International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

E.840

(06/2018)

SERIES E: OVERALL NETWORK OPERATION, TELEPHONE SERVICE, SERVICE OPERATION AND HUMAN FACTORS

Quality of telecommunication services: concepts, models, objectives and dependability planning – Models for telecommunication services

Statistical framework for end-to-end network performance benchmark scoring and ranking

Recommendation ITU-T E.840

International Telecommunication Union

ITU-T E-SERIES RECOMMENDATIONS

OVERALL NETWORK OPERATION, TELEPHONE SERVICE, SERVICE OPERATION AND HUMAN FACTORS

INTERNATIONAL OPERATION	
Definitions	E.100–E.103
General provisions concerning Administrations	E.104–E.119
General provisions concerning users	E.120–E.139
Operation of international telephone services	E.140–E.159
Numbering plan of the international telephone service	E.160–E.169
International routing plan	E.170–E.179
Tones in national signalling systems	E.180–E.189
Numbering plan of the international telephone service	E.190–E.199
Maritime mobile service and public land mobile service	E.200–E.229
OPERATIONAL PROVISIONS RELATING TO CHARGING AND ACCOUNTING IN THE INTERNATIONAL TELEPHONE SERVICE	
Charging in the international telephone service	E.230–E.249
Measuring and recording call durations for accounting purposes	E.260–E.269
UTILIZATION OF THE INTERNATIONAL TELEPHONE NETWORK FOR NON-TELEPHONY APPLICATIONS	
General	E.300–E.319
Phototelegraphy	E.320–E.329
ISDN PROVISIONS CONCERNING USERS	E.330–E.349
INTERNATIONAL ROUTING PLAN	E.350–E.399
NETWORK MANAGEMENT	
International service statistics	E.400–E.404
International network management	E.405–E.419
Checking the quality of the international telephone service	E.420–E.489
TRAFFIC ENGINEERING	
Measurement and recording of traffic	E.490–E.505
Forecasting of traffic	E.506–E.509
Determination of the number of circuits in manual operation	E.510–E.519
Determination of the number of circuits in automatic and semi-automatic operation	E.520–E.539
Grade of service	E.540–E.599
Definitions	E.600–E.649
Traffic engineering for IP-networks	E.650–E.699
ISDN traffic engineering	E.700–E.749
Mobile network traffic engineering	E.750–E.799
QUALITY OF TELECOMMUNICATION SERVICES: CONCEPTS, MODELS, OBJECTIVES AND DEPENDABILITY PLANNING	
Terms and definitions related to the quality of telecommunication services	E.800–E.809
Models for telecommunication services	**E.810–E.844**
Objectives for quality of service and related concepts of telecommunication services	E.845–E.859
Use of quality of service objectives for planning of telecommunication networks	E.860–E.879
Field data collection and evaluation on the performance of equipment, networks and services	E.880–E.899
OTHER	E.900–E.999
INTERNATIONAL OPERATION	
Numbering plan of the international telephone service	E.1100–E.1199
NETWORK MANAGEMENT	
International network management	E.4100–E.4199

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T E.840

Statistical framework for end-to-end network performance benchmark scoring and ranking

Summary

Recommendation ITU-T E.840 is the first in a series covering benchmarking of end-to-end network performance. Recommendation ITU-T E.840 presents a framework for the statistical analysis underlying performance benchmarking of networks and services. The framework describes benchmarking scenarios, use cases, as well as procedures and statistical techniques for ranking end-to-end key performance indicators (KPIs) or key quality indicators (KQIs). Recommendation ITU-T E.840 refers to mobile services and benchmarking campaigns performed using mobile agents (devices) in drive or walk tests, as well as fixed agents or devices placed at fixed locations (e.g., within shopping malls, office buildings or stadia).

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T E.840	2018-06-13	12	11.1002/1000/13621

Keywords

End-to-end performance, network performance benchmarking and ranking, statistical framework.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T E.840

Statistical framework for end-to-end network performance benchmark scoring and ranking

1 Scope

The Recommendation specifies a statistical framework, as well as benchmarking scenarios and conditions within which it can be applied, whose use is required by operators and regulators when qualifying and quantifying performance differences between end-to-end key performance indicators (KPIs) or key quality indicators (KQIs) affecting the user experience.

The need for this Recommendation arises because, in the intense race to satisfy increasingly demanding existent users while expanding customer bases at optimal cost, operators have improved network performance to such an extent that differences between them have become smaller and smaller.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T E.800] Recommendation ITU-T E.800 (2008), *Definitions of terms related to quality of service*.

[ITU-T E.804] Recommendation ITU-T E.804 (2014), *KPI aspects for popular services in mobile networks*.

3 Definitions

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

KPI Key Performance Indicator

KQI Key Quality Indicator

QoE Quality of Experience

MOS Mean Opinion Score

RF Radio Frequency

TCP Transmission Control Protocol

5 Conventions

5.2.1 StatScore: This represents statistical score, i.e., the relative overall quality across various networks or operators against the best performing network. StatScore is calculated per service.

5.2.2 GlobalNetScore: This represents global network score, i.e., the relative overall quality across various networks or operators against the best performing network. GlobalNetScore is calculated across all services.

5.2.3 StatDiff: This represents the statistically significant difference between two compared key performance indicators (KPIs) or key quality indicators (KQIs).

5.2.4 THrelv: This represents the minimum difference between values of two specific key performance indicators (KPIs) or key quality indicators (KQIs), which is relevant to the user of a service and above which statistical significance takes precedence.

6 Benchmarking scenarios

Network benchmarking generally has two main use cases: internal and competitive. Internal benchmarking is focused on continuous cost-efficient network performance assurance and improvements requiring evaluation on initial roll-out of a network as well as during its development, as well as new service and new device launches. Internal benchmarking is also performed in well-established and mature networks. In addition, regions with highways and cities, as well as areas of interest (e.g., workplaces, shopping malls, stadia and residential premises) require consideration during evaluation campaigns. Competitive benchmarking performed by operators themselves (or service companies on behalf of operators), as well as by regulators for checking out competition and self-ranking is generally used across regions, transport routes (highways, railways) and cities, even countries in the case of multinational operator groups and for mature networks.

A summary of these use cases, recommended types of tools and techniques are presented in Figure 1. Areas (such as shopping malls, stadia and workplaces, generally indoors) are often benchmarked using walk tests. Besides traditional route-based drive or walk testing campaigns, internal, and to some extent competitive benchmarking, as well as indoor scenarios benefit from fixed probe-based tools. The latter having the advantages of fast and remote scalability and of device independence. Therefore, these tools are very suitable for indoor test scenarios and new services launched in areas of interest, and cities to some extent. In addition, it can be seen that either *a posteriori* or *a-priori* analysis techniques can be applied. In the first case, mostly used in benchmarking, data are collected and statistical significance is used to evaluate and rank end-to-end KPI or KQI performance; the measurement accuracy is, by default, embedded in the statistical significance level. *A-priori* techniques involve advance calculation of the number of test probes needed for a specified statistical significance and measurement accuracy. This technique is generally used when test probes are costly or testing time is limited.

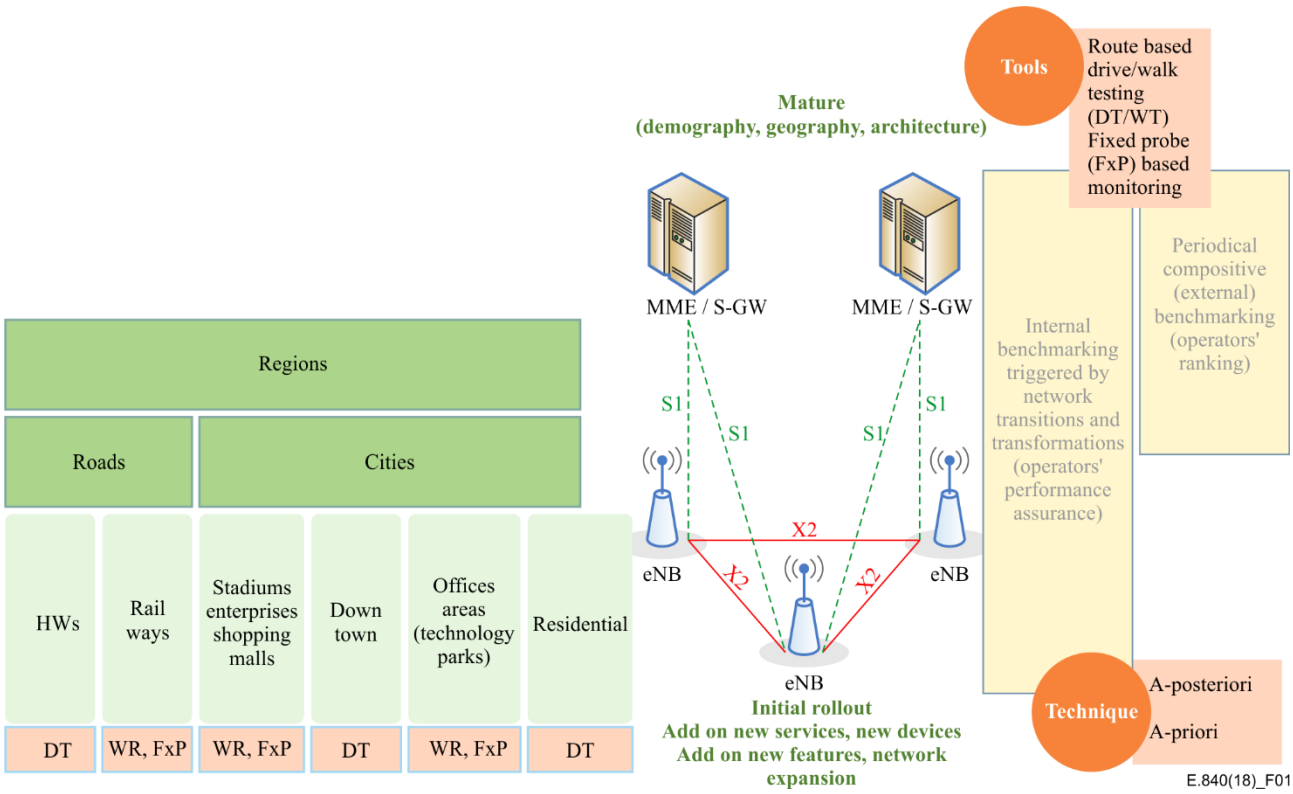

Figure 1 – Benchmarking use cases, recommended types of tool and technique

7 Benchmarking conditions

Regardless of the use case, the benchmarking framework needs to rely on a set of prerequisites that should ensure consistency, validity, reliability and repeatability. Table 1 presents these prerequisites for each benchmarking phase: equipment set-up, test configuration, data collection, data processing and analysis. It should be noted that Table 1 refers to the minimum required prerequisites in order to ensure a fully controlled test environment, as well as a valid statistical analysis. Specific details of measurements are given in other ITU-T Recommendations (e.g., [ITU-T E.804]).

Table 1 – Guidance on minimum required prerequisites

Benchmarking	Benchmarking prerequisites
Equipment set-up	Equipment set-up to be consistent across networks, platforms and devices; same device models to be used for *competitive comparative* ("like for like") benchmarking.
	Equipment to run in proper conditions, as described by the test equipment vendor (e.g., overheating avoidance since it can negatively impact device performance.
Test set-up and configuration	Test set-up to represent real user experience; guidance can be found in [b-ETSI TR 102 581].
	Test set-up settings to avoid artefacts that can artificially impact network performance: e.g., data server location and set-up with transmission control protocol (TCP) parameters should be verified to ensure good throughput for all compared operators. This test should be performed before proceeding with data collection.

Table 1 – Guidance on minimum required prerequisites

Benchmarking	Benchmarking prerequisites
	Test set-up scripting of various scenarios reflecting user behaviour (e.g., use of TCP, various file, video or voice call lengths) while minimizing intrusiveness for not artificially overloading the network.
Data collection	Collect measurement data to reflect user experience [e.g., mean opinion score (MOS) per service as well as the main network end-to-end KPIs or KQIs impacting it]. The measurements should be based on adequate metrics or measurement techniques according to guidelines provided by vendors or in the relevant Recommendations.
	Collect measurement data under various geographical or demographic conditions at different time windows (rush or non-rush hour, weekend or weekday, vacation or non-vacation season).
Data processing and analysis	Compare data collected with the same devices within the same area and during the same time window – "like for like" comparison.
	Use statistical significance to run meaningful comparisons.
	Analysis to be performed per KPI or KQI

.8 Benchmarked services

The list of traditionally benchmarked mobile services and their KPIs or KQIs along with their triggering points lie outside the scope of this Recommendation. See [ITU-T E.800] and [ITU-T E.804] for more details.

If the scope of mobile benchmarking is to perform a detailed comparative analysis per service, a task generally undertaken during the internal benchmarking use case (e.g., scenarios such as new device, new technology add-on), then it is recommended that a comprehensive set of end-to-end KPIs and KQIs be used for analysis (here, KQIs are measurements obtained by using quality-estimation models, such as [b-ITU-T P.863] for voice or [b-ITU-T P.1203] for video streaming). In addition, it is recommended that the main root causes of possible poor performance be analysed based on this set.

On the other hand, if end-to-end KPI or KQI performance ranking is the goal of mobile benchmarking, a task generally undertaken during the comparative benchmarking use case, as well as in some internal benchmarking scenarios (such as market comparisons, periodic market performance evaluation), then scoring and ranking can consider a smaller set of KPIs or KQIs impacting quality of experience (QoE; see [b-ITU-T P.10/G.100]) per service and across all benchmarked services.

This Recommendation refers to the latter case of competitive benchmarking. Other ITU-T Recommendations in the benchmarking series cover details of sets of such KPIs or KQIs.

9 Statistical framework

The recommended framework aims to score and rank network end-to-end performance from a user perspective and it can be used for both competitive and internal benchmarking. The framework defines procedures for data validation, statistical evaluation metrics and significance testing, as well as general guidelines for ranking and scoring.

9.1 Data cleansing

To ensure that benchmarking results are meaningful and accurate, a validation of the data used as input for the analysis is required. The validation mainly consists of data cleansing, meaning that any

measurement-specific artefacts and incomplete data are removed. It is recommended that missing data be replaced with newly collected data. If 5% or more of the data contain artefacts or are missing, then it is recommended that new data be collected. Measurement-specific artefacts include, but are not limited to: silent calls in the case of voice or video conversational services, periods of data or video server unavailability (e.g., server down) in the case of video streaming services, consistently unexpected very low MOS values for voice, video conversational or streaming. Here, any type of degradation that is not caused by the network itself, but rather by either the test equipment or the test device, should be considered as a measurement-specific artefact.

Benchmarking data need to be filtered from these artefacts, in order to ensure the validity of the assumptions made when applying the statistical models.

9.2 Measurement statistical distribution

Generally, the statistical distribution of the values of any measured KPI or KQI can be approximated by a Gaussian distribution, based on the central limit theorem [b-Shaum]; typically, the larger the number of samples, the better the accuracy of the approximation to the Gaussian distribution becomes. It is recommended that the statistical distributions of all KPI or KQI values be verified. This can be performed in two ways. One possibility is to generate distribution charts of the measurements for the KPI or KQI analysed and verify its normality by observation. Another possibility is to use goodness-of-fit tests for normality verification such as Kolmogorov-Smirnov, Anderson-Darling or Shapiro-Wilk [b-Mehta]. In addition, in the rare or extreme case of non-Gaussian distributions, non-parametric tests can be applied. Tests showing experimental distributions of various KPIs or KQIs, tests for verification of distribution normality, as well as special cases of non-Gaussian distributions, are discussed in other ITU-T Recommendations in the benchmarking series.

9.3 Statistical performance metrics, standard errors and statistical significance of the benchmarking results

9.3.1 Statistical performance metrics

Benchmarking analysis should be based on statistical performance metrics that reflect the average network performance (represented by mean values, m) or its consistency (represented by the probability Pth to be above a pre-defined threshold value). This Recommendation refers to the mean statistical performance metric, as an example. Similar techniques can be applied for consistency.

9.3.2 Standard error

The standard error at the 95% confidence level for the mean or Pth is calculated assuming a gaussian distribution of the KPIs or KQIs measured (see clause 9.2).

Therefore, depending on the type of KPI or KQI, continuous scores such as radio frequency (RF) parameters or MOS or discrete ones such as success to failure ratio (r), the standard error at the 95% confidence level is given by:

$$StdError(m)=z\_95\% * std/sqrt(N)= 1.96*std/sqrt(N)$$

$$StdError(r)= z\_95\% * sqrt (r*(1-r)/N) = 1.96 * sqrt (r*(1-r)/N)$$

If fewer than 30 samples are available, then the gaussian quantile z_95% should be replaced by the Student t_95% (N-1), tabulated value, where N represents the number of available samples.

NOTE – Standard errors represent the measurement accuracy. Therefore, if a specific accuracy is required and an estimate for the standard deviation is known, then the minimum required number of samples to meet that accuracy with a selected confidence level can be determined, based on the equations in paragraph 2. This can be used for an *a-priori* technique as shown in Figure 1 and is also used in [b-ITU-T E.802] for the calculation of the minimum number of samples.

9.3.3 Statistical significance

Confidence intervals describe the standard error margins for a given statistical confidence level, generally 95%. However, two compared KPIs or KQIs with close values and overlapping confidence intervals are not necessarily statistically the same. An accurate comparison between KPIs or KQIs should be based on statistical significance. This ensures that the chance of falsely rejecting the hypothesis that two KPI or KQI values are the same, when they are actually the same, remains at 5%.

Based on the statistical significance test, it can be concluded that one operator performs better than another (in the competitive benchmarking use case), as well as determining whether a new technology or feature brings significant improvement (in the internal benchmarking use case).

In addition to statistical significance, KPI- or KQI-specific relevant difference-thresholds must be used, whenever the differences are irrelevant or within the measurement accuracy of each KPI or KQI. KPI- or KQI-specific relevant difference-thresholds (THrelv) are defined in other ITU-T Recommendations in the benchmarking series.

An example of how KPI or KQI comparisons must be performed is presented in Table 2. To maintain the generality of this Recommendation, KQI1 and KQI2 are used as examples of metrics to be compared. Therefore, KQI1 and KQI2 can be any of the KQIs chosen for a particular service (KPIs can also be used). The mean and standard deviation values are calculated for KQI1 and KQI2. Using the number of available test samples and applying hypothesis test Equations A-1 and A-2, shows that the quality delivered by network 1 and 2 for KQI1 and KQI2 metrics is statistically the same. In addition, it can be seen that the differences for each of the KQIs, KQI1 and KQI2, remains below THrelv. The performance of network 1 and 2 is statistically different in the case of the third metric, KQI3. However the difference (0.02) remains below THrelv (0.025) and therefore it must not be concluded that the performances of network 1 and 2 are different.

Table 2 – Statistical significance example

Service	KQI	Network 1			Network 2			Statistics @95% CL (Z>1.96)	StatDiff	THrelv
		Mean	std	N	Mean	std	N			
Service 1	KQI1	3.27	0.3	287	3.35	0.6	212	1.78	No	0.09
	KQI2	0.02	0.14	12	0.015	0.12	10	0.09	No	0.006
	KQI3	0.93	0.26	69	0.91	0.29	71	2.04	Yes	0.025

This kind of analysis applied to detailed benchmarking results may be extended to various services as well as a larger set of KPIs or KQIs per service, as mentioned in clause 8. In addition, based on statistical significance results (Z statistics @ 95%CL, Table 2), individual KPIs or KQIs may be ranked across networks as described in clause 9.4.

It must be noted that claiming that one network or service configuration can be considered as "better" than another requires – besides statistical significance – a KPI- or KQI-specific relevant difference-threshold definition and measurement accuracy information, as defined in other ITU-T Recommendations in the benchmarking series.

9.3.4 Results reporting

The benchmarking statistical analysis and results must be reported along with the detailed description of the test scenarios and conditions used for benchmarking; otherwise the interpretation of results can be wrong and consequently meaningless.

9.4 End-to-end KPI or KQI scoring and ranking

Network end-to-end KPIs or KQIs can be scored and ranked across each of the areas considered in the benchmarking campaign. In addition, as already mentioned in clause 9.3.3, it needs to rely on statistical significance in order to be able to discriminate fine performance differences that tend to be

more and more frequent between current operator networks. Comparisons against KPI- or KQI-specific relevant difference-thresholds must be used if differences are irrelevant or within measurement accuracy as described in other ITU-T Recommendations in the benchmarking series.

The statistical scoring and ranking methodology for a set of KPIs or KQIs is presented in Annex B and an example is shown in Table 3.

Table 3 – Example of statistically significant end-to-end KPI or KQI scoring and ranking

	Network 1				Network 2				
	KPI/KQI	**std**	**N**	**StatDiff**	**KPI/KQI**	**std**	**N**	**StatDiff**	**THrevl**
KPI1/KQI1	0.95	0.22	87	0.05	0.97	0.17	69	0.00	0.018
KPI2/KQI2	0.93	0.26	87	0.00	0.91	0.29	69	0.23	0.019
KPI3/KQI3	3.89	0.50	2600	0.00	3.56	0.70	2070	17.15	0.31
KPI4/KQI4	105.00	5.00	435	42.67	70.00	15.00	350	0.00	34
KPI5/KQI5	1 200.00	300.00	87	0.00	1 800.00	275.00	69	12.31	596

Table 3 shows the aggregated values across the same area during the same time window, for "a like to like" comparison (see guidance in Table 1). Along with the performance values, the standard deviations are calculated and the number of test samples is shown. The StatDiff (as described in Annex B) is calculated for each KPI or KQI as a statistically significant difference against the best performing KPI or KQI (highlighted in yellow in Table 3) at the 95% confidence level (see Annex B). The lower the StatDiff, the closer the value to the best performing KPI or KQI; StatDiff = 0 represents the best performing KPI or KQI. In addition, it can be noted that in all cases, the differences between the KQIs coming from the two networks are higher than THrevl, meaning that the statistical significance analysis takes precedence.

Based on this analysis, the KPI or KQI ranking is shown in Table 4.

Table 4 – Example of statistical ranking of KPIs or KQIs

KPI/KQI	Network 1	Network 2
KPI1/KQI1	Rank 2	Rank 1
KPI2/KQI2	Rank 1	Rank 2
KPI3/KQI3	Rank 1	Rank 2
KPI4/KQI4	Rank 2	Rank 1
KPI5/KQI5	Rank 1	Rank 2

This ranking may be extended to a larger set of KPIs or KQIs, as can be considered for detailed benchmarking as well as for the internal benchmarking use case.

In some benchmarking cases, a network statistical score per service and over all services can be required. Annex A presents a possible technique of how this can be performed. However, it should be noted that this kind of technique has validity only if fully described and based on technically backed-up assumptions and conditions.

Annex A

Statistical significance to be applied in mobile networks benchmarking analysis

(This annex forms an integral part of this Recommendation.)

Benchmarking analysis refers to the comparison between KPIs or KQIs describing the performance of various operator networks. Meaningful comparison should rely on statistical significance tests (hypothesis tests) that depend on the types of KPIs or KQIs compared, continuous (e.g., MOS, RF parameters) and ratios (e.g., completion or failure ratios).

In the first case, Equation A-1 determines significant difference [ITU-T P.1401]:

$$Z = StatDiff/sqrt(std1^2/N1 + std2^2/N2) > Zth \quad (A\text{-}1)$$

where StatDiff denotes the difference between compared metrics, std1 and std2 their standard deviations and N1 and N2 the total numbers of samples used in the comparison for each metric. In other words, if Z is higher than Zth (based on a gaussian distribution for more than 30 samples, at CL% confidence level), then StatDiff is a statistically significant difference at CL% confidence level.

In the second case, the KPI or KQI ratio type is described by p number of successes or failures out of the total number of samples. The significant difference is given by Equation A-2 [b-ITU-T P.1401]

$$Z = StatDiff/sqrt(p1*(1-p1)/N1 + p2*(1-p2)/N2) > Zth \quad (A\text{-}2)$$

where p1 and p2 represent the numbers of successes or failures of each of the compared metrics.

Table A.1 shows the mapping between the significance thresholds Zth at different levels of confidence.

Table A.1 – Mapping between significance thresholds at different levels of confidence

CL%	90	95	96	97	98	99
Zth	1.64	1.96	2.05	2.17	2.33	2.58

If fewer than 30 samples are available, then a t-Student distribution should be used, in which t-Student (n) where n = N-1 is the number of degrees of freedom, with a total number N of test samples.

It should be noted that, along with statistical significance, KPI- or KQI-specific relevant difference-thresholds must be used, whenever the differences are irrelevant or possibly within each KPI or KQI accuracy of measurement. KPI- or KQI-specific relevant difference-thresholds (THrelv) are defined in other ITU-T Recommendations in the benchmarking series.

Annex B

Statistical scoring and ranking of the performance of a network

(This annex forms an integral part of this Recommendation.)

This annex describes the algorithm used to score and rank end-to-end performance of networks used in the calculations in Table 3.

– Calculate end-to-end KPIs (KQIs can also be used) for the service analysed for each network or operator:
  - KPI_1...KPI_i…KPI_N, i=1,n which can be either average or median or proportions (ratios).

– Build a benchmarking matrix for j=1,M networks (operators) per service described by N KPI or KQI metrics – see Table B.1.

Table B.1 – Benchmarking matrix

	Netwk_1……	**Netwk_j…….**	**Netwk_M**
KPI_1…	KPI_1,1	KPI_1,j	KPI_1,M
KPI_i..	KPI_i,1	KPI_i,j	KPI_i,M
KPI_N	KPI_N,1	KPI_N,j	KPI_N,M

– Calculate the statistical significance distance for each KPI_i,j in the matrix.

  For i=1,N
  - select the best value KPI_i,best, with "best" being the network j with the best network
  - calculate the statistical significance difference StatDiff_i,j per KPI_i,j against the best value KPI_i,best based on Equations B-1 and B-2 depending on the type of metric

  StatDiff_i,j = max {0,(KPI_i,best-KPI_i,j)/sqrt (std1^2/N1 +std2^2/N2) – Zth} (B-1)

  StatDiff_i,j = max {0,(KPI_i,best-KPI_i,j)/sqrt (p1*(1-p1)/N1 + p2*(1-p2)/N2) – Zth} (B-2)

NOTE – Zth is F(0.05, N1, N2), the statistical result at 95% significance with N1 and N2 degrees of freedom.

  End

– Determine rank 1 network or operator ("best performing") for the tested service with the area.

  Rank 1 is attributed to the network with the minimum statistical significant quality distance of the service under test across networks or operators.

  For j=1,M

  Rank 1 = Rank(j) if StatDiffQuality_min=min(j=1,M) {SUM (i=1,N) {StatDiff_i,j * w_i}}

  where StatDiff_i,j are as defined in Equation B-1 or B-2 and w_i represent pre-established weights per service; if it is desired to use weights. Otherwise, equal unitary weights can be used.

  End

  Determine the ranks for all other networks or operators considered in the benchmarking campaign.

  For j=1,M

  If Dist(j) =max (0,StatDiffQuality/StatDiffQuality_min – Zth) = 0

  Rank (j)= Rank 1

(determines all networks showing the same statistical performance with rank 1)

Else

Order ascendingly Dist(j)

Rank (j) = position in vector Dist(j)

End

It should be noted that the ranking must be applied based on the statistical significance along with the KPI- or KQI-specific relevant difference-thresholds THrelv, whenever the differences are irrelevant or within each KPI or KQI accuracy of measurement. KPI or KQI-specific relevant difference-thresholds (THrelv) are defined in other ITU-T Recommendations in the benchmarking series.

Appendix I

A possible technique for the statistical scoring and ranking of a network

(This appendix does not form an integral part of this Recommendation.)

Sometimes it may be desirable that an overall network statistical score be determined. To this end, network performance per service type is often used as an underlying criterion. Such a score should be reported for each area as well as aggregated across regions, as shown in Figure 1.

The statistical score per service may be defined by all considered end-to-end KPI_i or KQI_i (i=1,N) metrics affecting the overall end-to-end quality of the service analysed. Thus, the score can be defined by a weighted sum of the StatDiff_i (see Annex B) of each KQI versus those best performing, as described in clause 9.4. If available, the StatDiff_i value is corrected in terms of the relevant difference threshold values. The final outcome, the StatScore, describes end-to end performance of the compared networks against the best performing network.

$$StatScore = \Sigma(w\_i*StatDiff\_i)$$

Here, w_i is the weight allocated to each KPI or KQI metric contributing to the quality of the service. The lower the StatScore, the better the performance (or closer to the best performing network) and the corresponding rank.

Table I.1 is a new version of Table 3, in which some examples of weightings have been added; these are just informative, since weight definition lies outside the scope of this Recommendation. The weightings can also be unitary if it is decided that all KPIs or KQIs have equal importance in the overall statistical score of network performance. However, it should be noted that, even if equal unitary weights are considered, if the number of KPIs or KQIs is increased or decreased, the statistical score of the network can change and provide different statistical results.

Therefore, the statistical scoring and ranking of a network in this Recommendation is only valid with a detailed description and motivation of the selection and underlying selected weights of the KPIs or KQIs. Without this transparency, the statistical scoring and ranking of the network is not valid.

For the example in Table I.1, network 1 receives the best rank 1, with a minimum score of 2.15, based on the given weights. In addition, it can be noted that in all cases, the differences between the KPIs or KQIs coming from the two networks are higher than the appropriate THrevl, meaning that the results of the statistical significance analysis are valid.

Table I.1 – Example of statistical scoring and ranking

	Network 1				**Network 2**				
	KPI	std	N	StatDiff	KPI	std	N	StatDiff	THrevl
KPI1/KQI1	0.95	0.22	87	0.05	0.97	0.17	69	0.00	0.018
KPI2/KQI2	0.93	0.26	87	0.00	0.91	0.29	69	0.23	0.019
KPI3/KQI3	3.89	0.50	2600	0.00	3.56	0.70	2070	17.15	0.31
KPI4/KQI4	105.00	5.00	435	42.67	70.00	15.00	350	0.00	34
KPI5/KQI5	1 200.00	300.00	87	0.00	1 800.00	275.00	69	12.31	596
StatScore				2.15				5.83	
Rank				1				2	

A complete benchmarking campaign may target the calculation of a global statistical score for each network under test for all supported services (j=1,M number of supported services). It must be noted

that any such global score must be well motivated and respective service-weightings be made explicitly transparent to comply with this Recommendation.

Such a score may be calculated by adding all statistical scores per service calculated as in Annex B and shown in the example in Table I.1. In this case, the best performing network has a global statistical score of zero. The lower the score, the better is the corresponding network performance as defined by the respective instantiation of the per-service KPI- or KQI-weighting and summing. The score resulting from such an operation is often used to represent a global network ranking of the compared networks against that best performing:

$$GlobalNetScore = \Sigma(Wser\_i * StatScore\_j)$$

The definition of the weights Wserv_j for the GlobalNetScore lie outside the scope of this Recommendation. However, some guidance is provided.

Depending on operator policies or business focus, the GlobalNetScore can be calculated as a weighted sum, similar to the weights used for the calculation of StatScore. The weights, as well as how many or what services to be considered in the calculation of the GlobalNetScore can be at the choice or decision of operators, if the internal benchmarking scenario (Figure 1) is considered.

Alternatively, weights can be decided based on user statistics analysis (e.g., crowd sourcing); types of service used and usage percentage for various types of area.

However, the validity of the global score depends on the detailed description and motivation of KPI or KQI selection and their underlying selected weights per service as well as each service weighting in the global score. Without this transparency, the statistical global scoring and ranking of a network is not valid. Moreover, as stated in the main body of this Recommendation, the underlying StatDiff scores must be handled in light of the relevance of between-service differences for the various KPIs or KQIs and should be set to 0 if any difference may be significant, but this does not contribute to a better performance of the network from the user perspective.

Bibliography

[b-ITU-T E.802] Recommendation ITU-T E.802 (2007), *Framework and methodologies for the determination and application of QoS parameters*.

[b-ITU-T P.10] Recommendation ITU-T P.10/G.100 (2017), *Vocabulary for performance and quality of service*.

[b-ITU-T P.863] Recommendation ITU-T P.863 (2018), *Perceptual objective listening quality prediction*.

[b-ITU-T P.1203] Recommendation ITU-T P.1203 (2017), *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport*.

[b-ITU-T P.1401] Recommendations ITU-T P.1401 (2012), *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*.

[b-ETSI TR 102 581] ETSI TR 102 581, V1.2.1 (2015), *Speech processing, transmission and quality aspects (STQ); A study on the minimum additional required attenuation on the antenna path of the field test equipment*. https://www.etsi.org/deliver/etsi_tr/102500_102599/102581/01.02.01_60/tr_102581v010201p.pdf

[b-Mehta] Mehta, S. (2014). *Statistics topics*. CreateSpace. 160 pp.

[b-Shaum] Spiegel, M.R., Schiller, J.J., Srinivasan, R.A. (2013). Schaum's outlines: *Probability and statistics*, 4th edition. New York, NY: McGraw-Hill. 424 pp.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	**Overall network operation, telephone service, service operation and human factors**
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2018