periodica polytechnica

Social and Management Sciences
16/1 (2008) 21–32
doi: 10.3311/pp.so.2008-1.03
web: http://www.pp.bme.hu/so

RESEARCH ARTICLE

A review of management issues related to express line systems

Noémi Kalló / *Tamás* Koltai

Received 2008-09-14

Abstract

As time-based competition is gathering ground among leading companies, express checkouts become more and more widely used for waiting time reduction. Express line systems, however, can be formed and operated in several ways. This article reviews the most important management issues related to the analysis and improvement of waiting process at the checkouts of express line systems.

The most important parameter of express line systems is the limit value. This parameter determines the group of customers who are entitled to use the express checkouts. In this way, the limit value has significant effect on waiting experience as well. In express line systems, the classical objective of operations management, waiting time minimization, should be complemented by the consideration of other factors as well. In these systems, the variation of waiting, the perceived waiting time and the evaluation of waits are important issues as well. The operation of these systems should be evaluated only knowing the effects of introducing express checkouts on all of these parameters.

Keywords

queuing theory · express line systems · variance of waiting time · perception management · satisfaction · utility theory

Noémi Kalló

Department of Industrial Management and Economics, BME, Budapest, 1111, Műegyetem rkpt 9., Hungary
e-mail: kallo@mvt.bme.hu

Tamás Koltai

Department of Industrial Management and Economics, BME, Budapest, 1111, Műegyetem rkpt 9., Hungary
e-mail: koltai@mvt.bme.hu

1 Introduction

Companies, which are successful in cost- and quality-based competitions, are looking for other factors that can help them to gain further competitive advantage. Consequently, time-based competition spreads among leading companies. Time has turned into a strategic resource and, as a consequence, its importance has become equivalent to the significance of money, productivity and innovation [30]. While companies have to hold on in cost- and quality-based competitions, they also should become competitive in the area of time parameters. That is, competitiveness nowadays requires balancing cost, quality, and time.

Time parameters can appear in several forms. There are internal parameters, measurable only by the company, and external ones, visible to customers as well; in addition, parameters related to product development and production, distribution also can be differentiated [3]. Production companies have a choice which one to bring into focus. In services, however, the decisive time parameters mainly are external and related to "distribution", to serving customers. Customers in a supermarket – assuming acceptable supply and price – are interested in fast services: the market is sensitive to delivery time. That is, "process time-based competitor" strategic orientation is characteristic for this market [3].

In time-based competition environment, one of the main objectives of service managers is reducing customers' waits. Talking about waiting for a service, however, does not indicate the same to everybody. There are different types of waiting measures (related to wait duration or the number of waiting people). Some measures can cover different periods of service (waiting in the queue or in the system); and some can occur in different phases of the service (pre-process, in-process, and post-process waiting).

The number of waiting customers can have significant effect on customer decisions. For example, if people have to choose among similar, unknown service providers (e.g. restaurants), they tend to choose the service for which more people are waiting. When we are talking about waiting reduction as a competitive advantage, it is the duration of waiting that should be reduced. The number of waiting people, however, does not have

a direct linkage to waiting duration (especially when people's services require highly different amounts of time). For this reason, one of the most important objectives of service managers is to reduce the (average) waiting time of their customers.

People are tolerable to different extents when they are waiting and when they are served. People are generally less anxious when their own service lasts too long than when they must wait in the line longer because of the long service of other people. Moreover, before service people are nervous whether they will really receive the service for which they are waiting. Accordingly, decreasing of pre-process wait should get higher priority than reducing in- and post-process waitings [22]. Similarly, in storehouses, where people are waiting in queues before their service starts at a checkout, the waiting in queue is a more suggestive waiting parameter than the waiting time in system.

The simplest way for decreasing customers' waiting time in line is using additional servers [8]. This kind of waiting time reduction is, however, quite expensive. Therefore, search for the best configuration of waiting lines and service facilities has become a major concern of service managers [7].

A frequently used configuration of queuing systems for waiting time reduction is the application of express lines. When express lines are applied two customer groups are created: the first group involves customers buying only a few items, and all other customers belong to the second group. Customers buying more items than a certain amount have to use the regular checkouts, and only the others can join the express lines. The number of items that controls line-type selection is called limit value.

As most queuing problems, express line systems can be modelled both analytically and empirically. Analytical models are based on the results of queuing theory. In this way, some existing analytical models are used to approximate the operation of the queuing system. These models are quite simple but they give only rough estimation of the real operation. Simulation models are used for empirical studies. This kind of modelling is more sophisticated, at the same time, it is also more difficult to build and manage. To analyze express line systems both approaches should be used.

The operation of express line systems is the following: there are many checkouts, all with their own waiting lines; people buying few items use the express checkouts, other people use the other servers. All express line systems – where no snake lines are used – operate in this way. However, there are two decision points in this process, which can make a difference in the operation. First, a decision must be made about the group of customers who are entitled to use the express checkouts. This question is answered by the management with the value of limit parameter. Next, customers facing several waiting lines must decide on which line to pick. Answers given for the second question can make differences in the operation of express line systems. As this decision is beyond the management's control, its effect on operation should be known. The analyses of this kind of customer behaviour should also be a part of studies analyzing express line systems.

Our analyses revealed that one of the main parameters which influences waiting time and which can be controlled by the management is the limit value. With different limit values different waiting times can be achieved. An important management objective related to express line systems is to determine the limit value which optimizes the operation.

There is a limit value which minimizes the average waiting time in line of all customers. However, average wait should not be considered as the only measure describing customer waiting. Also from statistical point of view, alone the mean value of a parameter does not bear sufficient information [6]. Besides measures of central tendency, measures of statistical dispersion must be analyzed as well. In terms of waiting: it makes significant difference whether an average waiting time of 5 minutes varies between 4.5 and 6 minutes or alters between 1 and 20 minutes. The difference is made by uncertainty, unsteadiness. As people are generally risk-averse, it can happen that a longer mean waiting time with smaller variance is more desirable than a shorter average with larger variability. That is, the objective of waiting time reduction should be complemented with the intention to decrease the variance of waiting times as well.

As in service systems people are the objects of waiting, beside of the exact quantitative data, qualitative information should also be taken into consideration. With the help of operation management and queuing theory, the average waiting time and the variance of waits can be determined. However, the discipline of psychology and perception management should be invoked when human characteristics related to waiting experience are considered as well.

The measurable waiting times are experienced by the customers, and the observed waiting duration generally differs from the measured one. The explanation is that people do not observe time linearly [31]. Accordingly, the perceived waiting time and the actual wait duration are different. Waiting time is a strong determinant of customer satisfaction with service [34], [10]. At the same time, customers' reactions to waiting duration have a closer relationship with the perceived than with the actual waiting time [9]. Consequently, satisfaction maximization is rather an analogue to minimizing perceived waiting time than to the minimization of actual waits.

While time perception across line types (e.g. multi-server, snake, express lines) is not statistically different [9], the same cannot be said about satisfaction related to waiting. People's tolerance for waiting highly depends on the perceived value of the service for which they wait [22]. Actually this characteristic of customer tolerance has led to the formation of express checkouts. For evaluating the operation of express line systems, not only (actual and perceived) waiting times should be concerned but a more sophisticated approach should be used. Applying a kind of utility function these differences in (dis)satisfaction related to waiting can be considered and a more appropriate objective function can be formulated.

This paper provides a review of the most important operational issues related to express line systems. First, analytical and simulation models developed for analyzing express line systems are presented. With the help of simulation, customer behaviour issues, like waiting line selection, and its effect on optimal operation is analyzed as well. Next, different objective functions are analyzed: management can minimize the average waiting time, the variance of waits, the perceived waiting duration or the dissatisfaction generated by waiting as well. In all cases, the effect of limit value on the objective function and on its minimum point is analyzed. Finally, the main conclusions of the analyses are summarized.

2 Model selection

To analyze express line systems a model suitable to the queuing problem should be created, which can raise some issues. The first issue is related to the design of these queuing systems. They are generally applied in stores where every checkout has its own waiting line. This characteristic does not seem to be special but, from modelling point of view, it is a design hard to handle. Unfortunately, in those basic analytical models where there are many checkouts all of them serve a common waiting line. The next important issue is associated to the operation of these systems. The operation of express line systems is based on a rule controlling the checkout and waiting line access. However, elementary queuing models do not contain this type of regulations.

The operation of express line systems, fortunately, is built on simple rules: people who are entitled to use the special servers (i.e. express checkouts) use these ones (as the system is designed in that way that this results the most favourable waiting characteristics); the other customers use the other checkouts (as the regulation does not allow them to use the express lines).

Customers, however, tend to push and break rules. Specialties caused by their behaviour can be experienced in several forms: defections from the queue, jockeying among many waiting lines, balking before entering queues, bribing or cheating for queue position, etc. [15]. In our analyses, it is assumed that customers are patient (i.e. do not leave the store without service and do not jockey among waiting lines) and use the checkout type assigned to them.

To analyze express line systems, an analytical and a simulation model were created as well. As it will be shown later, analytical models can be used only as approximations for this queuing problem. In issues which cannot be modelled analytically the more sophisticated simulation model provides the only practical approach [8]. Accordingly, the fast analytical and the sophisticated empirical approaches complement well each other in the studies of express line systems.

For building the analytical and the simulation models, only such information and data are used which can be determined without actual introduction of express checkouts. Therefore, decisions needed to adopt express lines to the actual system can be made in advance – based on their effect on customer waiting.

In our analyses, the real data of a do-it-yourself superstore is used [16]. In this store, generally five checkouts operate. Using the data provided by the checkout information system, the arrival rates for the different days and for the different parts of the days was estimated. For all periods the Poisson arrival process is acceptable according to Kolmogorov-Smirnov tests. Based on Rényi's limiting distribution theorem and its generalizations, the distribution function of the time interval between two consecutive events remains the same after a rarefaction and coordinate transformation [28, 32, 33]. That is, the arrival processes of the two customer groups also can be approximated with Poisson processes.

The density function of the number of items bought by customers is also provided by the checkout information system, and for describing it a truncated geometric distribution with a mean of 3.089 is found acceptable by a chi-square test.

The service time of customers cannot be obtained from any information systems; therefore, it was measured manually. The linear relationship between the number of items bought and the service time was tested with regression analysis. A 0.777 correlation coefficient supported the assumption of linearity. According to the results of linear regression, service time has two parts. On the average the part independent of the number of items bought lasts 0.5463 minute, reading a bar code needs 0.1622 minute. With linear regression the standard deviation of these parameters and the service times of customers buying different amounts were determined as well (for details see [16]).

Results presented in this article are valid for a midday traffic intensity with an arrival rate most characteristic for the store (λ=95 customers/hour). According to the geometric distribution, customers buy generally only few items. Therefore, two of the 5 working checkouts was considered to be express servers (S=5, E=2).

2.1 Analytical approach

Express lines are generally used in supermarkets where many service facilities are located and each has its own separate waiting line. Analyzing this kind of queuing systems with the models of queuing theory presents difficulties because there is no existing analytical model which can properly describe such systems. In this case, two analytical models can be used as approximations: one consisting many service facilities with *one common waiting line* (snake line) and another containing many *independent* queuing systems in which there is *one service facility* with its own separate queue.

If analytical formulae have to be used for the whole queuing system – containing k checkouts and k waiting lines –, the following two approaches can be used:

1) *One-common-line approach*. In this case, the queuing system can be modelled as if all checkouts had one common queue. In this case, a G/G/k model can be applied. In the store in question, since the arrival rate follows Poisson distribution, the M/G/k model can be used. If there are E express and R

($R = S - E$) regular checkouts, then a model with $k = E$ and another with $k = R$ are required.

Modelling the checkout system as a queuing system with one common line for all checkouts is an overly optimistic approach. It underestimates the average waiting time by assuming optimally efficient queue selection of customers which minimizes their waiting times. Consequently, this approach provides a *best case* estimate of the operation of the queuing system.

2) *Independent-queuing-system approach*. In this case, k independent G/G/1 models are assumed. In the store in question, since the arrival rate follows Poisson distribution, the M/G/1 model can be used. If there are E express and R regular checkouts, then $E + R$ independent M/G/1 models are required.

Modelling the checkouts of a supermarket as independent queuing systems gives a pessimistic approach of wait. It overestimates the average waiting time by assuming a random selection of lines and by neglecting the waiting time reduction effect of line selection and jockeying. The independent-queuing-system approach provides a *worst case* estimate of the operation.

It depends on the system characteristics which of the presented approaches gives a more accurate approximation of actual operation. If customers estimate the workloads of servers before selecting waiting line, the application of the one-common-line approach gives a better approximation [29]. For some reasons, however, customers do not always selects the queue with the minimal workload, for example, when workloads cannot be observed. In this case, the independent-queuing-system approach is more suitable.

To create a numerical model for analyzing express line systems the approaches discussed formerly are used as best-case and worst-case estimations of waiting time. In Fig. 1, a part of the numerical model can be seen – with the data of the do-it-yourself superstore [16]. Introducing express lines into a queuing system can be carried out in several ways. If there is only one express line, the only question is the value of limit parameter. If there are many express lines, the number of express lines and their limit values must be determined jointly. To determine the limit value which optimizes operation, express line systems with all possible limit values must be analyzed. For this, characteristics of customer groups generated with all possible limit values must be determined. These main characteristics are the arrival rates, the average service times and the variances of service times.

For determining the service characteristics of different customers the relationship between the number of items bought and service times must be analyzed. By using this relationship, the average service time and the variance of service times can be determined for customers buying the same number of items. With the help of the distribution function of the number of items bought, the average arrival and service rates, and the variances of service times can be determined for all possible customer groups as well (for details see [16]).

The model in Fig 1 works in the following way. Based on the main characteristics of the existing queuing systems (in italics), the formerly mentioned special characteristics of the express line systems with different limit values are determined. With these parameters, using the formulae of M/G/1 and M/G/k queuing models, the average waiting times can be calculated (typed boldface). Knowing all possible waiting times, the smallest one must be selected (framed). This minimal average waiting time determines the optimal limit value. Analyses with different parameter values showed that the waiting time as a function of the limit parameter has a distinct minimum. That is, a limit value which optimizes the operation can easily be determined for any express line system.

2.2 Simulation approach

Analytical models cannot handle precisely queuing problems where many checkouts all with their own waiting lines should be modelled. The analytical model appropriate for describing express line systems assume either a certain customer distribution among accessible waiting lines (independent-system approach) or do not deal with line selection at all (one-common-queue approach).

For studying the operation of express line systems in more details, a simulation model was built. The simulation model using the block of Arena simulation software can be seen on Fig. 2.

The simulation model works in the following way. The CREATE block generates customers according to a stochastic distribution. The ASSIGN block, based on a formerly defined distribution function, determines the number of items bought by each customer. With this quantity, knowing the stochastic relationship of the two parameters, the service time of each customer is also calculated. The BRANCH block separates two customer groups: one of them can use the express checkouts, the other are directed to the regular lines. Customers entitled to use express checkouts buy no more items than the limit value. Customers in each group have to make a decision about which line to choose. Queue selection is controlled by the PICKQ blocks. Next, the customer joins the selected QUEUE, waits until the server becomes free and can be SEIZEd. At this point, the waiting process in queue ends. The waiting time is recorded by a TALLY block. The following BRANCH block is needed for data collection and statistical analyses. The customer's route continues along the solid lines while waiting time data of the same customer group are combined in TALLY blocks (along the dashed lines). The service needs a specific amount of time; therefore, the customer is DELAYEd. When service ends, the server is RELEASEd and made free for the next customer. At this point the sojourn time ends as well and it is also recorded by a TALLY block. With combining the different waiting time data, the customer can leave the system at the DISPOSE block.

In the store in question, express lines have not been used yet. Consequently, the real queuing system could not be used to validate the simulation model. Therefore, the analytical models

Arrival rate (λ) [cust./hour]	*95*
Average number of items bought (l)	*3.0890*
Fix element of service time (a) [min]	*0.54627*
Variable element of service time (b) [min]	*0.16222*
Number of lines (S)	*5*
Number of express lines (E)	*2*
Sample size (n)	*10000*

Limit parameter (L)	**0**	**1**	**2**	**3**	**4**	**5**
Density function (p_i)	*0.0000*	*0.3237*	*0.2189*	*0.1481*	*0.1001*	*0.0677*
Distribution function (P_i)	0.0000	0.3237	0.5427	0.6907	0.7908	0.8585
Service time when L item is bought (t_i)	0.0000	0.7085	0.8707	1.0329	1.1951	1.3574
Variance of service time (σ_i^2)		*0.0016*	*0.0012*	*0.0011*	*0.0012*	*0.0015*
Arrival rate to the express line (λ_E)	0.0000	0.5126	0.8592	1.0936	1.2522	1.3594
Service time in the express line (t_E)		0.7085	0.7739	0.8295	0.8758	0.9137
Service rate in the express line (μ_E)		1.4114	1.2921	1.2056	1.1419	1.0944
Variance of service time (σ_E^2)		0.0016	0.0078	0.0176	0.0303	0.0449
Arrival rate to the regular line (λ_R)	1.5833	1.0708	0.7241	0.4897	0.3312	0.2240
Service time in the regular line (t_R)	1.0474	1.2096	1.3718	1.5340	1.6962	1.8585
Service rate in the regular line (μ_R)	0.9548	0.8267	0.7290	0.6519	0.5895	0.5381
Variance of service time (σ_R^2)	0.1716	0.1718	0.1722	0.1728	0.1737	0.1748
Average waiting time in the line {M/G/1}						
Express line (t_{qE})		**0.0788**	**0.1952**	**0.3530**	**0.5525**	**0.7890**
Regular line (t_{qR})	**0.7486**	**0.5134**	**0.3706**	**0.2750**	**0.2072**	**0.1573**
All lines (t_q)	**0.7486**	**0.3727**	**0.2755**	**0.3289**	**0.4803**	**0.6997**
Average waiting time in the line {M/G/*k*}						
Express line (t_{qE})		**0.0121**	**0.0487**	**0.1102**	**0.1957**	**0.3023**
Regular line (t_{qR})	**0.1338**	**0.0670**	**0.0334**	**0.0162**	**0.0076**	**0.0035**
All lines (t_q)	**0.1338**	**0.0492**	**0.0417**	**0.0811**	**0.1563**	**0.2600**

Fig. 1. Numerical model

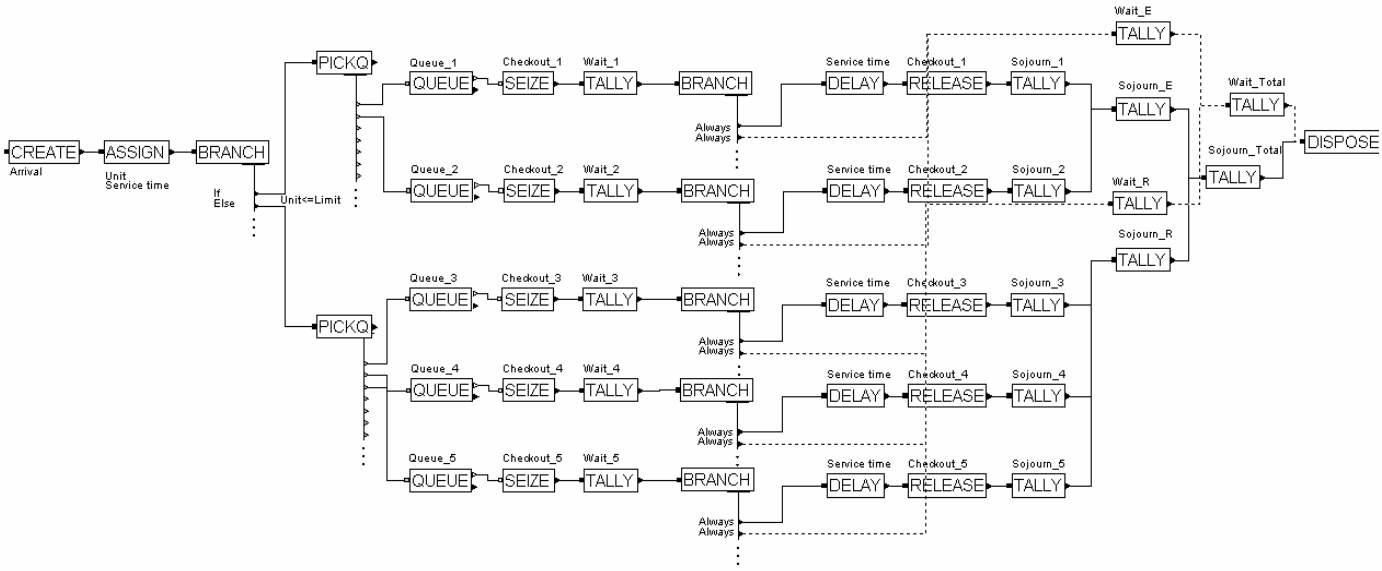

Fig. 2. Simulation model

were applied for checking the validity of results. For this, the fundamental simplifications applied in analytical models were introduced to the simulation model. In the M/G/k simulation model, there is *a common* line for customers entitled to use express checkouts and another one for customers buying many items. In the M/G/1 simulation model, there are *independent* arrival processes for all of the checkouts and their own waiting lines. The analytical and simulation results gained by the same type of models are quite close to each other; therefore. the simulation model can be considered valid [16]. The analytical and simulation results are listed later in Table 1.

3 Customer behaviour

The real queuing system, like the simulation model, contains many service facilities each with its own separate waiting line. Accordingly, customers have to make a decision about which line to pick. This decision is beyond the control of managers; therefore, its effects on the waiting characteristics should be studied as well. In this way, the optimal limit value can be determined considering the customers' behaviour as well.

As discussed before, it was assumed that customers are patient and use the checkout type assigned to them. It is also supposed, however, that they are speculative, try to select waiting lines in a way that minimizes their waiting time. To simulate this behaviour, different *line selection rules* can be applied. We analyzed whether this kind of customer behaviours has significant effect on the optimal operation or management can neglect this information while making decisions about the value of limit parameter.

Most simulation software can help to analyze the line selection behaviour of customers by having built-in line selection rules. For analyzing the effects of line selection on optimal operation the following of the available built-in line selection rules of Arena was considered as most suitable.

– RANdom (RAN) rule. Customers randomly choose among the available waiting lines. If customers do not make considerable effort to decrease their waiting time or if we have no information about their behaviour, applying this rule can be a good solution. This kind of behaviour, however, is not really characteristic for customers trying to reduce their waiting times.

– Smallest Number in Queue (SNQ) rule. This rule sends the customer to the shortest waiting line. It seems to be a realistic decision rule, supposing that empty queues occur rarely enough. However, if there are many checkouts and waiting lines in the store, then customers cannot compare all of the queues. It can also happen that the length of the waiting lines cannot be observed. In these cases this rule cannot be applied.

SNQ rule tries to minimize the waiting time which is, however, not always depending only on the number of customers. Express lines are generally formed in stores where there are significant differences among the amounts bought by customers. Because of these differences, the length of services has large variance. As a consequence, in express line systems the distinctions in customers' services is more important than in other systems. Therefore, it can be assumed that customers consider the differences of service times in their waiting line selection. For example, they do not judge waiting time based only on the number of people in queue but considers the amounts bought by customers in queue as well. To take into consideration this argument some other (user defined) queue selection rules were created.

– Minimal WorkLoad (MWL) rule. This rule determines the workload presented by each waiting line. Workload of a queue is defined as the number of items bought by all customers in that waiting line. Customers select the waiting line with the minimal workload. This rule is preferred to SNQ rule because it takes into consideration the different amounts bought by customers. This rule can be used, however, only when the number of items bought can be observed by customers.

– Smallest Number in System (SNS) rule. Another problem with the operation of the SNQ rule is that it does not take into account those customers who are being served. Consequently, this rule does not distinguish an idle *server* from a checkout serving a customer. The *queue* is empty in both cases, joining the empty server is, however, better than joining the one with a customer being served. The SNS rule differentiates such servers by reckoning the number of customers in system instead of the number of customer in line. If there are too many queues or they cannot be observed, this rule, similarly to the SNQ rule, cannot be applied.

– Fewest Last Items (FLI) rule. Counting the number of items bought by all customers in the queue is generally impossible. The number of items bought by the *last* customers in the different queues, however, can be easily observed. In these situations, customers can select a queue according to the FLI rule. This rule generally complements the SNQ or SNS rule and it is used only in cases when there are many identically short waiting lines. In this case, that shortest line will be chosen in which the last customer has the fewest items.

– Minimal Perceived Amount (MPA) rule. Even if customers can observe the amounts bought by other customers, it is hard to tell exactly how many items they have. However, it can be easily decided that there are few or many items in their baskets or in their shopping carts. MPA rule approximates this kind of customer behaviour: different weights are given to the different amounts. In our analyses, four customer groups (buying very small, small, large, and very large amounts) were used with exponentially increasing weights. The sum of these attributes in each line is calculated and the line with the smallest value is chosen.

– Shortest Service Time (SST) rule. It takes into account the number of items bought and the number of customers in queue as well. The waiting time has two parts: the first part depends on the number of items bought by previous customers, the other is independent of it. The independent part (caused by the time needed for payment) is influenced by the number of customers in queue. Customers, according to their experiences, can use different weights to calculate the weighted sum of number of customers and number of items, that is, to estimate their service times. If the weight of number of customers is near to zero, then the MWL rule is approximated. If the weight of number of items converges to zero, then the SNS rule is obtained. In our analysis a weight according to the service time data of the store in question were used.

These are only examples of line selection rules that can be used by customers in express line systems. These rules, from the more accurate through the simpler ones, however, cover well the range of rules based on which customers make their line selection decisions.

To approximate the average waiting time in an express line system two analytical model can be used. From the results listed in Table 1, however, it can be seen that if simulation models work according to the real system (that is, consider customer behaviour), the results are more similar to the analytical results of the independent M/G/1 models. There is only one rule providing similar data to the results of the M/G/k model. Consequently, the independent-queuing-systems approach seems to be a better approximation for the operation of express line systems. That is, customer speculation for reducing waiting time, in these systems, does not help to make customer distribution among cash desks more efficient but intensifies the deficiencies arisen from waiting line selection.

4 Waiting time minimization

In time-based competition environment, improving time related parameters should be a primary objective for service managers. To make a careful decision, first of all, the parameters influencing waiting time should be determined. In this issue, analytical models are superior to simulation as they unfold the cause-of-effect relationships within the system [8].

In express line systems, the following parameters influencing waiting time can be determined:

- Arrival rate,
- Average number of items bought (distribution of items bought),
- Service time (its part depending on the number of items bought and its independent part),
- Number of checkouts (all, express, and regular checkouts),
- Limit value.

All of these parameters have significant effect on waiting times. There are, however, only two which can be directly controlled and effortlessly changed by the management: the number of operating checkouts and the limit value. Changing the number of operating checkouts is highly expensive and the change of the ratio of express and regular servers (keeping the number of all checkouts constant) results too high variation in the operation. Consequently, the most important parameter which can be used for managing operation is the limit value. The change of this parameter requires no expenses and it is controlled completely by the management.

To analyze the detailed effect of limit parameter on customer waiting, the average waiting time as a function of the limit parameter was determined using all suggested models (results are listed in Table 1). Sensitivity analyses showed that the optimal limit value is very robust to the change of the major parameters of the system. Therefore, it can be presumed that the effect of the limit parameter on the average waiting time of an express line system has a general pattern [16]. This pattern can be seen in Fig. 3. The solid curve describes the average waiting time as a function of the limit parameter. The dashed line represents the average waiting time in the operating queuing system (which is independent of the limit value).

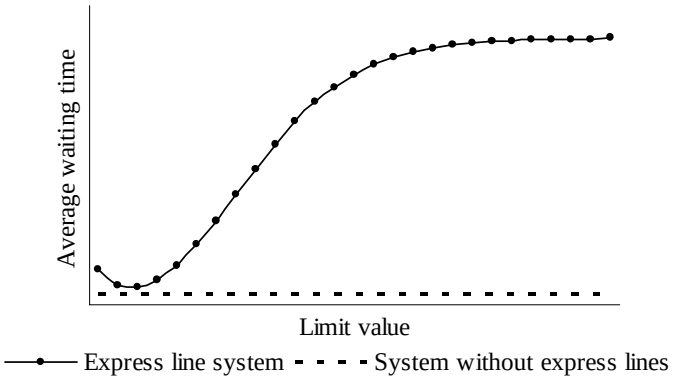

Fig. 3. Waiting time as a function of the limit parameter

Based on the curve in Fig. 3, it can be concluded that the limit value has great impact on the average waiting time. Applying a limit value which is close to the optimal one, however, does not increase waiting time significantly, that is, applying such a limit value can be reasonable as well (for example to avoid too frequent changes of limit value which could confuse customers). To identify limit values which do not increase customer waiting significantly, however, the optimal limit value must be determined. Consequently, it is one of the most important issues in express line systems to determine the optimal limit value – even in those cases when not the optimal value is applied.

In Fig. 3, it can also be seen that applying express checkouts does not necessarily reduce the average waiting time (it can happen that the average waiting time with express lines is shorter than the original one, but in these cases no significant difference can be experienced). In the system in question, where customers buy only few items, the average waiting time cannot be signifi-

Tab. 1. Waiting times using different models

			L=1	L=2	L=3	L=4
Analytical	M/G/k		0.0492	**0.0417**	0.0811	0.1563
	M/G/1		0.3727	**0.2755**	0.3289	0.4803
Simulation	Built-in	RAN	0.3403	**0.2570**	0.3125	0.4614
		SNQ	0.3638	**0.3219**	0.3366	0.3945
	User defined	MWL	0.3631	**0.3218**	0.3358	0.3911
		SNS	0.0688	**0.0561**	0.1002	0.1821
		FLI	0.3636	**0.3218**	0.3356	0.3909
		MPA	0.3632	**0.3218**	0.3359	0.3916
		SST	0.3632	**0.3217**	0.3356	0.3908

cantly smaller than waits in a system without express lines.

Applying express lines can reduce waiting only in systems where customers cannot effectively decrease their own waiting. For example, if customers cannot observe all waiting lines and, therefore, cannot choose the shortest one, or if customers cannot jockey among the queues because of waiting line configurations or big shopping carts.

However, if express lines do not decrease waiting and customers' objective is to minimize their waiting time, then customers buying few items will not necessarily use the express checkouts. In this case, the application of express lines is pointless since the operation of the system converges to the operation of the queuing system without express lines. Practice shows, however, that customers buying few items do use express checkouts; moreover, they generally welcome them. Since express lines are popular among customers, they must have some benefits. These additional effects are related to the *variance of waiting time*, the *perception* and *evaluation of waiting*. Creating two different groups of services – containing shorter and longer ones – decreases the standard deviation of service times. Consequently, the variation of waiting times is also reduced. The more calculable waiting times, with other psychological effects, reduce perceived waiting time – even if actual waiting time does not decrease [12, 18, 24]. In the following sections, these additional effects will be studied in details.

From the data listed in Table 1, it can be seen that the same optimal value of the limit parameter is obtained with different models. With different models, of course, different waiting times are experienced. It is a natural phenomenon and it can also be concluded that the real waiting times are closer to the ones obtained with the simulation models. Simulation generally gives a better estimate of the real operation. The minimal waiting time, however, can be found at the same limit value in each case. That is, the limit value minimizing average waiting time is the same independently of the model used for the analysis.

It has to be mentioned that there are situations when different models determine different optimal limit values. Our additional analyses showed, however, that these limit values are numbers next to each other and, in these cases, there are only slight differences between the waiting times. Therefore, even if analytical models do not provide optimal solution, they do not cause significant waiting time increase.

Finally, it must be emphasized that express line systems have different effects on the waiting in express lines and on the waits in regular lines. Small limit values reduce waiting in the express lines and increase waits in the regular ones. Large limit values have inverse effects. Customers buying larger amounts (and receiving more valuable services) are more tolerant; therefore, applying smaller limit values should be preferred by managers. Accordingly, one of the main benefits of express line systems might be the more appropriate allocation of waiting among customers and not the reduction of the average waiting of all customers [22].

5 Predictability of waiting

Reactions given to waits can be divided into two groups: one related to anger and another connected to uncertainty [34]. When customers have to wait, they, temporarily, do not get the service claimed. That is, waiting is an obstacle to service, blocking satisfaction of a need, and so it causes frustration and anger [19]. The duration of waiting is a stochastic value; therefore, its exact length cannot be determined in advance. In this way, time planning becomes more difficult for people [20]. As they cannot estimate the consequences of their waits, they will experience uncertainty with some of its associated feelings (e.g. uneasiness, anxiety) [22].

The management objective of minimizing waiting time implicitly tries to minimize anger generated by waiting. When focusing only on reducing waiting time and anger, benefits from decreasing uncertainty can be lost easily. Uncertainty can be reduced in several ways: information can be provided about expected waiting time, services can be standardized, etc. In these ways, customers can become sure that their wait will be ended within a certain time.

The application of express checkouts, as other specialization of servers, helps to make the services more standard. They do not necessarily reduce average waiting time but decrease the variation in the length of services and, consequently, reduce the variance of waiting time as well. It can happen that customers in an express line have to wait to the same amount as they would wait in a regular line (and feel the same amount of anger). The satisfaction even in this case will be higher, because customers do not have to worry that their wait will be too long because a customer in front buys huge amount (that is, the experienced uncertainty will be lower). At the same time, they also give up the opportunity of a short waiting duration. This is, however, a rational decision – in the terms of decision theory.

The explanation is that time can be considered as a resource. As a resource, like money, time can be gained and can be lost, that is, can be saved and can be wasted. In this sense, waiting time is a kind of loss. According to the prospect theory, people are risk-seeking in choices involving losses [11]. In decisions about waiting, however, people are risk-averse [20]. Therefore, for satisfying risk-averse customers, minimizing the variance of waiting time should be another important management objective.

Accordingly, the effect of the limit parameter on the variance of waiting time was also studied. Our analyses showed that the minimum of average waiting time and the minimum of the standard deviation of waiting time can be experienced with using the same limit value. This is in accord with the heavy-traffic distribution of waiting times (and express lines are generally applied in busy stores). In M/G/k systems, when utilization tends to unity from below, the waiting time is approximately exponentially distributed [14]. The mean value and the standard deviation of an exponentially distributed stochastic variable are equal [26]. Therefore, the minimum of the mean and the minimum of the standard deviation of waiting time can be experienced with using the same limit value. There are cases, however, in which the minimum of the average waiting time and the minimum of the variance of waiting time can be found at different limit values. These limit values are, however, numbers next to each other, and the moments of waiting time are not significantly different. Consequently, it practically does not make difference which limit value is applied. That is, the optimal limit value can be considered independent of whether the average waiting time or the variance of waits is minimized.

Table 2 lists the standard deviations observable in the different line types using RAN waiting line selection rule for some limit values. (Using other line selection rules similar results can be obtained.) It can be seen, that the standard deviation of waiting time is highly different in the express and in the regular lines. Based on the fact that express line systems are popular among customers, it can be concluded that customers buying few items are interested not only in waiting reduction but in the standardization of services and waits as well. Accordingly, it can be said that customers in express lines are more risk-averse than the average.

Tab. 2. The standard deviation of waiting time in different line types

	L=1	L=2	L=3	L=4
Express lines	0.2022	0.3619	0.5642	0.7730
Regular lines	0.8581	0.7113	0.5950	0.5465
Total	0.7374	**0.5535**	0.5757	0.7444

6 Perception of waiting time

Minimizing waiting time is a general management objective, which is appropriate in production systems. In services, however, where people have to endure waiting, waiting time minimization is not necessarily a proper objective function.

In services, two different waiting times can be distinguished. The first one is the *actual* waiting time which can be measured easily and which is registered by the managers. The second one is the *perceived* waiting time which is experienced and known only by the customers. As people do not percept time linearly, there can be significant differences between the two values.

The human time perception, according to the psychophysical literature, can be approximated by power functions [31]. To express the relationship of perceived and actual waiting times, the double-logarithmic specification suggested by Antonides et al. [2] can be used,

$$\ln \psi_j = \alpha_\mathrm{p} + \beta_\mathrm{p} \cdot \ln t_j + \gamma_\mathrm{p} \cdot X_j + \varepsilon_{\mathrm{p}j} \qquad (1)$$

This equation describes the actual waiting time of customer j (t_j) related to the perceived waiting time (ψ_j) with the double-logarithmic specification according to Stevens' law [31]. X_j stands for the values of waiting-time fillers (for example, music, information about waiting time) which significantly affect perceived waiting time. α_p is a constant adjusting the perceived time received after the transformation to the appropriate level, and reflecting the measure with which customers generally overestimate or underestimate actual waiting time. β_p and γ_p are coefficients for reflecting the effect sizes of the according variables. $\varepsilon_{\mathrm{p}j}$ is the normal error term.

To analyze the management objective of perceived waiting time minimization, some simplifications are used. The differences in time perception across the types of line (multi-server, snake, express lines) are not statistically significant [9]. Accordingly, the same parameter values can be used for each customer groups. The first part of (1) is independent of the actual waiting

time; consequently, it is independent of the limit parameter. It affects the minimal perceived waiting time, but does not influence the optimal limit value. The third part can be omitted as there is no filling during waiting in the store in question (this part can be omitted as well if the same filling is used in all waiting lines). As the average perceived waiting time is minimized, the error term can also be left out because during the calculation these parts (following normal distribution) will eliminate each other. Accordingly, by expressing the perceived waiting time in the terms of actual waiting time and using the simplifications mentioned earlier the objective function will be the following,

$$\min \psi_j = \min t_j^{\beta_p} \tag{2}$$

The actual value of the exponent (β_p), as all other parameters in (1), can be determined with linear regression. The value of β_p must be between 0 and 1 since the marginally decreasing perception of time [4], [5]. Based on studies which analyzed the exponent of subjective duration with different influencing parameters (e.g. practice, sensory modality, group differences, experimental effects), its average value approximates 0.9 [4]. To study the robustness of the optimal limit value to the value of the exponent (since there were no data available to determine the exact value of β_p for the store in question) analyses with different exponent values were performed.

Results obtained using RAN line selection rule are listed in Table 3. (Using other line selection rules similar results can be obtained). Perceived waiting times are determined using different limit values and different β_p values.

Tab. 3. Perceived waiting times with different limit values and exponents

	L=1	L=2	L=3	L=4
β_p=0.1	0.3914	**0.3716**	0.4222	0.5034
β_p=0.2	0.3225	**0.2959**	0.3508	0.4438
β_p=0.3	0.3106	**0.2784**	0.3332	0.4314
β_p=0.4	0.3070	**0.2691**	0.3235	0.4269
β_p=0.5	0.3066	**0.2626**	0.3168	0.4258
β_p=0.6	0.3086	**0.2581**	0.3124	0.4277
β_p=0.7	0.3130	**0.2554**	0.3098	0.4323
β_p=0.8	0.3197	**0.2544**	0.3091	0.4394
β_p=0.9	0.3288	**0.2550**	0.3100	0.4491

From the results listed in Table 3, it can be seen that the value of β_p influences the perceived waiting times but does not have significant impact on the optimal limit value. It can happen, however, that the value of the exponent influences the optimal limit value. In these cases, perceived waiting times determined by the different limit values are only slightly different (fewer than 3 percentages). That is, the optimal limit value can be considered independent of the way of perception of waiting time.

7 Evaluation of waiting

In the relationship between customer waiting and customer satisfaction, perceived waiting time acts as a mediator [27]. The primary objective of management should be the maximization of customer satisfaction. As mentioned before the perception of waiting time does not vary in the different lines. The evaluation of waiting is, however, not independent of the line types. That is, to formulate an objective function which maximizes satisfaction or minimizes dissatisfaction a refinement of the functions used in the earlier sections of this paper is needed.

Evaluations in general, and, consequently, the evaluation of waits, are influenced not only by perception but by expectation as well [1], [25]. Satisfaction, according to The First Law of Science, is determined as the difference of perceived and expected service levels [22]. When the actual waiting time is shorter than expected, customers are satisfied. When the actual waiting time is longer than customers' expectation, they become dissatisfied.

The application of express lines is based on the assumption that customers are more tolerant towards wait while they are waiting for a more valuable service. People buying only few items do not receive a service valuable enough to waiting too long. In this way, objective and perceived waiting time of customers buying few and many items could be equal but the dissatisfaction generated by waiting will not be identical. To model this behaviour, utility functions can be used.

Utility functions are often considered exponential [13]. To model the characteristic of risk-averse customers in the domain of time, a negative exponential function can be applied [20]. Based on the related literature, the following utility function can be applied [17],

$$\mathrm{E}\left[\mathrm{U}\left(W_\mathrm{x}, T_0\right)\right] = \mathrm{E}\left[-C \cdot \exp\left(-r \cdot \left(T_0 - W_\mathrm{x}\right)\right)\right] \tag{3}$$

In Eq. (3) C is a corrective parameter for modifying utility to the suitable level after the transformation. It can be interpreted as the customer's value of time and as the direct effect of the expected total waiting time (T_0). W_x is the actual waiting time – in our case following exponential distribution with a mean value of t_x (where x can be qE, qR, or q), and r is the measure of risk-aversion. There is only one stochastic parameter, the perceived waiting time; therefore, all constants can be moved outside the expectation operator. The mean value of the stochastic part is the following (based on [13]),

$$\mathrm{E}\left[\exp\left(r \cdot W_\mathrm{x}\right)\right] = \frac{1/t_\mathrm{x}}{1/t_\mathrm{x} - r} \tag{4}$$

Using (4) as a simplification of the expected utility model, a mean-variance (or a two-moment) decision model can be used [21, 23]. Accordingly, the expected utility, or satisfaction, will be the following,

$$S = \mathrm{E}\left[\mathrm{U}\left(W_\mathrm{x}, T_0\right)\right] = -C \cdot e^{-r \cdot T_0} \cdot \frac{1/t_\mathrm{x}}{1/t_\mathrm{x} - r} = -C \cdot e^{-r \cdot T_0} \cdot \frac{\mu_\mathrm{x}}{\mu_\mathrm{x} - r} \tag{5}$$

The formula for determining expected utility is valid for customers both in express and in regular lines since (5) is based on

usual characteristics of customers. There are differences, however, in the parameter values. Consequently, the expected utility in the express and in the regular lines can differ from each other more than it follows from the observable distinctions between the waits in the different line types.

Waiting time of customers in express lines is an increasing function of the limit parameter. People waiting at these checkouts are more risk-averse and less tolerable, that is, their expectation of the total waiting time is short. Accordingly, the value of parameter r will be higher; the value of T_0 will be lower. With a larger r value, transformation (5) will make the decreasing utility function of express lines steeper. With a smaller T_0 value, transformation (5) will shift the decreasing utility function of express lines upwards.

Waiting time of customers in regular lines is a decreasing function of the limit parameter. People waiting at these checkouts are less risk-averse and more tolerable, that is, their expectation of the total waiting time is relatively long. Consequently, the value of parameter r will be smaller; the value of T_0 will be larger. With a smaller r value, transformation (5) will make the increasing utility function of regular lines flatter. With a larger T_0 value, transformation (5) will shift the increasing utility function of regular lines downwards.

As the result of these two effects, the minimum point of the total utility function can differ from the minimum point of the actual and perceived waiting time functions, that is, different optimal limit value can be obtained. This value, however, cannot be larger than the formerly determined one. Furthermore, if customers' expectations are not too far from the real waiting times, and they are not extremely risk-averse, the maximum point of the satisfaction function does not differ from the minimum point of the other objective functions. The transformations (shifting and rotation) of the waiting time functions, in these cases, do not refine the optimum of the formerly reviewed objective functions; however, they can be used as justification of applying express checkouts. Based on an objective function which takes into consideration the evaluation of waiting as well, the popularity of express line systems may be explained since, in the term of satisfaction, higher service level may be offered in queuing systems with express checkouts than in systems without them.

8 Conclusion

Applying express checkouts is a widely used management tool for waiting time reduction. Their main parameter, the limit value, must be selected carefully, because introducing express lines with an improper limit value can ruin customers' waiting experience in high degree. Therefore, determining a limit value which minimizes average waiting time is one of the most important tasks of store managers operating express lines.

For determining optimal limit value special tools are required. Our analyses show, however, that simple analytical models provide results accurate enough. Although they give only rough approximation of operation but they are appropriate for determining optimal limit value. In this way, the time, money and knowledge needed for developing and running simulation models can be saved. If, however, managers are interested in a precise estimate of waiting times, then the application of simulation cannot be avoided.

The operation of express line systems can be optimized from several aspects. The classical operations management objective, waiting time minimization, should be complemented by the intention of reducing variation in waits as well. Furthermore the objective of perception management, minimizing perceived waiting time, should be developed into a more sophisticated objective function of utility (satisfaction) maximization.

To determine an exact utility function, the data needed for operations management is not sufficient. A thorough analysis of customer preferences (risk-awareness and tolerable waiting time) is needed. These values, however, can be highly dissimilar in the different customers groups (it can be influenced, for example, by nation, by the type of the store in question or by the applied limit value itself). To get a precise description about the operation of express line systems, more diversified analyses are needed than the ones classically used in operations or perception management.

References

1 **Anderson E W, Sullivan M W**, *The Antecedents and Consequences of Customer Satisfaction for Firms*, Marketing Science **12** (1993), no. 2, 125-143.

2 **Antonides G, Verhoef P C, Van Aalst M**, *Consumer Perception and Evaluation of Waiting Time: A Field Experiment*, Journal of Consumer Psychology **12** (2002), no. 3, 193-202.

3 **De Toni A, Meneghetti A**, *Traditional and Innovative Path Towards Time-based Competition*, International Journal of Production Economics **66** (2000), no. 3, 255-268.

4 **Eisler H**, *Experiment on Subjective Duration 1868-1975: A Collection of Power Function Exponents*, Psychological Bulletin **83** (1976), no. 6, 154-1171.

5 **Fraisse P**, *Perception and Estimation of Time*, Annual Review of Psychology **35** (1984), 1-36.

6 **Hanke J E, Reitsch A G**, *Understanding Business Statistics*, Irwin, Boston, 1991.

7 **Hill A V, Collier D A, Froehle C M, Goodale J C, Metters R D, Verma R**, *Research opportunities in service process design*, Journal of Operations Management **20** (2002), no. 2, 189-202.

8 **Hillier F S, Lieberman G J**, *Introduction to Operation Research*, McGraw-Hill Book Co., 1995.

9 **Hornik J**, *Subjective vs. Objective Time Measures: A Note on the Perception of Time in Consumer Behavior*, Journal of Consumer Research (11 June 1984).

10 **Hui M. K, Tse D K**, *What to Tell Consumers in Waits of Different Lengths: An Integrative Model of Service Evaluation*, Journal of Marketing (60 April 1996), 81-90.

11 **Kahneman D, Tversky A**, *Prospect Theory: An Analysis of Decision under Risk*, Econometrica **47** (1979), no. 2, 263-291.

12 **Katz K L, Larson B M, Larson R C**, *Prescription for the Waiting-in-Line Blues: Entertain, Enlighten, and Engage*, Sloan Management Review **32** (1991), no. 2, 44-55.

13 **Keeney R. L, Raiffa H**, *Decisions with Multiple Objectives*, Cambridge University Press, 2003.

14 **Kimura T**, *Diffusion Approximation for an M/G/m Queue*, Operations Research **31** (1983), no. 2, 304-321.

15 **Kleinrock L**, *Queueing Systems – Volume 1: Theory*, John Wiley & Sons, Inc., 1975.

16 **Koltai T, Kalló N, Lakatos L**, *Optimization of express line performance: numerical examination and management considerations*, Optimization and Engineering, posted on 2008, DOI 10.1007/s11081-008-9053-3, (to appear in print).

17 **Kumar P, Kalwani M. U., Dada M**, *The Impact of Waiting Time Guarantees on Customers' Waiting Experiences*, Marketing Science **16** (1997), no. 4, 295-314.

18 **Larson R C**, *Perspectives on Queues: Social Justice and the Psychology of Queueing*, Operations Management **35** (1987), no. 6, 895-905.

19 **Lawson R**, *Frustration*, The MacMillan Company, New York, 1965.

20 **Leclerc F, Schmitt B. H, Dubé L**, *Waiting Time and Decision Making: Is Time like Money?*, Journal of Consumer Research (22 June 1995), 110-119.

21 **Levy H, Markowitz H M**, *Approximating Expected Utility by a Function of Mean and Variance*, The American Economic Review **69** (1979), no. 3, 308-317.

22 **Maister D H**, *The Psychology of Waiting Lines*, The Service Encounter (Cziepel J A, Solomon M R, SURPRENANT C F, eds.), Lexington Books, Lexington, 1985, pp. 113-123.

23 **Meyer J**, *Two-Moment Decision Models and Expected Utility Maximization*, The American Economic Review **77** (1987), 421-430.

24 **Nie W**, *Waiting: Integrating social justice and psychological perspectives in operations management*, Omega **28** (2000), no. 6, 611-629.

25 **Oliver R L**, *A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions*, Journal of Marketing Research **17** (1980), 460-469.

26 **Pfaffenberger R C, Patterson J H**, *Statistical Methods for Business and Economics*, Irwin, 1987.

27 **Pruyn A, Smidts A**, *Effects of Waiting on the Satisfaction with the Service: Beyond Objective Time Measures*, International Journal of Research in Marketing **15** (1998), 321-334.

28 **Rényi A**, *A Poisson folyamat egy jellemzése (A possible characterization of the Poisson process)*, MTA Mat. Kut. Int. Közl. **1** (1956), 519-527.

29 **Rothkopf H. M, Reich P**, *Perspectives on Queues: Combining Queues is not Always Beneficial*, Operations Research **35** (1987), no. 6, 906-909.

30 **Stalk G jr**, *Time – The next Source of Competitive Advantage*, Harvard Business Review **66** (1988), 41-51.

31 **Stevens S S**, *On the Psychological Law*, The Psychological Review **64** (1957), no. 3, 153-181.

32 **Szántai T**, *On limiting distributions for the sums of random number of random variables concerning the rarefaction of recurrent processes*, Studia Scientiarum Mathematicarum Hungarica **6** (1971), 443-452.

33 _____, *On an invariance problem related to different rarefactions of recurrent processes*, Studia Scientiarum Mathematicarum Hungarica **6** (1971), 453-456.

34 **Taylor S**, *Waiting for Service: The Relationship between Delays and Evaluation of Service*, Journal of Marketing **58** (1994), 56-69.