

International Telecommunication Union

**ITU-T**

TELECOMMUNICATION  
STANDARDIZATION SECTOR  
OF ITU

**P.1301**

(07/2012)

SERIES P: TERMINALS AND SUBJECTIVE AND  
OBJECTIVE ASSESSMENT METHODS

Telemeeting assessment

---

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

Recommendation ITU-T P.1301



ITU-T P-SERIES RECOMMENDATIONS

**TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS**

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 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
<b>Telemeting assessment</b>	<b>Series</b>	<b>P.1300</b>
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400

*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 than two 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 it 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
1.0	ITU-T P.1301	2012-07-14	12

## 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 2013

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

	<b>Page</b>
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 .....	4
5 Conventions .....	4
6 General Recommendations concerning the subjective quality evaluation of multiparty telemeetings .....	5
7 Multiparty specific aspects in subjective quality evaluation .....	5
8 Guidance to suitable test methods .....	6
8.1 Test method decision criteria.....	6
8.2 Flow charts to be used when selecting test methods .....	7
Annex A – Set-up of a multiparty telemeeting assessment test.....	13
A.1 Assessment of conversational quality – Conversation tests .....	13
A.2 Assessment of non-interactive quality – Non-interactive tests .....	15
Annex B – Assessment of telemeetings with text-based communication and graphical information means (e.g., web conferencing) .....	17
Annex C – Assessment of video-only telemeetings.....	19
Annex D – Effect of transmission delays on telemeeting quality.....	20
D.1 Background.....	20
D.2 Existing test task Recommendations .....	20
D.3 Recommended test tasks.....	21
D.4 Set-up of a delay test .....	21
D.5 Test subjects .....	22
D.6 Training session.....	22
D.7 Instructions .....	22
D.8 Test questions .....	22
D.9 Objective measurements.....	23
D.10 Effects of delay.....	23
Annex E – Assessment of 3D audio and 3D video reproduction of multiparty telemeetings .....	25
Annex F – Assessment of asymmetric multiparty telemeetings .....	26
F.1 Overview .....	26
F.2 Interactions of different group sizes and different communication modes ....	26
F.3 Remarks concerning the experiment design.....	26
F.4 Remarks concerning scales.....	27

	<b>Page</b>
F.5 Remarks concerning data analysis.....	27
Annex G – Assessment of multiparty telemeetings with non-stationary quality.....	28
Annex H – Assessment of multiparty telemeetings using multi-dimensional scaling methods.....	29
Appendix I – Influential factors .....	30
Appendix II – Overview of multiparty non-interactive test stimuli and conversation test tasks .....	32
II.1 Non-interactive audio-only stimuli.....	32
II.2 Non-interactive video-only stimuli.....	32
II.3 Non-interactive audiovisual stimuli .....	32
II.4 Audio-only conversation tasks .....	32
II.5 Audiovisual conversation tasks .....	33
Appendix III – Examples of multiparty conversation test tasks (audio-only and audiovisual): Free conversation.....	34
Appendix IV – Examples of multiparty conversation test tasks (audio-only): Three-party conversation test scenarios (3CTs) .....	37
IV.1 Introduction .....	37
IV.2 Test scenario development .....	37
IV.3 Scenario validation .....	38
IV.4 Cultural aspects .....	39
Appendix V – Examples of multiparty conversation test tasks (audiovisual): Audiovisual multi-point tasks for three parties (Survival task, Leavitt task, Brainstorming task)..	40
V.1 Overview and most suitable task.....	40
V.2 Leavitt task .....	40
V.3 Brainstorming task.....	41
V.4 Survival task .....	41
Appendix VI – Additional proposals for multiparty conversation test tasks (audiovisual): Formal and informal multiparty video conferences.....	54
Appendix VII Overview of documents describing suitable test methods.....	55
Bibliography.....	57

## **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 a proper 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 than two, 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 not 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 types of equipment (asymmetric). 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 might be needed to execute.

**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	Audio	ITU T P.800, ITU-T P.880 ITU-R BS.1116, ITU-R BS.1285, ITU-R BS.1534
	Video	Video	ITU-T P.910 ITU-R BT.500, ITU-R BT.710, ITU-R BT.1788
	Audiovisual	Audio	ITU-T P.800, ITU-T P.880 ITU-R BS.1116, ITU-R BS.1285, ITU-R BS.1534
		Video	ITU-T P.910 ITU-R BT.500, ITU-R BT.710, ITU-R BT.1788
		Audiovisual	ITU-T P.911
	Text	Video	
	Graphical	Video	
	Audio	Audio	ITU-T P.800, ITU-T P.805
	Video	Video	–
	Audiovisual	Audio	ITU-T P.800, ITU-T P.805
Conversational/Interactive		Video	–
		Audiovisual	ITU-T P.920
	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.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.920] Recommendation ITU-T P.920 (2000), *Interactive test methods for audiovisual communications*.
- [ITU-R BS.1116] Recommendation ITU-R BS.1116-1 (10/97), *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*.
- [ITU-R BS.1285] Recommendation ITU-R BS.1285 (10/97), *Pre-selection methods for the subjective assessment of small impairments in audio systems*.
- [ITU-R BS.1534] Recommendation ITU-R BS.1534-1 (01/03), *Method for the subjective assessment of intermediate quality levels of coding systems*.
- [ITU-R BT.500] Recommendation ITU-R BT.500-13 (01.12), *Methodology for the subjective assessment of the quality of television pictures*.
- [ITU-R BT.710] Recommendation ITU-R BT.710-4 (11.98), *Subjective assessment methods for image quality in high-definition television*.
- [ITU-R BT.1788] Recommendation ITU-R BT.1788 (01/07), *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 communication mode:** The mode that the system under test is providing for the communication between telemeeting participants. It can be audio-only, video-only (for hearing-impaired) or audiovisual.

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

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

**3.2.4 multiparty:** More than two persons. Example: More than two persons are participating in a telemeeting, having a conversation, performing a test task together, etc. The term multiparty does not specify if the persons are distributed across two or more locations. If not explicitly stated differently, multiparty implicates that the persons are at two or more than two 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 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 may be 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.

**3.2.6 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.

**3.2.7 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.

**3.2.8 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 interlocutors are present, as there are more than two persons.

**3.2.9 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, though both provide the same communication mode "audio".

**3.2.10 single-party:** One person. Example: Single-party test task means a test in which individual persons are performing the test tasks on their own, e.g., viewing-only test. Note that only non-interactive tests can comprise single-party test tasks.

**3.2.11 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.

**3.2.12 telemeeting participant, interlocutor:** A person taking part in a conversation between people via a telemeeting system. In a conversation test, telemeeting participants and test participants are the same persons; 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 persons.

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

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

**3.2.15 two-party:** Two persons. Example: Two persons are participating in a telemeeting, having a conversation, performing a test task together, etc. If not explicitly stated differently, two-party implicates that the persons are at two locations.

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

## 4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

3CTS	Three-party conversation test scenario
MOS	Mean Opinion Score
QoE	Quality of Experience
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 the present Recommendation.

The selection and – if necessary – adaptation of an appropriate test method may be done using the guidelines provided in the present 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 (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 persons 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 possibility to identify individual speakers is important 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, also the rendering conditions can influence the importance of such aspects. Regarding the example mentioned, 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 and different network 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 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.

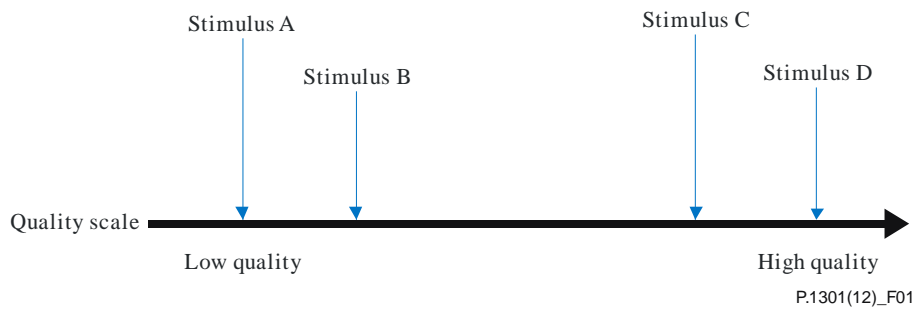
As guidance to suitable test methods, the following subclauses provide a list of decision criteria to be considered by the investigator and flowcharts using these criteria pointing at the appropriate text documents.

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 in the system's audio, video or audiovisual quality?
- 3) Type of quality: Is the main interest in 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 instance 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 size displays (e.g., smartphones, tablets) or normal and large size 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 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 (MOS)), or in multidimensional values (e.g., multidimensional scaling methods)?



Range of quality differences between stimuli:

- If an experiment uses stimuli A and B, or C and D, then the range of differences is small.
- If an experiment uses stimuli A and C, or B and D, then the range of differences is large.

Operation range on quality scale:

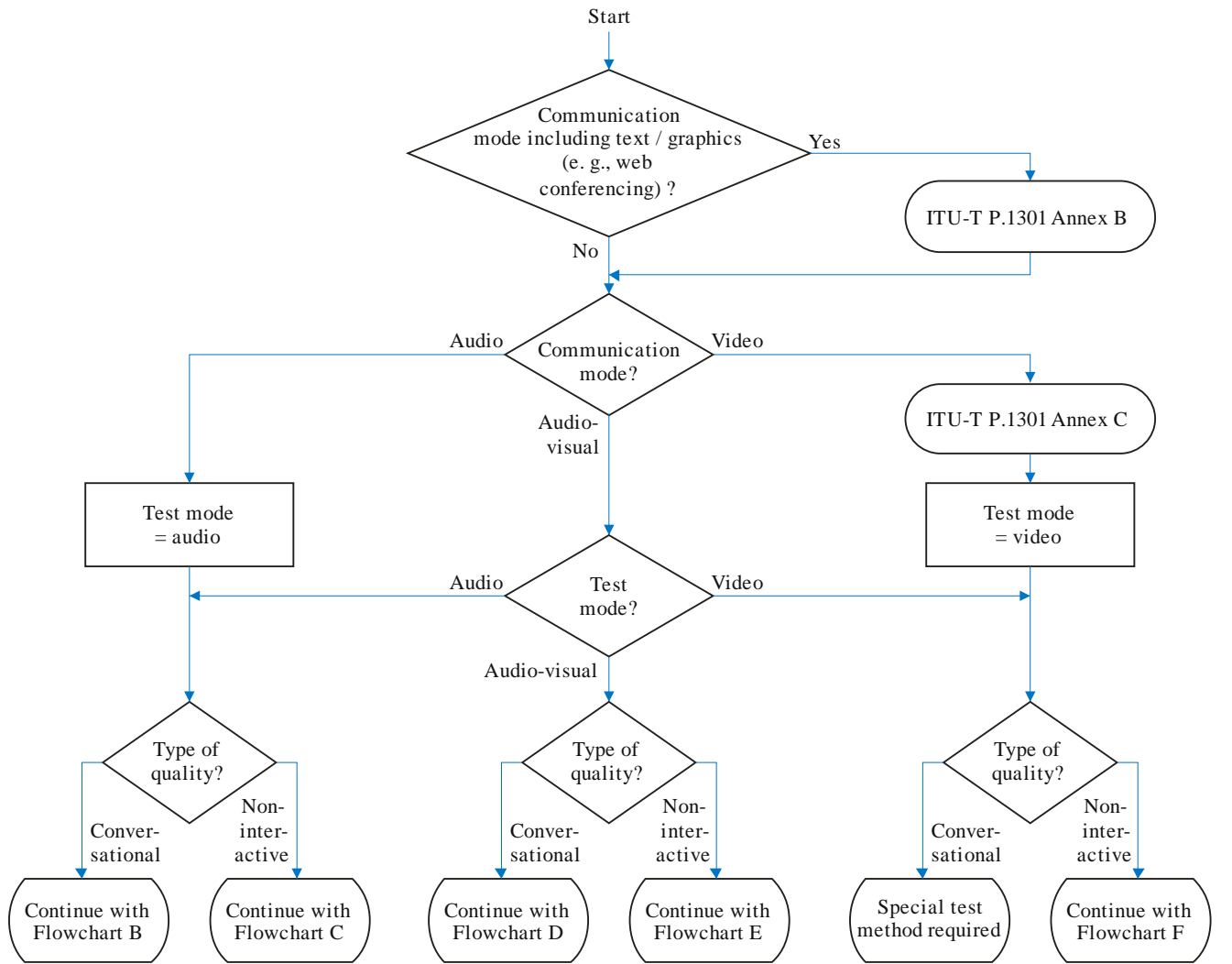
- If an experiment uses stimuli A and B, then the operation range of the experiment is on the lower part of the quality scale.
- If an experiment uses stimuli C and D, then the operation range of the experiment is on the higher part of the quality scale.

NOTE – For a small range of differences, the notion of an operation range is reasonable and can differ between experiments. For a large range of differences, the notion of an operation range is not reasonable since large parts of the scale are covered.

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

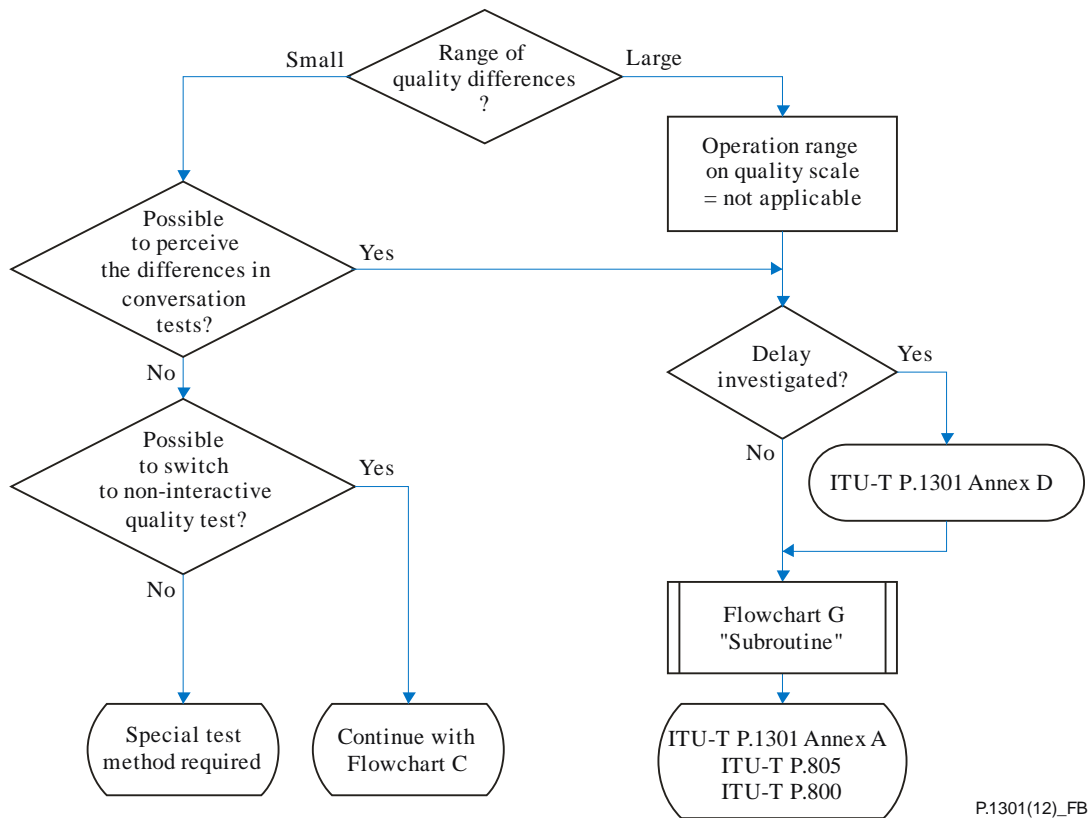
## 8.2 Flow charts to be used when selecting test methods

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

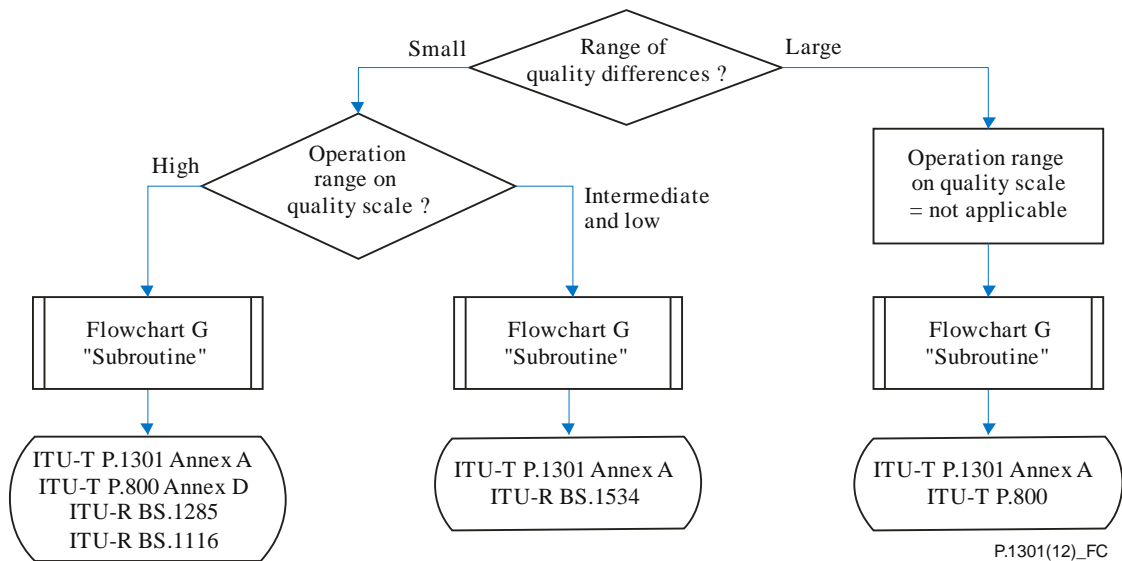


P.1301(12)\_FA

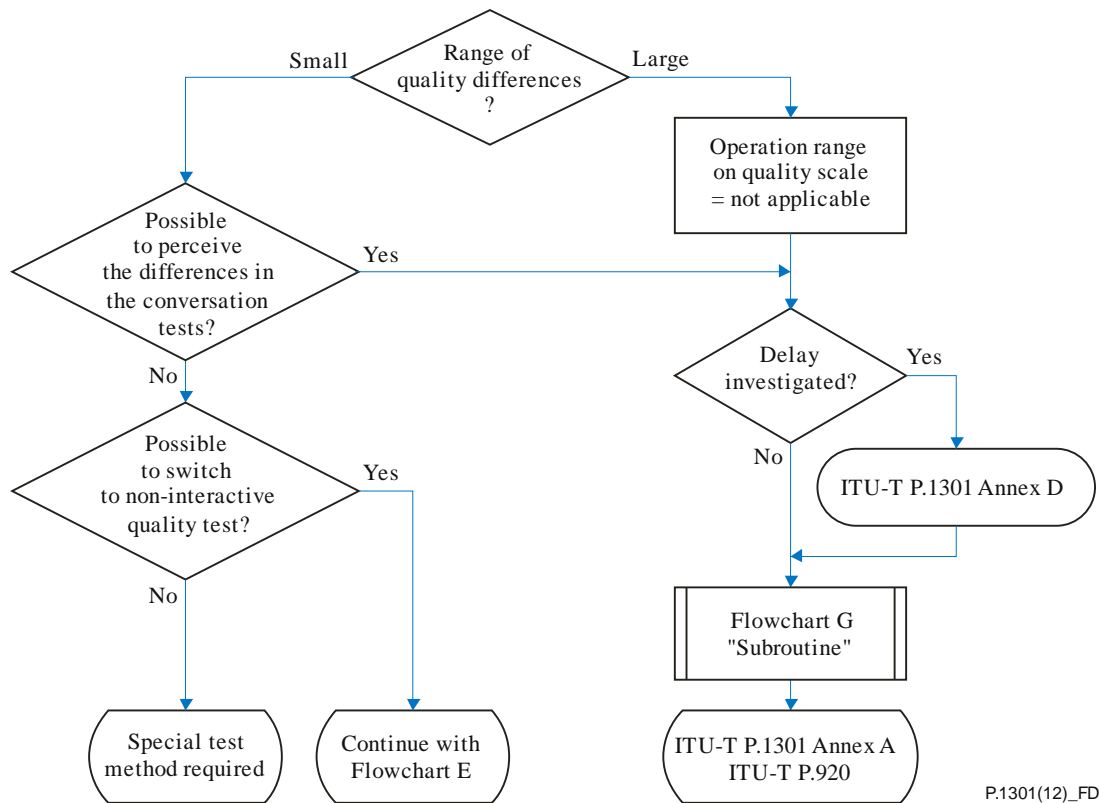
**Flowchart A – Start of the decision tree**



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

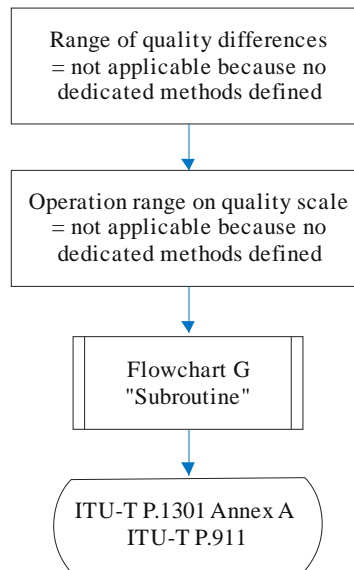


**Flowchart C – Continuation of decision tree for assessing non-interactive audio quality**



P.1301(12)\_FD

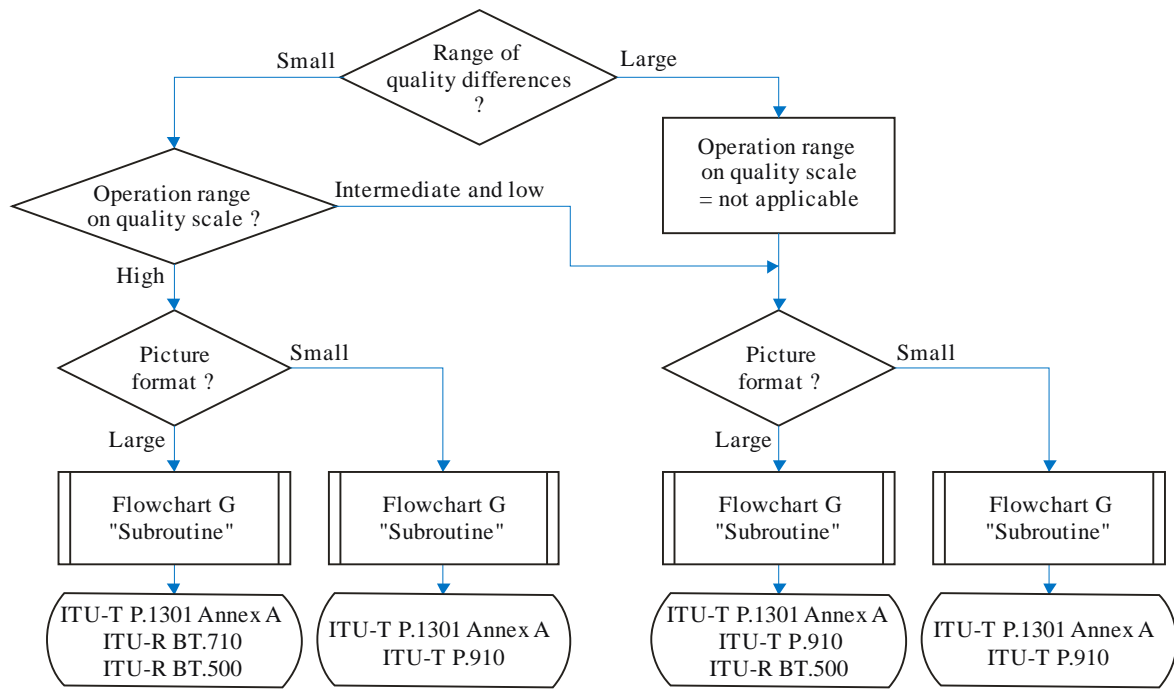
**Flowchart D – Continuation of decision tree for assessing conversational audiovisual quality**



P.1301(12)\_FE

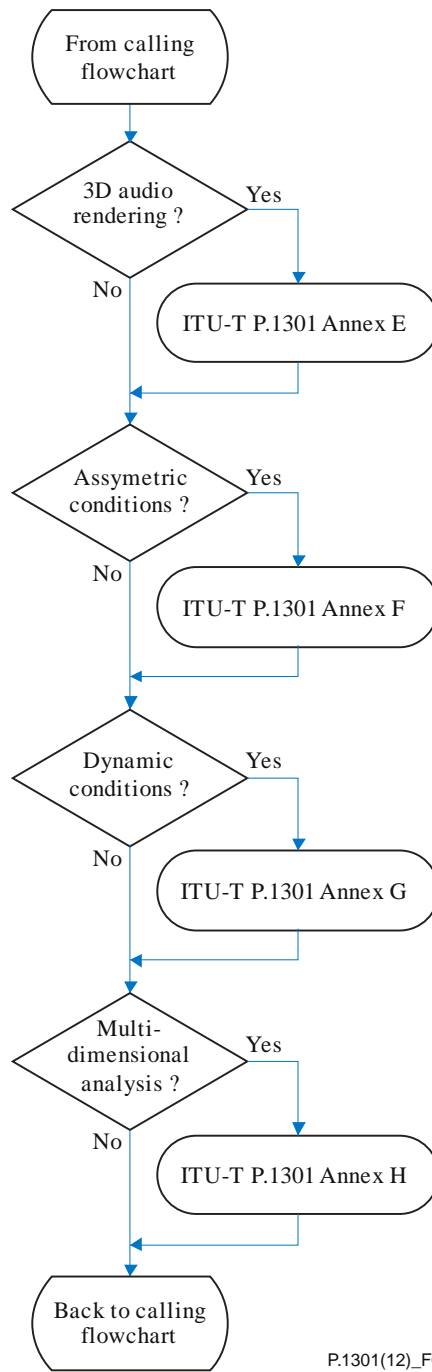
**Flowchart E – Continuation of decision tree for assessing non-interactive audiovisual quality**





P.1301(12)\_FF

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



P.1301(12)\_FG

**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 of this Recommendation) 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**

It is recommended to realize the test conditions according to the detailed descriptions provided in the Recommendations according to the decision tree in clause 8.2 of this Recommendation.

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

Concerning video, persons might 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 properly 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. Then the individual conditions should be documented.

###### **A.1.1.2 Call set-up**

There are many different possibilities to realize in the initial call set-up in terms of technology and user interface. Despite the differences in those actual realizations, the initial call set-up of a two-party conversation always follows the same process: caller invites callee, callee accepts the call. In multiparty conversations, different processes are possible for the initial call set-up: interlocutors are dialling 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 the different realization possibilities – 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, the 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 increase as well.

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 in respect of the total test time (see clause A.1.3, Experiment design).

Tasks that work fine 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 (for instance, as 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 would remain feasible for the interlocutors to follow.

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

### **A.1.3 Experiment design**

General aspects one should consider for a proper experimental design are, for instance, the presentation order of conditions, suitable length for a session in a test, the need for pauses and training phases. For details on these aspects of the experiment design, it is recommended to apply the Recommendations according to the decision tree in clause 8.2 of this Recommendation as well as to consult the ITU Handbook on Practical procedures for subjective testing [b-ITU-T HB-PPST].

However, as multiparty tasks creating the desired conversational situation require longer durations than conventional single- or two-party tasks, the experiment design needs a deliberate balancing between 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 persons 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 a possibility to get to know each other somewhat before the test. Another alternative is to invite persons 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 in turn can have a significant impact on results. If it is desired to minimize the influence of such gender-related differences, then subjects within the groups should have the same gender. If it is desired 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 had experience with multiparty systems, as inexperienced test participants might be overwhelmed by the system under test.

#### **A.1.5 Scales**

First of all, direct quality judgements may be collected by using established quality scales as they are described in the Recommendations (according to the decision tree in clause 8.2 of this Recommendation).

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

To check if the test was performed as intended, it is advised to give the test subjects the possibility to write comments during the test. This might help to interpret the voting results and might reveal other things that were not asked for in the quality questionnaires. After the test, the participants could be asked to describe their experience of the test situation either one-by-one on paper or orally in 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 to 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 maybe perform some kind of 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 pre-test) as part of the training phase.

Additionally, in the case of having free conversations as test task, another possibility to facilitate a free conversation is to show the names of either the video rooms or of 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 test goals.

#### **A.1.7 Data collection and analysis**

It is recommended to apply the Recommendations according to the decision tree in clause 8.2 of this Recommendation as well as to consult the ITU Handbook on Practical procedures for subjective testing [b-ITU-T HB-PPST].

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

#### **A.2.1 Test facilities**

It is recommended to follow the advice in clause A.1.1 of this annex.

## **A.2.2 Stimuli**

### **A.2.2.1 Technical production**

Concerning the technical production of stimuli, it is recommended to follow the detailed descriptions given in the Recommendations according to the decision tree in clause 8.2 of this Recommendation.

### **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 a production or selection of stimuli that sufficiently resemble the conversational situation. Therefore, non-interactive quality should be evaluated with contents that are 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 to sit in an actual meeting room and maybe change rooms during the observational test or to follow the conversations in a separate room, for instance by means of audio or audiovisual equipment.

## **A.2.3 Experiment design**

It is recommended to follow the advice in clause A.1.3 of this annex.

## **A.2.4 Test subjects**

Similar to conversation tests (clause A.1.4), investigators should deliberately consider using participant profiles when inviting people to the subjective experiments. In particular, the subject's experience with multiparty telemeeting systems might influence the results.

If the test requires a high sensitivity to differences between conditions, test participants should have had experience with multiparty systems as inexperienced test participants might be overwhelmed by the system under test irrespectively from the tested conditions.

## **A.2.5 Scales**

It is recommended to follow the advice in clause A.1.5 of this annex.

## **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, as one quality reference could be a multiparty face-to-face meeting and another one could be a two-party telemeeting.

## **A.2.7 Data collection and analysis**

It is recommended to follow the advice in clause A.1.7 of this annex.

## **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.)

Concerning telemeeting systems providing text-based communication and graphical information means, e.g., web conferencing services, many aspects can contribute to the quality of experience, which could be organized in the following categories:

1) Web browsing aspects (in the case of web conferencing applications)

Here, quality determining aspects are considered that can be perceived by the user until the web application is loaded. Such aspects would be, for instance, the time needed for loading the web application.

NOTE – At the time of writing this Recommendation, a new Recommendation concerning the method and procedures for subjective testing of user-perceived quality of web browsing is in preparation in ITU-T.

2) Aspects concerning the arrangement of conferencing elements

Often conferencing systems allow the 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, e.g., if the size of the video frame is changed. Hence, testing conferencing applications should either control for such effects during a subjective test, i.e., by disabling such arrangement functions for the users, or check for such effects after the subjective test, i.e., 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 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 here, 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 here, 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, or 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 split their attention between the different communication modes, a careful experimental design is required. Instructions, tasks/stimuli, and questions/scales should especially help test participants to 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.

Concerning interactive quality, new dedicated test methods need to be developed; concerning non-interactive quality, the video test methods cited in clause 8 of this Recommendation 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) and 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.

#### **D.1 Background**

Long transmission delays might cause several problems in multiparty conversations. The focus in 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 easily detectable and thus more disturbing in the case of longer delays, which in turn might 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, e.g., when a person thinks the latest speaking party is ready and starts to talk at the same time as that talker continues, or when two parties start to 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 questionnaire.

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]. As most recommended conversation tasks are for two persons, 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, e.g., like thinking about what to answer or searching for items, or they lack natural speaking behaviour, e.g., if 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 parts in the test tasks, (e.g., a table lookup), which might lead to short pauses that could mask a 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]:

### **"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 the 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 selection of test task will highly affect test result. 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 the possibility to evaluate the quality might be affected.

The goal here 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 here in the future.

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, e.g., instructions needed 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 of test methodology for free conversations can be found in Appendix III of the present Recommendation.

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

### **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. That 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 persons 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 evaluation of telecommunication qualities.

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

#### **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 could gather in one room at the beginning. This could make them know each other better, which probably makes the conversations more interactive. It is suggested to have a first short presentation face-to-face, and a round of quiz or another type of game to get the group to interact together before the test starts. They should also repeat the names of all participants until they know all names. That will make it easier to address a certain person during the test conversations. 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 delays 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 not to distract the subject's attention from the most important task.

For impairments that do not occur continuously (codec artefacts, packet loss, etc.) but are more unevenly distributed 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 quality distortions. For a naïve test subject, if a distortion occurs a limited number of times, it is considered easier to grade the amount of impairment rather than the quality.

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 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 possibility 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 asked for in the test might be commented.

## D.9 Objective measurements

If possible, the communication efficiency should be measured as time delays affects task efficiency. If the subjects are to perform a specified task the completion time can be measured, but too 'open' conversations make it impossible to measure communication efficiency. On the other hand, too structured communications 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 measure if the amount of speech varies depending on the conversational situation, for instance, if the subjects are sitting alone in a room or in a room together with other persons 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

Delays are not always noticed by the test participants but can anyway 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 is often adjusted automatically (slows down) when there are delays, without noticing the delay until it is large.

Test subjects that do not know each other are often polite. If two persons speak at the same time, often both stop talking. If the test persons know each other they have often 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 together over a connection than for a normal conversation. In a normal conversation you reflect on what the other person said, and think about what 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 persons that participate from a certain location. There is no delay between the participants in the same room, but there are always delays between different sites, which affect the conversations. Test participants that experience long delays can have more difficulties to enter a conversation, especially if the other participants experience less delay. The size of a group has also an influence on the dynamics of a conversation.

## **Annex E**

### **Assessment of 3D audio and 3D video reproduction of multiparty telemeetings**

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

Certain side aspects of a conversation can have an impact on the quality assessment. Considering systems providing 3D audio or video rendering conditions, the importance of such aspects may differ between such systems and conventional 2D systems.

For instance, if the possibility to identify individual speakers is important 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 for 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 3D video with 2D audio). In such cases, the actual viewing and listening positions are critical factors for the perceived quality that require careful control.

It is recommended to consider such effects by a careful experiment set-up (keywords: listening and viewing positions), a careful experiment design (keywords: presentation order and sessions) and a careful control of the side aspects (keywords: instructions, questionnaires and scales). An additional possibility is to conduct pilot tests to verify the impact of such aspects.

Additional information may be available in the future, as a Recommendation on quality evaluation of spatial audio meetings is planned in ITU-T.

## Annex F

### Assessment of asymmetric multiparty telemeetings

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

#### F.1 Overview

There are many factors that can be asymmetric in a telemeeting, for instance:

- 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. In principle, all those factors can be asymmetric in a telemeeting.

#### F.2 Interactions of different group sizes and different communication modes

Participants sitting alone in a room with audiovisual equipment probably look at the video screen for the main part of the time. Persons in a group room probably do not look as much at the screen, especially if there are many persons in the room. It matters in a video conference how the test participants are seated in relation to the screen. It is probably easier to include other participants in a discussion for participants in a group room if they turn at least partly towards the video screen and not only towards the other group members.

It is easy to forget to address participants with audio-only connections for persons 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 face expression.

Persons might perceive the video quality differently depending on the combination of screen size and viewing distance, which usually varies in a multiparty meeting. Also, considering audio quality some spots might have a better audio quality than other.

#### F.3 Remarks concerning the experiment design

To be able to judge the quality of a telemeeting it is important that the quality perceived by all participants in a telemeeting is taken into account. If possible, all test subjects should test all types of equipment in a test to get a balanced test.

If the participants have very different types of equipment it is important that there is a possibility to test these different qualities already in the training phase.



#### **F.4 Remarks concerning scales**

Multiple modes and multiple rendering conditions in asymmetric telemeetings have also 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 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 Remarks concerning data analysis**

At the time of writing this Recommendation, there is no specific Recommendation available on how to generate a common rating for an asymmetric telemeeting based on the assessment of the connections between individual telemeeting participants.

It is recommended, if possible, to ask for both an overall quality judgement about the telemeeting and for quality judgements about the connections between individual telemeeting participants. This may allow a proper interpretation of results and may enable conclusions on the contribution of individual connections on the overall quality.

## **Annex G**

### **Assessment of multiparty telemeetings with non-stationary quality**

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

Concerning non-interactive video quality, 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].

Concerning non-interactive audio quality, Recommendation [ITU-T P.880] deals with continuous evaluation of time-varying speech quality.

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

However, concerning conversational quality, no such recommendations are available, neither for conventional one-to-one conversations, nor for multiparty conversations.

Accordingly, it is recommended to conduct pilot testing to check if an adaptation of these methods might be necessary or if these methods can be directly applied in conversational multiparty tests.

## **Annex H**

### **Assessment of multiparty telemeetings using multi-dimensional scaling methods**

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

As 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.

At the time of writing this Recommendation, ITU-T is preparing a Recommendation to develop a subjective testing methodology that uses multiple rating scales for assessing speech quality.

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

Concerning multiparty assessment, multi-dimensional methodologies may also be used to assess the individual contributions of the multiparty specific aspects described in the present Recommendation, such as the influence of conversational situation, asymmetries, multiple modes, 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 desired, it is recommended to conduct pilot tests 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 with different types of equipment (terminal devices and connections) at the meeting. The telemeeting participants can use the same types 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 below mentioned factors 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.

#### 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 are transcodings). 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 devices. 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 amount 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 several 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 contents are recommended in [ITU-T P.910].

As 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 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 principle 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 and French.

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 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, as there is no need to read a written instruction during the test. In a free conversation other talkers can be interrupted spontaneously.

The recommended time length 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 diminished 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 an own topic.

The test persons should be instructed to totally 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 wanted 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 further down in this appendix.

In free conversations the amount of speech for every participant is not controlled by the test task. This is why the amount of speech can vary depending on the conversational situation. For instance, it can depend on the equipment used or on 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 analysed objectively according to [b-Hoeldtke].

If the aim is to create a business-like conversation, more controlled scenarios should be used because in a business meeting 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**

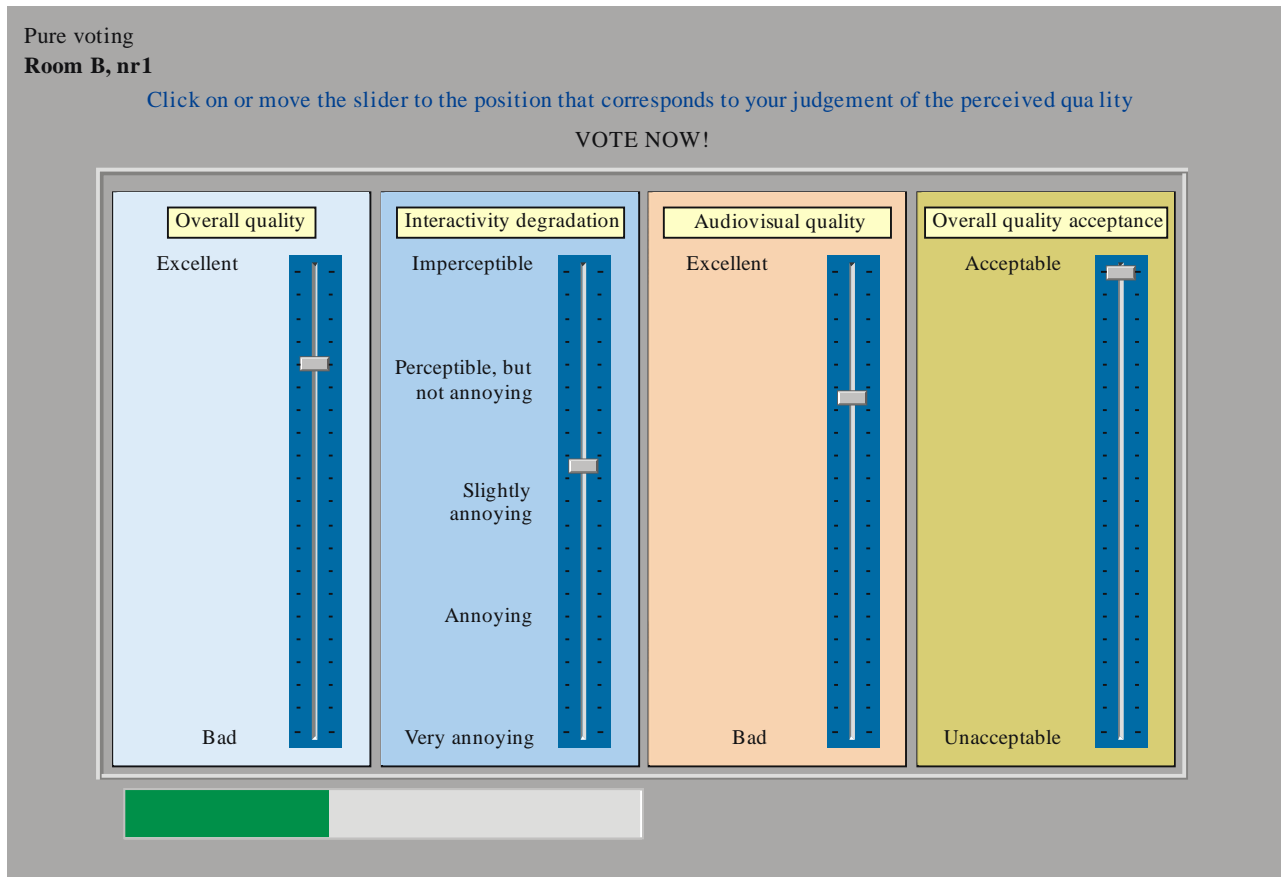
The test subjects are given the possibility to read the written instructions with the test question as soon as they arrive to the test. The instructions are also given orally with all participants in the same room, if possible. The test subjects should also have the possibility to ask questions regarding 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 become acquainted. Afterwards, the test persons move to the test rooms to be able 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 be present during the training session, as questions may arise.

The test persons should be instructed to totally 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 all.



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 supposed to finish the conversation and vote on the perceived quality. All test persons 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 – 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 persons 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 shortly interviewed after completing the entire test whether he or she had any comments or observed anything special that he or she wanted to mention. Finally, some complementary questions were asked to the whole test group.

## 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 and 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 are 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

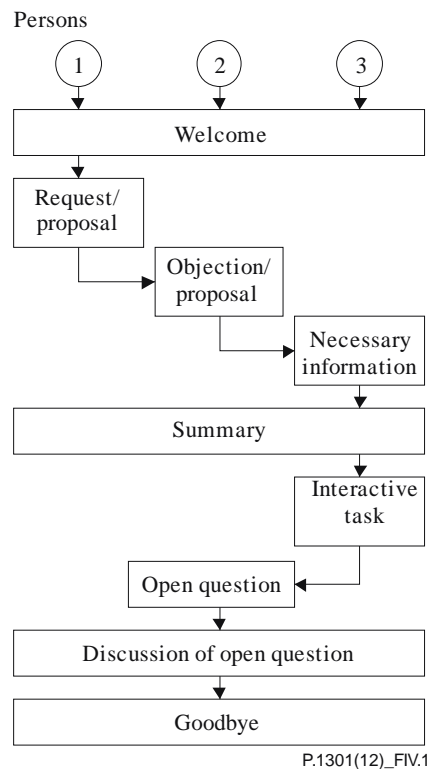
Following the SCT-development, the following set of requirements was used for the three-party conversation test scenarios (3CTs) developed here (see [b-Moeller, p. 75]):

- 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 -judgment 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 resulting 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, and any other cultural reference, need to be adopted according to the cultural context.

That means that if existing scenarios, such as the examples for 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 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, in itself, be 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 are not suited to assess audiovisual systems in multipoint configuration. Further, 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.

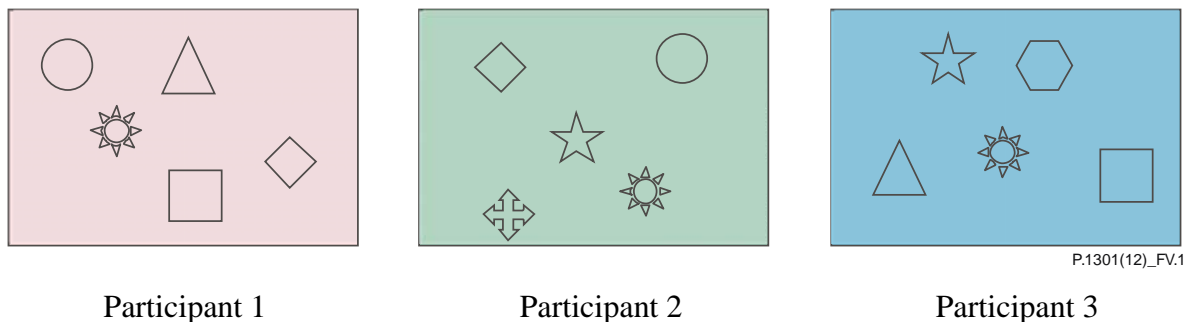
Three potential tasks are: Leavitt task, Brainstorming task, and Survival task. Though none of the three tasks was found to answer all mentioned requirements, the Survival task achieved the best compromise in terms of visual attention, number of speech turns, naturalness and satisfaction (easy, interesting,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

#### **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 below.

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 shape 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 ad infinitum, with 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 a 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 the 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 trousers 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	 <p>P.1301(12)_FSC1.1</p>
Quart of 85-proof whiskey	 <p>P.1301(12)_FSC1.2</p>

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

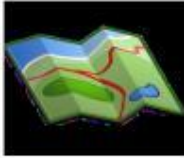




You are drifting in a private yacht in the South Pacific. A fire with unknown origin has destroyed much of the yacht, notably navigational and radio equipment. After having controlled the fire, you realize that the boat is sinking little by little. Your best estimate is that you are many hundreds of 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	 P.1301(12)_FSC2.1
A case of army rations	
15 feet of 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	 <small>P.1301(12)_FSC3.1</small>
------------------	---

Stellar map	
50 feet of nylon rope	
One case of dehydrated milk	 P.1301(12)_FSC3.2
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	

A 100 lb. tanks of oxygen	
A .45 calibre pistol	
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 kms 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	 <p>P.1301(12)_FSC4.1</p>






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 kms 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 kms 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	 P.1301(12)_FSC4.2

<p>.45 calibre pistol (loaded)</p>	
<p>Overcoat (for everyone)</p>	
<p>Two parachutes (red and white)</p>	

## Appendix VI

### **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 contribute equally 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 set-ups 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 VII

### Overview of documents describing suitable test methods

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

**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.

**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.

## 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*, (2011).
- [b-ITU-T H-Sup.1] ITU-T H-Series 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 Supplement 26, *Scenarios for the subjective evaluation of three-party audio telemeetings quality*.
- [b-Hall] Hall, J. and Watson, W.H. (1970), *The Effects of a Normative Intervention on Group Decision-Making Performance*. Human Relations 23, p. 299.
- [b-Hoeldtke] Hoeldtke, K. and Raake, A. (2011), *Conversation analysis of multi-party conferencing and its relation to perceived quality*, IEEE International Conference on Communications ICC.
- [b-Johnson] Johnson, David W. and Johnson, Roger T. (1994), *Learning together and alone: Cooperative, competitive, and individualistic learning*, 4th edition, Boston, Allyn and Bacon.
- [b-Leavitt] Leavitt, Harold J. (1960), *Task ordering and organizational development in the common target game*, Behavioral Science, Volume 5, Issue 3, pp. 233-239.
- [b-Moeller] Möller, S. (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., pp. 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] Raake A. (2006), *Speech Quality of VoIP – Assessment and Prediction*, Chichester, UK: John Wiley & Sons Ltd.
- [b-Raake2008] Raake A. and Schlegel, C. (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] Raake, A. et al (2010), *Listening and conversational quality of spatial audio conferencing*, in Proc. 40th AES Conference on Spatial Audio, Tokyo, Japan.

[b-Taylor]

Taylor D.W., Berry, P.C. and Block, C.H. (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	General tariff principles
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	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
<b>Series P</b>	<b>Terminals and subjective and objective assessment methods</b>
Series Q	Switching and signalling
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 and next-generation networks
Series Z	Languages and general software aspects for telecommunication systems