

International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.1301

(10/2017)

SERIES P: TELEPHONE TRANSMISSION QUALITY,
TELEPHONE INSTALLATIONS, LOCAL LINE
NETWORKS

Telemeeting assessment

**Subjective quality evaluation of audio and
audiovisual multiparty telemeetings**

Recommendation ITU-T P.1301

ITU-T



ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
Methods for objective and subjective assessment of speech and video quality	Series	P.800
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than speech and video	Series	P.1500

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

Recommendation ITU-T P.1301

Subjective quality evaluation of audio and audiovisual multiparty telemeetings

Summary

Recommendation ITU-T P.1301 concerns subjective quality assessment of telemeeting systems that provide multiparty communication between distant locations, using audio-only, video-only, audiovisual, text-based or graphical means as communication modes. The term multiparty refers to more than two meeting participants who can be located at two or more locations.

Evaluation of those systems can focus on audio-only, video-only or audiovisual quality aspects and non-interactive or conversational quality can be assessed.

This Recommendation gives an overview of relevant aspects that need to be considered for subjective quality evaluation of multiparty telemeetings and provides guidance to recommendations describing the details of applicable methods and procedures. Aspects in this Recommendation are also applicable to two-party telemeetings.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.1301	2012-07-14	12	11.1002/1000/11687
2.0	ITU-T P.1301	2017-10-29	12	11.1002/1000/13403

Keywords

Audio, audiovisual, asymmetric connections, conversation tests, guidance, multiparty conferencing, non-interactive tests, subjective quality evaluation, telemeetings, test methods, 3D audio and video.

* 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/>.

© ITU 2017

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

	Page
1 Scope.....	1
2 References.....	2
3 Definitions	3
3.1 Terms defined elsewhere	3
3.2 Terms defined in this Recommendation.....	3
4 Abbreviations and acronyms	5
5 Conventions	5
6 General recommendations concerning the subjective quality evaluation of multiparty telemeetings	5
7 Multiparty specific aspects in subjective quality evaluation	6
8 Guidance to suitable test methods	6
8.1 Test method decision criteria.....	7
8.2 Flow charts to be used when selecting test methods	8
Annex A – Set-up of a multiparty telemeeting assessment test.....	14
A.1 Assessment of conversational quality – Conversation tests	14
A.2 Assessment of non-interactive quality – Non-interactive tests	16
Annex B – Assessment of telemeetings with text-based communication and graphical information means (e.g., web conferencing)	18
Annex C – Assessment of video-only telemeetings.....	20
Annex D – Effect of transmission delays on telemeeting quality.....	21
D.1 Background.....	21
D.2 Existing test task recommendations	21
D.3 Recommended test tasks.....	22
D.4 Set-up of a delay test	22
D.5 Test subjects	23
D.6 Training session.....	23
D.7 Instructions	23
D.8 Test questions	23
D.9 Objective measurements.....	24
D.10 Effects of delay.....	25
Annex E – Assessment of spatial audio and 3D video reproduction of multiparty telemeetings	26
E.1 General remarks.....	26
E.2 Recommendations concerning spatial audio	26
E.3 Remarks concerning 3D video	26
Annex F – Assessment of asymmetric multiparty telemeetings.....	27
F.1 Problem statement	27

	Page
F.2 Translation of technical degradations into possible perceptual impairments	28
F.3 Influencing factors	30
F.4 Experiment design	31
F.5 Rating scales	32
F.6 Test stimuli	33
F.7 Data analysis	33
Annex G – Assessment of multiparty telemeetings with non-stationary quality	36
Annex H – Assessment of multiparty telemeetings using multi-dimensional scaling methods	37
Appendix I – Influential factors	38
Appendix II – Overview of multiparty non-interactive test stimuli and conversation test tasks	40
II.1 Non-interactive audio-only stimuli:	40
II.2 Non-interactive video-only stimuli	40
II.3 Non-interactive audiovisual stimuli	40
II.4 Audio-only conversation tasks	40
II.5 Audiovisual conversation tasks	41
Appendix III – Examples of multiparty conversation test tasks (audio-only and audiovisual): Free conversation	42
Appendix IV – Examples of multiparty conversation test tasks (audio-only): Three-party conversation test scenarios 3CTs	44
IV.1 Introduction	44
IV.2 Test scenario development	44
IV.3 Scenario validation	45
IV.4 Cultural aspects	46
Appendix V – Examples of multiparty conversation test tasks (audiovisual): Audiovisual multi-point tasks for three parties (Survival task, Leavitt task, Brainstorming task) ..	47
V.1 Overview and most suitable task	47
V.2 Leavitt task	47
V.3 Brainstorming task	48
V.4 Survival task	48
Appendix VI – Additional proposals for multiparty conversation test tasks (audiovisual): Extended survival tasks scenarios and celebrity name guessing task	61
VI.1 Overview and background	61
VI.2 Modification of survival scenarios from Appendix V	61
VI.3 Three additional survival scenarios	68
VI.4 Celebrity name guessing task	74
Appendix VII – Additional proposals for multiparty conversation test tasks (audiovisual): Formal and informal multiparty video conferences	75

	Page
Appendix VIII – Overview of documents describing suitable test methods	76
VIII.1 Baseline test methods on which the current Recommendation is based	76
VIII.2 Multiparty specific recommendations to adapt the baseline test methods	77
VIII.3 Further test methods, which are referred to in the annexes of this Recommendation	77
Bibliography.....	78

Recommendation ITU-T P.1301

Subjective quality evaluation of audio and audiovisual multiparty telemeetings

1 Scope

This Recommendation concerns subjective quality assessment of telemeeting systems that provide communication between multiple parties at remote locations. As multiparty telemeetings can differ in a large number of aspects, the assessment of such systems requires careful selection and control of the considered aspects and a precise description when reporting results.

The main aspects defining the scope of this Recommendation are:

- Number of participants and number of locations
The term multiparty refers to more than two meeting participants who can be located at two or more locations. Hence several multiparty situations are possible on which methodological aspects could depend: two sites with more than one person at at least one site (multiparty point-to-point), more than two sites with one person at each site (multiparty one-per-site), and more than two sites with more than one person at at least one site (multiparty multi-point).
- Communication mode and rendering conditions
The telemeeting systems considered in this Recommendation can provide audio-only, video-only (i.e., sign language or lip reading) or audiovisual communication. Telemeeting systems can render communication modes using different techniques such as mono channel vs. spatial sound reproduction or 2D video vs. 3D video display. Furthermore, web conferencing applications are considered as telemeeting systems that can provide additional text-based (chat, e-mail, etc.) and graphical information means (presentation slides, etc.).
- Evaluation mode and type of quality
Evaluation of multiparty telemeeting systems can focus on audio-only, video-only or audiovisual quality aspects and it can assess non-interactive or conversational/interactive quality. Hence, the assessment of multiparty telemeetings quality can be organized along five communication modes (audio, video, audiovisual, text-based, graphical), three test modes (audio, video, audiovisual), and two types of qualities (non-interactive, conversational/interactive).
- Controlled and non-controlled environments
Assessment tests can be conducted in a laboratory or in a real-life environment where the system is supposed to be used, for example, in a telepresence room. This Recommendation concerns testing in both controlled and non-controlled environments. Accordingly the test environments should be properly described and specified when reporting the test results.
- Symmetric and asymmetric set-ups
All telemeeting participants can be connected with the same type of equipment (symmetric) or different (asymmetric) equipment types. This Recommendation concerns testing of both symmetric and asymmetric telemeeting set-ups.

At the moment of writing this Recommendation, a number of ITU-T and ITU-R Recommendations are in force describing subjective quality evaluation methods; each of those methods focusing on individual communication modes, test modes or types of quality. Table 1 provides a corresponding overview. Note that in order to evaluate several quality aspects of a system intended for multiparty telemeetings several assessment tests may be required.

Table 1 – Focus of ITU-T and ITU-R Recommendations in terms of type of quality, communication mode and test mode

Type of quality	Communication Mode	Test Mode	ITU-T/ITU-R Recommendations
Non-interactive	Audio-only	Audio	P.800, P.806, P.880, BS.1116, BS.1285, BS.1534, P.1302, P.1310, P.1311
	Video-only	Video	P.910, BT.500, BT.710, BT.1788, P.915, P.916
	Audiovisual	Audio	P.800, P.880, BS.1116, BS.1285, BS.1534, P.913
		Video	P.910, BT.500, BT.710, BT.1788, P.913
		Audiovisual	P.911, P.1302, P.913
	Text	Video	–
	Graphical	Video	–
Conversational/Interactive	Audio-only	Audio	P.800, P.805, P.1305, P.1310, P.1312
	Video-only	Video	–
	Audiovisual	Audio	P.800, P.805, P.1310, P.1312
		Video	–
		Audiovisual	P.920, P.1305
	Text	Video	–
	Graphical	Video	–

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 P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.
- [ITU-T P.805] Recommendation ITU-T P.805 (2007), *Subjective evaluation of conversational quality*.
- [ITU-T P.806] Recommendation ITU-T P.806 (2014), *A subjective quality test methodology using multiple rating scales*.
- [ITU-T P.880] Recommendation ITU-T P.880 (2004), *Continuous evaluation of time-varying speech quality*.
- [ITU-T P.910] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.
- [ITU-T P.911] Recommendation ITU-T P.911 (1998), *Subjective audiovisual quality assessment methods for multimedia applications*.

- [ITU-T P.913] Recommendation ITU-T P.913 (2016), *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment.*
- [ITU-T P.915] Recommendation ITU-T P.915 (2016), *Subjective assessment methods for 3D video quality.*
- [ITU-T P.916] Recommendation ITU-T P.916 (2016), *Information and guidelines for assessing and minimizing visual discomfort and visual fatigue from 3D video.*
- [ITU-T P.920] Recommendation ITU-T P.920 (2000), *Interactive test methods for audiovisual communications.*
- [ITU-T P.1302] Recommendation ITU-T P.1302 (2014), *Subjective method for simulated conversation tests addressing speech and audiovisual call quality.*
- [ITU-T P.1305] Recommendation ITU-T P.1305 (2016), *Effect of delays on telemeeting quality.*
- [ITU-T P.1310] Recommendation ITU-T P.1310 (2017), *Spatial audio meetings quality evaluation.*
- [ITU-T P.1311] Recommendation ITU-T P.1311 (2014), *Method for determining the intelligibility of multiple concurrent talkers.*
- [ITU-T P.1312] Recommendation ITU-T P.1312 (2016), *Method for the measurement of the communication effectiveness of multiparty telemeetings using task performance.*
- [ITU-T P.1501] Recommendation ITU-T P.1501 (2014), *Subjective testing methodology for web browsing.*
- [ITU-R BS.1116] Recommendation ITU-R BS.1116-1 (1997), *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.*
- [ITU-R BS.1285] Recommendation ITU-R BS.1285 (1997), *Pre-selection methods for the subjective assessment of small impairments in audio systems.*
- [ITU-R BS.1534] Recommendation ITU-R BS.1534-1 (2003), *Method for the subjective assessment of intermediate quality levels of coding systems.*
- [ITU-R BT.500] Recommendation ITU-R BT.500-13 (2012), *Methodology for the subjective assessment of the quality of television pictures.*
- [ITU-R BT.710] Recommendation ITU-R BT.710-4 (1998), *Subjective assessment methods for image quality in high-definition television.*
- [ITU-R BT.1788] Recommendation ITU-R BT.1788 (2007), *Methodology for the subjective assessment of video quality in multimedia applications.*

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 telemeeting: A meeting in which participants are located at at least two locations and the communication takes place via a telecommunication system.

The term telemeeting is used to emphasize that a meeting is often more flexible and interactive than a conventional business teleconference and could also be a private meeting. The telemeeting could be audio-only, audiovisual, text-based or a mix of these modes.

3.2.2 test participant, test subject: A person taking part in an assessment test.

3.2.3 telemeeting participant, interlocutor: A person taking part in a conversation with one or more people via a telemeeting system.

In a conversation test, a telemeeting participant is also a test participant; in a non-interactive test with recorded conversations as stimuli, or in a test in which test participants observe a real-time telemeeting, telemeeting and test participants are different people.

3.2.4 multiparty: More than two people.

Example: More than two people participating in a telemeeting, having a conversation, performing a test task together, etc.

The term multiparty does not specify if people are distributed across two or more locations. If not explicitly stated otherwise, multiparty indicates that people are at two or more locations. When further specification is necessary, additional terms will be used (see point-to-point and multi-point) or the number of locations will be explicitly stated.

3.2.5 two-party: Two people.

Example: Two people participating in a telemeeting, having a conversation, performing a test task together, etc.

If not explicitly stated otherwise, two-party indicates that people are at two locations.

3.2.6 single-party: One person.

Example: Single-party test task means a test in which one person performs the test tasks on their own, e.g., a viewing-only test. Note that only non-interactive tests can comprise single-party test tasks.

3.2.7 multi-point: More than two locations.

Example: A multiparty multi-point telemeeting means that more than two interlocutors are taking part, and the interlocutors are located across more than two locations.

Multi-point does not specify if one or more than one interlocutor is present at each location.

In the special case that only one person is present at each location the term one-per-site will be used.

3.2.8 point-to-point: Two locations.

Example: A multiparty point-to-point telemeeting means that more than two interlocutors are taking part, and the interlocutors are at exactly two locations. That means that in one location more than one interlocutor is present, as there are more than two people.

3.2.9 one-per-site: One person per connected location.

Example: In a multiparty one-per-site telemeeting more than two sites are connected with only one person present at each site.

3.2.10 conversational quality: The perceived quality when two or more test participants have a conversation.

3.2.11 interactive quality: Synonym for conversational quality. This term might appear to be more appropriate when considering video-only communication, e.g., sign language or lip reading communication, while the term conversational quality is usually used in the referenced Recommendations.

3.2.12 non-interactive quality: The perceived quality when a person evaluates the listening-only, viewing-only, or listening-and-viewing-only quality of test stimuli. It can also refer to the quality of a conversation between telemeeting participants that is evaluated by observation in real-time.

3.2.13 type of quality: A term to differentiate between different types of qualities (here conversational and non-interactive quality).

3.2.14 communication mode: The mode that the system under test is providing for communication between telemeeting participants. It can be audio-only, video-only (for hearing-impaired) or audiovisual.

3.2.15 test mode: The mode that is investigated in the assessment test. It can be audio-only, video-only or audiovisual.

3.2.16 rendering condition: A term to differentiate between methods to render the audio or video content. The differentiation here is focused on spatial versus non-spatial rendering.

Example: A telemeeting system providing spatial audio reproduction and one providing (mono channel) non-spatial audio reproduction differ in the rendering condition, even though both provide the same communication mode "audio".

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ANOVA	Analysis of Variance
3CTS	Three-party conversation test scenario
ACR	Absolute Category Rating
EC	Extended Continuous
QoE	Quality of Experience
SCTS	Short Conversation Test Scenarios
SSCQE	Single Stimulus Continuous Quality Evaluation

5 Conventions

None.

6 General recommendations concerning the subjective quality evaluation of multiparty telemeetings

It is recommended to carry out the subjective quality evaluation of multiparty telemeetings as much as possible according to existing test methods recommended by ITU-T and ITU-R.

It is recommended to take the purpose of the test into account when selecting and – if necessary – adapting appropriate test methods. Test purposes differentiate for example by the granularity of the quality judgement (overall quality vs. individual quality aspects), the mode (audio, video, audiovisual) or by the type of quality (observational vs. conversational).

It is recommended to take multiparty-specific considerations into account when selecting and – if necessary – adapting appropriate test methods. Such considerations may be based on the multiparty specific aspects described in this Recommendation.

The selection and – if necessary – adaptation of an appropriate test method may be achieved using the guidelines provided in this Recommendation.

7 Multiparty specific aspects in subjective quality evaluation

Multiparty conversations differ in various aspects from one-to-one conversations, especially when the conversation takes place over a telecommunication medium. Those aspects comprise facets of human group communication and characteristics of the telemeeting systems under test. Two main differentiators between two-party and multiparty conversations via a telemeeting system are the conversational situation and the type of equipment used at each site (which can be different from site to site).

As multiparty telemeeting participants will communicate with more than one interlocutor simultaneously, the conversational situation has some implications, for example, on the required cognitive load or on aspects of group communication, which in turn can have an influence on quality judgements. Some typical multiparty conversational set-ups are group-to-group, one-to-group and other different combinations of single people and groups of different sizes.

As telemeeting systems can provide different communication modes and rendering conditions, the importance of certain side aspects of a conversation may differ between audio-only and audiovisual telemeetings and thus may have different impacts on the quality assessment. For instance, consider speaker identification as one such side aspect. If the ability to identify individual speakers is important in order to follow a telemeeting, then a system's ability for allowing good speaker identification would influence the quality judgement. However, speaker identification might be less important in audiovisual than in audio-only telemeetings because it is easier to understand who is talking if a video of that person can be seen and it is easier to understand who wants to speak next when video is present, compared to audio only.

In addition, the rendering conditions can influence the importance of such aspects. Regarding the previous example, speaker identification may be less important for an audio-only system providing spatial audio rendering than for an audio-only system providing only mono channel audio rendering.

As people are often connecting with different terminal devices over networks with different transmission qualities, the presence of asymmetric transmission chains is an important question in multiparty telemeeting assessment. Note that asymmetric qualities may be possible also for two-party conversations, e.g., certain impairments are present only in one of the two transmission directions. However, as long as the interlocutors do not discuss this, they do not – at least consciously – experience such asymmetries from their perspective, as they do not know how the other person is experiencing the connection. Asymmetry in the multiparty case goes beyond that as telemeeting participants may experience different qualities for different interlocutors and thus are able to directly perceive asymmetry.

8 Guidance to suitable test methods

In order to select and – if necessary – adapt an appropriate test method for the intended multiparty assessment experiment, this clause provides guidance to a set of relevant texts describing the necessary details to be considered. Those texts are the annexes of this Recommendation as well as existing ITU-T and ITU-R Recommendations.

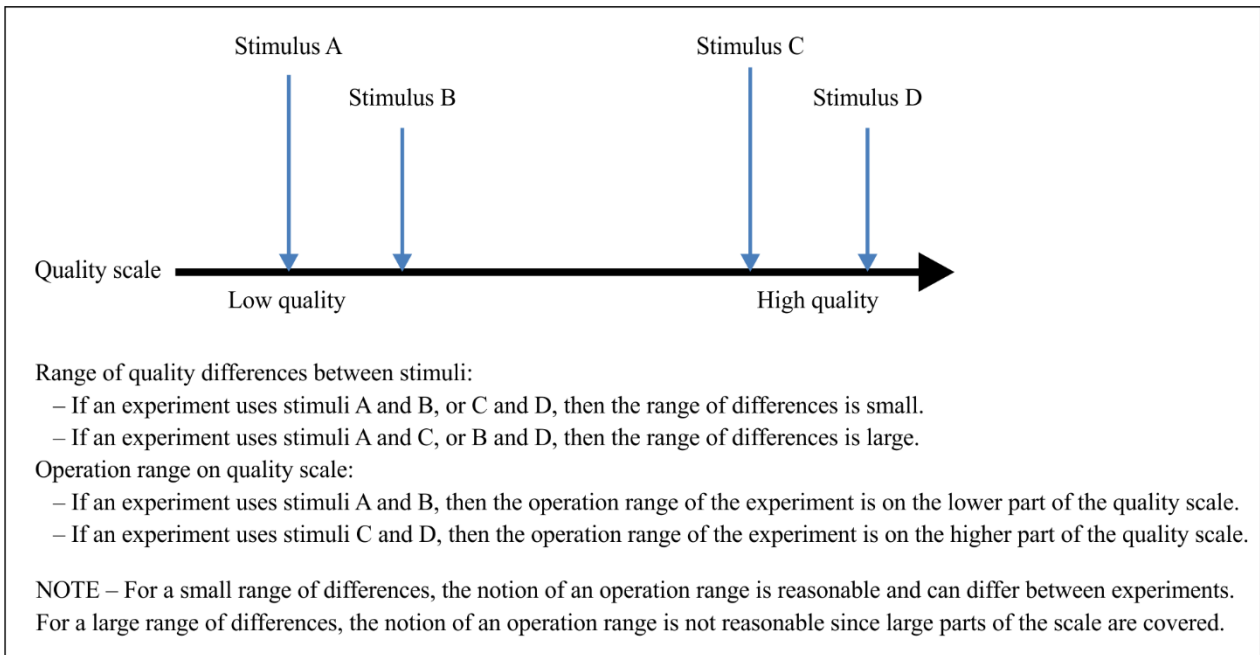
As guidance for suitable test methods, the following clauses provide a list of decision criteria to be considered by the investigator and flowcharts using these criteria with references to associated sources.

Notice that in cases in which enough knowledge is available, this Recommendation provides specific advice. In cases in which more knowledge needs to be built up, this Recommendation aims to generate awareness on specific aspects that a reader should consider.

8.1 Test method decision criteria

In order to select an appropriate testing method, investigators should consider the following decision criteria:

- 1) Communication mode: Does the system under test support audio-only, video-only or audiovisual communication?
- 2) Test mode: Is the main interest the system's audio, video, or audiovisual quality?
- 3) Type of quality: Is the main interest conversational or non-interactive quality?
- 4) Expected range of quality differences between stimuli (see also Figure 1): Can it be expected that people perceive large, medium, or small differences between the stimuli that we will use in the test?
- 5) Expected operation range on quality scale (see also Figure 1): Can it be expected that people will perceive the test stimuli on a certain area on the quality scale? Can it be expected that the majority of ratings are relatively high on the scale, as the tested systems are, for example, high-end systems? Or do ratings reflect intermediate or relatively low quality, as the systems use, for example, high compression rates or low bitrates? Or do they spread across the whole quality range, for example, if a mix of systems is considered in the test?
- 6) Delay under investigation: Does the test comprise different transmission delay times as experiment conditions? Or is delay a major characteristic of the system under test?
- 7) Picture format (for video): Do the considered devices have small sized displays (e.g., smart phones, tablets) or normal and large sized displays (e.g., PC screen, TV)?
- 8) Audio/video rendering conditions: Does the system under test support non-spatial (2D) or spatial (3D) video? Does it support non-spatial (mono) or spatial (multichannel, binaural, wavefield synthesis, etc.) audio reproduction?
- 9) Asymmetric conditions: Are terminal devices and transmission channels different or the same for all test participants? Does the system under test provide a mix of communication modes? Does it provide a mix of rendering conditions?
- 10) Dynamic conditions: Are there non-stationary system parameters in the test, or a non-stationary number of interlocutors (participants entering and leaving the telemeeting at different moments)?
- 11) Dimensionality of assessment: Is the main interest in scalar values (e.g., mean opinion score) or in multidimensional values (e.g., multidimensional scaling methods)?

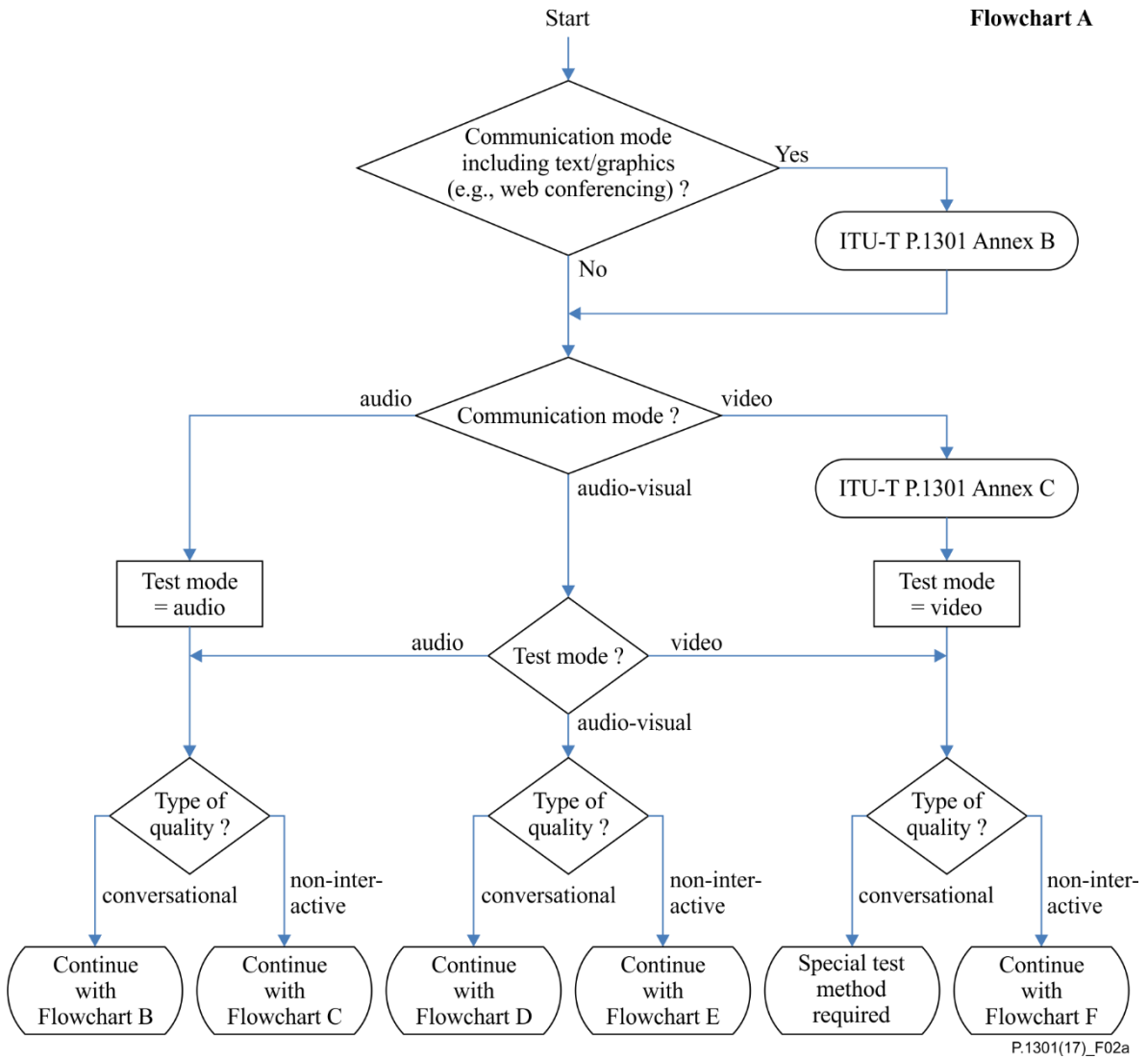


P.1301(17)_F01

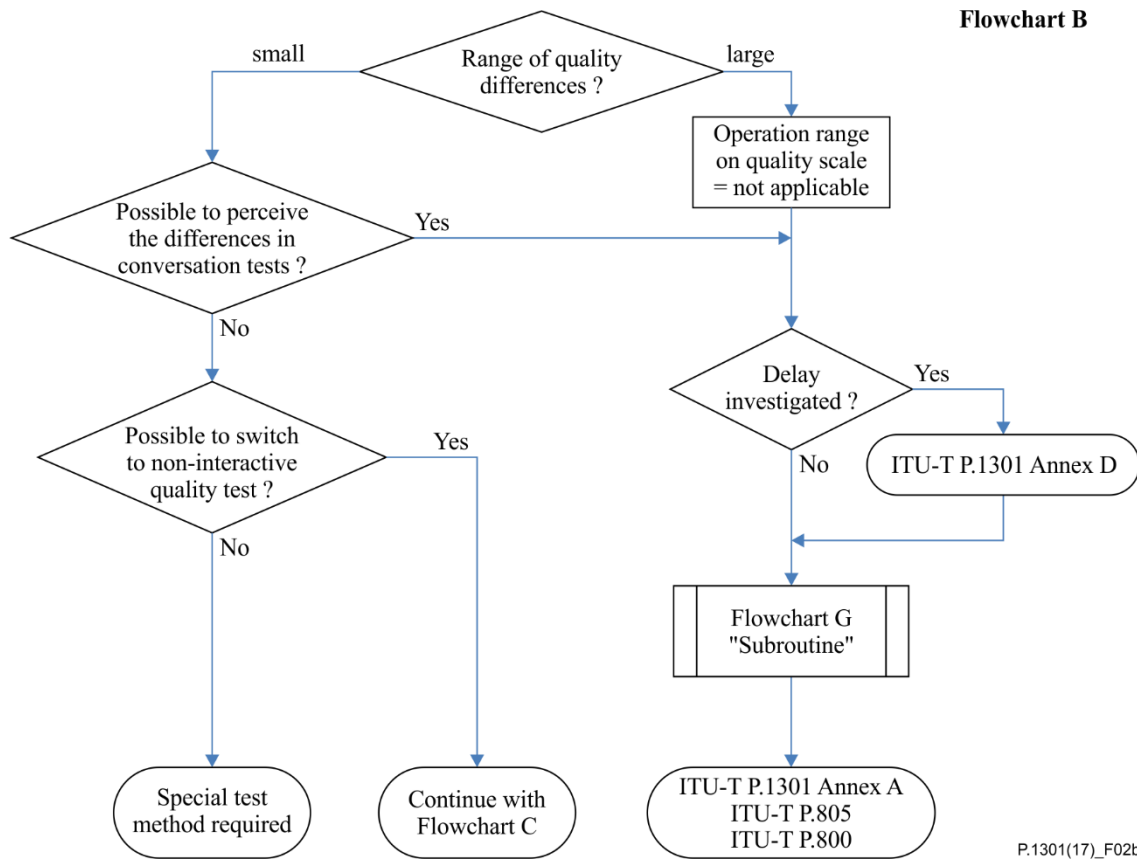
Figure 1 – Relation between "Range of quality differences" and "Operation range on quality scale"

8.2 Flow charts to be used when selecting test methods

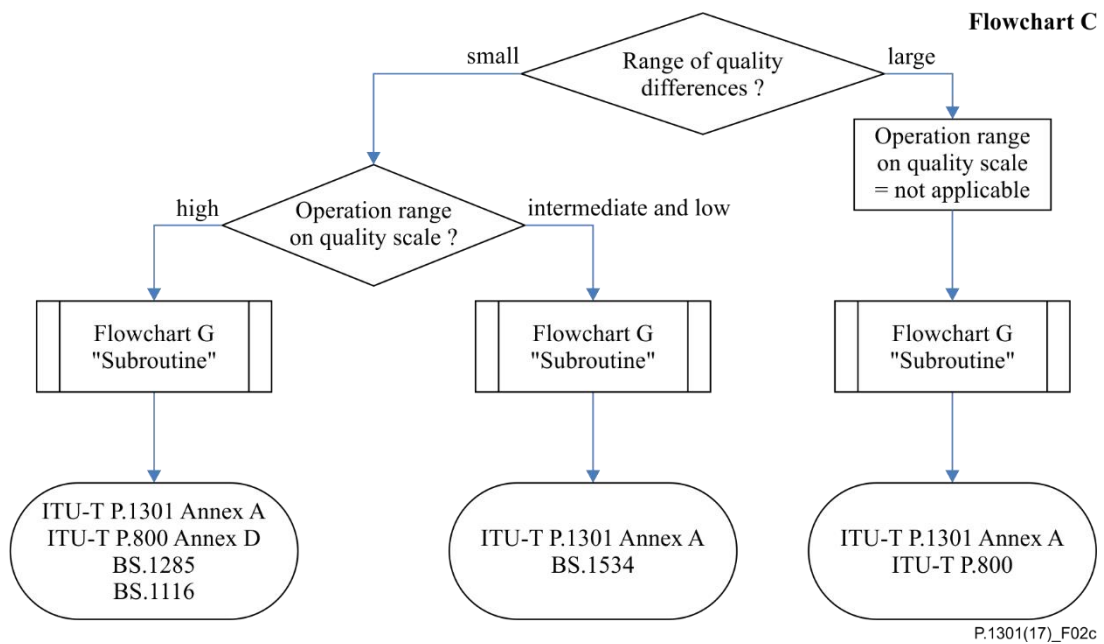
The following flow charts help to find the most appropriate test method and the corresponding documents according to the decision criteria described above.



Flowchart A – Start of the decision tree

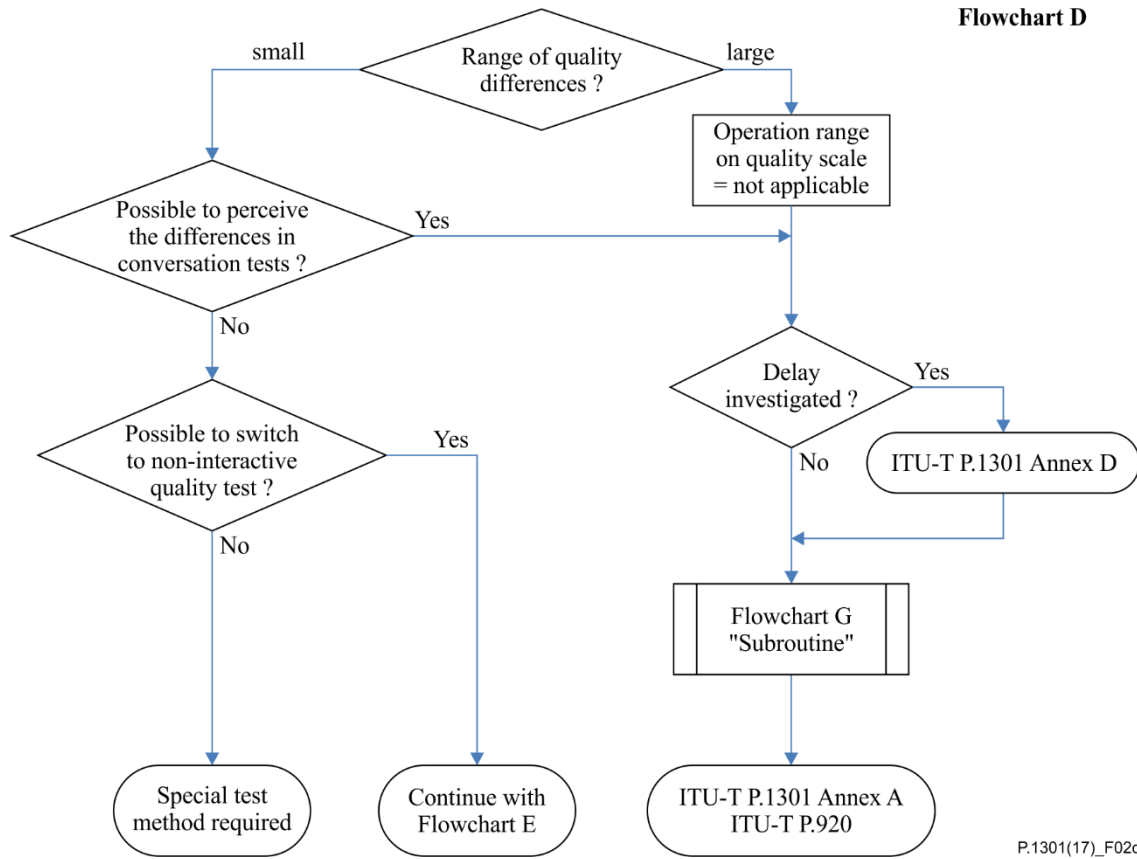


Flowchart B – Continuation of decision tree for assessing conversational audio quality



Flowchart C – Continuation of decision tree for assessing non-interactive audio quality

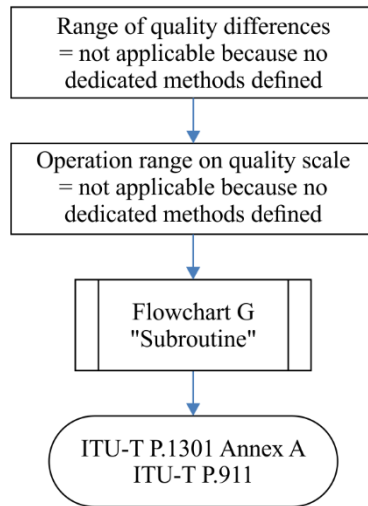
Flowchart D



P.1301(17)_F02d

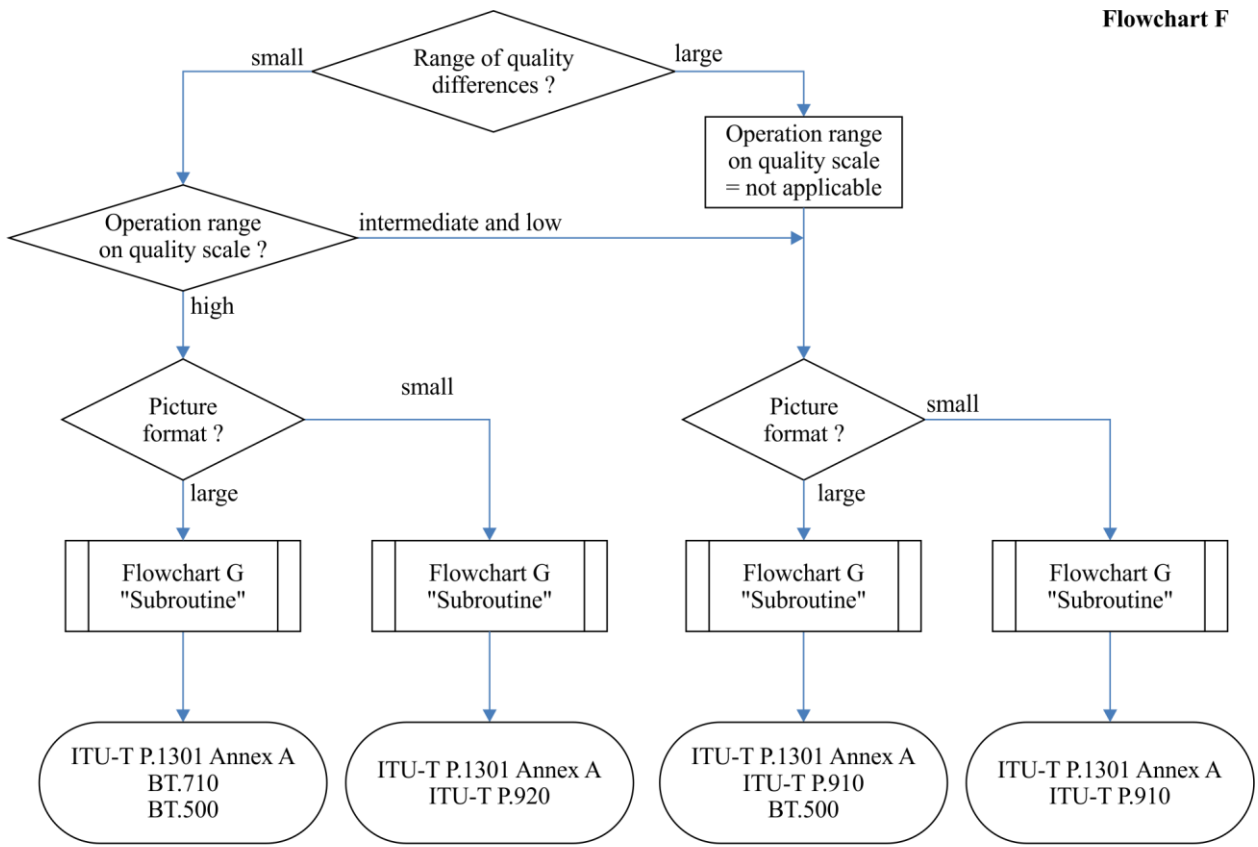
Flowchart D – Continuation of decision tree for assessing conversational audiovisual quality

Flowchart E



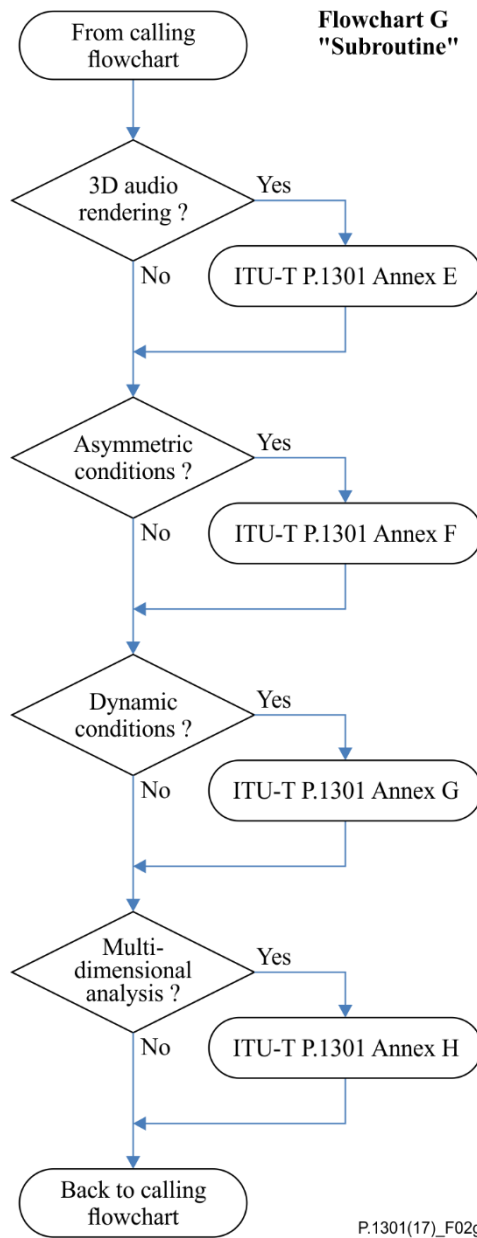
P.1301(17)_F02e

Flowchart E – Continuation of decision tree for assessing non-interactive audiovisual quality



P.1301(17)_F02f

Flowchart F – Continuation of decision tree for assessing non-interactive video quality



Flowchart G – Subroutine called from different branches in the flowcharts B to F

Annex A

Set-up of a multiparty telemeeting assessment test

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

This annex describes in more detail the set-up of a multiparty telemeeting assessment test. It follows the general structure of existing Recommendations by distinguishing between conversation tests and non-interactive tests. Test facilities, conversation tasks and non-interactive stimuli, experiment design, test subjects, scales, instructions and training phases, data collection and analysis are addressed. This annex explains in detail multiparty specific aspects that need to be considered, while it refers to appropriate recommendations (according to the decision tree in clause 8.2) for details that are not multiparty specific.

A.1 Assessment of conversational quality – Conversation tests

A.1.1 Test facilities

A.1.1.1 Physical test conditions

Test conditions should be generated in accordance with the detailed descriptions provided in the associated recommendations according to the decision tree in clause 8.2.

The requirements of the physical test conditions are generally valid for all test subjects in a conversation test and when several test subjects are located in the same room. For that reason test facilities may require careful calibrations of test equipment and test environments to ensure the conditions are the same between interlocutors. If variations between test subjects cannot be avoided, those differences should be noted (light, sound levels, distance to screen etc.).

Concerning video, people may perceive video quality differently depending on their distance to the screen, which usually varies in a multiparty meeting. If this is not accounted for, the viewing positions should be documented.

In case of tests in arbitrary locations, e.g., using real-life telemeeting systems, it may not be possible to take all these considerations, e.g., acoustics and lighting, into account. In this case the individual conditions should be documented.

A.1.1.2 Call set-up

There are many possible variations in the initial call set-up in terms of technology and user interface. Despite these differences, the call set-up process for a two-party conversation is always the same: caller invites callee, callee accepts the call. In multiparty conversations, different processes are possible for the initial call set-up: interlocutors can dial into a telemeeting on their own, interlocutors can be invited by other interlocutors to join the call, only one person (a chairperson) can invite the other interlocutors, etc.

Given this additional complexity in the call set-up process – on top of different technologies and user interfaces – the initial call set-up is an important aspect for the quality of experience (QoE) of a telemeeting system. This also holds for test situations, especially if the call initiation is done by the test participants.

Hence, test facilities may provide a proper call set-up by supporting dialling possibilities, ringtones, speech prompts, etc. Alternatively, if the call set-up is not in focus, it can be outside of the test so that the conference call is already set up for the test participant at the start of the test.

A.1.2 Conversation task

Given the importance of the conversational situation for multiparty telemeetings, conversation tests may require appropriate experimental tasks that put test participants into a proper conversational situation.

The number of interlocutors can have a major influence on the quality assessment. Hence, investigators should deliberately choose the number of interlocutors according to their needs. In addition, special attention may be required when comparing results between studies in which different numbers of interlocutors have been used.

With an increasing number of interlocutors, the complexity of the conversation tasks in terms of conversational structure, cognitive effort or difficulty for the test participants can also increase.

The time required for each conversation in a test needs to be increased as the number of interlocutors increases, to allow all participants time to be active in the conversations. This aspect has to be considered with respect to the total test time (see Experiment design in clause A.1.3).

Tasks that work well for two-party tests might become more difficult for test participants when they are adapted to multiparty tasks. Investigators are advised to pay special attention to such effects and consider appropriate adaptations. For example, in case investigators would like to use structured conversation test scenarios (such as those described in [ITU-T P.805]) for a multiparty test, an extension of such test scenarios to a larger number of interlocutors might require the introduction of a formal discussion leader in order to ensure that the structure of the new conversations remains feasible for the interlocutors to follow.

Furthermore, specific attention should be paid to the test modes when choosing the task. For example, the task for audiovisual tests should be designed such that, during their conversation, test participants primarily maintain their attention on the audiovisual terminal.

A.1.3 Experiment design

General aspects that should be considered for a proper experimental design include, the presentation order of conditions, suitable length for a session in a test and the need for pauses and training phases. For details on these aspects of the experiment design, the recommendations according to the decision tree in clause 8.2 should be applied, and the ITU Handbook on Practical procedures for subjective testing should be consulted [b-ITU-T HB-PPST].

However, multiparty tasks require the conversational situation to be of longer duration than that of conventional single- or two-party tasks. Therefore the experiment design needs to carefully balance task duration, number of tasks and overall experiment duration.

A.1.4 Test subjects

Due to the special conversational situation of multiparty telemeetings compared to conventional one-to-one conversations, a number of aspects regarding the test participants might influence the results. For that reason, investigators should deliberately consider using participant profiles when inviting people to the subjective experiments and, if applicable, when scheduling participants in groups.

If the test task is more formal, like booking a train or asking for information, it is not an advantage if the test participants know each other.

If it is important to provide a more fluent, natural conversational situation in a test, it is recommended that the test subjects have the opportunity to get to know each other to some extent before the test. Another alternative is to invite people that know each other before the test.

Investigators should consider that gender-related differences in the voice character, such as pitch, provide strong cues for speaker separation, and can thus influence the task difficulty for subjects, which can in turn have a significant impact on results. If it is desirable to minimize the influence of such gender-related differences, then subjects within the groups should have the same gender. If it is

desirable to have a more representative pool of subject groups, then both same-gender and mixed-gender groups should be considered in the experiment.

If the test requires a high sensitivity to differences between conditions, test participants should have prior experience with multiparty systems, since inexperienced test participants might be overwhelmed by the system under test.

A.1.5 Scales

Direct quality judgements may be collected by using established quality scales as described in the recommendations (according to the decision tree in clause 8.2).

The participants could be asked to assess how well the conference system operates for communication purposes with multiple parties, considering questions such as: Is it possible to communicate similarly well as in real life? Does the system facilitate or hinder your ability to interact during a conversation?

To check if the test was performed as intended it is advisable to give the test subjects the chance to write comments during the test. This might help to interpret the voting results and might reveal additional information that was not requested in the quality questionnaires. After the test, participants could be asked to describe their experience of the test situation either one-by-one on paper or orally within a group.

A.1.6 Instructions to subjects and training phases

To get reliable results, the instructions and training phases are important. In conversation tests one quality reference could be a face-to-face meeting. If the participants are to compare the interaction through the system with the interaction in a face-to-face meeting it is advised that they get the opportunity to experience a face-to-face meeting before the test. If they do not know each other a longer training phase is recommended where the subjects present themselves, memorize the names of the other participants and, possibly, take part in a game. It is easier to address someone if you know their name.

Furthermore, test subjects should try some test scenarios and the voting procedure (a pretest) as part of the training phase.

Another method to facilitate free conversation is to show the names of either the video rooms or the participants on the screen or on the background wall. In an audio conference the subjects could be instructed to say their names before they speak, if this is in line with the goals of the test.

A.1.7 Data collection and analysis

For details on data collection and analysis, the recommendations according to the decision tree in clause 8.2 of this Recommendation should be applied and the ITU Handbook on Practical procedures for subjective testing should be consulted [b-ITU-T HB-PPST].

A.2 Assessment of non-interactive quality – Non-interactive tests

A.2.1 Test facilities

For detailed recommendations regarding test facilities, see clause A.1.1.

In the case of tests in arbitrary locations, for example, real-life telemeeting systems or mobile scenarios, [ITU-T P.913] may be consulted for further information on non-interactive quality testing in arbitrary environments.

A.2.2 Stimuli

A.2.2.1 Technical production

For details on the technical production of stimuli, the recommendations according to the decision tree in clause 8.2 should be applied.

A.2.2.2 Stimuli content

Given the importance of the conversational situation for multiparty telemeetings, non-interactive tests may – similarly to conversation tests – require the production or selection of stimuli that sufficiently resemble the conversational situation. Therefore, non-interactive quality should be evaluated with content that is suitable for conversational applications, such as recordings of audiovisual and audio conferences.

A.2.2.3 Real-life observation as alternative to pre-produced stimuli

One possible type of observational test is to observe an ongoing telemeeting, either by sitting in an actual meeting room and changing rooms during the observational test or by following the conversations in a separate room using audio or audiovisual monitoring equipment.

A.2.3 Experiment design

For detailed recommendations regarding experiment design, see clause A.1.3.

A.2.4 Test subjects

As is the case with conversation tests (clause A.1.4), investigators should deliberately consider using participant profiles when inviting people to subjective experiments. In particular, any prior experience subjects have with multiparty telemeeting systems might influence the results.

If the test requires a high sensitivity to differences between conditions, test participants should have prior experience with multiparty systems, since inexperienced test participants might be overwhelmed by the system under test, irrespective of the tested conditions.

A.2.5 Scales

For detailed recommendations regarding ratings scales, see clause A.1.5.

A.2.6 Instructions to subjects and training phases

To get reliable results, the instructions and training phases are important. In the context of multiparty assessment, the quality reference on which subjects form their judgement should be addressed because one quality reference could be a multiparty face-to-face meeting and another could be a two-party telemeeting.

A.2.7 Data collection and analysis

For detailed recommendations regarding data collection and analysis see clause A.1.7.

Annex B

Assessment of telemeetings with text-based communication and graphical information means (e.g., web conferencing)

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

The quality of experience of telemeeting systems that provide text-based communication and graphical information, for example web conferencing services, can be influenced by many factors. These have been categorized as follows:

- 1) **Web browsing aspects (in case of web conferencing applications)**
Quality of experience can be impacted before the web application is loaded, for example, the time it takes for the application to load. For more information on the subjective testing of web browsing, it is recommended to consult [ITU-T P.1501].
- 2) **Aspects concerning the arrangement of conferencing elements**
Conferencing systems often allow users, or at least the session chairmen, to arrange the individual frames/elements of a conference application (video, presentation slides, chat, etc.). This user defined arrangement can have an impact on the quality of experience (for example, if the size of the video frame is changed). Hence testing conferencing applications should either account for such effects during a subjective test (e.g., by disabling such arrangement functions), or check for such effects after the subjective test (e.g., by tracking the arrangement and performing a post-analysis depending on the arrangement). In any case, it is recommended to be very stringent in detailed reporting of the test set-up, especially regarding the arrangement of the conferencing elements.
- 3) **Audiovisual communication aspects**
If the focus is on testing the audiovisual communication of a web conferencing system, it is recommended to select the testing method according to the guidelines given in clause 8 of this Recommendation.
- 4) **Text-based communication aspects**
As no standardized methods are available for the subjective evaluation of text-based communications, special test methods for both non-interactive and conversational quality need to be developed. However, as video quality would be the primary test mode in this case (e.g., picture quality of the displayed text) new methods may be developed based on the existing video methods cited in clause 8 of this Recommendation.
- 5) **Graphical information means**
As no standardized methods are available for the case of communicating with graphical information means (e.g., presentation slides), special subjective test methods for both non-interactive and conversational quality need to be developed. However, as video quality would be the primary test mode in this case (e.g., picture quality of the displayed graphical elements) new methods may be developed based on the existing video methods cited in the guidelines in clause 8 of this Recommendation.

6) Aspects concerning multiple communication modes at the same time

Web conferencing services and other telemeetings with text-based communication or graphical information means provide more than one communication mode at the same time (e.g., audiovisual and text-based). As subjects might divide their attention between different communication modes, a careful experimental design is required. Instructions, tasks/stimuli and questions/scales should be especially helpful in ensuring test participants focus their attention on the communication mode under investigation. For instance, test tasks could ensure that all test participants are switching between the different communication modes in similar patterns (e.g., by providing a fixed sequences of subtasks for each of the considered modes).

Annex C

Assessment of video-only telemeetings

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

The most typical scenarios for video-only telemeetings are conversations between hearing impaired people using sign language and/or lip reading. However, no ITU Recommendations on subjective quality testing of such video-only communication are available.

For interactive quality, new dedicated test methods need to be developed; for non-interactive quality, the video test methods cited in clause 8 may be adapted.

[b-ITU-T H-Sup. 1] gives suggestions for system requirements to provide a sufficient quality that is needed for efficient sign language and lip reading communication. Although the focus is not on subjective testing, this supplement might provide useful insights for developing video-only test methods, such as

- Experimental tasks: finger spelling, general signing and lip reading
- Target conditions: temporal resolution (frame rate, delay), spatial resolution (blur), synchronism (lip reading with audio)

Annex D

Effect of transmission delays on telemeeting quality

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

The intention of this annex is to recommend suitable evaluation methods to assess the impact of transmission delays in telemeetings. General considerations for multiparty conversation tests are described in clause A.1. Special considerations that should be taken into account regarding delay tests are described in this annex. For more details on the effects of delay on telemeetings and additional advice on the subjective testing and analysis methods of such effects, it is recommended to consult [ITU-T P.1305].

D.1 Background

Long transmission delays might cause several problems in multiparty conversations. The focus of this annex is on problems caused by transmission delays occurring when two or more participants communicate with each other, but there might also be quality impacts caused by delay occurring when only one participant talks (e.g., echo becomes more obvious and therefore more disruptive in the case of longer delays, which might in turn affect the conversational quality).

The situation when different parties interrupt each other and/or talk at the same time is a common problem that might be caused by long delays. This might happen when, for example, a person thinks that the party that is currently speaking is going to stop talking, so this person starts talking, but the other party carries on and so they talk at the same time. This can occur also when the delay is short, but the problem increases with longer delays.

To test and evaluate the quality when transmission delays are in focus requires a different test methodology compared to when no delay is present. This annex gives some guidance on how to evaluate the conversational quality taking long transmission delays into account. A methodology for a conversational multiparty test requires proper test design considering test task, test subjects, training session, instructions, and quality assessment questionnaires.

The focus here is to recommend methods suitable for evaluating multiparty conversational quality, not on methods to detect if there is a delay or not.

D.2 Existing test task recommendations

Test tasks for audio conversational tests are described in [ITU-T P.800] and [ITU-T P.805]. Test tasks for audiovisual conversational tests are described in [ITU-T P.920]. Since most recommended conversation tasks are for two people, corresponding tasks need to be designed for multiparty conversations. Concerning delay, most existing test tasks either require too high cognitive efforts to be delay sensitive (such as thinking about what to answer or searching for items) or they lack natural speaking behaviour (such as when the conversations are too structured). Predetermined scenarios, such as the short conversation test scenarios for two participants in [ITU-T P.805], lead to realistic conversations, but there are some information retrieval components to the test tasks (e.g., a table lookup) that might lead to short pauses that could mask delay. As another example, the interactive short conversation test scenarios for two participants, also in [ITU-T P.805], include quick exchanges of numbers and names. These scenarios are more interactive, but the fact that one person is supposed to reply makes the other participant wait for the answer without interrupting.

For a telemeeting scenario appropriate test tasks should ideally reflect a normal conversation, but also allow for high delay sensitivity. Concerning interactivity, [ITU-T P.920] states that lively audiovisual conversations can be stimulated if the test subjects know each other. Concerning naturalness of the conversations, [ITU-T P.805] recommends that test tasks should allow for interruptions from the subjects and should lead to both long and short utterances.

A list of tasks for evaluating the effects of delays is cited below from [ITU-T P.920]:

[ITU-T P.920] I.2 Tasks to evaluate the effects of speech delay on communication quality

In the following tasks the talk spurt increases from task 1) to task 6), whereas the conversation switching rate decreases.

- 1) take turns in counting;
- 2) take turns reading random numbers aloud as quickly as possible;
- 3) take turns verifying random numbers aloud as quickly as possible;
- 4) words with missing letters are completed with letters supplied by the other talker;
- 5) take turns verifying city names as quickly as possible;
- 6) determine the shape of a figure described verbally;
- 7) free conversation.

The previous tasks [with exception of task 1) and task 7)] cannot be used for audiovisual quality evaluations because most of them require the subjects to concentrate their attention on a sheet of paper and not on the screen."

D.3 Recommended test tasks

The perceived conversational quality will depend on the task of the conversation (e.g., if it is highly interactive or not). The test task used will affect the perceived quality and quality ratings given by test participants, so the selection of test tasks will significantly impact test results. Delays can more easily be detected if the interaction is fast and some kind of competition might motivate subjects to interact efficiently and may make them more aware of delays, but if the test task is too engaging it may adversely impact the ability to evaluate the quality.

The goal is to recommend test tasks suitable for evaluating the conversational quality, or the quality impact when delay is present, in different situations. Test tasks not included here might be suitable as well and it is important to describe the test task when analysing the results after a test.

There are different reasons to perform a test, e.g., to evaluate a system intended for specific types of multiparty communication. Therefore more tasks are expected to be included in future versions of this Recommendation.

Free conversation is a natural task and is recommended if the conversation is to be realistic and spontaneous. It is suitable for both audio and audiovisual tests since there is no need to read a written instruction during the test. During audiovisual conversations it is important that the task does not prevent the participants from looking at the video screen during the main part of the test, i.e., instructions that need to be read during the test should be limited. In a free conversation other talkers can be interrupted spontaneously, which might lead to natural double-talk situations. To stimulate a more interactive conversation the participants could be encouraged to debate, take opposite standpoints and not be too polite.

An example test methodology for free conversations can be found in Appendix III.

Other examples of test tasks suitable for audiovisual quality evaluations regarding delays, for example the Survival task, are described in Appendix V.

D.4 Set-up of a delay test

Generally, a multiparty conversational delay test should be set up according to the cited conversational methods in clause 8 of this Recommendation.

If possible the test should be balanced so that all participants experience all different delay conditions. This also balances out personal characteristics, for example if a person is more dominant and tends to talk more.

If possible the participants should experience all kinds of test situations (e.g., alone in own room or co-located with other people in a group room). The conversational behaviour of the test subjects might be different in different test scenarios.

D.5 Test subjects

In general, the test subjects should be naïve in the sense that they should not work with delay related tasks or the evaluation of telecommunication qualities.

It is an advantage if the test subjects know each other if the goal is to create fluent and natural conversations.

D.6 Training session

In a delay test it is an advantage if the conversation is interactive.

If this is requested and the test subjects do not know each other beforehand, a two phase training session is recommended. In the first part, all test participants gather in one room to help them get to know each other better and make the conversations more interactive. A short presentation face-to-face, followed by a round of quizzes or another type of game will also help to get the group to interact before the start of the test. They should repeat the names of all participants until they know all names making it easier to address a certain person during the test. After the face-to face part of the training session, test participants continue with the second part of the training session. In this part participants can go to their test rooms and try out conversations as they would do in the main test and vote on the perceived quality.

This prolonged face-to face part of the training session is not needed for groups that already know each other well.

D.7 Instructions

As usually recommended, nothing should be mentioned about the specific test conditions in the test, so delay should not be mentioned. If it is mentioned that the test is about delay, and/or the subjects were trained to recognize delay, the subjects will not be naïve in that sense and would probably be more sensitive to delay effects. The resulting judgements will probably not reflect a normal telemeeting situation.

D.8 Test questions

Several examples of scales that can be used are described in [ITU-T P.920], [ITU-T P.800] and [ITU-T P.805].

The test questions and scales described in Appendix III can also be used in a conversation test.

The most relevant questions should be chosen for each occasion. The number of questions should be reduced to avoid distracting subjects from the most important task.

For impairments that do not occur continuously (e.g., codec artefacts, packet loss, etc.) but occur unevenly in time (such as disturbances due to long delays, which depend on the conversational interactivity), an impairment scale might be more appropriate to capture the grade of distortions. For a naïve test subject it is considered easier to grade the amount of impairment rather than the quality if a distortion occurs a limited number of times.

Examples of questions suitable for evaluation of the effects of delay using scales for effort and impairment (discrete or continuous with labels) are shown below:

- 1) How would you judge the effort needed to interrupt the other party (or parties)?
 - No effort
 - Minor effort

- Moderate effort
- Considerable effort
- Extreme effort

2) Did you perceive any reduction in your ability to interact during the conversation?

- Imperceptible
- Perceptible but not annoying
- Slightly annoying
- Annoying
- Very annoying

In conversations with long delays people might adapt automatically to the delay, but they might also get slightly more irritated. It might be possible to capture this effect by making the test subjects more observant towards small conversation discrepancies.

An example of a quality evaluation question that might be used in a delay test is shown below (continuous scale with the labels "Bad" and "Excellent" at the endpoints):

1) How would you assess your ability to converse back and forth during the conversation (did you feel a vague irritation..?)

Bad ----- Excellent

The test subjects could be encouraged to write comments on paper during the test to make it easier to understand why they voted as they did. They could then note anything that affected the perceived quality, even if it was not applicable to any of the questions in the test. The comment fields might look like this, but with several lines:

Did something hinder the ability to communicate? In that case, what?

Other comments.....

After the main test it is recommended to ask the test subjects some questions about their experience of the test. Then other important aspects that were not explicitly covered in the test might be expressed through the comments.

D.9 Objective measurements

If possible the communication efficiency should be measured as delay affects task efficiency. If the subjects are to perform a specified task the completion time can be measured, but conversations that are too 'open' make it impossible to measure communication efficiency. On the other hand communications that are too structured do not leave room for the subjects to develop a natural conversation.

In free conversations the amount of speech for every participant is not controlled by the test task. This aspect makes it possible to determine whether the amount of speech varies depending on the conversational situation (for example, if the subjects are sitting alone in a room or in a room together with other people and whether they have an audiovisual or audio only connection). If the conversations can be recorded the flow of conversation can be examined objectively by analysing the recorded files.

D.10 Effects of delay

Delay is not always noticed by the test participants but can still influence the quality of a conversation. If there are long delays the conversation partner can be perceived as being unusually slow or not particularly interested in the interaction.

The pace of a conversation often changes automatically (slows down) when there are delays, without participants noticing the delay until it becomes lengthy.

Test subjects that do not know each other are often polite. If two people speak at the same time, they often both stop talking. If the test subjects know each other they often have a more spontaneous conversation.

It is important to note that the combination of echo and delay is much more disturbing than a pure delay. Delay is also much more disturbing for playing music or games that are time-critical over a connection than for a normal conversation. In a normal conversation you reflect on what the other person said, and think about how to reply, so there is usually a short natural delay before the response.

The sensitivity for delay might also be different depending on the number of people that participate from a certain location. There is no delay between participants in the same room, but there are always delays between different sites, which affects the conversations. Test participants that experience long delays can have more difficulties in entering a conversation, especially if the other participants experience less delay. The size of a group also has an influence on the dynamics of a conversation.

Annex E

Assessment of spatial audio and 3D video reproduction of multiparty telemeetings

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

E.1 General remarks

Certain side aspects of a conversation can have an impact on the quality assessment. Considering systems providing spatial audio (3D audio) or 3D video rendering conditions as well as virtual/augmented reality and 360° video, the importance of such aspects may differ between such systems and conventional 2D systems.

For example, if the ability to identify individual speakers is important in order to follow a telemeeting, then a system's ability for allowing good speaker identification would influence the quality judgement.

Furthermore, new questions arise when different rendering conditions are presented in the same experiment. For example, the presence of a spatial audio rendering condition among conditions with non-spatial audio impairments might reduce subjects' sensitivity to the non-spatial impairments if the 3D effect dominates the experience.

Another aspect concerning audiovisual communication is the agreement between the rendered audio scene and the visual scene, especially when it comes to mixtures between 2D and 3D techniques (2D video with 3D audio, or 2D audio with 3D video). In such cases the actual viewing and listening positions are critical factors for the perceived quality and must be carefully controlled.

E.2 Recommendations concerning spatial audio

It is recommended to consult [ITU-T P.1310] for more information on the quality testing of 3D audio technology in telemeeting systems.

[ITU-T P.1310] describes test aspects that need to be considered when testing 3D audio, and it provides a few test protocols for the subjective testing of 3D audio. In addition, [ITU-T P.1310] refers to further ITU-T stand-alone Recommendations, which describe further test methods appropriate for evaluating spatial audio telemeetings. Currently, this includes [ITU-T P.1311] and [ITU-T P.1312].

For test cases that are not covered in [ITU-T P.1310] or its related stand-alone methods, the effects described above should be accounted for as part of careful experiment setup (keywords: listening and viewing positions), careful experiment design (keywords: presentation order and sessions) and careful control of the side aspects (keywords: instructions, questionnaires and scales). Pilot tests may also be conducted to verify the impact of such aspects.

E.3 Remarks concerning 3D video

Currently, there are no 3D video test methods available that have been specifically developed for two-party or multiparty communication systems.

Note that with [ITU-T P.915] and [ITU-T P.916] non-interactive test methods for stereoscopic and multi-view monitors (i.e., 3DTV) are available. These documents may be consulted when the system under test uses such displays, especially since they address visual discomfort and visual fatigue (issues that are also relevant in communication scenarios). However, these methods have not yet been validated for telemeeting scenarios.

Annex F

Assessment of asymmetric multiparty telemeetings

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

F.1 Problem statement

In asymmetric cases, one technical degradation can have different effects for individual interlocutors. As an example, consider a three-party telemeeting visualized in Figure F.1. In this telemeeting, for two of the three interlocutors, IL1 and IL3, the technical quality of the connections is unaffected, but IL2 has a degradation in its send-direction, for instance a problem with the uplink to the network. This degradation has the effect that the other two interlocutors, IL1 and IL3, perceive an impairment of the signals from IL2 and no impairments from each other. IL2 on the other hand does not perceive any impairment as the degradation is in its send-direction and not in its receive-direction. Degradations are indicated in Figure F.1 by red arrows.

In such a case, subjective quality ratings collected from the three interlocutors will actually be based on two different "test conditions":

- a) IL1 and IL3 perceive a condition with an asymmetric impairment (one interlocutor with impaired signals, the other one with no impairments).
- b) IL2 perceives a condition with no impairments.

For that reason, the quality ratings need to be treated differently in order to avoid mixing quality ratings that actually belong to different test conditions.

This requires a method to analyze the technical situation in order to identify which "test conditions" result from that situation for each interlocutor, or in other words, a method to translate the technical degradations into possible perceptual impairments.

This annex provides such a method (clause F.2), extended with a number of suggestions on influencing factors (clause F.3), experiment design (clause F.4), ratings scales (clause F.5) and data analysis (clause F.6).

It should be noted that adhering to this annex is not necessary for symmetric test conditions, i.e., all interlocutors perceive the same degradations. However, it can be applied to symmetric cases as well, which enables both asymmetric and symmetric conditions to be tested with one common procedure.

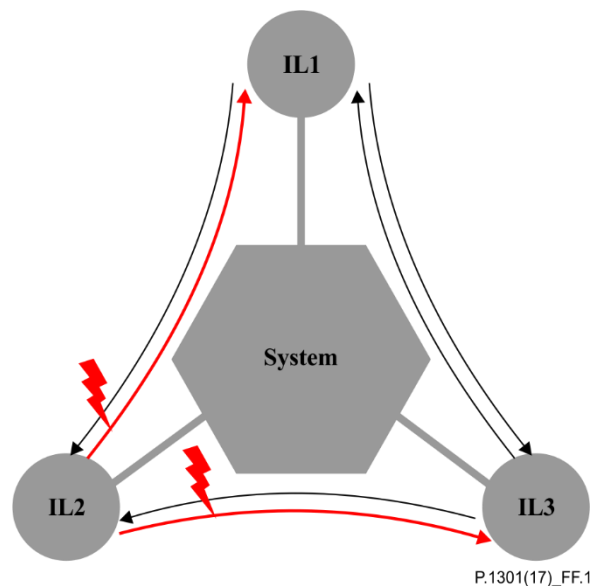


Figure F1 – Visualization of an example of a three-party telemeeting with an asymmetric technical situation [b-Skowronek2017]

F.2 Translation of technical degradations into possible perceptual impairments

This method should be seen as part of the test preparation in order to ensure proper treatment of asymmetric conditions in a subjective test. It should be clarified that this method helps to deduce the perceptual impairments that are theoretically possible given the technical situation. Thus it describes the test conditions from the perspectives of each interlocutor. The degree to which the impairments are actually perceived is the goal of the subjective test.

The general idea is to conduct a signal path analysis, starting with the points and type of the technical degradation and following the degraded signals to each interlocutor.

More specifically, the method consists of four steps:

- 1) Description of the technical situation
- 2) Identification of the Degradation-Types and Degradation-Points
- 3) Analysis of signal paths and deduction of possible perceptual impairments
- 4) Generation of a comprehensive representation for all interlocutors

Step 1: Description of the technical situation

To facilitate the path analysis in Step 3, the technical situation should be described in terms of the signal paths between all interlocutors. As the example in clause F.1 suggests, one important aspect is that the signal paths in both send direction and receive direction are considered separately.

Furthermore, including the topology of the telemeeting (i.e., central-bridge, peer-to-peer, or hybrid) in the description helps, particularly in Step 2.

A graphical visualization such as the one in Figure F.1 may be used for this purpose. Alternatively, other visual, textual, mathematical or algorithmic descriptions may be considered. An example for a visual approach can be found in [b-Skowronek2013] and [b-Skowronek2017].

Step 2: Identification of Degradation-Types and Degradation-Points

In this step the investigator needs to identify what kind of degradations – henceforth referred to as Degradation-Types – are occurring and how they impact the signals. Moreover, the investigator needs to identify at what points between sender and receiver –henceforth referred to as Degradation-Points – those degradations impact the signals.

If this method is applied in a subjective test where full control of the technical degradations is possible, then this step is rather simple: the investigator actually needs to select the Degradation-Types and Degradation-Points when he or she is deciding on the test conditions.

If this method is applied to a subjective test in which a system is assessed in real-life (or close to real-life) operation, then a technical method is required that is capable of measuring the occurrence of any such degradations and to locate them on the way from sender to receiver.

For instance, consider the situation in Figure F.1 with the problem in the uplink of interlocutor IL2. A concrete example of such a case is that the Degradation-Type is packet loss, and the Degradation-Point is on the send-side of the connection from IL2 to the network.

Appendix VII provides a comprehensive description of how these Degradation-Types and Degradation-Points can be determined with the presentation of a more sophisticated graphical representation. However, alternative approaches may also be used if desired.

Step 3: Analysis of signal paths and deduction of possible perceptual impairments

The purpose of this step is to conduct the actual path analysis and to deduce with which perceptual impairment I_{xy} each Interlocutor x is – theoretically – perceived from each Interlocutor y .

For that purpose the task is to proceed first along the individual signal paths from the send-side to the receive-side and to check which Degradation-Points the individual signals are passing by.

Then, the task is to apply technical and perceptual knowledge about the Degradation-Type in order to characterize how the signals are affected by that degradation, and to deduce how this leads to a theoretically possible perceptual impairment.

For example, the Degradation-Type packet loss at the Degradation-Point send-side on the connection from IL2 to the network leads to the following path analysis for all connections between the three interlocutors (Figure F.1):

- I_{11} : Interlocutor IL1 does not perceive any impairments from himself or herself, i.e., no echo or side tone problems or the like
- I_{12} : Interlocutor IL1 perceives some speech distortion (if the packet loss is moderate) from Interlocutor IL2, see corresponding red arrow in Figure F.1
- I_{13} : Interlocutor IL1 does not perceive any impairments from Interlocutor IL3, see corresponding black arrow in Figure F.1
- I_{21} : Interlocutor IL2 does not perceive any impairments from Interlocutor IL1, see corresponding black arrow in Figure F.1
- I_{22} : Interlocutor IL2 does not perceive any impairments from himself or herself, i.e., no echo or side tone problems or the like
- I_{23} : Interlocutor IL2 does not perceive any impairments from Interlocutor IL3, see corresponding black arrow in Figure F.1
- I_{31} : Interlocutor IL3 does not perceive any impairments from Interlocutor IL1, see corresponding black arrow in Figure F.1
- I_{32} : Interlocutor IL3 perceives some speech distortion from Interlocutor IL2, see corresponding red arrow in Figure F.1
- I_{33} : Interlocutor IL3 does not perceive any impairments from himself or herself, i.e., no echo or side tone problems or the like

Step 4: Generation of a comprehensive representation for all interlocutors

The fourth step is to summarize the signal path analysis in a convenient format, for example, a table, in which each table entry refers to one impairment I_{xy} and contains a simple code that characterizes the impairment.

As an example, Table F.1 shows the result for the above mentioned example analysis:

Table F.1 – Example representation for the path analysis according to Step 3

I_{11}	I_{12}	I_{13}	I_{21}	I_{22}	I_{23}	I_{31}	I_{32}	I_{33}
0	<i>pl</i>	0	0	0	0	0	<i>pl</i>	0

Table F.1 uses the following codes : 0 = no impairment, *pl* = speech distortion caused by packet loss.

By using an appropriate coding, such a table provides a quick overview of the dependencies of the impairments seen from the perspective of each interlocutor and the actual technical degradations characterized by the Degradation-Type and Degradation-Point.

This information can then be used to identify which subjects actually perceive the same situation (in the example: IL1 and IL3) or a different situation (in the example: IL2).

This needs to be considered for a proper data analysis (see also clause F.6) and – in some cases – may also be exploited in the experimental design (see clause F.4).

F.3 Influencing factors

So far, this annex concerns the impact and analysis of technical asymmetries. However, there are also many influencing factors that can, in turn, also be asymmetric in a telemeeting. For example:

- Different group sizes, different personalities (i.e., subject profiles)
- Different positions of participants in the rooms (viewing positions, distances from microphones, etc.)
- Different environments (room sizes, acoustics, lighting, interior, etc.)
- Different communication modes (audio, audiovisual) and rendering conditions (2D audio, 3D audio, 2D video, 3D video)
- Different transmission qualities (e.g., varying bit-rate, delay, and transmission error performance)
- Different qualities of capturing and reproduction equipment (e.g., loudspeakers, headphones, microphones, displays, cameras, etc.)

Note that a more detailed list of factors influencing the perceived quality in a telemeeting is suggested in Appendix I.

A further and particularly important influencing factor in terms of asymmetry concerns the interactions of different group sizes and different communication modes at each location.

Participants sitting alone in a room with audiovisual equipment probably spend the majority of time looking at the video screen. Persons in a group room probably do not look at the screen as much, especially if there are lots of people in the room. How the test participants are seated in relation to the screen also matters in a video conference. It is probably easier to include other participants in a discussion in a group room if they turn at least partly towards the video screen and not just towards the other group members.

It is easy to forget to address participants with audio only connections for people sitting in an audiovisual room.

Test subjects know that they cannot be seen without video equipment so they might feel more relaxed in a pure audio condition. They might listen more to the conversation than taking an active part in it.

It is easier to detect when people sitting in the same room want to say something, due to body language and facial expression.

People might perceive the video quality differently depending on the combination of screen size and viewing distance, which usually varies in a multiparty meeting. Also, the audio quality may be better in some parts of the room than others.

F.4 Experiment design

In general, it is recommended to aim for a balanced test, that is, all participants are exposed to all technical degradations under test. For asymmetric conditions this means multiple test calls for the same technical degradation are required because in each test call a different interlocutor of a test participant group is exposed to the technical degradation. Considering the example in Figure F.1, this means three test calls are required: in each call the degradation is in the uplink of another interlocutor.

To achieve this, investigators usually define the test conditions in terms of the technical properties and distribute those conditions across the test participants. This approach, however, may lead to sub-optimal designs for multiparty tests with asymmetric conditions. For example, the situation in Figure F.1 means that for this particular call, only the ratings of interlocutors IL1 and IL3 reflect a telemeeting call with a degradation, but not the ratings of interlocutor IL2 (for that person there are no perceivable technical degradations). Thus, the ratings of one of three interlocutors in such a scenario cannot be used to evaluate the impact of the degradation under test.

Such design inefficiencies can be overcome by *inverting* the design paradigm: Instead of distributing the technical degradations across the interlocutors, which may lead to different perceptual impairments, the investigator distributes those perceptual impairments, which may require smart combinations of technical degradations.

As an example, Table F.2 shows the advantage of such an approach for an efficient combination of two degradations in the end devices of the interlocutors: (a) packet loss in the send-direction and (b) signal distortion in the receive-direction, e.g., poor loudspeaker.

As show in the upper part of Table F.2, a conventional design requires six calls to obtain a balanced design. However, in calls 1 to 3, one interlocutor cannot perceive an impairment at all (call 1: $I_{11} = I_{12} = I_{13} = 0$, call 2: $I_{21} = I_{22} = I_{23} = 0$, call 3: $I_{31} = I_{32} = I_{33} = 0$). Furthermore, in calls 4 to 6, even two interlocutors cannot perceive any impairment at all.

In an *inverted* design, these two technical degradations can be combined as shown in the lower part of Table F.2, which reduces the number of required calls to three. This is possible if one considers that the perceptual impairments of calls 1 and 4 in the conventional design are mutually exclusive in the sense that call 1 leads to no impairments for IL1 but for IL2 and IL3 while call 4 leads to an impairment for IL1 but no impairments for IL2 and IL3.

Table F.2 – Comparison between conventional design (upper part) and inverted design (lower part)

Conventional design												
Call	D_1	D_2	D_3	I_{11}	I_{12}	I_{13}	I_{21}	I_{22}	I_{23}	I_{31}	I_{32}	I_{33}
1	PLS	0	0	0	0	0	pl	0	0	pl	0	0
2	0	PLS	0	0	Pl	0	0	0	0	0	pl	0
3	0	0	PLS	0	0	pl	0	0	Pl	0	0	0
4	DistR	0	0	0	dist	dist	0	0	0	0	0	0
5	0	DistR	0	0	0	0	dist	0	dist	0	0	0
6	0	0	DistR	0	0	0	0	0	0	dist	dist	0

**Table F.2 – Comparison between conventional design (upper part)
and inverted design (lower part)**

<i>Inverted design</i>												
<i>Call</i>	<i>D₁</i>	<i>D₂</i>	<i>D₃</i>	<i>I₁₁</i>	<i>I₁₂</i>	<i>I₁₃</i>	<i>I₂₁</i>	<i>I₂₂</i>	<i>I₂₃</i>	<i>I₃₁</i>	<i>I₃₂</i>	<i>I₃₃</i>
1	PLS + DistR	0	0	0	dist	dist	pl	0	0	pl	0	0
2	0	PLS + DistR	0	0	pl	0	dist	0	dist	0	pl	0
3	0	0	PLS + DistR	0	0	pl	0	0	pl	dist	0	dist

D_x with $x \in [1,2,3]$ refers to the presence of technical degradations in the end device of interlocutor IL_x . I_{xy} refers to the possible perceptual impairment that Interlocutor x perceives from Interlocutor y . Abbreviations used for the technical degradations in device D_x : PLS = packet loss in send-direction, DistR = signal distortion in receive-direction, e.g., poor loudspeaker. Abbreviations used for the perceived impairments: pl = signal distortion due to packet loss, dist = distorted speech signal due to poor loudspeaker.

This example suggests that up to 50% of the test calls can be saved. The actual number of saved calls, however, depends on two aspects: (a) how far the degradations under test can be combined without mutually influencing each other in terms of perceivable impairments, and (b) how far the system under test is technically capable of providing such combinations.

Note that the use of simulated systems allows the realization of independent degradations, which provides more options for combining degradations than the use of real-life systems.

F.5 Rating scales

F.5.1 General advice concerning rating scales

Multiple modes and multiple rendering conditions in asymmetric telemeetings also have implications on the use of rating scales. While most test methods recommended by ITU define only one rating scale, test set-ups with multiple modes or rendering conditions might require the use of several rating scales to catch all aspects of the quality perceived during the test.

As the number of rating scales used in one test should be limited for feasibility reasons, it is recommended to determine that number by means of pilot tests.

F.5.2 Rating scales for different aggregation levels of telemeeting quality

From a perceptual perspective, asymmetric conditions can trigger test participants to consider different *aggregation levels* of the telemeeting in the sense that test participants differentiate between the quality of the individual connections of the interlocutors and the overall *aggregated* quality of the whole telemeeting system.

In the example of Figure F.1, interlocutors IL_1 and IL_3 may differentiate between the quality of the degraded connection from IL_2 (e.g., "*The signal from that one person is distorted, so the quality of that connection is **bad**.*") and the overall system quality (e.g., "*There are some problems with one interlocutor, but for the rest the system is fine, so the quality is **fair**.*").

If it is necessary to account for such effects, it is recommended that participants are asked to rate both aggregation levels on individual scales, i.e., on one scale rate the quality they perceive from the whole telemeeting and on additional scales the quality they perceive from each individual connection of the interlocutors.

In order to help test participants in the differentiation of these two types of question, there are two possibilities that may be considered:

- (a) Use different types of rating scales for each aggregation level:

The advantage is that the different rating scales trigger participants to reflect on the two aggregation levels as two different aspects. The disadvantage is that a scale transformation between the two scales is necessary, if the results between both aggregation levels are to be compared or related.

An example that may be used: The rating scale for overall quality is the 5-point absolute category rating (ACR) scale according to [ITU-T P.800], the ratings scales for the individual connections use the extended continuous (EC) rating scale according to [ITU-T P.851]. There is also a scale transformation available from the EC scale to the ACR scale, which is available in [b-Koester2015]:

$$Q_{ACR} = -0.0262 \cdot Q_{EC}^3 + 0.2368 Q_{EC}^2 + 0.1907 Q_{EC} + 1$$

- (b) Present the rating scales for each aggregation level sequentially, with the overall quality rating first.

One advantage is that test participants do not see both types of rating scales simultaneously on one page, which avoids them forming a visual average across the individual connection ratings for the system as a whole. Another advantage is that the participants are first put into a more utilitarian rating mode (they give an intuitive rating of the overall quality) before they are put into a more analytical rating mode (they judge individual aspects – individual connections – of the system).

An optimal implementation is to use an electronic questionnaire that enforces the rating order and ensures the overall quality rating scale is shown first (and separately) followed by the individual connection rating scales.

F.6 Test stimuli

For running tests according to this annex, the test scenarios in Appendix IV, V and VI may be used.

F.7 Data analysis

As the individual interlocutors can have different perspectives of the same telemeeting call (clause F.1), the data analysis needs to be conducted in such a way that a mixing of ratings is avoided that actually belong – from the perspective of the interlocutors – to different test conditions. In other words, the data analysis may not simply aggregate the rating from all interlocutors in one test call. Instead, a more detailed treatment of the data is necessary which can become quite complex.

This may be avoided by conducting a systematic restructuring of the data before analysis. The task is to obtain sets of data observations that share the same combinations of perceptual impairments on the individual connections. This task consists of two essential steps:

Step 1: Define the sets in terms of "Set x is defined by the fact that one connection has an impairment I_a , one connection I_b , and so on."

Step 2: Assign the ratings per participant and test call to the sets by applying selection criteria that matches the set definitions, for example "Assign rating R to set x , if in the call the test participant had was affected by Impairment I_a on one connection, by Impairment I_b on another connection, and so on."

Practical experience shows that such a restructuring of the data may become laborious and error-prone if this is conducted manually. Therefore, the use of a software-based approach, such as dedicated self-written scripts or database functions, is recommended to perform this task.

The following example explains this procedure in more detail. First, consider the experimental design in Table F.3, which is a reproduction of the lower part of Table F.2. Further assume that each interlocutor ($IL1$, $IL2$, and $IL3$) gives ratings on both quality aggregation levels according to

clause F.5.2. This leads to the ratings $R_{i,x}$ ($i,x \in [1,2,3]$, i : interlocutor, x : call) for the overall quality and $R_{ij,x}$ ($i,j,x \in [1,2,3]$, i,j : interlocutors, x : call) for the individual connection qualities shown in Table F.4. The definitions of possible sets for this example and the corresponding assignment of ratings are shown in Table F.5.

Table F.3 – Example experimental design

<i>Experimental Design</i>												
<i>Call</i>	<i>Technical Degradations</i>			<i>Impairments for IL1</i>			<i>Impairments for IL2</i>			<i>Impairments for IL3</i>		
	D_1	D_2	D_3	I_{11}	I_{12}	I_{13}	I_{21}	I_{22}	I_{23}	I_{31}	I_{32}	I_{33}
1	PLS + DistR	0	0	0	dist	dist	pl	0	0	pl	0	0
2	0	PLS + DistR	0	0	pl	0	dist	0	dist	0	pl	0
3	0	0	PLS + DistR	0	0	pl	0	0	pl	dist	0	dist

D_x with $x \in [1,2,3]$ refers to the presence of technical degradations in the end device of interlocutor IL_x . I_{xy} refers to the possible perceptual impairment that Interlocutor x perceives from Interlocutor y . Abbreviations used for the technical degradations in device D_x : PLS = packet loss in send-direction, DistR = signal distortion in receive-direction, e.g., poor loudspeaker.

Table F.4 – Overview of ratings that would be collected in an experiment with a design according to Table F.3

<i>Call</i>	<i>Overall Ratings from</i>			<i>Individual connection ratings for</i>								
	$IL1$	$IL2$	$IL3$	I_{11}	I_{12}	I_{13}	I_{21}	I_{22}	I_{23}	I_{31}	I_{32}	I_{33}
1	$R_{1,1}$	$R_{2,1}$	$R_{3,1}$	$R_{11,1}$	$R_{12,1}$	$R_{13,1}$	$R_{21,1}$	$R_{22,1}$	$R_{23,1}$	$R_{31,1}$	$R_{32,1}$	$R_{33,1}$
2	$R_{1,2}$	$R_{2,2}$	$R_{3,2}$	$R_{11,2}$	$R_{12,2}$	$R_{13,2}$	$R_{21,2}$	$R_{22,2}$	$R_{23,2}$	$R_{31,2}$	$R_{32,2}$	$R_{33,2}$
3	$R_{1,3}$	$R_{2,3}$	$R_{3,3}$	$R_{11,3}$	$R_{12,3}$	$R_{13,3}$	$R_{21,3}$	$R_{22,3}$	$R_{23,3}$	$R_{31,3}$	$R_{32,3}$	$R_{33,3}$

Further, the ratings shown refer to the overall system quality as perceived by each interlocutor and to the quality of each individual connection, see clause F.5.2.

Table F.5 – Overview of the different sets of ratings that are possible with the example experiment according to Tables F.3 and F.4.

<i>Sets for overall ratings</i>			
<i>Set</i>	<i>Selection criterion</i>	<i>Interpretation</i>	<i>Assigned ratings</i>
1	$R_{i,x} / I_{ia} = 0 \ \& \ I_{ib} = dist \ \& \ I_{ic} = dist$ with $a = i$: own connection, $b, c \neq i$ connections to interlocutors	All ratings in which an interlocutor perceived no impairment of their own connection and the terminal-related impairment for the two interlocutors.	$R_{1,1},$ $R_{2,2},$ $R_{3,3}$
2	$R_{i,x} / I_{ia} = 0 \ \& \ I_{ib} = 0 \ \& \ I_{ic} = pl$ with $a = i$: own connection, $b, c \neq i$ connections to interlocutors	All ratings in which an interlocutor perceived no impairment of their own connection, no impairment on the connection of one interlocutor, and the packet loss-related impairment on the connection of the other interlocutor.	$R_{2,1}, R_{3,1},$ $R_{1,2}, R_{3,2},$ $R_{1,3}, R_{2,3}$
<i>Sets for individual connection ratings</i>			
<i>Set</i>	<i>Selection criterion</i>	<i>Interpretation</i>	<i>Assigned ratings</i>
1a	$R_{ia,x} / I_{ia} = 0 \ \& \ I_{ib} = dist \ \& \ I_{ic} = dist$ with $a = i$: currently considered connection is the own connection, $b, c \neq i$ connections to interlocutors	All ratings of their own unimpaired connection, under the constraints that the other two connections have the impairment <i>dist</i> .	$R_{11,1}$ $R_{22,2}$ $R_{33,3}$
1b	$R_{ia,x} / I_{ia} = dist \ \& \ I_{ib} = 0 \ \& \ I_{ic} = dist$ with $a \neq i$: currently considered connection is a connection to an interlocutor, $b = i$: own connection, $c \neq i$: connection to remaining interlocutor	All ratings of a connection to an interlocutor with the impairment <i>dist</i> , under the constraints that their own connection has no impairment and the remaining connection also has the impairment <i>dist</i> .	$R_{12,1}, R_{13,1}$ $R_{21,2}, R_{23,2}$ $R_{31,3}, R_{32,3}$
2a	$R_{ia} / I_{ia} = 0 \ \& \ I_{ib} = 0 \ \& \ I_{ic} = pl$ with $a = i$: currently considered connection is the own connection, $b, c \neq i$ connections to interlocutors	All ratings of their own unimpaired connection, under the constraints that one connection to an interlocutor has no impairment and the connection to the other interlocutor has the packet loss-related impairment <i>pl</i> .	$R_{22,1}, R_{33,1}$ $R_{11,2}, R_{33,2}$ $R_{11,3}, R_{22,3}$
2b	$R_{ia} / I_{ia} = 0 \ \& \ I_{ib} = 0 \ \& \ I_{ic} = pl$ with $a \neq i$: currently considered connection is a connection to an interlocutor, $b = i$: own connection, $c \neq i$: connection to remaining interlocutor	All ratings of a connection to an interlocutor with no impairment, under the constraints that their own connection has no impairment and the other connection has the packet loss-related impairment <i>pl</i> .	$R_{21,1}, R_{31,1}$ $R_{12,2}, R_{32,2}$ $R_{13,3}, R_{23,3}$
2c	$R_{ia} / I_{ia} = pl \ \& \ I_{ib} = 0 \ \& \ I_{ic} = 0$ with $a \neq i$: currently considered connection is a connection to an interlocutor, $b = i$: own connection, $c \neq i$: connection to remaining interlocutor	All ratings of a connection to an interlocutor with the packet loss-related impairment <i>pl</i> , under the constraints that their own connection has no impairment and the other connection has no impairment.	$R_{23,1}, R_{32,1}$ $R_{13,2}, R_{31,2}$ $R_{12,3}, R_{21,3}$

Annex G

Assessment of multiparty telemeetings with non-stationary quality

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

There are two evaluation paradigms to assess the non-stationary quality of communication systems: (a) subjects rate the quality continuously during the duration of a call and an overall rating is computed afterwards, and (b) subjects rate the *call quality* after the call, whereas the test protocol aims to ensure that the rating considers the whole call.

Concerning the continuous rating paradigm, non-interactive test methods are available for video and audio quality.

For video, a recommendation currently available on how to generate a common rating over time is the single stimulus continuous quality evaluation (SSCQE) method specified in [ITU-R BT.500]. For audio, [ITU-T P.880] deals with continuous evaluation of time varying speech quality.

Concerning the call quality paradigm, [ITU-T P.1302] describes test protocols to evaluate non-stationary audio-only or audiovisual calls.

As these methods have not yet been tested for the assessment of multiparty telemeetings, pilot testing is recommended to check if an adaptation of these methods is necessary or if these methods can be directly applied in non-interactive multiparty tests.

Note that there are no such recommendations available for the evaluation of conversational quality, neither for conventional one-to-one conversations, nor for multiparty conversations.

Annex H

Assessment of multiparty telemeetings using multi-dimensional scaling methods

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

Since overall quality can be considered as a multi-dimensional attribute, multi-dimensional assessment methods could provide detailed insights on the individual aspects that constitute the overall quality.

For speech quality, a non-interactive test method that collects multi-dimensional ratings from test participants is [ITU-T P.806]. However, this method has not yet been validated for multiparty situations.

The ITU Handbook on Practical procedures [b-ITU-T HB-PPST] for subjective testing lists a number of distortions that should be covered by multi-dimensional scaling methodologies. Hence the philosophy is to assess individual contributions of distortions to quality by means of multi-dimensional scaling methodologies.

Concerning multiparty assessment, the benefit of using such multi-dimensional methodologies is to be able to assess the individual contributions of the multiparty specific aspects described in the present Recommendation. Examples are the influence of conversational situation, asymmetries, multiple modes and multiple rendering conditions, etc.

A fixed set of dimensions for the assessment of multiparty telemeetings has not been investigated and determined yet.

In case multi-dimensional scaling techniques are required, pilot tests are recommended for identifying the most appropriate dimensions before running the full assessment test.

Appendix I

Influential factors

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

There are many factors that can affect telemeeting quality. They can be different for different participants also in a point-to-point meeting, but the situation gets more complicated if there are several participants using different types of equipment (terminal devices and connections). The telemeeting participants can use the same type of equipment (symmetrical) or different types of equipment (asymmetrical). Often some properties are symmetrical while others are not.

It can be expected that some of the factors outlined below interact with each other and that they affect the subjective quality in a complex way. Many factors also vary in importance depending on the circumstances. For example, the importance of some factors can be different for two-party versus multiparty conversations.

Several factors that can affect the perceived quality of telemeetings are listed below. Several of these factors can vary in existing systems, but can also be changed for test reasons.

More information on influencing factors can be found for instance in [b-Reiter2014].

Speech/audio

- Audio capturing: The quality of the audio capture devices. The placement and number of microphones affects the signal-to-noise-ratio and the sound quality.
- Type and quality of the codec (or codecs if there is transcoding). The audio bandwidths can vary (NB, WB, SWB, and FB) as well as the bit rate. There could be different types of noise cancellation or echo cancellation or no such cancellation. Voice activity detection and comfort noise can also be present.
- Talkers can have different gender, age, pitch and level range, voice timbre, and use different languages. Their speech can be more or less intelligible.
- Audio levels can vary.
- Audio rendering: Number of, type of and placement of loudspeakers for mono, stereo or spatial audio. Stereo rendering can be used in a multiparty call even if the participants are captured in mono, e.g., the participants could be spread spatially to improve the possibility to distinguish between them. A mobile phone with or without headsets could also be used in a telemeeting.
- Different room acoustics, reverberation, background noise characteristics

Video

- Video capturing: Different quality of the video capture device. Placement of cameras in relation to the telemeeting participants.
- Codec type, bit rate, rate control (fixed or variable), frame rate, video resolution, video content
- Different video background (colours on wall and clothes) and viewing room colour.
- Video rendering: Different types of screens: Multiple screens – one screen – no screen – 3Dscreens. The layout of the videos of the participants on the screens can be different. Is it possible to see all participants? How fast is the switching of the current talker, if implemented in the system, and how does that influence the quality perception?
- Viewing distance
- Room illumination

- Agreement between the rendered audio scene and the visual scene
- Synchronization of audio and video

Communication channel/Network quality

- A communication channel can have different capacity and transport properties. There could be different types of transmission impairments such as packet loss and jitter. The end-to-end delay can be different for audio and video and lead to bad audio-video synchronization.
- Equipment from different vendors may require different hardware set-up and network connections with different properties.

Participants

- There can be one or several participants at two or more sites. Group dynamics can influence the quality perception. The situation is dependent on the degree of acquaintance of the participants. There could be a discussion leader in a structured meeting.
- Personality, for instance the tendency to dominate a conversation, and the current state of an interlocutor, for instance to be in a particularly good mood.
- Previous experience and expectations. Different cultures. Personal opinion.
- Participants can have different hearing and viewing abilities. They can be trained or naïve listeners and/or viewers.

Usability

- Ease of use
- Efficiency
- Connection time when establishing a call
- Different collaboration equipment can be possible to use.
- Participants can have different hearing and viewing abilities. They can be trained or naïve listeners and/or viewers.

Service quality

- Availability
- Reliability
- Security

Context

- Laboratory environment or field test setting
- Business or private use-case

Price

- The price sets an anchor for the expected quality. For a more expensive service people most likely expect a higher quality or additional features in the service.

Appendix II

Overview of multiparty non-interactive test stimuli and conversation test tasks

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

This appendix gives a brief overview on available multiparty stimuli for non-interactive tests and multiparty conversation tasks for conversation tests.

II.1 Non-interactive audio-only stimuli:

For the assessment of listening-only speech quality, unrelated sentences from various speakers are recommended in [ITU-T P.800].

As it is discussed and recommended in the present Recommendation, material consisting of recorded multiparty telemeetings might better address the multiparty-specific situation for non-interactive listening tests. Such recordings could stem from real-life telemeetings or from conversations based on the different conversation tasks mentioned below.

Accordingly, the recording set-up defined in [ITU-T P.800] needs to be properly adapted to allow the recording of multiple interlocutors.

II.2 Non-interactive video-only stimuli

For the assessment of viewing-only quality, different sequences with various content are recommended in [ITU-T P.910].

As it is discussed and recommended in the present Recommendation, material consisting of recorded multiparty telemeetings might better address the multiparty-specific situation for non-interactive viewing tests.

Such recordings could stem from real-life audiovisual telemeetings or from conversations based on the different audiovisual conversation tasks mentioned below.

Accordingly the recording set-up defined in [ITU-T P.910] needs to be properly adapted to allow the recording of multiple interlocutors.

II.3 Non-interactive audiovisual stimuli

For the assessment of non-interactive audiovisual quality, no specific content is recommended in [ITU-T P.911].

As it is discussed and recommended in the present Recommendation, material consisting of recorded multiparty telemeetings might better address the multiparty-specific situation for non-interactive audiovisual tests.

Such recordings could stem from real-life audiovisual telemeetings or from conversations based on the different audiovisual conversation tasks mentioned below.

Accordingly the recording set-up defined in [ITU-T P.911] (by referring to [ITU-T P.800] and [ITU-T P.910]) needs to be properly adapted to allow the recording of multiple interlocutors.

II.4 Audio-only conversation tasks

[ITU-T P.805] provides examples for a number of conversation tasks for two-party conversations: short conversation test scenarios, Richard's task, random number verification, interactive short conversation test scenarios. These could in principal be extended to multiple parties.

Appendix IV of the present Recommendation provides advice to generate conversation test scenarios for three parties. [b-ITU-T P-Sup. 26] gives examples for such three-party conversation tests (3CTs) in English, French and German.

Free conversations as described in Appendix III are also suitable for audio-only conversations.

II.5 Audiovisual conversation tasks

Most test tasks for audiovisual conversations that are suggested in [ITU-T P.920] are made for two-party conversations, not multi-party telemeetings. If the same test task is to be used for both audio and audiovisual telemeetings, the subjects should be able to keep their focus on the screen and should not need to read from a paper during a large part of the test.

The free conversation test task, described in Appendix III makes it possible to look at the screen during the conversation.

Appendix V presents three more possible tasks (Survival task, Leavitt task, Brainstorming task), from which the Survival task is suggested to be most appropriate and feasible.

Appendix VI presents additional scenarios of the Survival task and presents one more possible task, a celebrity name guessing game.

Appendix VII presents short descriptions of additional scenarios covering formal and informal telemeetings as well as distance learning scenarios.

Appendix III

Examples of multiparty conversation test tasks (audio-only and audiovisual): Free conversation

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

Free conversation can be used for both audio and audiovisual tests, since there is no need to read a written instruction during the test. In a free conversation other talkers can be interrupted spontaneously.

The recommended duration of a conversation varies depending on the number of test participants. If there are only a few participants, about one and a half minutes should be added per participant to get a feasible test length. If there are many participants the time per person might need to be reduced in order to keep the total test length reasonable. One minute per participant could be suitable for around six participants.

To facilitate the discussion topic a paper with topic suggestions could be handed out. Nevertheless, participants should still be free to choose their own topic.

The test participants should be instructed to completely avoid talking about the quality and properties of the system under test as this could influence their assessment of the system. They should also be told to try to divide the speech activity equally between themselves if that is the required type of conversation.

The test questions can be similar to the questions suggested in [ITU-T P.920] but might vary slightly depending on the test set-up and purpose of the test. An example of test questions is shown at the end of this appendix.

In free conversations the amount of speech for every participant is not controlled by the test task. The amount of speech can therefore vary depending on the conversational situation, for instance, be dependent on the used equipment or the number of interlocutors that are in the same room together with a person. If the audio of a meeting can be recorded, those files can be analyzed objectively according to [b-Hoeldtke].

If the aim is to create a business-like conversation, more controlled scenarios should be used because in business meetings there is usually an agenda and subjects are prepared for what they intend to contribute.

Example of test methodology for a free conversation test with six participants

The test subjects are given the possibility to read the written instructions with the test question as soon as they arrive for the test. The instructions are also given orally with all participants in the same room, if possible. The test subjects should also have the possibility to ask questions regarding the test methodology. It is recommended that the test subjects give a short presentation of themselves if they do not know each other. Some kind of game could be played to make the test participants more familiar with each other. Afterwards, the test participants move to the test rooms to try out the test equipment. Every test participant gets the opportunity to hear all participants through the conference system (and see them if it is an audiovisual conversation test). Participants can get familiar with the test procedure in a training session. It is recommended that the test leader is present during the training session as questions may arise.

The test participants should be instructed to completely avoid talking about the quality and properties of the system as this could influence their assessment of the system. They should also be asked to try to divide the speech activity equally between them.

At the beginning of every conversation one person might start a timer. After six minutes (in the case of six participants) the test subjects are required to finish the conversation and vote on the perceived quality. All test participants should be able to see the timer limiting the time to give a vote on a particular scale. After the voting a new conversation with different settings can be started.

In this example four questions were used in the test. The display of the voting terminal can be seen in Figure III.1 below:

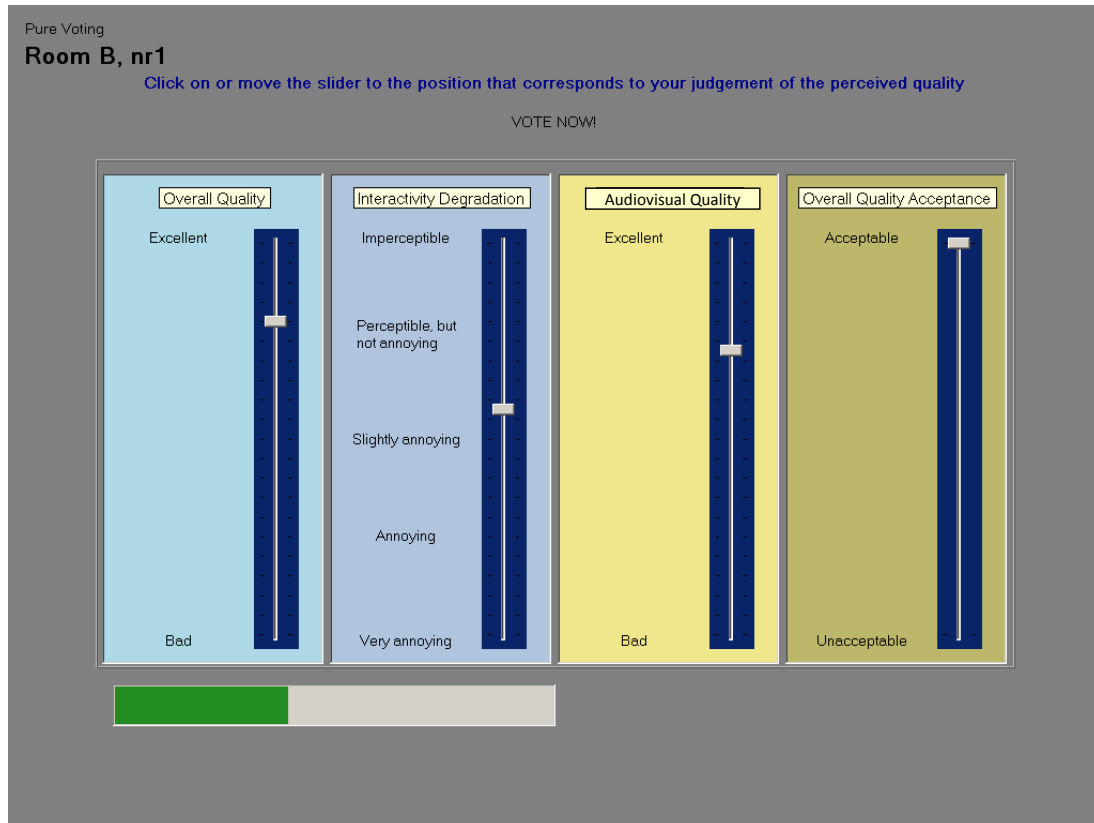


Figure III.1 – The voting terminal display for an audiovisual test

These questions were to be answered at the end of each test conversation:

- 1) How do you judge the overall quality of the communication?
Continuous voting scale with the labels "Bad" and "Excellent" at the endpoints.
- 2) Did you perceive any reduction in your ability to interact during the conversation?
Continuous voting scale with the labels "Imperceptible", "Perceptible but not annoying", "Slightly annoying", "Annoying", and "Very annoying" marked.
- 3) How do you judge the audiovisual quality during the conversation?
Continuous voting scale with the labels "Bad" and "Excellent" at the endpoints.
- 4) Was the quality acceptable or unacceptable?

The answer alternatives were only two, acceptable or unacceptable

The test participants were asked to mark their ratings on the rating scales on the voting terminals after every conversation. They were also asked to write comments on paper during the test. Furthermore, each participant was interviewed shortly after completing the entire test and asked whether they had any comments or observed anything special that they wanted to mention. Finally, the whole test group was asked some complementary questions.

Appendix IV

Examples of multiparty conversation test tasks (audio-only): Three-party conversation test scenarios 3CTs

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

IV.1 Introduction

To assess the conversational quality of telemeetings, it is necessary to involve the conversation partners in an appropriate conversation task. For classical two-person conversations, different types of conversation tasks have been proposed (see [b-Raake2006] for a summary). Tasks range from free conversations, interactive games, jointly solving certain military tasks, to finding locations on city maps, identifying differences between two versions of pictures, proofreading of texts, and the rapid exchange of random numbers. The main shortcomings of many of such test scenarios is that they either reduce the naturalness of the assessment situation or they lack a common conversational structure enhancing comparability between conversations.

To overcome these shortcomings short conversation test scenarios (SCTs) were developed for two-party conversations [ITU-T P.805]. Inspired by these scenarios this appendix describes a procedure to develop corresponding conversation test scenarios for three-parties.

IV.2 Test scenario development

Following the SCT-development, the following set of requirements was used for the 3CTs developed here (see [b-MOELLER, pp. 75]):

- Naturalness (subject and environment), i.e., natural conversation tasks, a natural beginning and end of the conversation, and a natural, limited distraction from the quality-perception and -judgement task.
- Balance (conversation flow), i.e., no fixed sender- and receiver-roles, short periods of monologues, realistic amount of double- or triple-talk, same repartition of speech activity between participants, and a limited overall duration.
- Comparability (between scenarios), i.e., similar instructions, comparable dialogue-structures, similar overall durations.

The intended dialogue flow is schematically depicted in Figure IV.1.

To develop the conferencing scenarios, a four-step procedure may be used that was found to be feasible and resulted in properly balanced scenarios:

- 1) Identify appropriate conferencing topics, by asking a group of subjects who have experience with telemeetings to list at least three of their most frequent topics during business conferences, and another three for conferences that they considered realistic for a spare-time conferencing application (note that the latter type of conferences are only slowly coming up, e.g., with free IP-based audio- and video-conferencing tools). This could be done by emailing or brainstorm sessions. Then conduct a workshop, in which participants with sufficient experience in telemeeting usage as well as the investigator rate the collected topics in terms of their suitability for the envisaged goals.

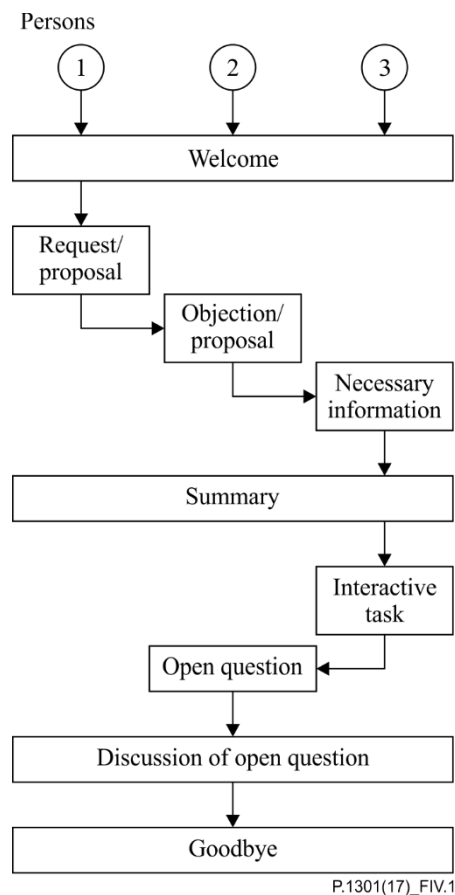


Figure IV.1 – Targeted conversation structure

- 2) Formulate actual scenarios for the identified topics.
- 3) Ask – in different sessions – volunteers to carry out conversations using the scenarios, to test their principle usability and to identify problems of dialogue flow or comprehension. However, instead of conversing over an actual technical system, have the participants seated in the same test-room, with card box wall separators between them to avoid visual interaction or distraction. Ideally two supervisors observe conversation flow and duration.
- 4) Refine and simplify the scenarios according to the pre-test results in step 3. The layout of the scenarios follows that of the two-person SCTs [b-Moeller, pp. 75]. In the case of the 3CTs, each scenario is captured by two sheets per interlocutor. The first sheet is identical for all participants and briefly outlines the overall situation in which the conversation takes place, the actual topics to be discussed, and the roles and names of the participants. The second sheet is individual for the three interlocutors, and consists of a mix of pictograms that indicate the type and function of the information to follow, short instructions, and tabulated data. The participants dispose of complementary information necessary to complete the conversation task. Example topics for the business scenarios are the planning of a business meeting, selection of titles for a new music CD compilation and the organization of an arts exhibition.

IV.3 Scenario validation

To validate the generated scenarios, either in a pilot before or as a post-analysis after the actual conversation test, the following validation methods may be applied to evaluate for overall duration per scenario and per conferee group:

- Duration deviations of scenarios: Error bar plots (means and 95% confidence intervals) showing the conversation durations for the individual scenarios
- Deviations between interlocutor groups: Error bar plots showing conversation durations as a function of the subject groups that participated in the text

- Duration and interlocutor groups: Two-factorial analysis of variance (ANOVA) using interlocutor group and the scenario as fixed factors
- Experiment conditions: One-factorial ANOVA with condition as fixed factor

Additional conversational analyses may be performed, for instance as described in [b-Hoeldtke].

IV.4 Cultural aspects

In order to ensure that test participants experience these scenarios as naturally as possible, the scenarios should fit to their cultural environment. Hence, scenario themes, items to be discussed, names of people, objects and locations, conventions regarding addresses and telephone numbers, any other cultural reference, needs to be adopted accordingly to the cultural context.

That means if existing scenarios, such as the examples for Germany, France and the U.S. in [b-ITU-T P-Sup.26] are to be translated into other languages, not only literal translation, but also the adaptation of such cultural references is required.

Appendix V

Examples of multiparty conversation test tasks (audiovisual): Audiovisual multi-point tasks for three parties (Survival task, Leavitt task, Brainstorming task)

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

V.1 Overview and most suitable task

To define a suitable task, the following guidelines are provided in [ITU-T P.920]:

- 1) The task should be designed so that, during their conversation, the subjects primarily maintain their attention on the audiovisual terminal;
- 2) The task must resemble real-life audiovisual communication to a sufficient degree.
- 3) It is preferable that the task is, in itself, sufficiently rewarding for the subjects. This has several advantages: the subjects learn the task faster and they are less susceptible to fatigue and loss of motivation.
- 4) Familiarity between pairs of conversing participants is highly desirable, if not essential.
- 5) A wide range of subjects should be able to perform the task.
- 6) In addition, for each tested condition, the conversation should last at least five minutes.

Ideally, the conversation resulting from the task achievement should be highly interactive to be sensitive to delay as well as rich in terms of audio and video content so that impairments on speech and/or video (coding artefacts, packet losses, desynchronization, etc.) can be perceptible.

[ITU-T P.920] describes a number of tasks but they are defined for only two people and they are not suited to assess audiovisual systems in multipoint configuration. Generally the tasks do not resemble real-life audiovisual communication.


The ideal audiovisual multiparty task would result in natural conversations that are highly interactive so as to be sensitive to delay. It should also be rich in terms of audio and video content so that impairments on speech and/or video can be perceptible, and it should be possible to maintain visual attention on screen so that video impairments can be viewed.

Three potential tasks are: Leavitt task, Brainstorming task, and Survival task. Though none of the three tasks was found to answer all mentioned requirements, the Survival task achieved the best compromise in terms of visual attention, number of speech turns, naturalness and satisfaction (easy, interesting, etc.).

Therefore the Survival task is recommended to assess the audiovisual quality of videoconferencing system in multipoint configuration (three people) and in a natural situation of discussion.

Four survival test scenarios can be found in clause V.4. Additional and modified survival test scenarios can be found in the Appendix VI.

V.2 Leavitt task

The Leavitt task [b-Leavitt], initially planned for five participants, has been modified for three participants. In the modified version, participants have sheets on which appear five shapes chosen among thirteen. The objective is to find the common shape, as illustrated in Figure V.1. The common shape in the given example is  .

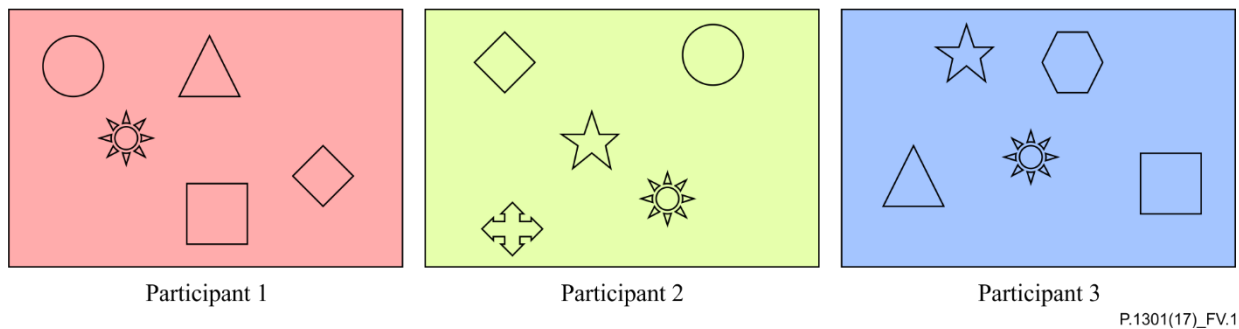


Figure V.1 – Example of set of sheets used for the Leavitt task

In order to avoid weariness and learning, four versions with thirteen different sheets for each participant have been prepared. All in all, fifty-two sets with different shapes combinations are available (with a total of 156 different sheets, one set being constituted of three sheets).

The task ends after five minutes or when the thirteen sheets given for one conversation are finished.

This Leavitt task is supposed to be a simple task that can be reproduced at infinity, with a quite light written material.

V.3 Brainstorming task

In the Brainstorming task, the objective is to produce the maximum of ideas on the proposed subject, respecting the four rules defined by Osborn [b- Osborn] to reduce social inhibitions among group members, stimulate idea generation and increase overall creativity of the group:

- you must give the maximum of ideas;
- you must give unusual ideas;
- you must not criticize the ideas of other participants;
- you must improve the ideas of other participants.

Four subjects of brainstorming are chosen: the tourists (give ideas to encourage American tourists to visit Europe [b-Taylor]), the additional thumb (imagine that people who will be born after 2050 will have an additional thumb on the opposite side of the first one. What would be the consequences – benefits or difficulties- for these people? [b-Taylor]), the box (give all the ideas you have to use a box?), the environment (give all the ideas you have to protect the environment).

The task ends after 5 minutes or when there is no idea left.

The Brainstorming task is similar to a free conversation proposed by [ITU-T P.920]. Without any written material, this task seems to be favourable to a visual attention focused on the screen.

V.4 Survival task

The Survival tasks were developed to explore the performance characteristics of a decision-making group. In their initial version [b-Hall], participants are invited to imagine themselves in a survival situation based on an accident (plane, space rocket, etc.). They have a list of fifteen items left intact and undamaged after landing. First, participants are asked to rank order them in terms of their importance for the crew. Secondly, they are asked to do it together in order to find a consensual ranking. Finally, they are asked to individually do the ranking again, in order to compare the individual rankings, before and after the group discussion. The initial version lasts one or two hours. In order to reduce the task duration, the task was limited to a group discussion with the goal to select six objects useful for the group survival. In addition, the initial 15-item or 12-item list was divided into three 5-item or 4-item lists, one for each participant, in order to avoid a long list per participant (that could require him/her to read the list many times during the discussion) and to force all

participants to speak. The 5-item or 4-item lists were also illustrated with photographs to help participants to identify some uncommon objects and to speed up the memory recall (to avoid that people take too much time to look at their sheet).

The task naturally ends when the list of the six objects is approved by all participants and recapitulated. The Survival task has the advantages of creating natural turn exchanges, with a quite light written material. Examples are given below, based on four survival tasks: in winter [b-Johnson], at sea [b-Nemiroff], on the moon [b-Hall] and in desert [b-Johnson]. Pictures are given as examples.

First instruction to participants





You are going to achieve a decision task with your partners. You will find a brief description of the context which you are in, you and your partners, as well as a list of objects. You have to choose six objects in the list that will help you to survive. Be careful, your lists are different. So share your objects, then discuss with your partners and come to an agreement on the objects to be selected, by justifying your choice. The group has agreed to stick together.

Scenario 1: Survival Task in winter

Participant 1

You have just crash-landed in the North of Canada. The small plane in which you were travelling has been completely destroyed except for the frame. The pilot and co-pilot have been killed, but no one else is seriously injured.

You are in a wilderness area, snow-covered and made up of thick woods broken by many lakes and rivers. The pilot announced shortly before the crash that you were eighty miles northwest of a small town that is the nearest known habitation. It is mid-January. The last weather report indicated that the temperature would reach minus twenty-five degrees in the daytime and minus forty at night. You are dressed in winter clothing appropriate for city wear—suits, pantsuits, street shoes, and overcoats. While escaping from the plane, your group salvaged the items listed below.





Ball of steel wool	
Extra shirt and pants for each survivor	
A little axe	
A strong sheet (6 m × 6 m)	

Scenario 1: Survival task in winter

Participant 2

You have just crash-landed in the North of Canada. The small plane in which you were travelling has been completely destroyed except for the frame. The pilot and co-pilot have been killed, but no one else is seriously injured.

You are in a wilderness area, snow-covered and made up of thick woods broken by many lakes and rivers. The pilot announced shortly before the crash that you were eighty miles northwest of a small town that is the nearest known habitation. It is mid-January. The last weather report indicated that the temperature would reach minus twenty-five degrees in the daytime and minus forty at night. You are dressed in winter clothing appropriate for city wear – suits, pantsuits, street shoes, and overcoats. While escaping from the plane, your group salvaged the items listed below.





Loaded .45-calibre pistol	
Sectional air map made of plastic	
Margarine in a big iron box	
Quart of 85-proof whiskey	

Scenario 1: Survival task in winter

Participant 3

You have just crash-landed in the North of Canada. The small plane in which you were travelling has been completely destroyed except for the frame. The pilot and co-pilot have been killed, but no one else is seriously injured.






You are in a wilderness area, snow-covered and made up of thick woods broken by many lakes and rivers. The pilot announced shortly before the crash that you were eighty miles northwest of a small town that is the nearest known habitation. It is mid-January. The last weather report indicated that the temperature would reach minus twenty-five degrees in the daytime and minus forty at night. You are dressed in winter clothing appropriate for city wear – suits, pantsuits, street shoes, and overcoats. While escaping from the plane, your group salvaged the items listed below.

Newspaper (one per person)	
Compass	
Cigarette lighter without the fluid	
Family-sized chocolate bar (one per person)	

Scenario 2: Survival task at sea

Participant 1






You are drifting in a private yacht in the South Pacific. A fire with unknown origin has destroyed much of the yacht, notably navigational and radio equipment. After having controlled the fire, you realize that the boat is sinking little by little. Your best estimate is that you are many hundreds of miles from the nearest landfall. You and your friends have managed to save 15 items, undamaged and intact after the fire. In addition, you have salvaged a four man rubber life craft and a box of matches.

A sextant	
A small transistor radio	
A shaving mirror	
20 square feet of Opaque plastic sheeting	
A quantity of mosquito netting	

Scenario 2: Survival task at sea

Participant 2

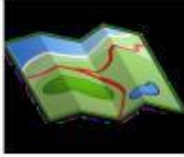




You are drifting in a private yacht in the South Pacific. A fire with unknown origin has destroyed much of the yacht, notably navigational and radio equipment. After having controlled the fire, you realize that the boat is sinking little by little. Your best estimate is that you are many hundreds of miles from the nearest landfall. You and your friends have managed to save 15 items, undamaged and intact after the fire. In addition, you have salvaged a four man rubber life craft and a box of matches.

A 20 litre container of water	
One bottle of 160 per cent proof rum	
A case of army rations	
15ft nylon rope	
A can of shark repellent	

Scenario 2: Survival task at sea

Participant 3






You are drifting in a private yacht in the South Pacific. A fire with unknown origin has destroyed much of the yacht, notably navigational and radio equipment. After having controlled the fire, you realize that the boat is sinking little by little. Your best estimate is that you are many hundreds of miles from the nearest landfall. You and your friends have managed to save 15 items, undamaged and intact after the fire. In addition, you have salvaged a four man rubber life craft and a box of matches.

A map of the Pacific Ocean	
2 boxes of chocolate bars	
A floating seat cushion	
A fishing kit	
A 7 litre can of oil/petrol mixture	

Scenario 3: Survival task on the moon

Participant 1






You are a member of a space crew originally scheduled to rendezvous with a mother ship on the lighted surface of the moon. However, due to mechanical difficulties, your ship was forced to land at a spot some 200 miles from the rendezvous point. In addition to your space suit, your crew has managed to save items left intact and undamaged after landing. Your task is to take the items which allow you to reach the mother ship.

Food concentrate	
Stellar map	
50 feet of nylon rope	
One case of dehydrated milk	
Portable heating unit	

Scenario 3: Survival task on the moon

Participant 2






You are a member of a space crew originally scheduled to rendezvous with a mother ship on the lighted surface of the moon. However, due to mechanical difficulties, your ship was forced to land at a spot some 200 miles from the rendezvous point. In addition to your space suit, your crew has managed to save items left intact and undamaged after landing. Your task is to take the items which allow you to reach the mother ship.

Magnetic compass	
Three signal flares	
Box of matches	
Parachute silk	
Solar-powered FM receiver-transmitter	

Scenario 3: Survival task on the moon

Participant 3





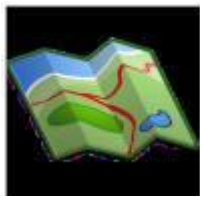
You are a member of a space crew originally scheduled to rendezvous with a mother ship on the lighted surface of the moon. However, due to mechanical difficulties, your ship was forced to land at a spot some 200 miles from the rendezvous point. In addition to your space suit, your crew has managed to save items left intact and undamaged after landing. Your task is to take the items which allow you to reach the mother ship.

First aid kit	
A torch	
Two 100 lb. tanks of oxygen	
Two .45 calibre pistols	
20 litres of water	

Scenario 4: Survival task in desert

Participant 1






You have just crash-landed in the Sonora desert in the south-west of United States. The pilot and co-pilot have been killed in the crash. However, the pilot announced that before impact you were approximately 110 km off the course of the flight plan. He also indicated that that you were 113 km southwest of a mining camp which is the nearest known habitation. The surrounding desert is made up of sand dunes and seems dry except for some cactus. The last weather report indicated that the temperature at the ground level will be about 45°C. All of you are dressed in light clothes – cotton shirts, trousers, socks, and soft shoes. Before the crash, your group was able to save some items.

Torch with 4 battery-cells	
Bottle of 1000 salt tablets	
Folding knife	
1 litre of water per person	
Air map of the area	

Scenario 4: Survival task in desert

Participant 2






You have just crash-landed in the Sonora desert in the south-west of United States. The pilot and co-pilot have been killed in the crash. However, the pilot announced that before impact you were approximately 110 km off the course of the flight plan. He also indicated that that you were 113 km southwest of a mining camp which is the nearest known habitation. The surrounding desert is made up of sand dunes and seems dry except for some cactus. The last weather report indicated that the temperature at the ground level will be about 45°C. All of you are dressed in light clothes – cotton shirts, trousers, socks, and soft shoes. Before the crash, your group was able to save some items.

Plastic raincoat (large size)	
A cosmetic mirror	
Magnetic compass	
Sunglasses (for everyone)	
A book entitled 'Desert Animals That Can Be Eaten'	

Scenario 4: Survival task in desert

Participant 3

You have just crash-landed in the Sonora desert in the south-west of United States. The pilot and co-pilot have been killed in the crash. However, the pilot announced that before impact you were approximately 110 km off the course of the flight plan. He also indicated that that you were 113 km southwest of a mining camp which is the nearest known habitation. The surrounding desert is made up of sand dunes and seems dry except for some cactus. The last weather report indicated that the temperature at the ground level will be about 45°C. All of you are dressed in light clothes – cotton shirts, trousers, socks, and soft shoes. Before the crash, your group was able to save some items.

First-aid kit	
2 litres of 180 proof liquor	
.45 calibre pistol (loaded)	
Overcoat (for everyone)	
Parachute (red and white)	

Appendix VI

Additional proposals for multiparty conversation test tasks (audiovisual):

Extended survival tasks scenarios and celebrity name guessing task

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

VI.1 Overview and background

Appendix V proposes four scenarios based on the survival task. In order to allow quality evaluation tests with more than four conditions per participant group, similar additional scenarios are required. This appendix provides three additional scenarios, originally developed in [b-Skowronek2017] for the German language.

In addition, all resulting seven scenarios were harmonized in the sense that the number of items for each participant in each scenario is equalized to four items – which was not the case of the scenarios in Appendix V. Also a few items were replaced that sometimes triggered confusion with test participants, thus such minor modifications are intended to avoid comments such as "This item does not make sense here". Therefore this appendix provides modified item lists for the four scenarios of Appendix V.

Furthermore, this appendix proposes another conversation task: Multiparty celebrity name guessing task. It was originally proposed for two-party testing in [b-Schoenenberg2014] for the subjective testing of transmission delay, and extended in [b-Skowronek2017] for the multiparty case.

This new task is proposed to augment the survival scenarios in order to test more than seven conditions with one subject group, while at the same time limiting the number of calls for each group with the same type of task (survival or celebrity name guessing). The motivation is to avoid negative effects on the mood of test participants, such as boredom or depressive feelings when running too many survival tasks, or frustration when running too many name guessing rounds if one is not particularly good in guessing those names. Successful results were achieved in [b-Skowronek2017] when test participants conducted the two tasks in an interleaved manner, which resulted in positive feedback of the test participants.

In this context, it needs to be clarified that the conversation structure of the two tasks is quite different. Thus a combination of both tasks in one test may only be applied if the conversational structure is not a critical factor of the test at hand, which for instance would be the case of delay.

If an investigator decides to have the two tasks in a test, a data analysis needs to be conducted (e.g., a t-test) in order to check for significant differences in the quality ratings between the two tasks. If there are no differences found, then a common data analysis may be considered, otherwise the data of the two tasks need to be analysed separately or the conversation task need to be properly taken into account (e.g., as fixed factor when conducting ANOVAs).


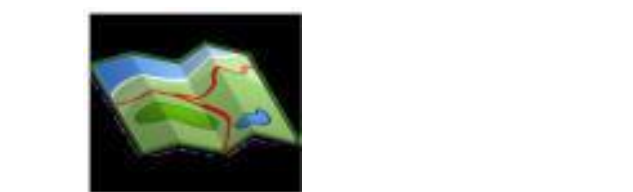


VI.2 Modification of survival scenarios from Appendix V

The description of the scenarios in Appendix V are not changed. If the test requires a better harmonization of the list of items of Appendix V, the following lists are proposed.





Scenario 1 Survival task in winter, Participant 1

Extra shirt and pants for each survivor	
A little axe	
A strong sheet (6 m × 6 m)	
Binoculars	





Scenario 1 Survival task in winter, Participant 2

Loaded .45-calibre pistol	
Sectional air map made of plastic	
Quart of 85-proof whiskey	
A camping kettle with mounting	





Scenario 1 Survival task in winter, Participant 3

Newspaper (one per person)	
Compass	
Cigarette lighter without the fluid	
A thermos bottle	

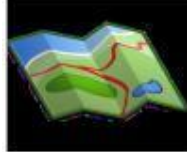



Scenario 2 Survival task at sea, Participant 1

A sextant	
A small transistor radio	
A shaving mirror	
20 square feet of opaque plastic sheeting	


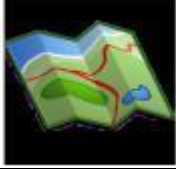


Scenario 2 Survival task at sea, Participant 2

A 20 litre container of water	
One bottle of 160 per cent proof rum	
A case of army rations	
15ft nylon rope	





Scenario 2 Survival task at sea, Participant 3

A map of the Pacific Ocean	
2 boxes of chocolate bars	
A fishing kit	
A 7 litre can of oil/petrol mixture	





Scenario 3 Survival task on the moon, Participant 1

Food concentrate	
Stellar map	
50 feet of nylon rope	
One case of dehydrated milk	





Scenario 3 Survival task on the moon, Participant 2

One battery set	
Three signal flares	
4m ² Parachute silk	
Solar-powered FM receiver-transmitter	





Scenario 3 Survival task on the moon, Participant 3

First aid kit	
A torch	
Two 100 lb. tanks of oxygen	
20 litres of water	





Scenario 4 Survival task in desert, Participant 1

Bottle of 1000 salt tablets	
Folding knife	
1 litre of water per person	
Air map of the area	

Scenario 4 Survival task in desert, Participant 2

Plastic raincoat (large size)	
A cosmetic mirror	
Magnetic compass	
A book entitled 'Desert Animals That Can Be Eaten'	

Scenario 4 Survival task in desert, Participant 3

First-aid kit	
.45 calibre pistol (loaded)	
Overcoat (for everyone)	
Parachute (red and white)	

VI.3 Three additional survival scenarios





Scenario 5, Survival task in the mountains

Scenario description (for each participant)





On a mountain hike, an avalanche occurred that destroyed most of your base camp. You and two fellow travellers could climb out of the snow. Unfortunately the avalanche carried away your mountain guide and you could not find him. Your base camp is located a three day's walk from the next village, on a height which does not require oxygen supply. The weather is good, but you need to expect temperatures below -15°C . You are all wearing winter cloths and mountain boots.

Not all of the camping gear was buried, you could find some items.





List of items for participant 1

Camping tent	
A case of army rations	
Climbing gear with reserve helmet	
A Camping kettle with mounting	

List of items for participant 2

A mountaineer's axe	
An animal trapping	
Binoculars	
A thermos bottle	

List items for participant 3

A sleeping mat	
Folding knife	
Box of matches	
A fishing kit	

Scenario 6, Survival task in the swamp

Scenario description (for each participant)

On a boat expedition on the Amazonas, there was an explosion in the boat's engine room in the early morning. Engine and controls are totally destroyed and there is a significant leak in the hull. Fortunately the boat was quickly taken ashore overnight and nobody was injured, as the cabins were at the end of the boat away from the engine room. You are in an inhabited area of the Amazonas, you past by the last village 1.5 days ago. You are wearing appropriate clothing for the jungle and it is rainy season. Most of the equipment was thrown overboard, but you could find some items.



List of items for participant 1

.45 calibre pistol (loaded)	
Plastic raincoat (large size)	
A funnel	
A 7 litre can of oil/petrol mixture	

List items for participant 2

Box of matches	
Machete	
One bottle of 160 per cent proof rum	
A pair of rubber boots	

List item for participant 3

A quantity of mosquito netting	
A set of filter paper	
Three towels	
A camping stool	

Scenario 7, Survival task in the cave labyrinth





Scenario description (for each participant)

You were on a tour through a large cave labyrinth with underground lakes for several days now. A tunnel crashed down, you and two fellow travelers were separated from the tour guide and the rest of the group. You also lost all possibilities to contact the others. Fortunately no one is injured and your protective clothes and helmet lamps are in good condition. The evening before, the tour guide explained that there are several exits from the cave, but it takes several days to reach those and one needs to cross some of the underground lakes. On your side of the collapsed tunnel you could find some items.





List of items for participant 1

Folding shovel	
Hammer	
20 square feet of opaque plastic sheeting	
Food concentrate	

List of items for participant 2

<p>Portable rubber boat</p>	
<p>Rope ladder</p>	
<p>A map of the cave labyrinth</p>	
<p>Three signal flares</p>	

List of items for participant 3

<p>Mine lamp</p>	
<p>Climbing gear with reserve helmet</p>	
<p>A simple diving equipment</p>	
<p>Spray can with fluorescent paint</p>	

VI.4 Celebrity name guessing task

A) Game instructions

Goal

In each round, each player is a celebrity and each player needs to find out which celebrity he or she is. The player that can guess most celebrities wins the game.

Game preparation

In front of you, you find large and small paper cards.

The large paper cards show the initials of the celebrity that you are and whose name you need to find out. There is also a small number written on the card which you can share with the other players, so they can look for the right answer.

Please fix these cards on your clothes directed to the camera, such that your game partners can see them well on their screens.

The small cards show the numbers and names of the celebrities, which your game partners are playing. Each game partner has a different colour.

Game play

In order to guess who you are, you may only ask questions that can be answered with *Yes* or *No*. You may ask the questions as long as the answer is *Yes*.

In case of a *No*, you need to stop and it is the next player's turn.

If you have found out your name, you get a point and can continue with a new large card, i.e., the initials of a new celebrity.

Hint:

In case the game partners, who have the answers, do not know that celebrity, you need to continue with another card/celebrity, which everybody knows.

End of game

Please play the game for 4-5 minutes until the test supervisor signals the end of the game and stops the call.

Please keep the cards as they are so that you may continue with the game in case there is another call.

B) Remarks on the choice of celebrities for the game

The celebrities should be chosen so that there is a high chance that participants know these people.

Therefore the celebrities should be taken from the cultural background of the test participants. This may also include international stars, as long as they are well known in the cultural background (e.g., country) of the test participants. Furthermore, the age of the participants should also be taken into account.

For that reason, a fixed list of celebrities cannot be given here.

Concerning the required number of celebrities, successful experiments have been conducted using 20 celebrities per test participant for tests with seven calls.

Appendix VII

Additional proposals for multiparty conversation test tasks (audiovisual):

Formal and informal multiparty video conferences

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

Additional conversation tasks for audiovisual telemeetings may contain all kinds of scenarios, including those described in [b-ITU-T F.733] for multimedia conferences and those described in [b-ITU-T F.742] for distance learning.

Accordingly the following test scenarios are proposed:

1) Formal multiparty video conferences

There are different kinds of formal conferences. In one kind there are participants who speak more than the others due to a specifically important function, such as in teacher-centric distance learning. In another kind all participants are equally contributing to the conversation and interactive communication is more important. The conference durations are often comparatively long. Possible examples for such scenarios combined with different set-ups are:

- a) A telemeeting scenario with one dominant interlocutor, similar to a face-to-face conference presentation:
One person is giving a presentation, while the others are first following and then asking questions.
- b) A telemeeting scenario similar to a face-to-face panel discussion:
Steered by a formal discussion leader, participants equally contribute to the telemeeting.
- c) Scenarios a) or b) with additional data media such as movie sequences, chat etc.
- d) Scenarios a) or b) with a large number of locations (e.g., 20) attending the telemeeting.

Such scenarios could serve as use cases for testing professional conference rooms with high-quality equipment and very good conditions of communication channels and networks. Or the scenarios could be tuned to variable communication channels and networks, in which some of the participants are in video conference rooms while others might be in normal rooms using PC or mobile equipment.

2) Informal multiparty video conferences

Compared to a formal conference that is often scheduled, informal conferences can begin more spontaneously. These cases could serve to test setups in which the communication channels and networks and terminals are variable. These test scenarios should pay more attention to the interactive discussion and the feelings of participants. Possible examples of such scenarios are:

- a) Temporary business discussion
In this scenario, participants coming from different locations can use PCs as conference terminals in their offices. The focus is the issues discussed, speaker identification and eye-contact. Sometimes exchange of additional data media is needed.
- b) Friends/family members chatting
In this scenario, participants often share their photos or video information during chatting.

Appendix VIII

Overview of documents describing suitable test methods

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

The purpose of this Recommendation is to conduct a multiparty test by selecting the most appropriate standardized baseline test method, applying the most appropriate multiparty-specific modifications, and consulting additional documents for specific test cases. This appendix gives a corresponding overview of all cited baseline test methods (see clause VIII.1), multiparty modifications (see clause VIII.2), and additional documents (see clause VIII.3).

VIII.1 Baseline test methods on which the current Recommendation is based

Recommendation ITU-T P.800 "describes methods and procedures for conducting subjective evaluations of transmission quality". This Recommendation addresses audio-only communication between two interlocutors located at two sites for both non-interactive and conversational quality, though the focus is on non-interactive quality.

Recommendation ITU-T P.805 complements ITU-T P.800 by providing more detailed information for conducting two-party conversation tests to evaluate audio-only communication quality.

Recommendation ITU-R BS.1116 "is intended for use in the assessment of audio systems which introduce impairments so small as to be undetectable without rigorous control of the experimental conditions and appropriate statistical analysis."

Recommendation ITU-R BS.1534 describes a "method for the subjective assessment of intermediate audio quality. This method mirrors many aspects of Recommendation ITU-R BS.1116" and is specifically designed to assess intermediate impairments at the lower end of the quality scale, while ITU-R BS.1116 "is used for the evaluation of high quality audio systems having small impairments."

Recommendation ITU-T P.920 describes "interactive test methods for audiovisual communications". In general ITU-T P.920, addresses point-to-point and multipoint scenarios. However, most of the suggested tasks are more suitable for two-party communications.

Recommendation ITU-T P.911 "describes non-interactive subjective assessment methods for evaluating the one-way overall audiovisual quality for multimedia applications".

Recommendation ITU-R BT.710 concerns the subjective assessment methods for image quality in high-definition television and recommends "that subjective assessment of image quality of high-definition television systems should be made following the general methodology given in Recommendation ITU-R BT.500"

Recommendation ITU-R BT.500 "provides methodologies for the assessment of" television "picture quality including general methods of test, the grading scales and the viewing conditions." Although the document is targeted to a very specific use case, it is often referred to even for other than TV-type video quality assessment and can be used – as ITU-R BT.710 suggests – for high-definition television systems. Hence it is applicable to non-interactive video quality of telemeeting systems.

Recommendation ITU-T P.910 "describes non-interactive subjective assessment methods for evaluating the one-way overall video quality for multimedia applications such as videoconferencing, storage and retrieval applications, telemedical applications, etc." This Recommendation concerns only non-interactive experiments with video-only stimuli. In the context of conferencing assessment, it is suited for instance for dedicated quality tests on the video signal quality of a telemeeting system.

Recommendation ITU-R BS.1285 "is based on Recommendation ITU-R BS.1116." ITU-R BS.1285 introduces "a pre-selection methodology that can reliably reject systems introducing large impairments" in order to avoid carrying out the stringent tests of ITU-R BS.1116 unnecessarily.

Recommendation ITU-R BT.1788 "specifies non-interactive subjective assessment methods for evaluating the video quality of multimedia applications. These methods can be applied for different purposes including, but not limited to: selection of algorithms, ranking of audiovisual system performance and evaluation of the video quality level during an audiovisual connection." Hence it complements ITU-R BT.500.

VIII.2 Multiparty specific recommendations to adapt the baseline test methods

Recommendation ITU-T P.1301 Annex A gives a general description for a set-up of a multiparty telemeeting assessment test and serves as the main guideline.

Recommendation ITU-T P.1301 Annexes B to H address specific topics concerning multiparty telemeeting assessment: text- and graphics-based communication (e.g., web conferencing), video-only communication, assessment methods for the influence of delay, 3D audio and 3D video reproduction, asymmetries, non-stationary quality, and multi-dimensional methods.

VIII.3 Further test methods, which are referred to in the annexes of this Recommendation

Recommendation ITU-T P.806 describes a non-interactive test method to obtain multidimensional ratings for speech stimuli.

Recommendation ITU-T P.913 describes non-interactive test methods when collecting audio, video or audiovisual quality ratings from tests that are conducted in arbitrary environments.

Recommendation ITU-T P.915 describes a non-interactive test method for 3D television devices.

Recommendation ITU-T P.916 describes a non-interactive test method for measuring visual discomfort and visual fatigue when watching 3D television.

Recommendation ITU-T P.1302 describes a "method for assessing the quality of simulated speech or audio-visual telephony calls with time-varying transmission conditions".

Recommendation ITU-T P.1305 describes the effect of delay on telemeetings and provides guidelines for the analysis of that effect. In particular, this document augments ITU-T P.1301 Annex D.

Recommendation ITU-T P.1310 describes a few test protocols on the subjective quality assessment of spatial audio telemeeting systems. It also refers to additional test methods that are suitable for spatial audio systems, which are at the time of writing the current version of this Recommendation ITU-T P.1311 and ITU-T P.1312.

Recommendation ITU-T P.1311 presents a method to obtain an objective measure of how well a telemeeting system allows users to follow a conversation when talk spurts of several talkers coincide. The method comprises a listening-only test that involves test participants listening to several concurrent talkers, identifying one of them, and reporting what that talker said.

Recommendation ITU-T P.1312 Recommendation ITU-T P.1312 describes a test method for quantifying the effectiveness of telemeeting systems in conveying information in multiparty conversation scenarios. This method utilizes a predefined set of tasks designed to provoke rapid turn-taking and concurrent talking among participants. The method measures the rate at which multiple participants exchange information to assess the effectiveness of communication systems compared to face-to-face communication.

Recommendation ITU-T P.1501 "describes subjective testing methods for assessing the user perceived quality for web browsing in browser-based applications of different device classes". Thus, this Recommendation "can be used to identify the impact of several different factors that influence the user perceived quality of web browsing".

Bibliography

- [b-ITU-T F.733] Recommendation ITU-T F.733 (2005), *Service description and requirements for multimedia conference services over IP networks*.
- [b-ITU-T F.742] Recommendation ITU-T F.742 (2005), *Service description and requirements for distance learning services*.
- [b-ITU-T HB-PPST] ITU Handbook on Practical procedures for subjective testing, International Telecommunications Union, ISBN 92-61-13791-1, 2011.
- [b-ITU-T H-Sup.1] ITU-T H-Series Recommendations – Supplement 1 (1999), *Application profile – Sign language and lip-reading real-time conversation using low bit rate video communication*.
- [b-ITU-T P-Sup.26] ITU-T P-Series Recommendations – Supplement 26 (2017), *Scenarios for the subjective evaluation of audio and audiovisual multiparty telemeeting quality*.
- [b-Hall] Hall, J., Watson, W.H., *The Effects of a Normative Intervention on Group Decision-Making Performance*. Human Relations 23, 299, 1970.
- [b-Hoeldke] Hoeldke, K., and Raake, A., *Conversation analysis of multi-party conferencing and its relation to perceived quality*, IEEE International Conference on Communications ICC, 2011.
- [b-Johnson] Johnson, David W., Johnson, Roger T. (1994), *Learning together and alone: Cooperative, competitive, and individualistic learning*, 4th edition, Boston, Allyn and Bacon.
- [b-Koester2015] F. Köster, D. Guse, M. Wältermann, S. Möller (2015), *Comparison between the Discrete ACR Scale and an Extended Continuous Scale for the Quality Assessment of Transmitted Speech*. In: Fortschritte der Akustik (DAGA2015) – 41. Jahrestagung für Akustik (German Annual Conference on Acoustics). Nürnberg, Germany, Mar.
- [b-Leavitt] Harold J. Leavitt (1960), *Task ordering and organizational development in the common target game*, Behavioral Science, Volume 5, Issue 3, pages 233-239.
- [b-Moeller] S. Möller (2000), *Assessment and Prediction of Speech Quality in Telecommunications*, USA–Boston, Kluwer Academic Publishers.
- [b-Nemiroff] Nemiroff, P.M. and Pasmore, W.A. (1975), *Lost at Sea*. The 1975 Annual Handbook for Group Facilitators, University Associates, Inc., 28-30.
- [b-Osborn] Osborn, Alex Faickney (1940), *Applied imagination: Principles and procedures of creative problem solving*, New York, NY, Charles Scribner's Sons.
- [b-Raake2006] A. Raake (2006), *Speech Quality of VoIP – Assessment and Prediction*, Chichester, West Sussex, UK: John Wiley & Sons Ltd.
- [b-Raake2008] A. Raake and C. Schlegel. (2008), *Auditory assessment of conversational speech quality of traditional and spatialized teleconferences*, in Proc. 8th ITG Conference Speech Communication, to appear, DE-Aachen.
- [b-Raake2010] A. Raake, C. Schlegel, K. Hoeldtke, M. Geier, J. Ahrens (2010), *Listening and conversational quality of spatial audio conferencing*, in Proc. 40th AES Conference on Spatial Audio, Tokyo, Japan.

- [b-Reiter2014] Ulrich Reiter, Kjell Brunnström, Katrien De Moor, Mohamed-Chaker Larabi, Manuela Pereira, Antonio Pinheiro, Junyong You and Andrej Zgank (2014), *Factors Influencing Quality of Experience*. In: *Quality of Experience – Advanced Concepts, Applications, Methods*. Ed. by Sebastian Möller and Alexander Raake. Springer, pp. 55-72.
- [b-Schoenenberg2014] Schoenenberg, K., Raake, A., Lebreton, P. (2014), *Conversational Quality and Visual Interaction of Video-Telephony under Synchronous and Asynchronous Transmission Delay*, Sixth International Workshop on Quality of Multimedia Experience (QoMex).
- [b-Skowronek2013] J. Skowronek, J. Herlinghaus, A. Raake (2013), *Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments*, in Proc. 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), Lyon, France, Aug., pp. 2604–2608.
- [b-Skowronek2017] J. Skowronek, *Quality of Experience of Multiparty Conferencing and Telemeeting Systems – Methods and Models for Assessment and Prediction*, Chapter 6 "Methodology for the Perceptual Assessment of Telemeeting Quality", PhD Thesis, TU Berlin, Germany.
https://depositonce.tu-berlin.de/bitstream/11303/6252/5/skowronek_janto.pdf
- [b-Taylor] D.W. Taylor, P.C. Berry & C.H. Block (1958), *Does group participation when using brainstorming facilitate or inhibit creative thinking*, *Administrative Science Quarterly*, Vol. 3, pp. 23-47.

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