewhat less than under the present service conditions (i.e., 3.75 minutes), and the student idle time of 20.25 minutes is approximately the same for both models.

Three other alternative simulation models were considered. One model would place a professional librarian at stage one of a two-stage service channel with a student assistant at stage two. This model has the disadvantage of requiring the librarian to field all questions and then referring the patron to a student who could answer some of the directional/locational questions. This idea was dismissed as being even more impractical than the current system.

A second model would place two professional librarians, in parallel, at a service point to simulate two single-stage service channels. This model would indeed afford users better service: total delay time for twenty-five arrivals was only 1.5 minutes. Librarian idle time, however, increased to 42.5 minutes. The increased effectiveness, however, must be examined in relationship to the corresponding increase in costs. This proposal will be discussed further in the next section.

The final model, a further variation of the previous models, involves the simulation of two single-stage channels, i.e., a librarian and a student in parallel rather than in tandem. Through observation and further discussion with reference department personnel, it was determined that 60 percent of the users arriving at this information center would first approach the librarian. As in earlier simulations, it was estimated that the student would be able to answer 80 percent of the .5-minute questions and 50 percent of the 1.0-minute questions. It was assumed, of course, that if one of the two staff members were engaged, the users would approach the other. If the student was unable to answer a question, the patron would then be referred to the professional.

This simulation produced a total delay time of 71.25 minutes, or an average of 2.85 minutes per patron. Eleven patrons, however, encountered no delay at all. The delays, then, were caused by several longer reference questions coming in succession. Idle time in this model was 27 minutes, with the student assistant experiencing 22.5 minutes of that idle time and the librarian 4.5 minutes. The librarian was obviously busy during the simulated period, but that staff member's efforts were primarily directed to longer reference questions, rather than directional questions.

In summing up these models, they range from the present service on one end to the two-librarian model on the other end. The first model is obviously the least costly, yet incapable of any expansion, while the last model, which is very efficient for the patrons' time, is extremely costly to the library. The next section attempts to compare these models using a common denomination of cost effectiveness.

Cost Analysis

As a basis for quantitative analysis, a common means of comparing the four models was developed. Obviously none of the individual factors—costs (c), idle time (i), or delay time (d)—would suffice alone, since the selected measure must incorporate each of the factors. A simple cost-effectiveness formula was developed with the effectiveness measure (r) derived as follows:

$$r = \frac{\dfrac{1}{d+i}}{c}$$

The formula has two conditions that may not be obvious but that must be mentioned here. First, the formula assumes that delay time for patrons and idle time for staff are of equal weight to the library. In other words, it is as important not have patrons waiting as it is not to have staff idle. Second, the formula cannot derive r in all circumstances; r is indeterminable when both $i = 0$ and $d = 0$. It may be argued, however, that this is the ideal situation, since every patron is completely serviced without delay and no staff member is ever idle.

A summary table was compiled, which includes the total idle time, total delay time, and total costs for each model (see table 2). The costs and idle times were divided, when appropriate, between professionals and students. In order to make the formula sensitive to the different costs and idle times for librarians and students, it needed to be revised for the student/professional model:

$$r = \frac{\dfrac{1}{\left(\dfrac{d}{2}\right) + i_1}}{c_1} + \frac{\dfrac{1}{\left(\dfrac{d}{2}\right) + i_2}}{c_2}$$

where i_1 represents the idle time for professionals;
i_2 represents the idle time for students;
c_1 represents professional costs;
c_2 represents student costs.

Finally, since it is immaterial for measuring the cost effectiveness of a system whether the patron is awaiting service from a student or a professional, the total delay time is arbitrarily divided across both of the effectiveness measures. Substituting the values in table 2, the following effectiveness measures are derived:

Model	*Effectiveness Rating*
1. Present	$r = .0007$
2. Student/Professional (in tandem)	$r = .0159$
3. Professional/Professional (in parallel)	$r = .0082$
4. Student/Professional (in parallel)	$r = .0090$

Although the simulated model of the present reference service is the least costly to the library, it is also by far the least effective, as measured by the cost-effectiveness formula above, of any of the models presented. Because of the model's low effectiveness rating, some change would seem appropriate. Because of its high cost, the professional/professional model could not be seriously considered vis-à-vis either of the

TABLE 2

SUMMARY TABLE FOR DELAY TIMES, IDLE TIMES AND COSTS.

Model	Total Delay Time	Student Idle Time	Professional Idle Time	Student Costs	Professional Costs
				(Dollars per hour)*	(Dollars per hour)*
1. Present Model	146.5	—	3	—	$10
2. Student/Professional Model	20.75	21.75	7.75	$3	$10
3. Professional/Professional Model	1.5	—	42.5	—	$20
4. Student/Professional (in parallel)	71.25	22.5	4.5	$3	$10

*The costs represent approximate average hourly costs for professional and student employees at the time of the study.

student/professional models.

Of these two remaining models, the student/professional model working in parallel causes the patron much more simulated waiting time. Thus the single-channel, two-stage service provides the highest cost-effectiveness rating. Since a student could be employed part time during the peak periods, this model would significantly decrease waiting lines, while only modestly increasing reference service costs.

IMPLEMENTATION

The library that was studied here moved to a new building between the time the data were gathered and first analyzed and the date of publication. Reference services changed from a highly centralized operation to a very decentralized arrangement. In fact, the library changed to a subject division arrangement, including reference services. There are now ten reference points.

Because of the move, direct implementation of any of the models was not possible. The library, however, does utilize a variation of the student/professional model and another approach as well. In theory, one professional librarian is always on duty on each floor. A graduate assistant is stationed in each reference area in the evenings when the librarian is not on duty. These graduate assistants answer those questions that they can and refer the others to the librarian on duty on the floor. During the day a library clerk is on duty on each floor to help direct and refer patrons and to answer the telephone and make the necessary referrals. The first contact with a patron is made by a nonprofessional employee, and the principles of a multistage, single-channel reference service have been employed in this reference department.

REFERENCES

1. W. M. Shaw, Jr., "Computer Simulation of the Circulation System of a Library," *Journal of the American Society for Information Science* 25:271–79 (1975).
2. Irving E. Stephens, "Computer Simulation of Library Operations: An Evaluation of an Administrative Tool," *Special Libraries* 61:280–83 (July/Aug. 1970).
3. Phyllis A. Richmond, "Classification from PRECIS: Some Possibilities," *Journal of the American Society for Information Science* 27:240–47 (1976).
4. Manual D. Lopez, "Academic Reference Service: Measurement, Costs and Value," *RQ* 12:234–42 (Spring 1973).
5. "Symposium on Measurement of Reference," *RQ*, 14:8 (Fall 1974).
6. John M. Dawson, "Not Too Academic," *College & Research Libraries* 27:39 (Jan. 1966).
7. Rutherford Rogers and David C. Weber, *University Library Administration* (New York: Wilson, 1971), p.210.
8. Marty Bloomberg, *Introduction to Public Services for Library Technicians* (Littleton, Colo.: Libraries Unlimited, 1977), p.86.
9. Laura M. Boyer and William C. Theimer, Jr., "The Use and Training of Nonprofessional Personnel at Reference Desks in Selected College and University Libraries," *College & Research Libraries* 36:195 (May 1975).