Revista Argentina de Clínica Psicológica
2026, Vol. XXXV, N°1, 67-86
DOI: 10.69980/03276716.2026.07
Received: 12 June 2026,
Acceptance: 03 August 2026

Modelling and Optimizing Psychological Service Systems Using Retrial Queue Models with Setup Times and Unreliable Servers

N. Meenatchi[1*], S. Pavai Madheswari[2], P. Sukania[3], P. Suganthi[4]

Abstract

In this paper, a single-server M/M/1 retrial queueing system is studied, which incor-porates therapist idle-state management, pre-session setup delays, and stochastic service disruptions with Bernoulli recovery, motivated from the operational dynamics of psychological service system like counseling centres, therapy clinics and psychiatric outpatient units. When there isn't a waiting room, clients who are not served enter a virtual orbit and will" retry" after an exponentially distributed time period, consistent with the tendency to help seek behavior seen in mental health settings. To ensure it operates with minimal energy waste, the server is switched off when not in use, and activates after a period of set-up in preparation for service, similar to the pre-session preparation phase of a therapist. A server breakdown model is developed to model service disruptions that occur when the practitioner's resources are exhausted, compassion fatigue occurs or the service is interrupted due to administrative issues, and a Bernoulli repair mechanism is added to distinguish between perfect and imperfect recovery. The steady-state probability distributions are derived by developing a continuous-time Markov chain framework to be used with probability generating functions. Closed form expressions for key performance metrics, such as the average orbit size and probabilities of the server state are derived. The findings demonstrate that higher repair effectiveness and reduced setup delays contribute significantly to system performance and lower overall operational cost, with direct implications for the design of responsive and sustainable mental health service systems.

Key words: *Retrial Queueing System, Psychological Service Systems, Setup Time, Unreliable Server, Bernoulli Repair Mechanism, Therapist Burnout, Cost Optimization.*

INTRODUCTION

The need for energy efficiency and reliability has become a key design concern in modern service systems, especially in environments where resource limitations and service downtimes lead to considerable economic loss and degradation of service performance. Systems such as cloud servers, automated production lines, and automated inventory replenishment systems often remain idle, require non-negligible setup times to restart, and may fail unpredictably, all of which directly affect congestion levels and operational costs. Classical queueing models generally assume continuously available service facilities, which is increasingly unrealistic for energy-conscious and failure-prone systems. Retrial queueing models provide a realistic frame-work for capturing such operational conditions. In these systems, customers who do not receive immediate service leave the service area and retry after a random delay, a feature that is highly applicable in scenarios where no physical waiting space exists. Retrial queues, coupled with server set-up delays, breakdowns and repairs, provide a valuable analytical tool to investigate service systems that operate in an active and energy-saving manner. It is therefore essential to understand how these set up times relate to repair quality and energy saving strategies to enhance the performance and sustainability of the system.

REVIEW OF LITERATURE

Dimitriou [10] developed a two-class retrial queue with a class-dependent general service time distribution and constant retrial policy, and considered the influence of

[1*]Department of Mathematics, R.M.K. Engineering College, Tiruvallur, Tamil Nadu, India. E-mail: prithiv.meena@gmail.com
[2] Department of Mathematics, R.M.K. Engineering College, Tiruvallur, Tamil Nadu, India. E-mail: pavai.arunachalam@gmail.com
[3]Department of Mathematics, R.M.K. Engineering College, Tiruvallur, Tamil Nadu, India. E-mail: psukanianandakumar@gmail.com
[4]Department of Mathematics, R.M.K. Engineering College, Tiruvallur, Tamil Nadu, India. E-mail: suganthirajkannu@gmail.com

customer heterogeneity on the performance of the queue. Later, Abdollahi and Salehi Rad [1] introduced batchness, feedback and admission control in this framework. Sundarapandiyan and Nandhini [28] also ex-tended the model by taking into consideration non-Markovian service times, customer reneging, delayed repair, and server failures and validated the model through detailed sensitivity analysis. To better reflect the actual service interruptions, Yang and Wu [37] added working vacation policies and starting failures.

More recent studies have shifted attention toward customer decision-making and optimization in unreliable systems. Yang et al [36] studied equilibrium balking in unobservable queues with vacation policies, while Kanavetas and Logothetis [17] investigated strategic customer behavior with abandonment. Xu et al [35] incorporated optimal pricing strategies into an unreliable M/M/1 retrial queue with delayed repair and breakdown deterioration. In addition, optimization and steady-state characteristics of queueing systems with vacation policies and breakdown conditions were investigated by He et al. [12], Jayamani et al [14] and Das and Pradhan [9]. Jeganathan et al. [15] emphasized integrated and realistic service system design by extending retrial modeling to interconnected queueing inventory systems with marked Markovian arrivals.

Retrial queueing models provide a realistic framework for capturing such operational conditions. Customers who are not immediately served leave the service area and retry after some delay, which is highly applicable in environments where there is no physical waiting space. Such retrial mechanisms are also relevant in discrete-time and deterministic queueing environments [27,31]. Combined with server setup delays, breakdowns, and repairs, retrial queues offer a powerful mechanism for analyzing service systems that dynamically switch between active and energy-saving states. It is thus imperative to have knowledge of the relationship between setup times, repair quality and energy-saving measures in order to improve performance and sustainability.

The study of the effects of setup time in retrial queueing systems is extensive. Phung-Duc [25] demonstrated that there is a significant increase in blocking and congestion in M/M/1/1 retrial queues if the setup delays are not small. Building on this, Chang and Wang [6] investigated unreliable retrial queues with setup time and showed that even further degradation of system performance can occur as a result of server breakdowns. Previous research such as Krishna Kumar et al. [18], Atencia and Moreno [3] and Wang and Zhang [31] highlighted the need to include start-up failures, negative customers, and disaster-type failures to truly model the service behavior.

More work was done on extending retrial queue models to incorporate repair mechanism, vacation policies, and customer feedback. Choudhury and Ke [7] studied the problem of unreliable retrial queues with delayed repair and Bernoulli vacation structure, and they demonstrated that repair duration and vacation structure have significant impact on the stability of the system. Gao et al. [11] and Jain and Kaur [13] investigated the case of unreliable retrial queues with delayed repair and Bernoulli vacation structure. Pavai Madheswari et al. [22,23], Narasimhan and Rajadurai [24] further examined flexible Bernoulli and hybrid vacation policies, proving their efficient combination of maintenance and service availability.

Much has been written about customers' behaviors in retrial queues. Chang et al. [5] analysed models with balking, impatience, reneging and strategic decision-making; Wang et al. [32], Arizono and Takemoto [2] studied models including balking and impatience; Xu and Xu [33] focused on reneging and strategic decision-making; Zhang [38] studied models with balking and impatience; and Liu et al. [21] investigated models including balking, impatience, reneging and strategic decision-making. It also identified that perceived congestion, service reliability and server vacation or breakdown states are significant factors affecting customers' decisions from these studies, which highlights the importance of considering the dynamic behavior of both customers and servers in the analysis.

Optimization and economic aspects have become more and more a focus of recent research. Kumar and Jain [19] applied cost functions to retrial queue models, Vijayashree and Pavithra[30] combined pricing strategies into retrial queue models, Xu et al. [34, 35] introduced cost functions, pricing strategies, and preventive maintenance policies into retrial queue models, Zhou et al. [39] applied preventive maintenance policies, and Singh et al. [27] combined cost functions with retrial queue models. The contributions reflect the increasing importance of cost-aware' modelling in the design of viable and sustainable service systems.

Queueing models have gained growing research interest in healthcare and psychological service systems. Using waiting line theory, Tran et al. [29] explored how the different components of community mental health clinics, such as waitlists, wait times, and treatment engagement, are connected and emphasized the operational burden of managing a mismatch between the demand and supply of mental health services. Complementing this, Compton et al. [8] showed that appropriate system changes in psychiatric service delivery can significantly shorten waiting periods and no-show rates for patients, lending credence to the usefulness of analytical models for capacity and scheduling optimisation. Bakker et al. [4] conducted a systematic review of prevalence research on burnout prevention measures for counselors or therapists, highlighting recovery quality and workload management as key factors in ensuring service availability.

In general, optimization of the healthcare system through queueing theory modeling has been examined by many previous studies: For example, Li et al. [20] studied patient queues in an Emergency Department (ED) using Markov Decision Process (MDP) and particle swarm optimization for the purpose of optimizing the use of medical resources in emergency situations with unreliable services.

After extensive research on retrial queues consisting of set up times, server failures, vacation policies and customer behavior, few studies have simultaneously studied energy-sensitive server shutdown policies and probabilistic repair quality within a single framework. The relationship of set-up failures, energy-saving idle states and Bernoulli repair mechanisms that allow for capturing both repair failure and successful repair is not well studied in a single-server retrial system.

In order to overcome this lack, the present study introduces a queuing model with an *M/M/*1 queue, switching off the server during idle periods, with a set-up time for the server to resume its operation, and with breakdowns occurring during both set-up and service phases. A Bernoulli repair mechanism is included to realistically model both imperfect and perfect repair in a single model. The steady state performance measures are computed by means of a continuous time Markov chain and probability generating functions. In addition, cost optimization analysis and numerical experiments are carried out to investigate the impact of setup rates, repair effectiveness and service rates on congestion and operational cost.

The detailed review of literature discussed in Section 2 is followed by the subsequent parts: The proposed model is discussed in Section 3 for its applicability to real world psychological service systems such as counselling centres, therapy clinics, psychiatric outpatient units etc. The mathematical model of *M/M/*1 retrial queue with setup time, server breakdowns and Bernoulli repair is given in Section 4 along with the derivation of the steady-state distributions. Section 5 introduces numerical examples to assess the effect of important system parameters and the special cases of the model are also discussed. The cost optimization problem is formulated and analyzed in Section 6 and then conducted a sensitivity analysis in Section 7. Finally, Section 8 ends the paper and suggests directions for future research.

APPLICATIONS TO REAL-WORLD PSYCHOLOGICAL SERVICE SYSTEMS

The one-server retrial queue model proposed in this study provides a solid analytical model for the study of the operational dynamics of psychological and psychiatric health service systems, e.g. counseling centres, therapy clinics, crisis helplines, psychiatric outpatient units, etc. The intermittent demand, limited-service capacity, therapist unavailability, and pre-session preparation in these settings reflect the natural fit with the structural components of the proposed model. The framework explicitly models server idleness, set up time, failure periods, and probabilistic recovery, allowing for practical insights into single-practitioner or single-desk mental health service settings.

Client Help-seeking behavior and the retrial mechanism: In psychological service contexts, clients may seek consultation immediately and when the therapist is not available, they do not stop needing psychological help altogether. Rather than that, they try to reach out again after some time, whether through a phone call, a visit to the clinic or an online appointment request. This behavior is just modeled by the retrial mechanism: clients not served are added to a virtual orbit, and have exponentially distributed waiting times before they reattempt service. The retrial rate, ϑ, indicates how often clients re-engage, and is shaped by psychological need urgency, the client's knowledge of the availability of the service, and personal persistence. This applies especially in campus counseling centers and outpatient psychiatric units, where walk-in appointments are a frequent occurrence and individuals may try to access services more than once before receiving treatment.

In the proposed model, the server is turned off when it is idle to save resources, which reflects in a psychological service system when there are no clients present, the therapist would go into a non-consulting state. The therapist can have low engagement operational states during these periods such as documentation, supervision, continuing education, and rest. The act of deactivating the server followed by reactivation, closely parallels the transition of a mental health professional from the passive state to readiness for service.

For a therapist to provide service to a new client or a returning client, a non-trivial setup time is required before the session. This could involve reading through case notes, setting up the consultation room, setting up assessment tools or doing required paperwork from previous consultations. This (pre-session) preparation time is captured by the exponentially distributed set-up time with rate α in the proposed model. A higher setup rate α indicates a faster response from the therapist to become service ready, and thus, efficient case management and streamlined administrative procedures. On the other hand, a lower α indicates a delay in starting therapeutic services, due to the actions of the bureaucracy or the organization, this delay can lead to a longer waiting time for the clients and accumulation of orbits.

Therapists and counselors are particularly prone to disruptions in their function that temporarily limit their ability to provide service. These interruptions can come from burnout or compassion fatigue, a sudden health event, an administrative event like emergency supervision or peer review, or an unplanned event such as an emergency in a member's personal life. In the model proposed, such episodes are modeled as a server breakdown that happens with probability 1-p during the setup or service phase. The reliability parameter p thus represents the probability that a therapist will not be incapacitated or otherwise unavailable during a session, or a preparation period. The higher p value, the more resilient the practitioner, or the more supported the clinical environment; the lower p value, the more vulnerable to disruption of services the practitioner and/or clinical environment.

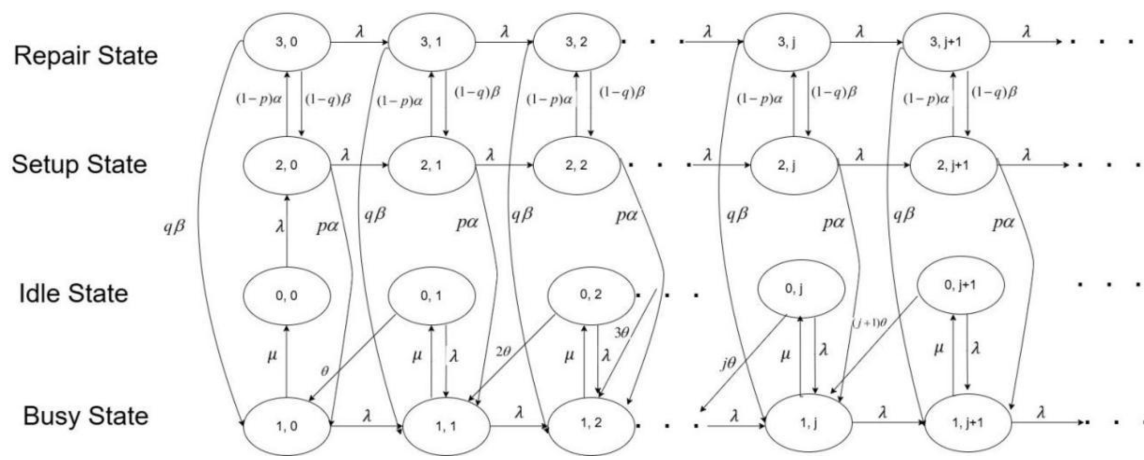
If a therapist's service is interrupted, there are two possibilities: complete or partial recovery, depending on the nature and degree of the interruption, which is the case of Bernoulli Repair as Probabilistic Therapist Recovery. The result can be a full recovery and the return of the therapist to full function, though an acute burn out or psychological distress may result in only partial recovery with the possibility of repeated unavailability. This distinction is reflected in the Bernoulli repair mechanism in the proposed model with parameter q. But if q=1 then the repair is perfect, and the therapist is restored to service after disruption. If q=0 then the repair is imperfect, and the therapist might have some failures, so this would result in longer service outages and longer waiting times in orbit for the client. This probabilistic approach allows for an analysis of the impact of mental health maintenance and support programs on the availability of mental health services and system congestion.

Managerial Implications for Mental Health Service Design: The proposed model is rigorous, and thus can be used for evaluating and optimizing the design of mental health service systems. Key performance measures, such as mean number of clients in the orbit, and probabilities of being idle, in set up, in session, or under recovery, provide measurable goals for services. The parameter α takes the form of the setup rate and can be directly measured by using electronic health records, administrative support, or by ensuring that sessions are prepared beforehand. Likewise, in the same way, investing in programs which improve therapist well-being, clinical supervision, and structured recovery protocols will result in the improvement of the reliability parameter p and the repair effectiveness parameter q. The size of the retrial orbit is a proxy for unmet psychological need; reducing the size of this orbit by joint optimization of set-up efficiency, service capacity and therapist support is a central management goal. These findings have implications for public mental health organizations, university counseling centers, services delivered through telecommunication services and private practice environments where the individual provider is the main service provider.

MODEL DESCRIPTION

In this paper, the arrivals are considered to be a Poisson process of rate λ and the service time to be exponentially distributed with mean $1/\mu$. The study is based on $M/M/1/1$ retrial queueing model with server failures and repair process along with setup time. Upon service completion, the server is deactivated if no customers remain in the system. However, if customers are present, the server remains idle, awaiting either an external arrival or a customer from the retrial orbit. When the server is deactivated, it requires an exponentially distributed setup time with mean $1/\alpha$ to restart. Upon reactivation, the server gives priority to serving the customer who triggered the restart. In the case of an unreliable server, failures occur with probability 1-p. If the server is busy, under setup, or undergoing repair, arriving customers are directed to the retrial orbit. Customers in the retrial orbit attempt to obtain service again after an exponentially distributed retrial time with mean $1/\vartheta$. The total retrial rate is $n\vartheta$, which is proportional to the number of customers n present in the orbit. The server may fail during the setup phase, preventing successful activation. In such cases, the system requires a repair time, governed by a repair rate β with probability q. A value of q=0 indicates the possibility that the server fails again immediately after repair completion, whereas q=1 guarantees immediate resumption of service once the repair process is completed. Figure 1 represents the steady state transition diagram of the proposed model.

Figure 1. **Transition diagram**



In this section, we derive analytic solutions using probability generating functions. Let *J(t)* denote the state of the service unit and *H(t)* denote the number of customers in the orbit.

The state of the service unit is defined as

$$J(t)=\begin{cases}0\text{ , the service unit is idle}\\1,\text{ the service unit is busy}\\2,\text{ the service unit is in set up process}\\3\text{ , the service unit is in repair process}\end{cases}$$

It is easy to observe that the stochastic process

$$X(t)=\{(J(t),H(t)),t\geq 0\}$$

forms a continuous-time Markov chain.

Let $\pi(i,j)$ denote the steady-state probability that the system where i=0,1,2,3 represents the server state and j denotes the number of customers in the orbit.

We define the probability generating functions for each server state as

$$G_0(\mathrm{z})=\sum_{j=0}^{\infty}\pi_{0,j}z^j \qquad G_1(\mathrm{z})=\sum_{j=0}^{\infty}\pi_{1,j}z^j$$

$$G_2(\mathrm{z})=\sum_{j=0}^{\infty}\pi_{2,j}z^j \qquad G_3(\mathrm{z})=\sum_{j=0}^{\infty}\pi_{3,j}z^j$$

Based on the transition diagram, the steady-state balance equations for states i=0,1,2,3 are given as follows.

Idle State (i=0)

$$\lambda\pi(0,0)=\mu\pi(1,0) \tag{1}$$

$$j\theta\pi(0,\mathrm{j})+\lambda\pi(0,\mathrm{j})=\mu\pi(1,\mathrm{j}),\quad \mathrm{j}\geq 0 \tag{2}$$

Busy State (i=1)

$$(\lambda+\mu)\pi(1,0)=\theta\pi(0,1)+p\alpha\pi(2,0)+q\beta\pi(3,0) \tag{3}$$

$$(\lambda+\mu)\pi(1,\mathrm{j})=\lambda\pi(1,\mathrm{j}-1)+p\alpha\pi(2,\mathrm{j})+q\beta\pi(3,\mathrm{j})+\lambda\pi(0,\mathrm{j})+(j+1)\theta\pi(0,\mathrm{j}+1),\quad j\geq 1 \tag{4}$$

Setup State (i=2)

$$\alpha(1-p)\pi(2,0)+\lambda\pi(2,0)+\alpha p\pi(2,0)=\lambda\pi(0,0)+(1-q)\beta\pi(3,0), \tag{5}$$

$$\alpha(1-p)\pi(2,\mathrm{j})+\lambda\pi(2,\mathrm{j})+\alpha p\pi(2,\mathrm{j})=\lambda\pi(2,\mathrm{j}-1)+(1-q)\beta\pi(3,\mathrm{j}),\quad \mathrm{j}\geq 1 \tag{6}$$

Repair State (i=3)

$$q\beta\pi(3,0)+(1-q)\beta\pi(3,0)+\lambda\pi(3,0)=(1-p)\alpha\pi(2,0) \tag{7}$$

$$\beta q\pi(3,\mathrm{j})+\beta(1-q)\pi(3,\mathrm{j})+\lambda\pi(3,\mathrm{j})=\alpha(1-p)\pi(2,\mathrm{j})+\lambda\pi(3,\mathrm{j}-1),\quad \mathrm{j}\geq 1 \tag{8}$$

Multiplying the above equations by z^j and summing up, which yields

$$z\theta G_0{}'(z)+\lambda G_0(z)=\mu G_1(\mathrm{z}) \tag{9}$$

$$\left(\lambda+\mu\right)G_1(z)=\alpha pG_2(z)+\lambda zG_1(z)+\beta qG_3(z)+\theta G_0'(z)+\lambda G_0(z)-\lambda\pi(0,0) \tag{10}$$

$$G_2(z)\left(\lambda+\alpha\right)=G_2(z)\lambda+\beta G_3(z)(1-q)+\lambda\pi(0,0) \tag{11}$$

$$\left(\lambda+\beta\right)G_3(z)=\lambda zG_3(z)+\alpha G_2(z)(1-\mathrm{p}) \tag{12}$$

By solving equations (9), (10), (11), and (12), we get

$$G_1(z)=\frac{1}{\mu}\left[\theta\, z\, G_0'(z)+\lambda G_0(z)\right] \tag{13}$$

$$G_2(z)=\frac{\beta+\lambda(1-z)}{[\lambda(1-z)+\beta][\lambda(1-z)+\alpha]-(1-\mathrm{q})\beta(1-p)\alpha}\lambda\pi(0,0) \tag{14}$$

$$G_3(z)=\frac{-\alpha(1-p)}{\beta(1-\mathrm{q})\alpha(1-p)-[\lambda(1-z)+\beta][\lambda(1-z)+\alpha]}\lambda\pi(0,0) \tag{15}$$

$$G_0'(z)=\frac{\lambda\mu\pi(0,0)}{\theta\left[\mu-\lambda\,\mathrm{z}\right]}\left[\frac{\lambda^2(1-z)+\alpha\lambda(1-p)+\beta\lambda}{[\lambda(1-z)+\beta][(1-z)\lambda+\alpha]-\beta(1-\mathrm{q})\alpha(1-p)}\right]+\frac{\lambda^2}{\theta\left[\mu-\lambda\,\mathrm{z}\right]}G_0(z) \tag{16}$$

To solve equation (16), we begin by solving the corresponding homogeneous equation,

$$G_0'(z)=\frac{\lambda^2}{\theta\left[\mu-\lambda\,\mathrm{z}\right]}G_0(z) \tag{17}$$

Further solving and differentiating we get equation

$$G_0(z)=A(z)\frac{1}{\left[\theta(\mu-\lambda\,\mathrm{z})\right]^{\lambda/\theta}} \tag{18}$$

$$G_0'(z)=A(z)\lambda^2(\mu\theta-\lambda\theta\,\mathrm{z})^{-\lambda/\theta-1}+(\mu\theta-\lambda\theta\,\mathrm{z})^{-\lambda/\theta}A'(z) \tag{19}$$

By substituting (18), (19) in (16) and integrating w.r.t *z*, we have

$$A'(z)=\lambda\mu\pi(0,0)\frac{1}{\left[\theta\left(\mu-\lambda\,\mathrm{z}\right)\right]^{(1-\lambda/\theta)}}\left[\frac{\beta\lambda+\alpha(1-p)\lambda+(1-z)\lambda^2}{[\lambda(1-z)+\beta][\lambda(1-z)+\alpha]-\beta(1-\mathrm{q})\alpha(1-p)}\right] \tag{20}$$

$$A(z)=A+\lambda\mu\pi(0,0)\int_0^z\frac{\left[\lambda^2(1-x)+\alpha\lambda(1-p)+\beta\lambda\right]\left[\theta(\mu-\lambda x)\right]^{\lambda/\theta-1}}{[\lambda(1-x)+\beta][\lambda(1-x)+\alpha]-\beta(1-\mathrm{q})\alpha(1-p)}\,dx \tag{21}$$

$$\pi(0,0)=\pi_0(0)=A(0)(\mu\theta)^{-\lambda/\theta}$$

Therefore,

$$A=\pi(0,0)(\mu\theta)^{\lambda/\theta}$$

Hence,

$$A(z)=\pi(0,0)(\mu\theta)^{\lambda/\theta}+\lambda\mu\pi(0,0)\int_0^z\frac{\left[\lambda^2(1-x)+\alpha\lambda(1-p)+\beta\lambda\right]\left[\theta(\mu-\lambda x)\right]^{\lambda/\theta-1}}{[\lambda(1-x)+\beta][\lambda(1-x)+\alpha]-\beta(1-\mathrm{q})\alpha(1-p)}\,dx \tag{22}$$

and

$$G_0(z)=\theta(\mu-\lambda z)^{-\lambda/\theta}\left[\mu\lambda\eta(z)+\mu^{\lambda/\theta}\theta\right]\pi(0,0)$$

where

$$\eta(z)=\int_0^z\frac{1}{(\mu-\lambda x)^{\left(1-\frac{\lambda}{\theta}\right)}}\left[\frac{\left[\lambda\alpha(1-p)+\beta\lambda+(1-x)\lambda^2\right]}{[\lambda(1-x)+\beta][\lambda(1-x)+\alpha]-\beta(1-\mathrm{q})\alpha(1-p)}\right]dx$$

Using the normalization condition,

$$G_0(1)+G_1(1)+G_2(1)+G_3(1)=1$$

We get

$$\pi(0,0)=\frac{\theta(\mu-\lambda)^{1+\lambda/\theta}\left[\alpha(p+q-pq)\beta\right]}{\left[\mu(p+q-pq)\beta\alpha\right]\left(\theta\mu^{\lambda/\theta}+\lambda\mu\eta(1)\right)+\dfrac{\lambda\theta}{(\mu-\lambda)^{-\left(1+\lambda/\theta\right)}}\left[\alpha(1-p)+\beta\right]}$$

where

$$\eta(1)=\int_0^1\frac{1}{(\mu-\lambda x)^{1-\lambda/\theta}}\left[\frac{\left[(1-x)\lambda^2+\beta\lambda+\alpha\lambda(1-p)\right]}{[\lambda(1-x)+\beta][\lambda(1-x)+\alpha]-\alpha(1-p)\beta(1-\mathrm{q})}\right]dx$$

The estimated number of customers in the system is equal to the sum of the average number of customers in the queue and the average number currently being served. Thus,

$$L=G_0'(1)+G_1'(1)+G_2'(1)+G_3'(1)+G_1(1)+G_2(1)+G_3(1)$$

Special Cases

Case 1

If q=0 then our result coincides with Jingwei Chang and Jinting Wang [6] with imperfect repair.

$$A(z)=\pi(0,0)(\mu\theta)^{\lambda/\theta}+\lambda\mu\pi(0,0)\int_0^z\frac{1}{\left[\theta(\mu-\lambda x)\right]^{1-\lambda/\theta}}\left[\frac{\left[(1-x)\lambda^2+\beta\lambda+\alpha\lambda(1-p)\right]}{[\lambda(1-x)+\beta][\lambda(1-x)+\alpha]-(1-p)\beta\alpha}\right]dx$$

After a failure, the server may still fail to activate even after being repaired. This leads to longer waiting times, larger orbit sizes, and degraded system efficiency. The performance metrics, such as the mean queue length, are worse compared to the perfect repair scenario, and the probability of successful service depends on the repair success probability p.

Case 2

If q=1 then our result coincides with Jingwei Chang and Jinting Wang [6] with perfect repair.

$$A(z)=\pi(0,0)(\mu\theta)^{\lambda/\theta}+\lambda\mu\pi(0,0)\int_0^z\frac{1}{\left[\theta(\mu-\lambda x)\right]^{1-\lambda/\theta}}\left[\frac{(1-x)\lambda^2+\beta\lambda+\alpha\lambda(1-p)}{[\lambda(1-x)+\beta][\lambda(1-x)+\alpha]}\right]dx$$

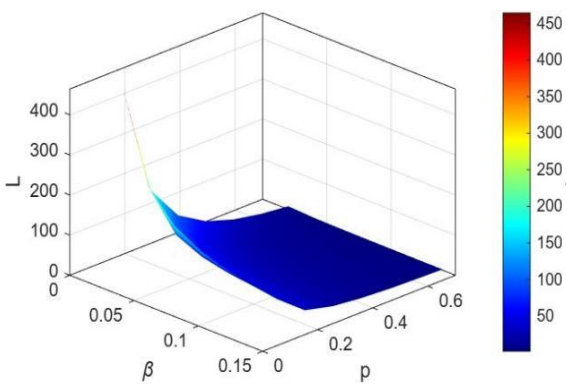
Here the server is completely rebooted once a failure has occurred and immediately it comes back into service to the pending customers. This saves waiting time, mean number of customers per system, and increases system throughput. Complete repair will make it more reliable and allow the running of the services.

Statistical Illustrations

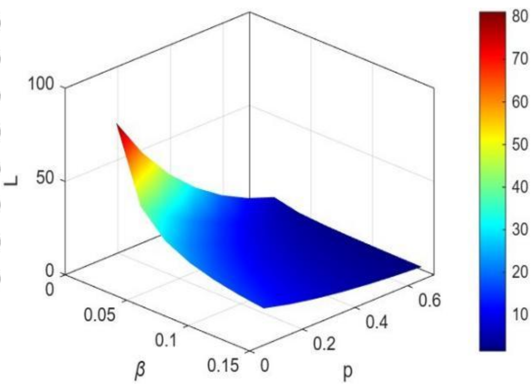
The numerical investigation examines the influence of the repair rate β, the reliability parameter p, and the repair effectiveness factor q on the performance of the proposed

$M/M/1/1$ retrial queue with setup, failures, and repairs, with particular emphasis on the mean queue length L and the idle probability $\pi(0,0)$. The findings given in table 1 and figure 2 demonstrate that at all values of q, the average queue length declines in large proportions as the repair rate β grows, which implies that higher speed of the repair process minimises the downtime of the server and facilitates the prompt completion of customers in the retrial orbit. This somewhat diminishing trend of L will be witnessed when the reliability parameter p is higher as the number of server failures will become such that there will be less service interruption. The effect of the repair effectiveness parameter q is more pronounced: Even if q equals 0, the probability of failure immediately after the repair causes significant congestion; if q equals 0.5, the congestion is significantly reduced; if q equals 1, the maximum congestion occurs to be smallest for all the parameter combinations.

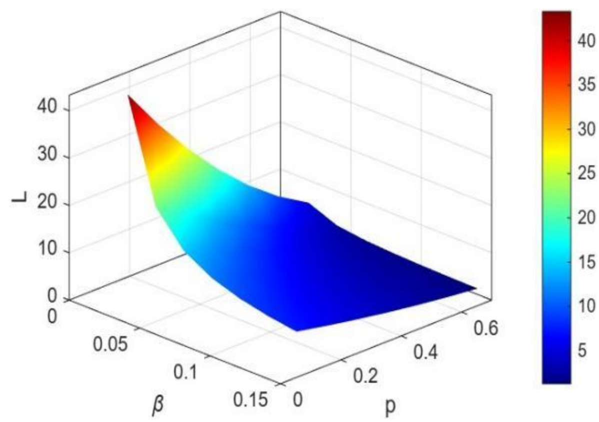
Figure 2. **Queue length (L) versus p and β for q=0, q=0.5, q=1**



(a) $q = 0$



(b) $q = 0.5$



(c) $q = 1$

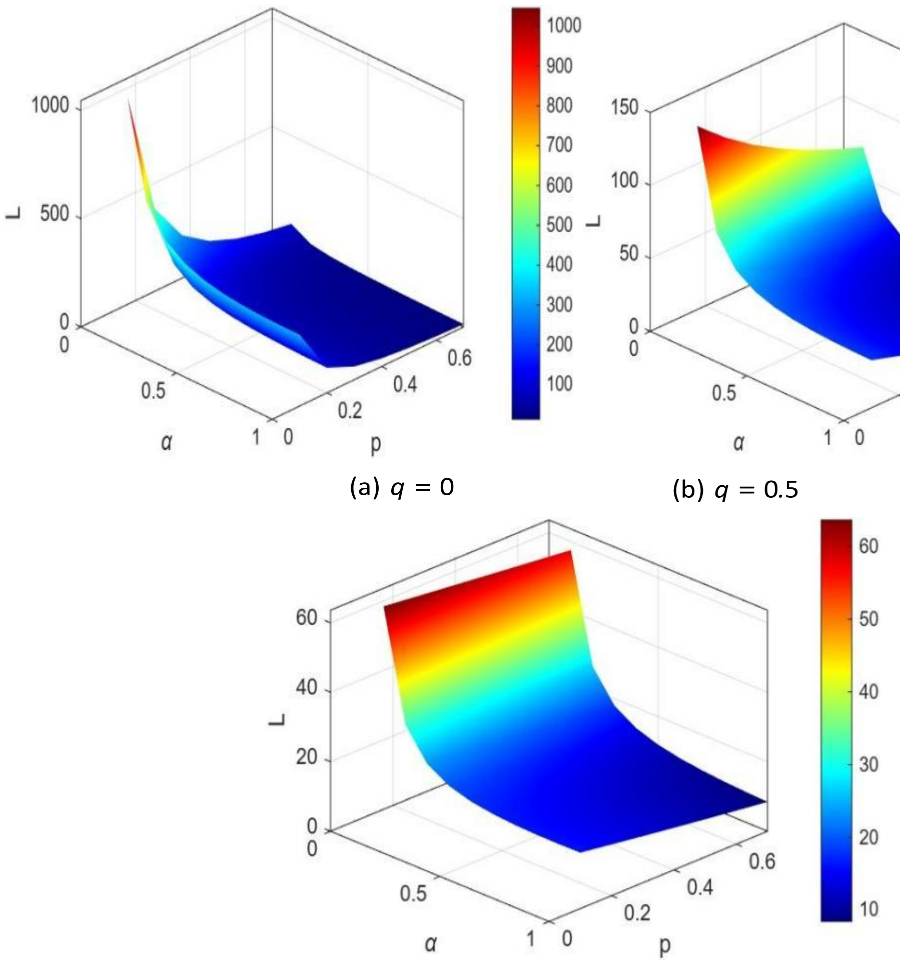
The three-dimensional surface plots reinforce these trends further showing that the steep falls in the queue length caused by higher β, as well as more smooth surfaces by higher q are evidence of improved system stability. Also, the idle probability $\pi(0,0)$ increases with increasing values of β, p and q, that is, the higher the speed of repair the higher the reliability, the higher the effectiveness of repair, the faster the system is cleared and the less the congestion. Overall, the numbers and tables results clearly demonstrate the need for successful and effective repair procedures in the management of congestion and the enhancement of the performance of unreliable retrial queueing systems [26].

Table 1. **Performance measures for different values of p, β when q=0, q=0.5 and q=1**

p	β	q=0		q=0.5		q=1	
		$\pi(0,0)$	L	$\pi(0,0)$	L	$\pi(0,0)$	L
0.2	0.02	0.003195	182.8835	0.009217	59.07213	0.014865	34.5633
0.4	0.02	0.004768	51.63345	0.008152	28.92436	0.011402	19.89439
0.6	0.02	0.004818	15.60326	0.006345	11.55054	0.007838	9.126468
0.2	0.04	0.00641	96.18412	0.017991	30.4641	0.028418	17.59574
0.4	0.04	0.009542	27.32947	0.016045	15.11975	0.022128	10.29701
0.6	0.04	0.009732	8.423011	0.012699	6.181103	0.015555	4.845463
0.2	0.06	0.009659	67.48818	0.026495	21.05709	0.041172	12.06099
0.4	0.06	0.014352	19.31406	0.0238	10.58622	0.032454	7.161518
0.6	0.06	0.014763	6.070027	0.019114	4.425681	0.02325	3.45009
0.2	0.08	0.012949	53.26814	0.034811	16.42114	0.053362	9.352846
0.4	0.08	0.019209	15.36041	0.031474	8.358686	0.042508	5.628574
0.6	0.08	0.019914	4.92129	0.025608	3.57078	0.030962	2.772751
0.2	0.1	0.016281	44.83003	0.042989	13.68316	0.065131	7.76402
0.4	0.1	0.024121	13.02803	0.039097	7.049397	0.052366	4.731998
0.6	0.1	0.025184	4.253558	0.032187	3.075164	0.03871	2.381498
0.2	0.12	0.019657	39.27926	0.051061	11.88938	0.076571	6.729639
0.4	0.12	0.029089	11.50482	0.046692	6.197356	0.062078	4.151459
0.6	0.12	0.030569	3.826129	0.038853	2.758844	0.046505	2.132818
0.2	0.14	0.023077	35.37661	0.059052	10.63259	0.087751	6.009327
0.3	0.14	0.029837	18.3384	0.057134	7.816438	0.079864	4.823888
0.4	0.14	0.034114	10.44329	0.054272	5.605626	0.071681	3.750375
0.6	0.14	0.036064	3.535986	0.045605	2.544846	0.054353	1.965392

The mathematical outcomes obtained in the proposed single-server retrial queue are enough to support its use for psychological service systems with preparation delays, unavailability of practitioners, and constraints on the management of the sessions. The planned scheduled disconnect and reconnect of the server during periods of inactivity and activation during a setup phase are logical representations of an inactive and active state of a therapist, respectively. From the numerical analysis it is clear that the higher the repair rate θ the more significant the decrease in the average orbit size, and this in the case of a psychological service would directly correspond to the waiting time of their clients for mental health care. In the same way, a lower value of the reliability parameter p means less service interruptions, meaning better therapeutic availability due to strong practitioners' support frameworks. The re-attempt behavior of a client following a period of time after an initial attempt, when they find a therapist unavailable, is common in outpatient mental health and counseling settings, and is replicated in the retrial mechanism. The numbers seen for the impact of the repair effectiveness parameter q have direct clinical significance: The lower q=0 leaves the therapist available to fewer clients, and the clients waiting longer, reflecting the consequences of inadequate support or insufficient recovery from therapist burnout, as well as the importance of therapist wellness and supervision programs; The higher $q \to 1$ leads to stable service delivery with little accumulation of orbits, highlighting the value of comprehensive therapist wellness and supervision programs. The setup rate α is the rate of pre-session preparation and the results of the numerical simulations indicate that the higher the rate of pre-session preparation, the less the congestion, which is consistent with the operational advantages of expedient administrative procedures and electronic case management in clinical environments. The overall findings in both the numerical and graphical results are consistent with the proposed model as an analytical tool for assessing the trade-offs between the system preparation efficiency, practitioner reliability and system recovery quality in the design of responsive mental health service systems that are sustainable.

Figure 3. **Queue length (L) versus setup rate α for different values of repair effectiveness**



(a) $q = 0$ (b) $q = 0.5$

(c) $q = 1$

The numerical results shown in the tables and figures (Table 2 and figure 3) support the use of the proposed retrial queueing model for psychological service systems with preparation delays, reliability constraints and practitioner recovery policies. The therapist's setup rate α is important for the buildup of clients in the retrial orbit, as it indicates how quickly the therapist is able to get ready for a new session after a period of inactivity or absence. The figures clearly show that the greater the level of practitioner reliability p and the level of repair effectiveness q is, the lower the level of decrease to the average orbit size L will be for a higher value of α with the faster the pre-session work, the more the service delivery will respond.

The quick setup time in practical psychological service settings translates to more efficient prerequisites in the session, including the ability to access to the client's notes through electronic means, automatic appointment booking, administrative assistance etc., and minimize the congestion of time between client contact and the start of service, thus increasing the number of services that can be undertaken without the need to wait in line. The negative impact of α is particularly noticeable at the extreme end of low values of q=0; slow setup times with imperfect recovery by the practitioners can lead to very large orbit sizes, making systems that have both recovery delay and inadequate recovery extremely vulnerable to this negative effect. The numerical results demonstrate that congestion decreases significantly as q approaches 0.5 and 1 (the optimal values), which validates the effectiveness of the recovery protocols in reducing the negative effects of preparation delays. The higher reliability parameter p the better the system performance the more resilient the practitioner the fewer interruptions in service and the more the therapist can maintain a smooth flow in therapeutic sessions. The retrial mechanism also seems to be an accurate reflection of the fact that clients re-engage in contact with the counselling service after it is no longer available, as is observed in help-seeking in mental health settings. The trends are also verified by the three-dimensional surface plots that show large decreases in L along the α-axis, and then surfaces that are more and more stable as the recovery quality is increased. As a whole, the mathematical analysis has shown that the efficiency of preparation, practitioners' reliability, and the effectiveness of recovery are key factors in determining the responsiveness of the services, the time spent by the clients awaiting them, and the sustainability of single-practitioner psychological service systems.

Table 2: **Performance measures for different values of p, α when q=0, q=0.5 and q=1**

p	α	q=0		q=0.5		q=1	
		$\pi(0,0)$	L	$\pi(0,0)$	L	$\pi(0,0)$	L
0.2	0.1	0.033535	466.6438	0.076177	121.5277	0.105545	62.61517
0.4	0.1	0.056497	195.7932	0.08324	96.43004	0.104195	60.58417
0.6	0.1	0.074083	113.6044	0.089499	78.41143	0.102829	58.57602
0.2	0.2	0.061465	254.3144	0.125733	61.55664	0.16631	31.26644
0.4	0.2	0.096022	98.74571	0.134125	47.18972	0.162667	29.40714
0.6	0.2	0.119849	54.32603	0.141024	37.07695	0.158927	27.58375
0.2	0.3	0.087026	188.0006	0.167301	44.04966	0.216245	22.50126
0.4	0.3	0.129525	69.46688	0.175749	32.94441	0.209857	20.64026
0.6	0.3	0.156816	36.93001	0.182031	25.20589	0.20321	18.82817
0.2	0.4	0.11131	157.1683	0.205142	36.26862	0.261488	18.73197
0.4	0.4	0.159995	56.04819	0.213071	26.59427	0.252044	16.80864
0.6	0.4	0.189503	29.03833	0.218108	19.89108	0.242095	14.94752
0.2	0.5	0.134838	140.2157	0.24093	32.1428	0.304255	16.80392
0.4	0.5	0.188704	48.6664	0.248009	23.18206	0.291522	14.78939
0.6	0.5	0.219718	24.69167	0.25139	16.99436	0.27795	12.85067
0.2	0.6	0.157892	130.078	0.275480	0.275480	0.345572	15.7425
0.4	0.6	0.216308	44.18641	0.281489	21.16004	0.329365	13.62054
0.6	0.6	0.248363	22.02384	0.282911	15.23437	0.311906	11.58822
0.2	0.7	0.180641	123.7765	0.309238	28.33448	0.385985	15.15366
0.4	0.7	0.243182	41.31204	0.314023	19.89841	0.366156	12.91413
0.6	0.7	0.275944	20.27425	0.313246	14.09285	0.344582	10.77825
0.2	0.8	0.203189	119.8426	0.342471	27.49047	0.425809	14.84937
0.4	0.8	0.269557	39.41468	0.345916	19.09535	0.402239	12.48544
0.6	0.8	0.302775	19.07814	0.342743	13.32268	0.376355	10.2393
0.2	0.9	0.225605	117.4707	0.375347	27.02395	0.465243	14.72911
0.4	0.9	0.295584	38.15413	0.377365	18.58911	0.43783	12.23593
0.6	0.9	0.329059	18.23996	0.371629	12.79193	0.407466	9.874768
0.2	1	0.247938	116.1825	0.407977	26.81608	0.504413	14.73449
0.4	1	0.321364	37.33088	0.408499	18.28552	0.473073	12.10847
0.6	1	0.354934	17.64589	0.400057	12.42408	0.438078	9.628746

Optimization of Cost Function

This section aims at minimizing the total cost per unit time on the proposed retrial queueing system using the service rate μ as the variable in the decision-making process. The incentive is based on server farm and data-centre operation where the responsiveness aspect is enhanced by the rise in the rate of service which in turn raises the energy usage and the operation cost. It is therefore important to determine the best service rate that can yield performance and power consumption to economically design a system.

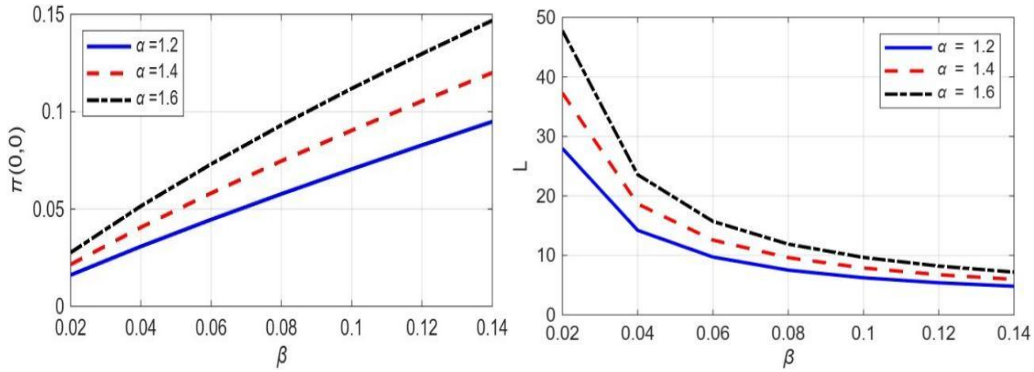
We now define S_c as the setup cost per server time to activate the unit. The breakdown cost per unit of time for a failed server is expressed by F_c. The cost per unit of time for keeping the server running at a service rate μ is indicated by B_c. It is crucial to remember that S_c changes as μ does.

More precisely, $B_C(\mu) = k\mu + C_0$ where both k and C_0 are constant positive parameters. Here, it is noted that S_c and F_c are not equal because repairing a failed server incurs higher costs. As a result, breakdown costs must be treated separately from setup costs. The cost of keeping customers in the retrial queue can be disregarded because their influence on the server is negligible. Similarly, because hibernation uses a lot less power than active operation, hibernation expenses are eliminated. Hence, our cost function is

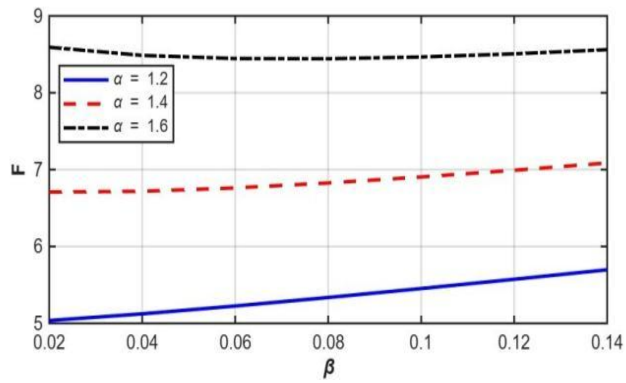
$$F(\mu) = B_C(\mu)G_1(1) + S_C G_2(1) + F_C G_3(1)$$

Our goal is to minimize this cost per unit of time. Since the probabilities of the busy state, retrial, and setup depend on μ, it is a complex function; deriving analytical results is challenging, if not infeasible. Therefore, we conduct statistical computations to optimize.

Figure 4. **Repair rate β on π(0,0) on L and F**



(a)Idle probability$\pi(0, 0)$ (b)Average number of customers L

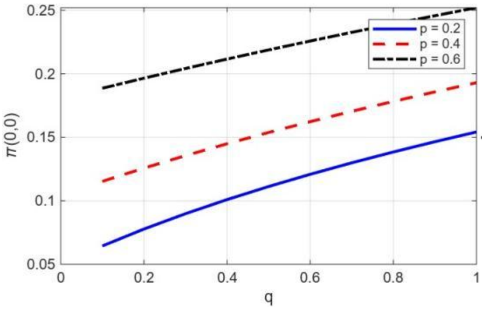


(c) Cost function F

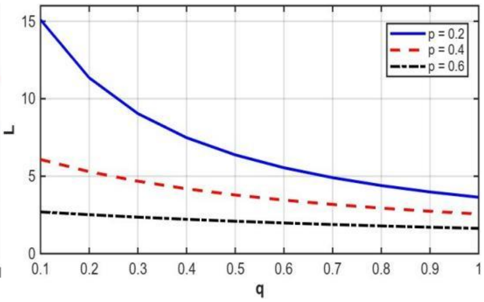
The repair rate β affects the probability of an empty system $\pi(0,0)$, the expected total cost function F as well as the average number of customers L of varying setup rates α for fixed parameter $\lambda = 1, p = 0.4, \theta = 1, \mu = 12, q = 1, k = 4, S_c = 2, F_c = 5, C_0 = 3$ are showed in figure 4. With higher β, the probability $\pi(0,0)$ tends to increase which is a reflection of the quicker recovery of the system and higher availability of servers. This growth causes a massive reduction of the overall cost F that is largely due to reduced costs to downtime and reduced costs to wait.

The average number of customers L is that the average growth is displayed with respect to β, i.e., the more reliable the system applied, the more the system will be in a position to cater to the demand without excessive overloads. Further, the effect of increasing the setup rates α is always the same which is that it increases the F and L which highlights the negative effect on the performance of the operational efficiency of long activation and preparation. This increase results in a significant decrease of the overall cost F, which is mainly caused by less downtime costs and less penalties related to waiting. The means of the number of customers L indicates that an average growth is exhibited with regards to β, i.e., the more reliability of the system used, the more the system will be able to serve the demand without too much overload. Moreover, increasing the setup rates α always leads to an increase in both F and L, which emphasise the negative impact on the performance of the operational efficiency of long activation and preparation. The observations point to the fact that a higher rate of service and repair increases the stability of the system, but the effects of high service and repair rates are heavily counterbalanced by high setup delays, thus supporting the necessity of joint optimization of the speed of service, repair efficiency, and setup operations in order to reduce the total system cost.

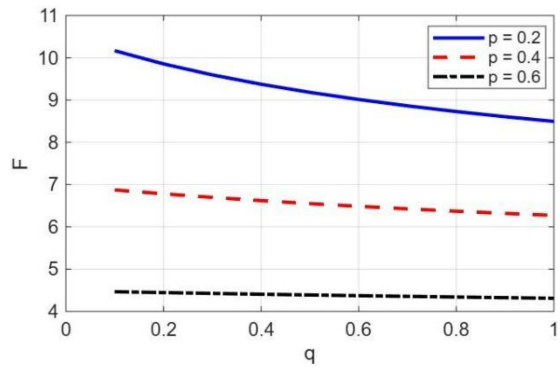
Figure 5. **Effect of repair probability q on π(0,0), L, and F**



(a)Idle probability $\pi(0,0)$



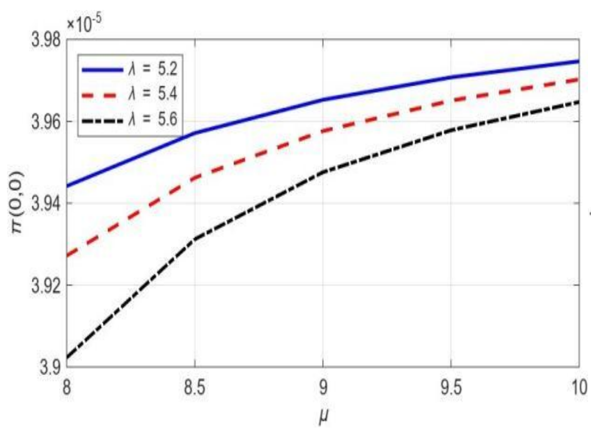
(b) Average number of customers L



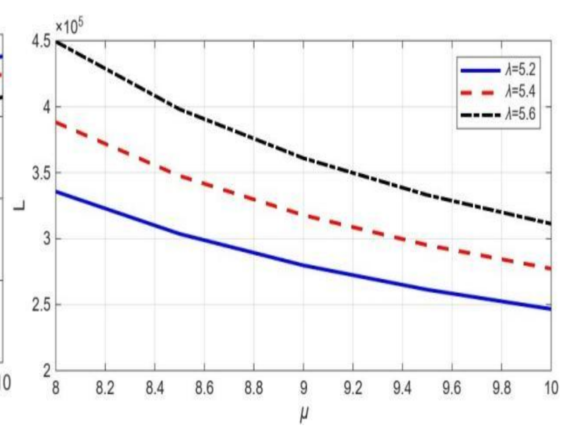
(c) Cost function F

Figure 5 illustrates the effect of the repair probability q on the probability of an empty system $\pi(0,0)$, the average number of customers L, and the expected total cost function F for fixed parameters $\lambda=1,\beta=0.4,\theta=1,\ \mu=12,\alpha=1,k=4,S_c=2,F_c=5,C_0=3.$ Numerical findings show that the more probable the probability of repair q, the higher the probability of an empty system $\pi(0,0)$, which represents more rapid and more dependable system recovery following failures. Simultaneously, the mean number of customers and the overall expected cost F decline considerably, due to the lower repair-related downtimes and a short waiting time of the customers. This illustrates that maintenance policies are effective and can enhance performance and cost effectiveness of the system.

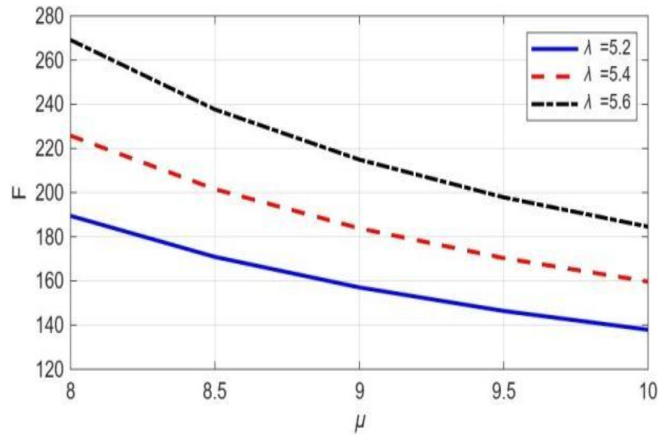
Figure 6. **Effect of arrival rate λ on π(0,0), L and F**



(a) Idle probability $\pi(0,0)$



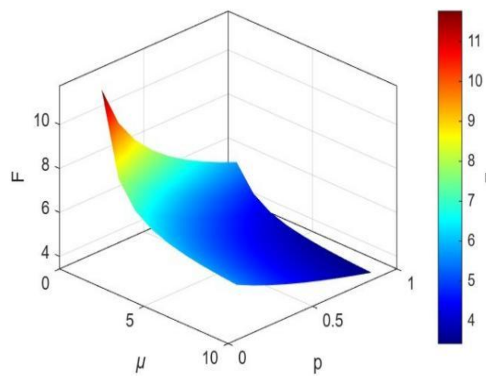
(b) Average number of customers L



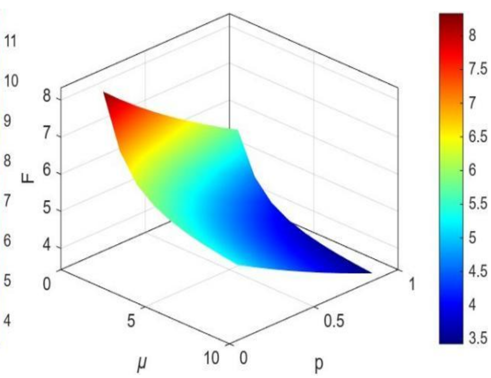
(c) Cost function F

The arrival rate λ on queue length L, the cost function F and $\pi(0,0)$ for $p=0.01, \beta=0.2, \theta=2, q=0, \alpha=0.1, k=2, S_c=3, F_c=5, C_0=4$ are represented in figure 6. It is observed that as the service rate μ increases, the probability of an empty system $\pi(0,0)$ increases due to faster service. The average number of customers in the system L decreases with increasing μ, indicating reduced congestion. The cost F also declines with increase of service rates as the waiting costs and the holding costs are reduced. Given a constant service rate, an increase in arrival rate λ will decrease $\pi(0,0)$ and will increase L and F.

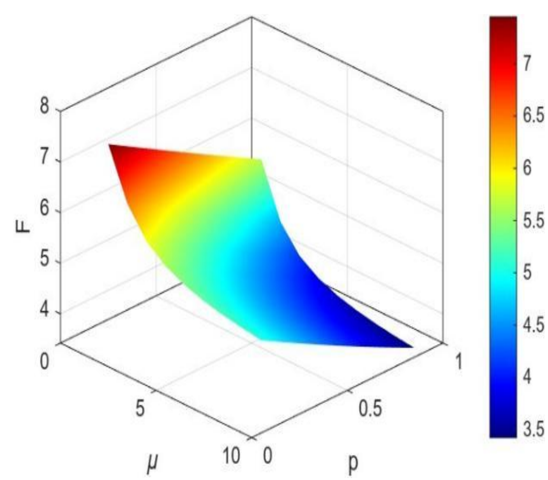
Figure 7. **Cost function F versus service rate μ for different values of q**



(a) $q = 0$



(b) $q = 0.5$



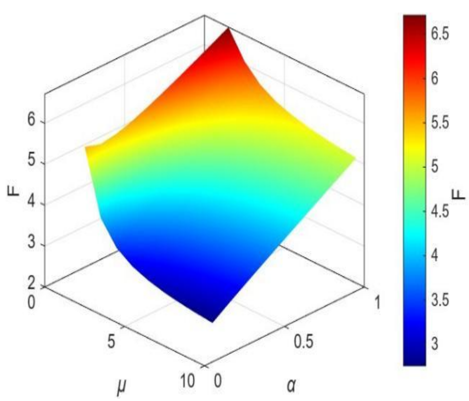
(c) $q = 1$

Table 3. **Cost function (F) for different μ and p when q=0, q=0.5 and q=1**

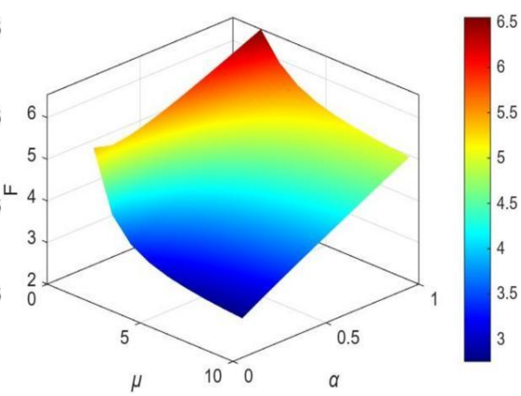
μ	p	$F(q=0)$	$F(q=0.5)$	$F(q=1)$	μ	p	$F(q=0)$	$F(q=0.5)$	$F(q=1)$
1.5	0.2	7.918767	7.918767	7.217394	5.5	0.5	4.595763	4.595763	4.523829
1.5	0.5	6.853338	6.853338	6.536367	5.5	0.8	3.919112	3.919112	3.903093
1.5	0.8	5.964091	5.964091	5.872221	6.5	0.2	5.223434	5.223434	5.056769
2.5	0.2	6.581959	6.581959	6.214243	6.5	0.5	4.472293	4.472293	4.407367
2.5	0.5	5.666099	5.666099	5.508156	6.5	0.8	3.809016	3.809016	3.794991
2.5	0.8	4.874522	4.874522	4.832882	7.5	0.2	5.123548	5.123548	4.967312
3.5	0.2	5.919274	5.919274	5.664095	7.5	0.5	4.384097	4.384097	4.323843
3.5	0.5	5.085047	5.085047	4.979667	7.5	0.8	3.730376	3.730376	3.717666
3.5	0.8	4.355492	4.355492	4.329705	8.5	0.2	5.048735	5.048735	4.899957
4.5	0.2	5.573014	5.573014	5.365444	8.5	0.5	4.318003	4.318003	4.261071
4.5	0.5	4.780496	4.780496	4.697023	8.5	0.8	3.671444	3.671444	3.659660
4.5	0.8	4.083845	4.083845	4.064497	9.5	0.2	4.990634	4.990634	4.847441
5.5	0.2	5.363382	5.363382	5.181175	9.5	0.5	4.266652	4.266652	4.212196
					9.5	0.8	3.625657	3.625657	3.614559

Figure 7 and table 3 observes the service rate μ on the cost function F in the system for $\lambda=1,\ \theta=1,\ \alpha=0.2$ and $\beta=0.7$ and illustrates the variation of the cost function F with respect to the service rate μ for different values of q and p. As μ increases, the cost function decreases steadily, indicating that higher service rates reduce overall system cost.

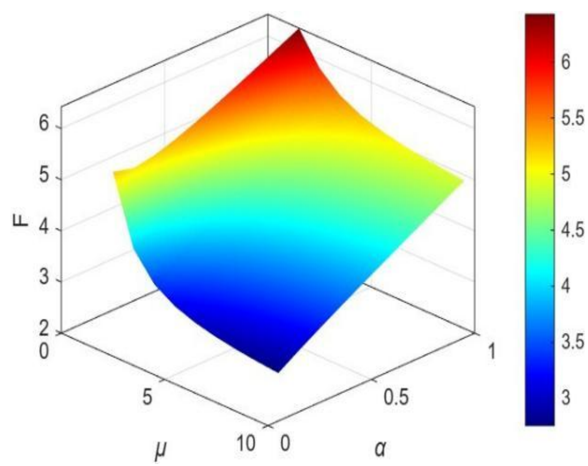
Figure 8. **Cost function F versus service rate μ for different values of α**



(a) q=0



(b) q = 0.5



(c) q=1

Table 4. **Cost function (F) for different μ and α when q=0, q=0.5 and q=1**

μ	α	$F(q=0)$	$F(q=0.5)$	$F(q=1)$	μ	α	$F(q=0)$	$F(q=0.5)$	$F(q=1)$
1.5	0.2	5.395563	5.248955	5.134686	5.5	0.6	4.325172	4.270624	4.225275
1.5	0.6	5.924876	5.777085	5.665523	5.5	1	5.325038	5.221909	5.138178
1.5	1	6.714073	6.549626	6.426697	6.5	0.2	3.194138	3.183697	3.174533
2.5	0.2	4.198839	4.149670	4.108362	6.5	0.6	4.226088	4.175621	4.133473
2.5	0.6	5.132355	5.034863	4.957043	6.5	1	5.227308	5.128764	5.048399
2.5	1	6.075451	5.932523	5.820735	7.5	0.2	3.123727	3.114837	3.107014
3.5	0.2	3.696379	3.670688	3.648600	7.5	0.6	4.154575	4.106891	4.066937
3.5	0.6	4.705785	4.633105	4.573777	7.5	1	5.156043	5.060729	4.982750
3.5	1	5.689298	5.567644	5.470568	8.5	0.2	3.071239	3.063419	3.056522
4.5	0.2	3.443732	3.426662	3.411829	8.5	0.6	4.100585	4.054911	4.016548
4.5	0.6	4.471149	4.410107	4.359707	8.5	1	5.101839	5.008916	4.932790
4.5	1	5.466858	5.356769	5.269795	9.5	0.2	3.030617	3.023575	3.017355
5.5	0.2	3.293447	3.280584	3.269337	9.5	0.6	4.058404	4.014246	3.977088
					9.5	1	5.059251	4.968167	4.893326

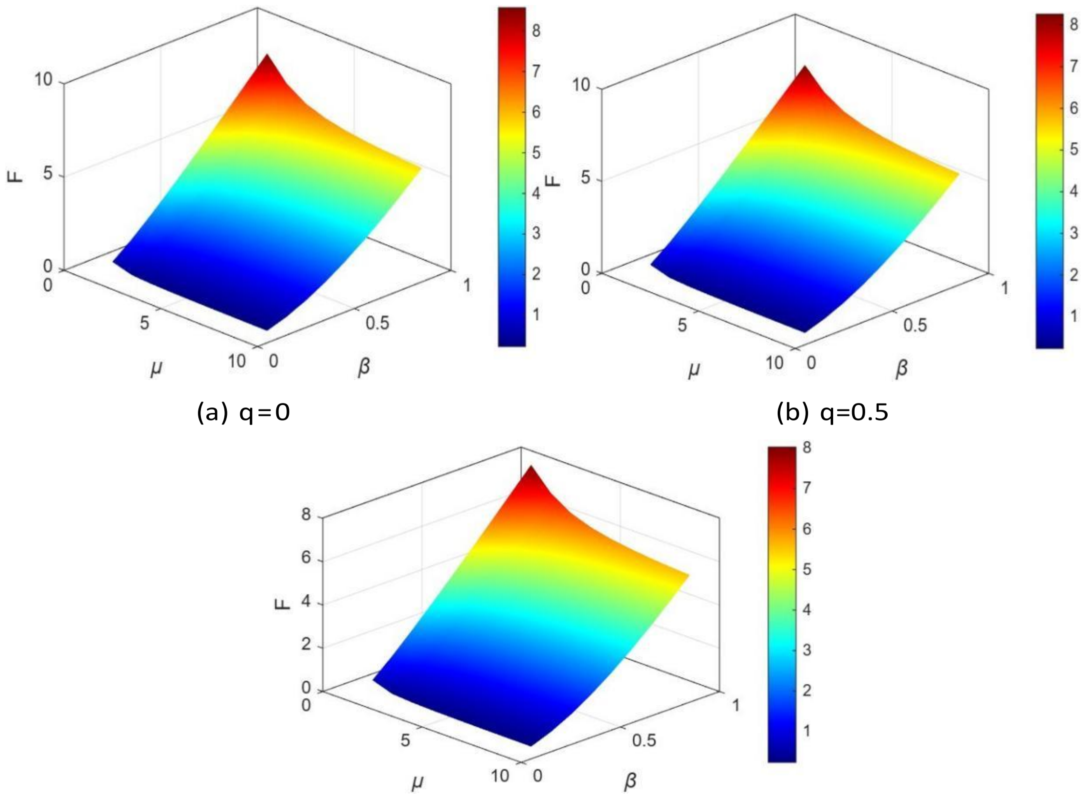
The service rate μ on the cost function F in the system for $\lambda=1,\ \theta=1,\ p=0.7$ and $\beta=0.6$ and it presents the behaviour of the cost function F with respect to the service rate μ for different values of q and α are presented in figure 8 and table 4. It is observed that as μ increases, the cost function decreases consistently. The influence of α is significant at lower service rates, while at higher μ the effect becomes less dominant.

Table 5. **Cost function (F) for different values of μ and β when q=0, q=0.5 and q=1**

μ	β	$F(q=0)$	$F(q=0.5)$	$F(q=1)$	μ	β	$F(q=0)$	$F(q=0.5)$	$F(q=1)$
1.5	0.2	1.471413	1.477757	1.483056	5.5	0.4	1.767807	1.778646	1.788347
1.5	0.4	3.356097	3.301130	3.257299	5.5	0.6	3.293447	3.280584	3.269337
1.5	0.6	5.395563	5.248955	5.134686	6.5	0.2	0.608579	0.619033	0.628720
2.5	0.2	0.885807	0.901387	0.915419	6.5	0.4	1.707716	1.718693	1.728551
2.5	0.4	2.357161	2.358966	2.360525	6.5	0.6	3.194138	3.183697	3.174533
2.5	0.6	4.198839	4.149670	4.108362	7.5	0.2	0.592118	0.602152	0.611464
3.5	0.2	0.735769	0.740995	0.761291	7.5	0.4	1.665556	1.676556	1.686458
3.5	0.4	2.019951	2.028735	2.036483	7.5	0.6	3.123727	3.114837	3.107014
3.5	0.6	3.696379	3.670688	3.648600	8.5	0.2	0.580035	0.589755	0.598787
4.5	0.2	0.669479	0.681398	0.692379	8.5	0.4	1.634359	1.645338	1.655237
4.5	0.4	1.860218	1.870589	1.879824	8.5	0.6	3.071239	3.063419	3.056522
4.5	0.6	3.443732	3.426662	3.411829	9.5	0.2	0.570791	0.580268	0.589081
5.5	0.2	0.632316	0.643358	0.653568	9.5	0.4	1.610348	1.621290	1.631167
					9.5	0.6	3.030617	3.023575	3.017355

The service rate μ on the cost function F in the system for $\lambda=1,\ \theta=1,\ p=0.7$ and $\alpha=0.2$ and illustrates the variation of the cost function F with respect to the service rate μ for different values of q and β are showcased in figure 9 and table 5. It is observed that F decreases as μ increases, reflecting the cost benefits of faster service rates.

Figure 9. **Cost function F versus service rate μ for different values of β**



(a) q=0 (b) q=0.5

(c) q=1

A comparative analysis of figures 7–9 gives relevant information on cost optimization behavior of the system. Figure 7 demonstrates that the overall cost will decrease when the reliability parameter p is increased, given a constant service rate μ because the higher the probability of a successful server activation the higher the availability of the system and the less the costs that are incurred during downtime. Figure 8 indicates that the total cost decreases with increasing setup rate α, as larger α corresponds to shorter setup durations and faster system reactivation. Likewise, figure 8 presents the fact that increased rates β of repairs will result in reduced operational costs as the time of repair will decrease and service continuity will be enhanced. In addition, the numerical findings indicate that the optimum service rate μ equals 0.5 for all repair effectiveness levels q=0, 0.5 and 1, and therefore, with increased repair effectiveness, the system can be run at lower service rates economically. These observations show that both high or low service rates can cause inefficiencies, hence the service rate should be at an optimal level to ensure that the total operating cost is minimised.

Sensitivity Analysis of the Model

Sensitivity analysis is conducted to analyses how the changes in the important system variables affect the performance measures of the proposed retrial queueing model. The analysis gives important information of the strength of the system and the parameters that influence greatly congestion and cost of operation.

(i) Probability of Successful Repair (q) and Reliability Parameter (p): The findings show that the greater the probability of successful repair and server reliability, the less is the average queue length and the cumulative costs. In the case where the repair is not perfect (q=0), the server is highly likely to fail again upon repair, which will lead to a long downtime, a significant increase in congestion, and a significant increase in operation cost. Conversely, perfect repair (q=1) guarantees timely and efficient restoration of service which greatly enhances stability of the system and the waiting time spent by customers is minimised. These observations show the relevance of good maintenance measures and stable server functionality towards reducing system congestion and cost.

(ii) Setup Time (α): The factor of setup time is the time that is needed to restart or activate server after it has been idle. The sensitivity analysis demonstrates that a decrease in the set-up time reduces the average queue length and enables the server to resume service more rapidly, thereby improving the overall performance of the system.

(iii) Service Rate (μ): Service rate is very important in the balancing of system efficiency and cost. As analysed, the service rate reduced the overall cost of operation, hence appropriate selection of service rates is essential. High service rates can cause unnecessary use of energy and upsurge of operating costs, whilst low service rates can cause longer queues and more costly waiting times. Hence, to have an efficient operation of the system at the lowest possible cost, it is important to ensure the service rate is maintained at an optimal range.

CONCLUSION

A new model of an energy conscious single-server with setup delays, stochastic server failure and Bernoulli repair process is offered to improve the knowledge on retrial queueing systems. The model incorporates realistic operational characteristics typical of real-life service systems such as psychological service systems like counseling centres, therapy clinics, crisis helplines and psychiatric outpatient units, which enhances the bridge between queueing theory and applied decision-making. A generating function method is used to obtain steady-state probability distributions and important performance measures for both perfect and imperfect repair models. The results show clear trade-offs between set-up efficiency, service capacity, and repair quality. In the psychological service system, a greater setup rate α means that the therapist builds his orbit more efficiently before the session, which results in more rapid service to the client and less orbit accumulation. The reliability parameter p, represents the probability that the therapist is functional at the time of the therapy or preparation; the greater the value of p (the more resilient the therapist or better supported the clinical setting), the less congestion in the system. The repair effectiveness parameter q reflects the quality of therapist recovery after a period of burnout, compassion fatigue, or administrative disruption: if q=1 the therapist will be immediately available to begin another period of therapy, but if q=0 the therapist will be unavailable for another period of therapy and the system's performance will be degraded. The findings emphasize the importance of optimizing the efficiency of the pre-session preparation, the programs for the well-being of the practitioners and the scheduling of services in single-practitioner mental health service settings to ensure cost-effective and reliable operation. The size of the retrial orbit is a reasonable indicator of unmet psychological needs and the minimization of the orbit by co-optimizing the parameters of the model is a real-world management goal for mental health services.

The proposed framework can be extended to multi-server configurations, such as team-based therapy or group counselling, by including non-exponential session duration distributions, and adaptive scheduling and prioritisation policies for urgent psychological care in the future. The addition of client impatience, drop-out behavior and the mental health service channels of telehealth would make this retrial queueing framework more realistic and applicable in designing sustainable and responsive mental health service systems.

REFERENCES

Abdollahi S, Salehi Rad M R. Analysis of a batch arrival retrial queue with two-phase services, feedback and admission[J]. *Bulletin of the Iranian Mathematical Society,* 2022, 48(3):791-804. https://doi.org/10.1007/S41980-021-00547-1

Arizono I, Takemoto Y. An analysis of steady-state distribution in M/M/1 queueing system with balking based on concept of statistical mechanics[J]. *RAIRO-Operations Research,* 2021, 55:S327-S341. https://doi.org/10.1051/RO/2019064

Atencia I, Moreno P. The discrete-time Geo/Geo/1 queue with negative customers and disasters[J]. *Computers & Operations Research,* 2004, 31(9):1537-1548. https://doi.org/10.1016/S0305-0548(03)00107-2

Bakker A, Bhullar N, Sherwood P. Effective burnout prevention strategies for counsellors and other therapists: A systematic review and meta-synthesis of qualitative studies[J]. *Counselling and Psychotherapy Research,* 2024, 24(2):526-555.

Chang F M, Liu T H, Ke J C. On an unreliable-server retrial queue with customer feedback and impatience[J]. *Applied Mathematical Modelling,* 2018, 55:171-182. https://doi.org/10.1016/J.APM.2017.10.025

Chang J, Wang J. Unreliable M/M/1/1 retrial queues with set-up time[J]. *Quality Technology & Quantitative Management,* 2018, 15(5):589-601. https://doi.org/10.1080/16843703.2017.1320459

Choudhury G, Ke J C. An unreliable retrial queue with delaying repair and general retrial times under Bernoulli vacation schedule[J]. *Applied Mathematics and Computation,* 2014, 230:436-450. https://doi.org/10.1016/J.AMC.2013.12.108

Compton M, Shim R. Eliminating the wait for mental health services: Lessons from systematic service delivery reform[J]. *Administration and Policy in Mental Health and Mental Health Services Research,* 2024, 52(2):346-363.

Das S, Pradhan S. Steady-state analysis of batch-size-dependent service queue with server breakdown, close-down time and multiple vacations[J]. *RAIRO-Operations Research,* 2025, 59(2):743-770. https://doi.org/10.1051/RO/2025013

Dimitriou I. A two-class queueing system with constant retrial policy and general class dependent service times[J]. *European Journal of Operational Research,* 2018, 270(3):1063-1073.

https://doi.org/10.1016/J.EJOR.2018.03.002

Gao S, Wang J, Van Do T. A repairable retrial queue under Bernoulli schedule and general retrial policy[J]. *Annals of Operations Research,* 2016, 247(1):169-192. https://doi.org/10.1007/S10479-015-1885-6

He Y, Tang Y, Yu M, et al. Performance and optimization analysis of a queue with delayed uninterrupted multiple vacation and N-policy[J]. *Methodology and Computing in Applied Probability,* 2024, 26(2):21. https://doi.org/10.1007/S11009-024-10090-1

Jain M, Kaur S. Bernoulli vacation model for MX/G/1 unreliable server retrial queue with Bernoulli feedback, balking and optional service[J]. *RAIRO-Operations Research,* 2021, 55:S2027-S2053. DOI:10.1051/RO/2020074

Jayamani V N, Madheswari S P, Suganthi P, et al. An N-policy M/M/1 queueing model for energy saving mechanism in networks[J]. *Sustainable Computing: Informatics and Systems,* 2024, 44:101016. https://doi.org/10.1016/J.SUSCOM.2024.101016

Jeganathan K, Selvakumar S, Srinivasan K, et al. Two-types of service facilities in interconnected stochastic queueing and queueing-inventory system with marked Markovian arrival process[J]. *Ain Shams Engineering Journal,* 2024, 15(9):102963. https://doi.org/10.1016/J.ASEJ.2024.102963

Kalaiselvi J, Saravanarajan M C. Exploring non-Markovian dynamics: Stochastic analysis of a single-server retrial queue with recurrent clients and balking behavior under extended Bernoulli vacations[J]. Heliyon, 2024, 10(13). https://doi.org/10.1016/J.HELIYON.2024.E33785

Kanavetas O, Logothetis D. Unreliable service system with strategic customers: Effect of abandonments and repair time[J]. Annals of Operations Research, 2024, 1-38. https://doi.org/10.1007/S10479-024-06297-7

Kumar B K, Madheswari S P, Vijayakumar A. The M/G/1 retrial queue with feedback and starting failures[J]. Applied Mathematical Modelling, 2002, 26(11):1057-1075. https://doi.org/10.1016/S0307-904X(02)00061-6

Kumar A, Jain M. Cost optimization of an unreliable server queue with two stage service process under hybrid vacation policy[J]. Mathematics and Computers in Simulation, 2023, 204:259-281. https://doi.org/10.1016/J.MATCOM.2022.08.007

Li Z, Wang H, Chen X. A Markov decision optimization of medical service resources for two-class patient queues in emergency departments via particle swarm optimization algorithm[J]. Scientific Reports, 2025, 15:2867. https://doi.org/10.1038/S41598-025-86806-7

Liu X, Xu X, Liu M. Strategy analysis of retrial queue with parallel customer and standby server[J]. Methodology and Computing in Applied Probability, 2024, 26(3):27. https://doi.org/10.1007/S11009-024-10098-7

Madheswari S P, Suganthi P. Retrial queue with unreliable server and second optional service under K types of Bernoulli vacations[J]. International Journal of Services and Operations Management, 2021, 40(4):502-533. https://doi.org/10.1504/IJSOM.2021.120058

Madheswari S P, Krishna Kumar B, Suganthi P. Analysis of M/G/1 retrial queues with second optional service and customer balking under two types of Bernoulli vacation schedule[J]. RAIRO-Operations Research, 2019, 53(2):415-443. https://doi.org/10.1051/RO/2017029

Sundararaman M, Narasimhan D, Rajadurai P. Performance analysis of two different types of waiting queues with working vacations[J]. Journal of Mathematics, 2024, 2024(1):5829171. https://doi.org/10.1155/2024/5829171

Phung-Duc T. M/M/1/1 retrial queues with setup time[C]. International Conference on Queueing Theory and Network Applications, 2015:93-104. https://doi.org/10.1007/978-3-319-22267-7_9

Shanmugam B, Saravanarajan M C. Unreliable retrial queueing system with working vacation[J]. AIMS Mathematics, 2023, 8(10):24196-24224. https://doi.org/10.3934/MATH.20231238

Singh S K, Acharya S K, Cruz F R, et al. Estimation of traffic intensity from queue length data in a deterministic single server queueing system[J]. Journal of Computational and Applied Mathematics, 2021, 398:113693. https://doi.org/10.1016/J.CAM.2021.113693

Sundarapandiyan S, Nandhini S. Sensitivity analysis of a non-Markovian feedback retrial queue, reneging, delayed repair with working vacation subject to server breakdown[J]. AIMS Mathematics, 2024, 9(8):21025-21052. https://doi.org/10.3934/MATH.20241022

Tran Q D. Going beyond waitlists in mental healthcare[J]. Community Mental Health Journal, 2024, 60(4):629-634. https://doi.org/10.1007/S10597-024-01233-2

Vijayashree K V, Pavithra P. Performance analysis and cost optimization of an M/M/1 queue with N-policy and working breakdown[J]. Methodology and Computing in Applied Probability, 2025, 27(3):68. https://doi.org/10.1007/S11009-025-10196-0

Wang J, Zhang P. A discrete-time retrial queue with negative customers and unreliable server[J]. Computers & Industrial Engineering, 2009, 56(4):1216-1222. https://doi.org/10.1016/J.CIE.2008.07.010

Wang J, Zhang X, Huang P. Strategic behavior and social optimization in a constant retrial queue with the N-policy[J]. European Journal of Operational Research, 2017, 256(3):841-849. https://doi.org/10.1016/J.EJOR.2016.06.034

Xu B, Xu X. Equilibrium strategic behavior of customers in the M/M/1 queue with partial failures and repairs[J]. Operational Research, 2018, 18(2):273-292. https://doi.org/10.1007/S12351-016-0264-7

Xu F, Tian R, Shao Q. Optimal pricing strategy in an unreliable M/G/1 retrial queue with Bernoulli preventive maintenance[J]. RAIRO-Operations Research, 2023, 57(5):2639-2657. https://doi.org/10.1051/RO/2023146

Xu F, Tian R, Shao Q. Optimal pricing strategy in an unreliable M/M/1 retrial queue with delayed repair and breakdown deterioration[J]. Methodology and Computing in Applied Probability, 2024, 26(2):11. DOI:10.1007/S11009-024-10080-3

Yang D Y, Dai Y Y, Wu C H. Equilibrium balking strategies in unobservable queues with multiple vacations and an optional service[J]. RAIRO-Operations Research, 2024, 58(3):2569-2589. https://doi.org/10.1051/RO/2024083

Yang D Y, Wu C H. Performance analysis and optimization of a retrial queue with working vacations and starting failures[J]. Mathematical and Computer Modelling of Dynamical Systems, 2019, 25(5):463-481. https://doi.org/10.1080/13873954.2019.1660378

Zhang Y. Strategic behavior in the constant retrial queue with a single vacation[J]. RAIRO-Operations Research, 2020, 54(2):569-583. https://doi.org/10.1051/RO/2019016

Zhou M, Liu L, Chai X, et al. Equilibrium strategies in a constant retrial queue with setup time and the N-policy[J]. Communications in Statistics-Theory and Methods, 2020, 49(7):1695-1711. https://doi.org/10.1080/03610926.2019.1565779