Robustness of Comparison Sequential Test for the Piloting in Service Delivery

Yefim Haim Michlin
Technion – Israel Institute of Technology
Technion City, Haifa, 32000, Israel

Genady Ya. Grabarnik
Dept. of Math and CS, St. John's University,
Queens, NY
USA

Larisa Shwartz
IBM T.J. Watson Research Center,
Yorktown, NY
USA

Ofer Shaham
Technion – Israel Institute of Technology
Technion City, Haifa, 32000, Israel

Abstract— **This paper discusses a piloting methodology that supports a complicated decision-making process of evaluating an Information Technology process or function relative to the business value it generates. We use sequential comparison testing as a direct tool for performance evaluation of the operational innovation in volatile Service Delivery processes. Sequential tests are exceptionally suitable for finding the best process or function, and also for statistically significant measurement of the process parameters. Since the characteristics of a service delivery process vary widely, on the one hand that evaluation should be comparative, i.e. matched against that of another group working in parallel and serving as the reference. On the other hand, there is a question of the correctness of these measurements under changing parameters, i.e. of the robustness of piloting. We studied robustness of the proposed sequential comparison tests in the context in question and demonstrated that it only depends on the coefficients of variation of the compared processes and not on the other parameters of their distributions. We established that the test is indeed robust in the settings of IT Service delivery environment. We also show that it is superior to the alternative fixed sample size test and pairwise sequential test. This study provided a statistically rigorous background for application of proposed test to pilot based measurements.**

Keywords—direct experimentation; piloting; robustness; sequential comparison; Wald's test, SPRT

I. INTRODUCTION

The variability and volatility of Information Technology (IT) services make it difficult to predict the value of improvement prior to deployment and even when deployed in a production environment. Piloting or direct experimentation is both necessary and possible for introduction of operational innovations into service delivery. Well-designed piloting minimizes the effect of the interpretive lenses through which the service business views the change. Piloting enables a business to see clearly how operational change performs in a production environment and aids in assessment of its potential benefit. Adopting the design-of-experiment approach for that of a pilot can provide a sound foundation for piloting in an IT service delivery environment. However, in manufacturing the traditional methods for measuring the benefit of a change are based on the supposition that the environment is relatively static, which is not the case for IT service providers.

One of the approaches to dealing with the high volatility of measurements in IT service provider's environments is to use the comparison-testing. Applicability of the testing may be checked on a sample metric that identifies the effect of change in the process or staffing of the business. A sequential probability ratio test (SPRT), first described by Wald [1], consists in triggering a decision point at each step of the test when a decision is immediately made – to end the test with a hypothesis accepted; end it with the hypothesis rejected; or continue it for further refinement.

In our prior work we proposed a comparison SPRT (CSPRT) for service processes and provided a number of reasons for considering robustness of the test based on the service requests' processing study. We ascertained the general statement of test robustness under Weibull type distributions: for a medium/large support group we used the theoretical results due to Palm / Khinchin to establish the approximate distribution, and for a small group – a simulation study.

By analyzing incoming request data and request processing times, we concluded that although the latter may be approximated by exponential resolution time, approximation by distributions of more complex nature, like Weibull's, could have an extended set of parameter combinations. This indicated a need to understand how the characteristics of the test will be affected by changes in the parameters of the Weibull distribution and if the distribution of the request time changes from exponential to other types of distribution like say log-normal. The question arises, what is the influence of the fact that the original distribution of resolution times is close in a certain sense, but different from the exponential. The analysis of test's robustness answers the question. The robustness is critical for practical applications of the test.

In addition, piloting requires the testing to conclude fast, and in some cases additionally within a certain amount of time or number of experiments. The requirement for finishing testing fast calls for the sequential test, with the theoretical property of being the shortest on average, us established by Wald [1]. The additional requirement results in the so called truncated test [1].

In this paper we show that this type of pilot is stable or robust for non-exponential distribution for fulfillment times typical of service delivery, and established a methodology for quantitative estimates. We determined the test's parameter estimates for various distributions of resolution times in the streams under comparison. The parameter most significantly affecting the robustness for these event stream distributions is the coefficient of variation. We studied the influence of the

978-3-901882-50-0

compared processes' distribution on the deviation of the test characteristics from those of a reference test. Finally, we evaluated the results for CSPRT against two other tests – the fixed sample size test (FSST) and the pairwise comparison SPRT (PCSPRT), and showed the superiority of CSPRT to both of them. While the CSPRT shows robustness, the PCSPRT behaves in a non-robust manner. While the FSST also possesses robustness, the duration of the test significantly exceeds that of the CSPRT.

The paper is organized as follows. This section describes the motivation and area of application of the paper. Section II provides a background and describes the existing state of the art. Section III introduces the three tests under consideration: the CSPRT, the PCSPRT and the FSST. Section IV focuses on the robustness of the Operating Characteristic (OC) and the Average Sample Number (ASN) of the CSPRT for the most frequently used distributions of the resolution time: Weibull, gamma, lognormal. The section also introduces the coefficient of variation as the main contributor to variation in a test. Section V extends the results of Section III to the FSST and PCSPRT. Conclusions and future research directions are presented in Section VI.

II. Background and Related Work

An experiment in the service delivery environment can be described by a number of steps. A service request arrives at triage ("dispatch") and is distributed among support teams who are working in parallel and using new or reference systems. The outputs from all groups are combined in a merge-type switch. A "merge" switch processes the requests on a first-come-first-serve basis. Once the request is resolved by at least one group, the next request is distributed. For the purpose of this study we assume that the group which received the request could resolve it within a finite time and that the support groups are fully occupied. In these circumstances the system considered is fully utilized and has no idle time.

In [2] we proposed to use the comparison-testing approach as the least affected by the high volatility of IT service provider's environments. We demonstrated the applicability of this novel approach on a sample business via a metric that identifies the effect of change in the process or staffing of the business. In a comparison test the two different processes are considered and the main hypothesis claims a relation between the processes' parameters [1,3,4]. In [2] we also proposed a CSPRT for service processes and provided a number of reasons for considering robustness of the test based on the service requests' processing study.

Here we consider resolution times of incoming service requests. We define the term 'event' as a point in time when a request is resolved. We denote by TBE the time between these events. In order to evaluate the robustness of the comparison binomial SPRT in a service delivery environment, we consider a reference test with characteristics associated with an exponential distribution of the resolution times. We study the influence of the compared processes' distribution on the deviation of the test characteristics from those of the referenced test. Two main characteristics are chosen for evaluation purposes: the OC and the ASN. The OC is the distribution of a parameter value, in this case Φ. The ASN is the expectation of the number of samples needed to conclude the experiment.

The subject of test robustness has been studied for many years, and we provide below a brief description of the state of the art.

The robustness of an SPRT was first approached by Wald in [1], where he characterized the test as having probabilities of the errors in their alpha and beta values not exceeding certain targets. A specific type of robustness was considered by Lorden in [5], where he established asymptotic properties of the test. This paper is concerned more with the effect of the parameters' variation in not making the error probabilities significantly worse.

The next step after Wald was taken by Harter and Moore [6], who ran a computer experiment to conclude that for the Weibull distribution, the SPRT does not make the values of the parameters much worse when the shape factor exceeds one, the exponential distribution with shape factor one being the limiting case.

Robustness of the SPRT under small perturbation of the distribution was considered by Quang [7].

Theoretical treatment of the case was done by Montagne and Singpurwalla [8], with the hazard function (event rate) as the defining characteristic of robustness. Those authors cautioned against the effect of the alpha and beta error probabilities' reduction on the interpretation of the test and its possible practical effects. To establish robustness they used monotonicity of the hazard function. Note that those results are not applicable for the frequently used lognormal distribution that has a non-monotonic hazard function.

In this work we investigate the dependence of robustness on the coefficient of variation C_V. (the ratio of the standard deviation and expectation). This parametrization allows to describe robustness in terms of one value instead of a whole function (like the hazard function), and there is no need to require monotonicity of the hazard function.

Further extension of the result to the more general, so-called life distribution was done by Chaturvedi, Tiwari, and Tomer [9]. There, the expressions for the domain of double-sided robustness (i.e. robustness that controls both significant decrease and increase of the test) become less tractable due to generality of the addressed distributions. By contrast, in our paper we are concerned with domains of robustness where error probabilities do not increase significantly, although we allow those values to drop, and we give simple criteria based on the coefficient of variation for finding the domain of robustness.

More theoretical consideration and characterization of double-sided robustness was done by Gordienko, Novikov, and Zaitseva [10], where a sufficient condition for the double-sided robustness is expressed under a mild additional condition in terms of restricted complete variation. Once again, we are more concerned with one-sided robustness and suggest characterization of the robustness domain in terms of the coefficient of variation.

Robustness of an SPRT under small convex perturbation by noise was considered by Kharin, Kishylau [11] . This type of perturbation, as a rule, leads to distributions outside the parameterized family. By contrast, we are interested in robustness under perturbation of the parameters of the distribution.

On practical grounds, we consider a test to be robust, if it can be concluded sooner or not significantly later than the main test. This definition differs from the mathematical notion of robustness, corresponding to mere continuity of the test's parameters under a small perturbation of the distribution (see for example Gordienko, Novikov, Zaitseva, [10]). Besides, we consider areas of robustness in the distribution's parameter space. We show that the robustness area is well described in terms of the coefficient of variation.

III. Description of the Considered Comparison Tests

The goal of this section is description of the main tests under consideration, and a methodology and formulas for use in either analytical analysis or simulation study of the robustness.

A. Comparison Sequential Probability Ratio Test (CSPRT)

1) Description of CSPRT

The CSPRT procedure was suggested in [12] and developed in papers [13- 15, 2].

The purpose of the test is to verify the hypothesis H_0 about the ratio Φ of Mean TBEs (MTBE) θ for the new (marked ***new***) and the reference (marked ***ref***) systems:

$$\Phi = \theta_{ref} / \theta_{new} \tag{1}$$

$$\begin{aligned} H_0 &: \Phi \geq \Phi_0, \quad P_a(\Phi_0) = 1-\alpha \\ H_1 &: \Phi < \Phi_0, \quad P_a(\Phi_1) = \beta \end{aligned} \tag{2}$$

where $P_a(\Phi)$ – the acceptance probability of H_0 at given Φ (the OC of the test), and

$$\Phi_1 = \Phi_0 / D \tag{3}$$

where D is the discrimination ratio of the test; Φ_0, D, α, β –are fixed.

During the CSPRT two compared systems are tested simultaneously (Fig. 1) [12], [13]. When an event occurs with one of the systems, it immediately goes into the initial state. At this point, the decision is made either to stop the test, or to accept/reject the hypothesis H_0, or to continue the test until the next event.

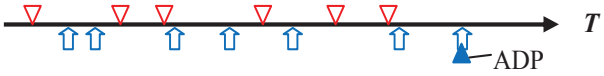

Fig. 1. Scheme of test course in time domain (upward marks – events of the reference system; downward marks – those of the new system; T – time line, common to both systems); ADP - Accept Decision Point and stopping of the test (see Fig. 2).

In [13] it is shown that for the test in question with exponential TBE, the estimate of Φ is time-invariant and changes only at the moment of an event with one of the systems. The probability that the next event will occur with the reference system is [13].

$$P_R(\Phi) = 1/(1+\Phi) \tag{4}$$

This permits presentation of the tests in binomial form and their reduction to the well-known SPRT [1, 16].

Fig. 2 shows the test space presented in discrete coordinates (n, r), new and reference system number of events respectively. The test begins at point (0,0) and with each event in either system, moves one step to the right (reference system) or upward (new system). Probability of an upward step, towards the reject boundary irrespective of the point's coordinates, is given by (4).

The test stops when it leaves the continue zone, bounded by parallel oblique boundaries and by truncation lines parallel to the coordinate axes. H_0 is accepted when the lower and right-hand boundary is crossed at points denoted ADP – Accept Decision Point and is rejected when the upper and left-hand boundary is crossed at RDP – Reject Decision Point.

The boundaries are plotted according to principles outlined in [1]. This methodology was studies by us in detail in earlier works [2, 17, 18].

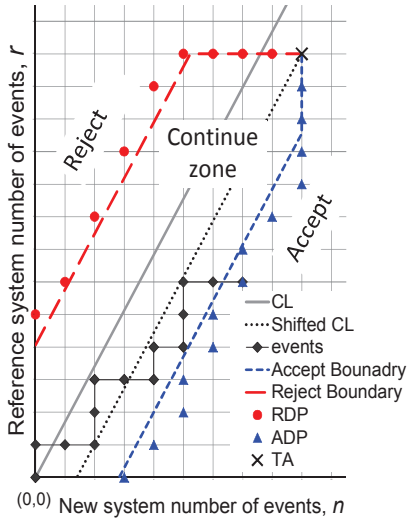

Fig. 2. Truncated test plan and example of test course in events domain. This example corresponds to the example in Figure 1. The test characteristics: α=0.2, β=0.1, D=3.

With the ADP and RDP given, the test characteristics (OC and ASN) are obtainable as follows.

The probability of hitting any point with coordinates (n, r) within the boundaries is obtainable recursively as

$$P(n,r) = P(n,r-1)\cdot P_R + P(n-1,r)\cdot(1-P_R) \tag{5}$$

The probability $P_a(\Phi)$ for given test boundaries is the sum of those of hitting ADP:

$$P_a(\Phi) = \sum_r P_{ADP(r)}(\Phi) \tag{6}$$

where $ADP(r)$ is the n coordinate of the ADP for given r,

$P_{ADP(r)}(\Phi)$ – the probability of hitting the above at given Φ.

Accordingly, α_{real} and β_{real} are calculated, see [15], as

$$\begin{aligned} \alpha_{real} &= 1 - P_a(\Phi_0) \\ \beta_{real} &= P_a(\Phi_1) \end{aligned} \tag{7}$$

The test ASN is calculated [15] as

$$\begin{aligned} ASN(\Phi) = &\sum_n \left[RDP(n) + n \right] \cdot P_{RDP(n)}(\Phi) + \\ &+ \sum_r \left[ADP(r) + r \right] \cdot P_{ADP(r)}(\Phi) \end{aligned} \tag{8}$$

The Average Test Duration (ATD) in terms of θ_{ref} is given by

$$ATD(\Phi) = \theta_{ref} \cdot ASN(\Phi) \big/ (1 + \Phi) \tag{9}$$

Wald proposed a methodology for a non-truncated binomial SPRT and established analytical expressions approximating OC и ASN [1].

We denote by p the probability of a move upward in the state space of the test in Fig. 2. The hypotheses in (2) are given by

$$\begin{aligned} &H_0 : p \le p_0, \quad P_a(p_0) = 1 - \alpha \\ &H_1 : p > p_0, \quad P_a(p_1) = \beta \end{aligned} \tag{10}$$

where

$$p_0 = P_R(\Phi_0); \quad p_1 = p_0 d; \quad d = 2D/(1+D) \tag{11}$$

For example, for D=2 and Φ_0=1, using expressions (4) and (11) we find p_0=0.5, p_1=0.667, d=1.33.

The binomial form for the OC of this test is the function $P_a(p)$ which has the parametric form [1]:

$$P_a(\eta) = \left(A^\eta - 1 \right) \Big/ \left(A^\eta - B^\eta \right) \tag{12}$$

$$p(\eta) = \left[1 - \left(\frac{1 - p_0 d}{1 - p_0} \right)^\eta \right] \Big/ \left[d^\eta - \left(\frac{1 - p_0 d}{1 - p_0} \right)^\eta \right] \tag{13}$$

where η – the construction parameter;

$$A = (1 - \beta_{real}) / \alpha; \quad B = \beta / (1 - \alpha) \tag{14}$$

Similarly, ASN for a non-truncated test has a parametric form which we denote by $ASN_{nonTr}(p)$:

$$ASN_{nonTr}(\eta) = \frac{P_a(\eta) \cdot \ln B + (1 - P_a(\eta)) \cdot \ln A}{p(\eta) \cdot \ln d + (1 - p(\eta)) \ln \left((1 - p_0 d) / (1 - p_0) \right)} \tag{15}$$

where $p(\eta)$ – by (13).

In the case when the distribution of the TBEs is not exponential, the probability of an upward step depends on the path of the test prior to this step (see Fig. 2). The test is then no longer binomial and the above formulations (4)-(5) do not apply, hence the necessity of a robustness check.

2) Methodology of CSPRT robustness estimation

We used the Monte Carlo method to establish robustness of the CSRPT for non-exponential distributions of TBE_{ref} and TBE_{new}. We considered an CSPRT with known Accept и Reject lines, hence with known OC and ASN, for exponential TBEs. We applied the obtained test to the TBEs corresponding to a non-exponential distribution belonging to one of the frequently used families. Note that in the general case the probabilities P_R of a step up depend on the time elapsed since the last steps up and to the right. Hence the test requires that the whole set of TBEs be considered. This was achieved as follows.

Simulation was implemented as shown in Fig. 1. The time intervals between the steps for the reference and new systems TBE_{ref} and TBE_{new} were generated using given distributions. Moving from T=0 along the T-axis, each point representing an event in the reference system (upward marks, Fig. 1) matched an upward step in Fig. 2, and one representing an event in the new system (downward marks) – a step to the right. And so on until the Accept or Reject boundary was crossed. At this juncture the test was stopped and the final point recorded.

The results from a large number of simulation runs yielded the statistical probability and the characteristics (6)-(9) of the test.

Note that the specific matrix capabilities of MATLAB minimized the computation time for a vast number of runs (10^5) due to the parallel simulations. This made for high accuracy and smoothness of graphical representation of our results.

B. Pairwise comparison SPRT (PCSPRT)

1) Description of PCSPRT

We use the PCSPRT to evaluate our results. In this test each pair of an compared systems is tested until an event occurs in one of them. On that event, the test is stopped and, after verifying hypothesis (2), a decision is made on acceptance of H_0 or continuation of the test. If case that test continues, a new pair of systems is chosen.

We chose this test since most of the tests are memoryless when the underlying TBE distributions are exponential. Perturbed tests, or those with non-exponential TBE distributions, are dependent on the past, or for the binomial test on the path to the current state. The PCSPRT, due to the fact that both systems are renewed simultaneously, remains memoryless, which makes for simplified evaluation, dispensing with simulation over all possible paths to the current state.

Special feature of Sample Number (*SN*) concept for PCSPRT. For this test (as well as for the CSPRT) the *SN* represents the total number of events undergone by the compared systems. As in this case each event means renewal of both systems, the number of systems (*IN*) participating in the test is double the *SN*. (*IN*=2SN), and if the TBEs of the

compared elements are exponentially distributed, the test is identical with the CSPRT.

2) Methodology of PCSPRT robustness estimation

When the above TBEs are non-exponential, formula (4) for P_R – probability of an upward step – does not hold. The upward step in question (see Fig. 2) is towards the Reject boundary, hence its probability is also that of the TBE for the reference system being less that for its new counterpart, and in the general case has the form

$$p(\Phi) \equiv P_R(\Phi) = \int_0^\infty f_{ref}(t)\left[1 - F_{new}(t,\Phi)\right]dt \qquad (16)$$

where $f_{ref}(t)$ – probability density of TBE for reference system for

given θ_{ref};

$F_{new}(t,\Phi)$ – cumulative function of TBE for new system for given Φ and θ_{ref}.

(For the proof, see e.g. [19]).

For a non-exponential distributions the test remains binomial, and formulae (5)-(8) remain valid for its characteristics.

Formulae (12)-(16) permit assessment of the influence of the TBEs on the test characteristics, using the following algorithm.

(a) Calculation of the boundary parameters and characteristics of the PCSPRT for exponential TBEs of the compared systems and for given Φ_0, D, α, β. In that case, calculation of $P_R(\Phi_0)$ as per (4) of A and B as per (14); of p_0, p_1 and d as per (11).

(b) Calculation of the parameters and characteristics for non-exponential TBEs, as per (16), (12)-(15).

(c) Comparison the $P_a(\Phi)$ and $ASN_{nonTr}(\Phi)$ obtained under (a) and (b), which will provide an indication of the test's robustness.

We used the algorithm to establish that the PCSPRT is not robust. Details of the analysis can be found in Subsection V.A.

C. Comparison Fixed Sample Size Test (FSST)

1) Description and calculation methodology for its parameters

Mace [20] describes a test for checking the hypotheses (2), which continues up to preset sample size (number of events), namely r and n for the reference and new systems respectively, not necessarily equal. When these sizes have been reached, a decision is taken on acceptance/rejection of the null hypothesis. The test is uniformly most powerful for given sizes, i.e. in a certain sense it is optimal [4, Subsection 1.3]. For this reason, it often serves for comparative assessment of the test efficacy [21, Section 6].

Let us denote by T_{new} and T_{ref} the total working times of the respective systems, up to stopping of the test. When the TBE distribution is exponential, T_{ref} and T_{new} have an χ^2-distribution with $2r$ and $2n$ degrees of freedom respectively, and the $(T_{ref}/2r)/(T_{new}/2n)$ ratio obeys an F-distribution with the same degrees of freedom.

The null hypothesis (2) is accepted when (17) is satisfied, and rejected in the opposite case:

$$F < c \qquad (17)$$

where

$$F = \left(T_{ref}/2r\right)\Big/\left(T_{new}/2n\right) \qquad (18)$$

c – critical value of the test statistic,

$$c = \Phi_0 q_F(\alpha, 2r, 2n) \qquad (19)$$

$q_F(\alpha, 2r, 2n)$ – quantile of F-distribution with $2r$, and $2n$ degrees of freedom at probabilities α.

The necessary n and r are obtainable us per

$$D \cdot q_F(\alpha, 2r, 2n) = q_F(1-\beta, 2r, 2n) \qquad (20)$$

This calculation requires that a ratio be set between n and r, e.g. on the basis of the expected rates of events arrived from the compared systems [17]. If the rates are the close, it is reasonable to set $n=r$.

2) Methodology of FSST robustness evaluation

For $2r \to \infty$, the χ^2-distribution tends to the normal and respectively $T_{ref}/2r$ tends to the normal with expectation 1 and standard deviation $\left(1/\sqrt{r}\right)$

$$\left(T_{new}/2r\right) \sim N\left(1, 1/\sqrt{r}\right) \qquad (21)$$

The χ^2_{2r} -distributed random value can be presented as the sum of $2r$ random values. It is usually accepted that for $2r>30$ ($r>15$), the approach to the normal is sufficiently close.

For an exponential distribution, $C_V=1$. For other distributions, C_V can differ from 1 and accordingly

$$\left(T_{new}/2r\right) \sim N\left(1, C_V/\sqrt{r}\right) \equiv N\left(1, 1/\sqrt{r_{eff}}\right)$$

where r_{eff} – the effective number of events

$$r_{eff} = r/C_V^2 \qquad (22)$$

All the above holds for T_{new} and n_{eff}, hence for $(r>15)\cap(n>15)$ the robustness of the FSST can be evaluated through α_{real} and β_{real} as follows:

- Calculating r, n by (20) for specified α, and β.
- Calculating r_{eff}, n_{eff} by (21)-(22) for specified C_V.
- Calculating α_{real}, and β_{real} by (23) for the r_{eff}, n_{eff}, c found above.

$$\alpha_{real} = F_F\left(c/\Phi_0, 2r_{eff}, 2n_{eff}\right),$$
$$\beta_{real} = 1 - F_F\left(c/\Phi_1, 2r_{eff}, 2n_{eff}\right) \qquad (23)$$

where $F_F\left(c/\Phi_0, 2r_{eff}, 2n_{eff}\right)$ – cumulative function of F-distribution with $2r_{eff}$, and $2n_{eff}$ degrees of freedom.

When C_V<1 for both input event streams, the degrees of freedom in (23) increase in accordance with (22) hence α_{real}, and β_{real} being less their nominal counterparts. In other words, the FSST is robust at C_V≤1 for both streams.

Subsection V.B presents a calculation example illustrating this conclusion.

IV. Results and Discussion. Robustness of the CSPRT for Various Distributions of TBEs

The goal of the section is to evaluate the CSPRT based on simulation of the events with one of the most frequently used life distributions [22-24] like Weibull, gamma, lognormal.

Bear in mind that in the case of a perturbed distribution of the TBE, the test result starts to depend on the time that event spent in the system, i.e. the test is not memoryless. We use the Monte Carlo method to simulate possible TBEs.

Note that for the Weibull and gamma distributions, the hazard function is monotonic. In this case our results are similar to those [8] concerning robustness of non-comparison tests. For the lognormal distribution, the hazard function is not monotonic, and the methods of [8] are not applicable even in the case of non-comparison tests.

We show that C_V is a major external factor of test robustness.

A. Weibull-distributed inputs

Figs 3-4 present a calculation example of the test characteristics (α_{real}, $ASN(\Phi_0)$) for input with Weibull-distributed TBE and different shape factors. The nominal characteristics (i.e. those for exponential TBEs) were as follows:

$$\Phi_0=1,\quad D=1.5,\quad \alpha_{real}=0.10,\quad \beta_{real}=0.10 \qquad (24)$$

In the above figures these values are reached at $WeibShape_{new}$= $WeibShape_{ref}$=1. The behavior of β_{real} is analogous to that of α_{real} in Fig .3.

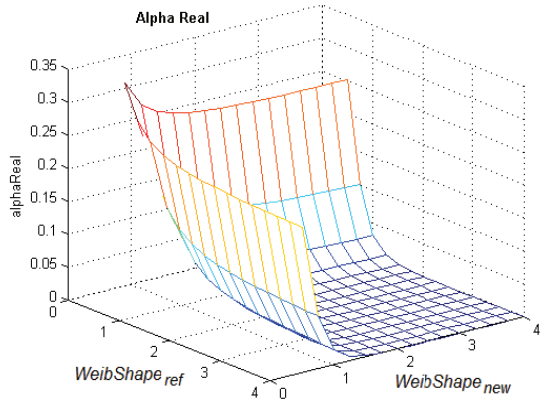

Fig. 3. α_{real} of CSPRT vs. shape factors of Weibull-distributed TBEs of new and reference systems for the test with nominal characteristics (24).

Figs 3 and 4 indicate that deviations of the TBE distributions from the exponential have a strong effect on the test characteristics. At the same time, increase of the shape factor above 1 (the situation of greatest practical interest) makes for a substantially improved OC (higher α_{real}, and β_{real}). A slight reduction below 1 in one of the shape factors, combined with an increase in the other above 1, does not cause deterioration of the OC versus the nominal. The test ASN is reduced by an increase above 1 in both factors and slightly increased by a reduction below 1, while the maximal test duration remains the same. Thus, practicewise, the test is suitable for most applications without the risk of the probability of a wrong decision exceeding the planned.

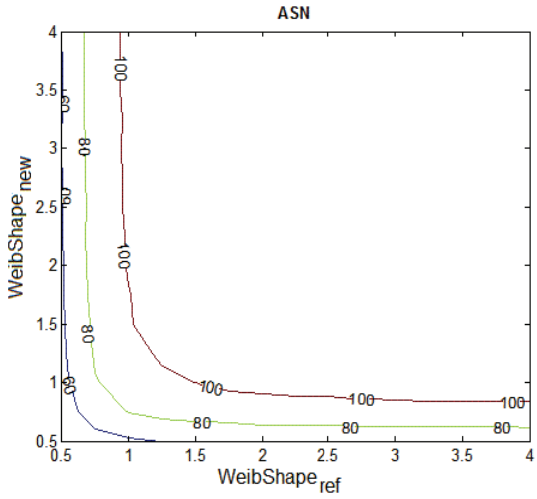

Fig. 4. Contour plot for $ASN(\Phi_0)$ (same conditions as Fig. 3)

B. Lognormal-distributed TBEs

Fig. 5 presents a calculation example for the characteristics of the same test as in Subsection A, but with the TBEs being lognormal-distributed with different shape factors, represented here by the standard deviation. Unlike the Weibull, the lognormal remains distinct from the exponential for all shape factors, and its hazard function is not monotonic. Even so the conclusions regarding the CSPRT robustness coincide completely with those of the Weibull case.

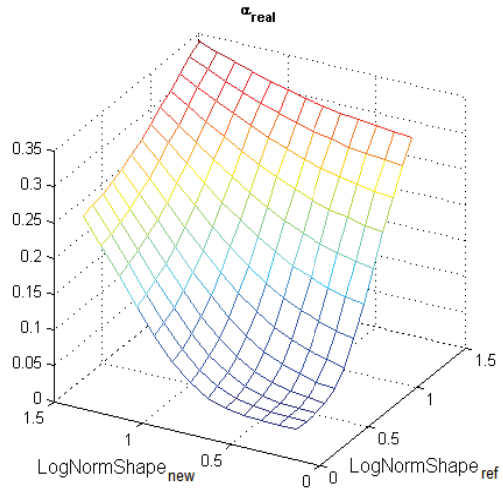

Fig. 5. α_{real} of CSPRT vs. shape factors of lognormal-distributed TBEs of new and reference systems for the test with nominal characteristics (24).

V. Results and Discussion. Robustness of Alternative Tests

In this section we evaluate the robustness of the PCSPRT and FSST. Note that for those tests we do not need to use simulation of the TBE.

Indeed, the PCSPRT is memoryless by virtue of its design, and hence there is no need to simulate TBE. Its state space consists of pairs of integers, and transition probabilities are the same for each pair. This significantly simplifies robustness evaluation of the test.

For the FSST we are able to provide a good approximation and a closed-form solution based on it. We observe that the C_V of the TBE distributions is a major factor of the test's robustness.

A. Pairwise comparison SPRT (PCSPRT)

Fig. 6 shows a calculation example for the characteristics of a PCSPRT (α_{real}) at Weibull-distributed TBEs and different shape factors. The methodology is presented in Subsection III.B.2. The test in question had the nominal characteristics (24). In the figure they are reached at $WeibShape_{new}$= $WeibShape_{ref}$=1.

The above leads to the following conclusion: In the alternative PCSPRT the characteristics exceed the permissible limits under different combinations of the TBE distributions of the compared streams; hence it cannot be rated as robust.

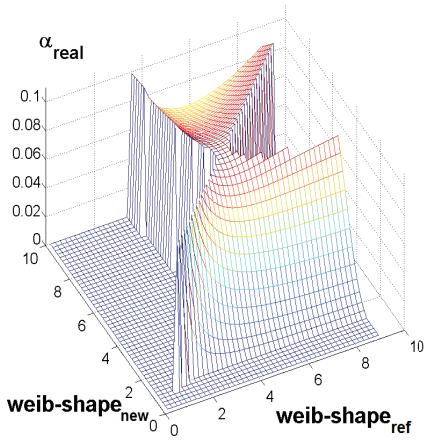

Fig. 6. α_{real} for the range shape factors of the TBE distributons in which the PCSPRT is robust, i.e. $(\alpha_{real} \leqslant \alpha) \cap (\beta_{real} \leqslant \beta)$. Nominal characteristics by (24).

B. Comparison Fixed Sample Size Test (FSST)

The methodology presented in Subsection III.C.2 yielded the parameters for an FSST with characteristics (24).

As per (19)-(20), it was obtained:

$$r=n=81,\ c=0.816 \qquad (25)$$

Fig. 7 presents the results for the relevant α_{real}, β_{real} vs C_V, which is the same for both TBE_{ref} and TBE_{new}. Accordingly, it was found that α_{real}=β_{real} (FSST curve). It is seen that α_{real}, and β_{real} are less than (i.e. superior to) their nominal counterparts at C_V <1; in other words, under these conditions the FSST is robust.

The figure contains also the data for the CSPRT with characteristics (24) and with Weibull-distributed TBEs. This test is described in details in Subsections III.A and IV.A. It is seen that the tests are practically equivalent in terms of robustness, but the ASN of the CSPRT is substantially less than the SN of the FSST (SN=r+n=162). Thus the CSPRT is substantially shorter on the average than the FSST.

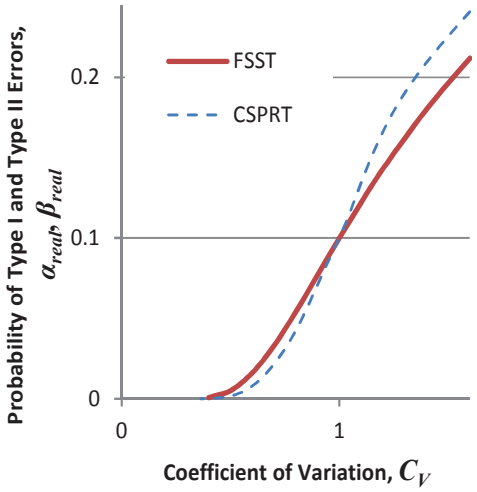

Fig. 7. α_{real}, β_{real} vs. C_V of TBEs of both compared streams.

VI. Conclusion

1. We studied robustness of the proposed Comparison SPRT (CSPRT) and the alternatives under various distributions of the compared TBEs.
2. It was shown that the main change factor in the test characteristics is the coefficients of variation (C_V) of the TBEs. The effect of the C_V on these characteristics is not connected to other parameters of the TBE distributions.
3. The deviations of the TBE distributions from the exponential strongly affect the test characteristics.
4. For the proposed CSPRT, reduction of the TBEs' C_V to less than 1 makes for drastic improvement in its OC (reduced α_{real}, β_{real}).

   Under a slight increase above 1 in the C_V of one of TBEs and reduction in the other below 1, the test OC does not deteriorate versus the nominal.

   The ASN of the proposed CSPRT decreases with increase of the C_V's above 1 and slightly increases with their reduction below 1, but the maximal test sample number remains the same.

   Thus, practicewise, there is no risk of the wrong-decision probability exceeding the planned one. In other words, the CSPRT can be rated as robust in a one-sided sense.
5. In the alternative Pairwise Comparison SPRT (PCSPRT) the characteristics exceed the permissible limits under

different combinations of the TBEs' distributions; hence it cannot be rated as robust.

6. The Comparison FSST manifests robustness like the CSPRT, but its sample number is substantially larger than the ASN of the latter.

ACKNOWLEDGMENT

The authors are indebted to Mr. E. Goldberg for editorial assistance. The research project was supported in part by the Israel Ministry of Absorption and the Planning and Budgeting Committee of the Israel Council for Higher Education.

REFERENCES

[1] A. Wald, *Sequential Analysis*. NY: John Wiley & Sons, 1947, pp. 221.

[2] G. Ya. Grabarnik, Y. H. Michlin, L Shwartz, Designing pilot for operational innovation in IT service delivery, (NOMS), 2012 IEEE, 1343-1351

[3] H. Chernoff, Sequential analysis and optimal design. Vol. 8. *SIAM*,1975

[4] B. K. Ghosh, "Brief History of Sequential Analysis," in *Handbook of Sequential Analysis*, B. K. Ghosh and P. K. Sen, Eds. NY: Marcel Dekker, 1991, pp. 1-19.

[5] G. Lorden, "Structure of sequential tests minimizing an expected sample size." *Probability Theory and Related Fields*, 51, no. 3, 1980, pp. 291-302.

[6] L. Harter, and A. H. Moore, "An Evaluation of Exponential and Weibull Test Plans," *IEEE Transactions on Reliability*, Vol. R-25, No. 2, 1976, 100-104.

[7] Quang, Pham Xuan. "Robust sequential testing." *The Annals of Statistics*, 1985, 638-649.

[8] E. R. Montagne, and N. D. Singpurwalla. "Robustness of sequential exponential life-testing procedures." *Journal of the American Statistical Association* 80, no. 391, 1985: 715-719.

[9] A. Chaturvedi, N. Tiwari, and S. K. Tomer, "Robustness of the sequential testing procedures for the generalized life distributions," *Brazilian Journal of Probability and Statistics,* 16, 2002: 7-24.

[10] E. Gordienko, A. Novikov, and E. Zaitseva, Stability estimating in optimal sequential hypotheses testing, *Kybernetika*, 45(2), 2009, pp. 331-344.

[11] A. Kharin, and Kishylau, D. Robust sequential testing of hypotheses on discrete probability distributions. *Austrian journal of statistics,* V. 34, No 2, 2005, 153–162.

[12] Y. H. Michlin, and R. Migdali, "Test duration in choice of helicopter maintenance policy." *Reliability Engineering & System Safety*, 86.3, 2004: 317-321.

[13] Y. H. Michlin and G. Ya. Grabarnik, "Sequential testing for comparison of the mean time between failures for two systems," *IEEE Trans. Reliab*., vol. 56, no. 2, pp. 321-331, 2007.

[14] Y. H. Michlin, G. Ya. Grabarnik, and E. Leshchenko, "Comparison of the mean time between failures for two systems under short tests," *IEEE Trans. Reliab*., vol. 58, no. 4, pp. 589-596, 2009.

[15] Y. H. Michlin and G. Ya. Grabarnik, "Search boundaries of truncated discrete sequential test," J. Appl. Stat., vol. 37, no. 5, pp.707-724, 2010.

[16] Y. H. Michlin, and O. Shaham, "Planning of truncated sequential binomial test via relative efficiency," *Quality and Reliab. Engineering* Int. (In press) DOI:10.1002/qre.1387

[17] Y. H. Michlin, D. Ingman, and Y. Dayan, "Sequential test for arbitrary ratio of mean times between failures," *Int. J. of Operations Research and Information Systems*, vol. 2, no. 1, pp. 66-81, 2011.

[18] Y. H. Michlin, D. Ingman, and V. Kaplunov, "Sequential testing for two exponential distributions at arbitrary risks," *Int. J. of Quality and Reliability Management*. 2012, Vol.29 No 4, pp. 451-468

[19] K.C. Kapur and L.R. Lamberson, *Reliability in Engineering Design*. NY: Wiley, 1977, pp. 342-363

[20] A. E. Mace, *Sample Size Determination*. NY: Robert E. Krieger Pub. Co., 1974, pp. 226.

[21] B. Eisenberg and B. K. Ghosh, "The sequential probability ratio test," in *Handbook of Sequential Analysis*, B. K. Ghosh and P. K. Sen, Eds. NY: Marcel Dekker, 1991. pp. 47-66.

[22] N. Gans, G. Koole, and A. Mandelbaum. "Telephone call centers: Tutorial, review, and research prospects." *Manufacturing & Service Operations Management* 5, no. 2 (2003): 79-141

[23] G. Koole, "*Optimization of business processes: An introduction to applied stochastic modeling*." Lecture notes (2009).

[24] M. Hollander and F. Proschan, "Testing whether new is better than used," *The Annals of Mathematical Statistics*, vol. 43, no. 4, pp. 1136-1146, 1972.