ITU出版物
建议书

国际电信联盟
无线电通信部门

ITU-R BT.500-15 建议书

(05/2023)

BT系列：广播业务（电视）

电视图像质量的主观评价方法

前言

无线电通信部门的职责是确保卫星业务等所有无线电通信业务合理、平等、有效、经济地使用无线电频谱，不受频率范围限制地开展研究并在此基础上通过建议书。

无线电通信部门的规则和政策职能由世界和区域无线电通信大会以及无线电通信全会在研究组的支持下履行。

知识产权政策（IPR）

国际电联无线电通信部门（ITU-R）的IPR政策述于ITU-R第1号决议中所参引的《ITU-T/ITU-R/ISO/IEC的通用专利政策》。专利持有人用于提交专利声明和许可声明的表格可从http://www.itu.int/ITU-R/go/patents/zh获得，在此处也可获取《ITU-T/ITU-R/ISO/IEC的通用专利政策实施指南》和ITU-R专利信息数据库。

ITU-R系列建议书

（也可在线查询https://www.itu.int/publ/R-REC/zh）

系列	标题
BO	卫星传送
BR	用于制作、存档和播出的录制；电视电影
BS	广播业务（声音）
BT	**广播业务（电视）**
F	固定业务
M	移动、无线电定位、业余和相关卫星业务
P	无线电波传播
RA	射电天文
RS	遥感系统
S	卫星固定业务
SA	空间应用和气象
SF	卫星固定业务和固定业务系统间的频率共用和协调
SM	频谱管理
SNG	卫星新闻采集
TF	时间信号和频率标准发射
V	词汇和相关问题

说明： 该ITU-R建议书的英文版本根据ITU-R第1号决议详述的程序予以批准。

电子出版
2024年，日内瓦

© 国际电联 2024

版权所有。未经国际电联书面许可，不得以任何手段复制本出版物的任何部分。

ITU-R BT.500-15 建议书

电视图像质量的主观评价方法[1]

（ITU-R第102-4/6号课题）

（1974-1978-1982-1986-1990-1992-1994-1995-1998-1998-2000-2002-2009-2012-2019-2023年）

范围

本建议书提供了图像质量的评价方法，包括通用测试方法、评价期间使用的等级量表和实施评价时建议的观看条件。本建议书由三个部分构成。

– 第1部分描述了实施所述电视图像评价的总体要求，以及关于特定方法的使用环境的导则。

– 第2部分描述了在实施主观图像质量评价时可采用的各类建议的评价方法。

– 第3部分描述了针对以第1和2部分给出的规范为基础的图像制式和应用的方法。

关键词

主观评价，图像评价

国际电联无线电通信全会，

考虑到

a) 已经收集了关于在各个实验室中使用的图像质量评价方法的大量资料；

b) 对这些方法的考察表明，在不同的实验室之间，在主观测试方法的诸多方面存在着相当程度的一致性；

c) 采用标准的评价方法，对于在各个实验室之间交换信息极为重要；

d) 某些负责监测的工程师，在例行或特殊运行期间按照五级质量量表和五级损伤量表对图像质量和/或损伤做例行或运行评价时，也能利用为实验室评价推荐的方法的某些方面；

e) 新电视信号、信号处理和新的或增强的电视业务的不断引入，都可能需要不同的实施主观图像评价的方法；

f) 这类处理、信号和业务等的引入，使信号链中每一段信号的性能都变得更可能依赖于信号链中之前各部分所进行的处理，

建议

1 在实验室实验中，应采用第1部分所述的图像质量评价的通用测试方法、等级量表和观看条件，且凡有可能，也应在操作评价中采用；

---

[1] 应提请ITU-T第12研究组注意本建议书。

2 尽管存在可替代方法并会开发一些新方法，但仍应在适用时采用第2部分所述的那些方法；

3 在实验室实验中，应采用第3部分所述的针对特定图像制式或应用的图像质量评价的通用测试方法、等级量表和观看条件，且凡有可能，也应在操作评价中采用；

4 为便于在不同的实验室之间交换信息，应按照第2部分所述，遵循所选测试方法的要求；

5 为便于在不同的实验室之间交换信息，应按照第1部分附件2详述的统计技术处理收集到的数据；

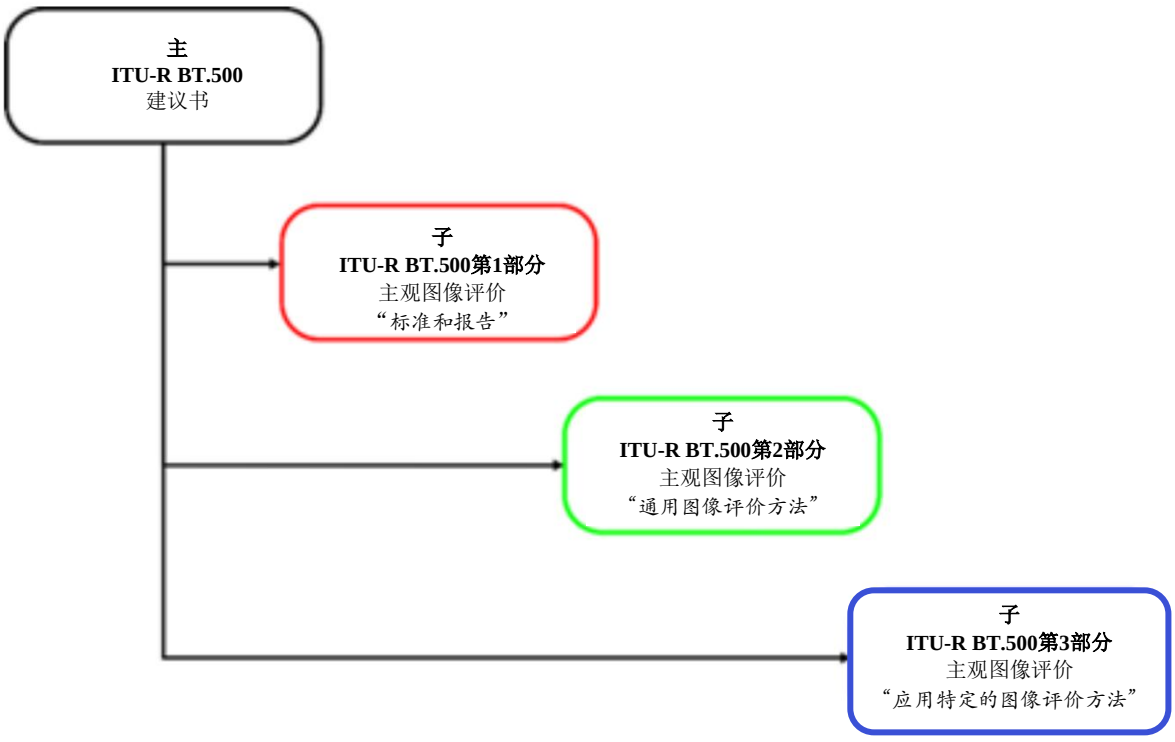
6 鉴于确定主观图像评价的基础很重要，在所有测试报告中应给出测试配置、测试素材、观察者和所用方法尽可能最全面的描述。

关于本建议书的结构和使用的说明（资料性）

ITU-R BT.500建议书由这一主建议书下的三个半独立部分构成，如图1所示。

图1

ITU-R BT.500建议书结构



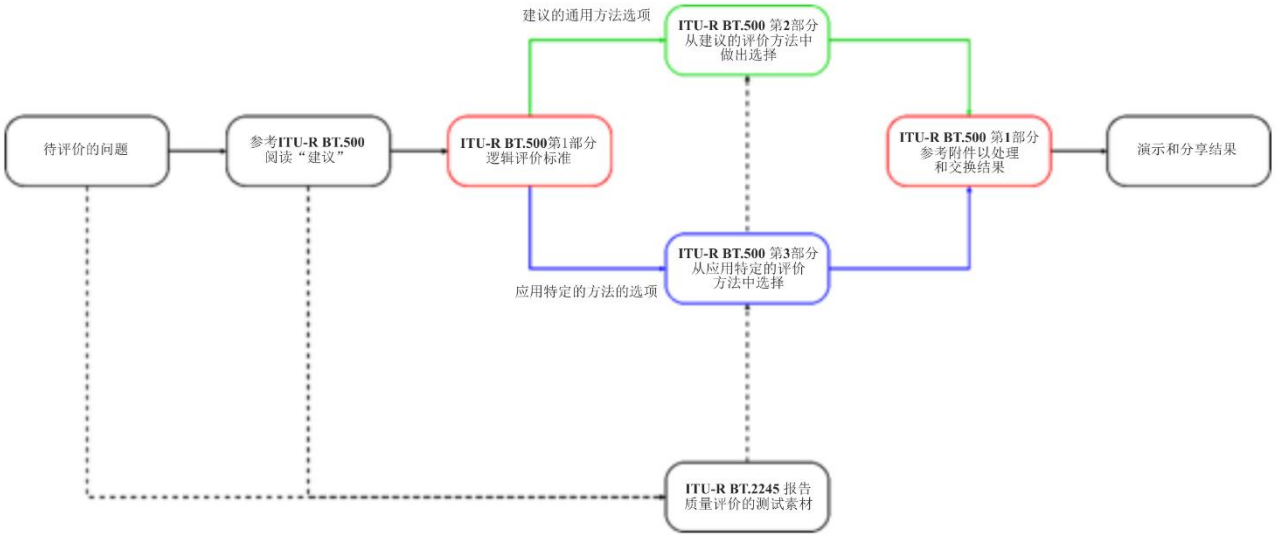
建议希望实施主观图像评价的实验室参考上述建议，之后采用第1部分详述的标准，以理解用于其评价流程的最适合的方法。第2部分提供了可使用的多个建议的主观图像评价方法的概述。第3部分提供了可协助准备相关主观图像评价流程的一些附加应用的特定方法的信息。

关于如何使用ITU-R BT.500建议书的建议

图2描绘了使用ITU-R BT.500建议书的可能工作流程。

图2

使用ITU-R BT.500建议书



理由

此版本的ITU-R BT.500建议书的分部结构使向现有的主观图像评价方法中添加新方法和修订而无须增加新建议书成为可能。增加新建议书要在多个文档中重复信息或对不需要更改的部分发布修订。

其他图像评价建议书

以下建议书涉及图像质量的客观测量，可能提供使用某些ITU-BT.500评价标准的其他应用特定的图像评价方法。

ITU-R BT.1683 建议书	存在全参考时标准清晰度数字广播电视的客观感知视频质量测量技术
ITU-R BT.1866 建议书	存在全参考时用于采用低清电视的广播应用的客观感知视频质量测量技术
ITU-R BT.1867 建议书	存在降低的带宽参考时用于采用低清电视的广播应用的客观感知视觉质量测量技术
ITU-R BT.1885 建议书	存在降低的带宽参考时标清数字广播电视的客观感知视频质量测量技术
ITU-R BT.1907 建议书	存在全参考信号时使用HDTV的广播应用的客观感知视频质量测量技术
ITU-R BT.1908 建议书	存在降低的参考信号时使用HDTV的广播应用的客观视频质量测量技术

第1部分

主观图像评价要求概述

1 引言

主观图像评价方法用于确定电视系统的性能，采用的测量能够更直接地预测可能观看在测系统的人的反应。就此而言，可以认为用客观方法可能无法全面地描述系统性能的特性；因此，有必要用主观测量作为客观测量的补充。

总体而言，主观评价分为两大类。第一类是确定在最佳条件下系统的性能的评价。这类评价通常称为质量评价。第二类是确定在与传输或发射有关的非最佳条件下系统维持一定质量的能力的评价。这类评价通常称为损伤评价。

为开展最适宜的主观评价，首先需要从不同选项中选择最符合特定环境和图像评价目标要求的方法。

为帮助做出这一选择，应考虑第2节中详述的一般特性，以理解哪些是与要评价的问题或流程相关的最适宜的选项。

一旦理解了这些选项，第1部分第3节提供了建议的图像评价方法的概述，可用于协助选择对要评价的问题或流程来说最适宜的方法，虑及采用的评价者类型和评价环境的情况。

虽然如此，选择最适宜的方法由待测系统所针对的业务目标决定。因此，在第2部分和其他ITU-R建议书中给出了特定应用的完整评价流程。

2 通用评价特性

在此给出主观评价的通用观看条件。特定系统的主观评价所用的特定观看条件在相关方法中给出。

注 – 在主观地评价高动态范围图像时，建议参考在适用章节参引的其他文件[2]。

2.1 通用观看条件

实验室观看环境旨在提供对系统进行检验的严格条件。第2.1.1节给出了实验室环境中主观评价的通用观看条件。

家庭观看环境旨在为电视链的消费型一侧提供质量评价的手段。第2.1.2节中的通用观看条件再现了家庭环境。之所以选择这些参数，是为了规定一个比典型家庭观看状况稍许严格的环境。

---

[2] 随着高动态范围进一步工作和经验的获得，建议本建议书纳入附加导则。

2.1.1 实验室环境中主观评价的通用观看条件

应如下设置评价者的观看条件：

a)	室内照明：	低
b)	背景色度：	D_{65}
c)	峰值亮度[3]:	70-250 cd/m^2（见第2.1.6.5节）
d)	显示器对比度：	≤ 0.02（见第2.1.6.4节）
e)	图像显示器背景亮度与图像峰值亮度之比：	≈ 0.15

2.1.2 家庭环境中主观评价的通用观看条件

a)	屏幕的环境照度（由周围环境在屏幕上形成的入射光，应在屏幕的垂直方向测量）：	200 lux
b)	峰值亮度	70-500 cd/m^2（见第2.1.6.4节）
c)	未激活屏幕亮度与峰值亮度显示器对比度之比：	≤ 0.02（见第2.1.6.4节）

2.1.3 观看距离

观看距离以屏幕尺寸为依据，并可根据两项独特条件加以选择：即首选观看距离（PVD）和设计观看距离（DVD）。在两个标准当中作出选择取决于研究的宗旨。

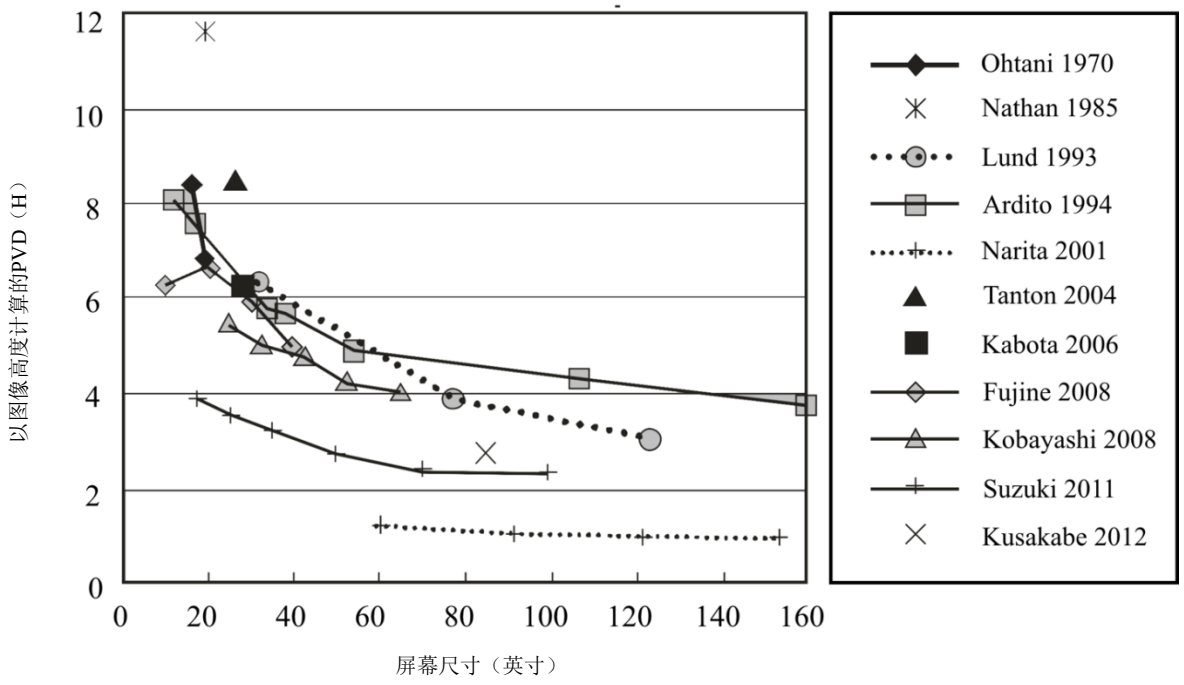
2.1.3.1 首选观看距离

首选观看距离（PVD）是以收视者根据经验确定的偏好为依据的。PVD（以屏幕尺寸的函数计算）见图1-1，其中包括采自现有来源的一系列数据集。可参照此项信息设计主观评价测试。

[3] 应根据室内照明调整峰值亮度。

图1-1

以屏幕尺寸函数计算的首选观看距离



BT.0500-01-1

2.1.3.2 设计观看距离

数字系统的设计观看距离（DVD）或最佳观看距离是两个邻近像素对观看者眼睛形成1弧分角度的距离；而最佳水平观看角度则是在最佳观看距离看到的图像的角度。

表1-1报告了以图像高度倍数表达的多个图像分辨率系统的最佳观看距离（和最佳水平观看角度）。

表1-1

以图像高度（H）计算的最佳水平观看角度、最佳观看距离

图像制式	参考文件	宽高比	像素宽高比	最佳水平观看角度	最佳观看距离
720 × 483	ITU-R BT.601	4:3	0.89	11°	7 *H*
640 × 480	VGA	4:3	1	11°	7 *H*
720 × 576	ITU-R BT.601	4:3	1.07	13°	6 *H*
1 024 × 768	XGA	4:3	1	17°	4.5*H*
1 280 × 720	ITU-R BT.1543及BT.1874	16:9	1	21°	4.8 *H*
1 400 × 1 050	SXGA+	4:3	1	23°	3.3*H*
1 920 × 1 080	ITU-R BT.709	16:9	1	31°	3.2*H*
3 840 × 2 160	ITU-R BT.2020	16:9	1	58°	1.6*H*
7 680 × 4 320	ITU-R BT.2020	16:9	1	96°	0.8*H*

注：当图像评估涉及分辨率时，7 680 × 4 320和3 840 × 2 160制式应采用观看距离的较低值。当不评估分辨率时，任何在范围以内的观看距离（对于3 840 × 2 160制式：图像高度1.6到3.2；对于7 680 × 4 320制式：图像高度0.8到3.2）都可以使用。

2.1.4 观察角

应限制相对于正常角的最大观察角，以便不使观察人员看到屏幕上的重构颜色中的偏差。还应虑及受测图像制式的最佳水平观察角，以确定观察角。进一步详细信息见ITU- R BT.2129报告第1.8节。

2.1.5 房间环境色彩方案

显示背景的色彩应与参考白点相同；对于其他的房间表面，应使用暗哑光表面，目标是最大程度减少在显示屏上的杂散光。

2.1.6 显示器

采用不同特性的显示器可能产生不同的主观图像质量。因此，强烈建议事先检查使用的显示器的特性。在采用专业FPD显示器进行主观评价时，可引证ITU-R BT.1886建议书 – HDTV演播室制作中使用的平板显示器的参考光电转换功能和ITU-R BT.2129报告 – 用户对周围HDTV节目制作环境中主显示器的平板显示器（FPD）的需求。

ITU-R BT.2390报告提供了关于用于评价高动态范围（HDR）图像的实验室和家庭显示器和观看环境的信息。

2.1.6.1 显示器处理

一旦采用图像缩放、帧速率转换、图像增强器等显示器处理程序，就应当设法避免伪像的干扰。应采用适合在测的或在评价期间使用的HDR系统的HDR处理程序。对于消费型环境或分布评价，可包含适宜的静态或动态元数据的使用。此类元数据的完整详细信息应被纳入评价说明中，从而使其他实验室能够准确地重复该评价。

在使用消费型显示屏用于主观图像评价时，重要的是禁用所有图像处理选项（除非此类图像处理的影响是评价对象）。

当访问隔行扫描图像时，测试报告应说明是否使用了解隔行扫描器。最好是不用解隔行扫描器就可以显示隔行扫描信号。

2.1.6.2 显示器分辨率

专业显示器的分辨率通常遵守其亮度操作范围内主观评价所要求的标准。

可以提议对最大和最小分辨率（屏幕中心和四角）进行检验和报告。

如果采用消费型FPD电视机显示器用于主观评价，强烈建议在使用的亮度值上对最大和最小分辨率（屏幕中心和四角）进行检验和报告。

目前，最实用的系统可用于主观评价实施者，以检查显示器或消费型电视机的分辨率，是使用电子生成的扫描测试模式。

2.1.6.3 显示器调整

应根据ITU-R BT.814建议书利用PLUGE波形，在环境亮度下调整显示器的亮度和对比度。

对于标准动态范围（SDR）图像评价，显示器对比度应根据ITU-R BT.815建议书来测量。在评价HDR图像时，应参考ITU-R BT.2390报告。

2.1.6.4 显示器对比度

对比度可能会受到环境照度的强烈影响。

专业显示器很少采用技术措施提高高照度环境下的对比度，因此若在高照度环境下使用，就有可能不符合要求的对比度标准。

消费型显示器通常采用技术措施获得高照度环境下更好的对比度。

2.1.6.5 显示器亮度

在调整LCD显示器亮度时，最好采用背景光强度控制，而不是采用信号电平缩放以保持比特精度。如使用不采用背景光的其他显示技术，应采用非信号电平缩放的方法调整白电平。请注意，PDP通过光辐射数量控制亮度，而如果设置的亮度较低，色调复制会出现退化。

2.1.6.6 显示器运动伪像

显示器不应产生具体显示技术形成的运动伪像。另一方面，应在显示器上显现包括在输入信号中的运动效应。在使用消费型显示器时，禁用全部运动处理选项至关重要。

2.1.6.7 宽屏16:9宽高比显示器的安全区

ITU-R BT.1848建议书对16:9显示器的安全区作了规定。

2.2 源信号

源信号提供直达基准图像，并作为待测系统的输入。对于所用的电视标准而言应是最佳质量的。在所演示的一对图像中，基准部分无缺陷是得到稳定结果的关键。

以数字方式存储的静态图像和视频序列是最能再现的源信号，所以它们是优选的信号源。它们还可以在实验室之间交换，以使得系统的比较更有意义。

早期阶段完成的任何处理所产生的效果都可能影响在测系统的性能，因此常常会需要考虑这种影响是如何产生的。因此，在希望检查信号链上分步处理引起的损伤是如何累积的时，凡是在信号链中有可能引入处理失真的段上完成的测试，即便处理失真不可见，最终信号最好也应透明地记录下来，然后提供给顺流的后续测试。这种记录应保存在测试素材库中，将来根据需要使用，这些记录还应附上已录信号形成过程的详细说明。如有需要，35毫米（mm）幻灯片扫描器可作为静态图像的一个来源。所得到的分辨率对于常规电视评价来说是足够的。胶片的色度和其他特性可能会给出与演播室摄像机图像不同的主观印象。如果它会影响结果，应使用演播室直达信号源，不过这样做常常不太方便。一般来说，为了得到尽可能高的主观图像质量，幻灯片扫描器应逐个图像进行调节，因为实际情况将会如此。

顺流处理能力的评价常常是用背景调色来进行的。在演播室的工作中，背景调色对演播室的灯光特别敏感。所以评价宁愿使用特殊的背景调色幻灯片对，这样始终能给出高质量的结果。如果需要，可在前景幻灯片中引入运动。

2.3 测试素材的选择

确定电视评价中所需的测试素材的种类有好几种方式。不过在实践中，要解决特定的评价问题，应采用特定种类的测试素材。表1-2给出了对典型评价问题的调查结果，以及对解决这些问题所用的测试素材的调查结果。

表1-2

测试素材的选择*

评价问题	所用的素材
采用普通素材的总体性能	通用的，“严格但并不过分严格”
容量，严格应用（例如馈给，后期处理等）	一定范围的，包括对待测应用来说极为严格的素材
“自适应”系统的性能	对于所用“自适应”方案来说极为严格的素材
识别出弱点和可能的改进措施	某种属性的严格素材
识别出影响系统出现可见变化的因素	范围广泛、内容丰富的素材
不同标准之间的转换	对于不同之处（例如场频）来说严格的素材

* 可以认为，所有测试素材都可能是电视节目内容的一部分。关于选择测试素材的进一步导则，见附件3和4。

某些参数可能会对大多数图像或序列引起相似的损伤等级。在这些情况下，以非常少的图像或序列（例如2个）所得到的结果仍然可能提供一种有意义的评价。

但是，新系统常常会产生某种在很大程度上取决于场景内容或序列内容的影响。在这种情况下，对于整个节目时间而言，将存在一种损伤概率的统计分布和图像内容或序列内容的统计分布。一般情况下，不知道这种分布的形式，必须仔细选择测试素材和整理分析得到的结果。

通常，纳入临界素材是很重要的，因为在分析结果时可能要考虑这种情况，而从非临界素材推断结果则是不可能的。在场景内容或序列内容影响到结果的情况下，应选择对于受试系统来说是“临界但不过界”的素材。“不过界”一语指这些图像仍可能形成正常节目时间的一部分。在这种情况下，至少要使用四个素材。例如，其中一半肯定是临界的，另一半是适度临界的。

2.3.1 ITU-R测试序列

一些组织已经开发了测试静止图像和序列的方法。ITU-R BT.2245报告 – 用于评价图像质量的包括HDR-TV在内的HDTV和UHDTV测试素材，给出了可被用于主观评价的HDTV和UHDTV测试素材的详细信息。关于选择测试素材的进一步见解在本建议书第1部分附件1和附件2中给出。

2.4 条件的范围和锚定

由于评价方法对可见条件的范围和分布很敏感，判断阶段应考虑变化因素的整个范围。但可以将此逼近为一个更为严格的范围，与此同时体现量表中极值处的某些条件。这些极值要么可由例子来表示并被确定为最大极值（直接锚定），要么分布在整个判断阶段内并被确定为非最大极值（间接锚定）。

2.5 观察者

根据评估的目标，观察者可能是专家或非专家。专家观察者对测试系统可能引入的图像伪像具有专长。非专家（“无知”）观察者对测试系统可能引入的图像伪像不具备专长。无论怎样，观察者不应直接参与或曾经参与所研究的系统的开发，即足以掌握具体和详细的情况。

2.5.1 观察者数量

除非所选的方法另有规定，否则应使用至少15位观察者。所需评价者的数目取决于所用测试程序的灵敏性和信度，并取决于所评估的影响的预期范围。对于在一定范围内开展的研究，例如，探索性研究，可使用少于15位的观察者。在这种情况下，应将研究确定为“非正式”性质。观察者评价电视图像质量的专业化水平应体现在报告中。

2.5.2 观察者筛选

在测试阶段开始之前，通常应通过斯内伦（Snellen）视力表或朗多（Landolt）环形视力表筛选具有（校正至）正常的视敏度，并采用专门选定的表（例如Ishihara检查表）筛选具有正常的彩色视觉的观察者。

A1-2.3和A1-2.4节详细介绍了可应用于各种测试方法的不同观察者筛选场景。当实验室或不太正式的测试作为多地点或组织测试计划的一部分进行时，重要的是应交换观察者筛选方法和标准的完整细节，并将其作为发布结果的一部分。

一般来说，尽可能详细地了解评估小组的特征，其中可能包括职业类别（例如广播组织员工、大学生、办公室工作人员……）、性别和年龄范围。

注 – 对不同实验室得出的结果之间的一致性的研究表明，不同实验室得出的结果之间可以存在系统性差别。在为提高某项实验的灵敏性和信度而综合若干不同实验室的结果时，这种差别将显得尤为重要。

对不同实验室之间的这种差别有一种可能的解释，也就是不同的评价者小组之间可能存在不同的熟练程度。必须进一步探索，以评价这一假设的有效性，并在得证的情况下对这一因素引起的变化进行量化。

2.5.3 评价须知

应向评价者仔细介绍评价方法、容易产生的损伤类型或质量因素、分级量表、序列及定时。应采用训练序列说明要评价的损伤的范围和类型，所用图像不同于测试中要用的图像，但具有类似的灵敏性。对于质量评价的情况，可以规定质量由具体的可感知属性构成。

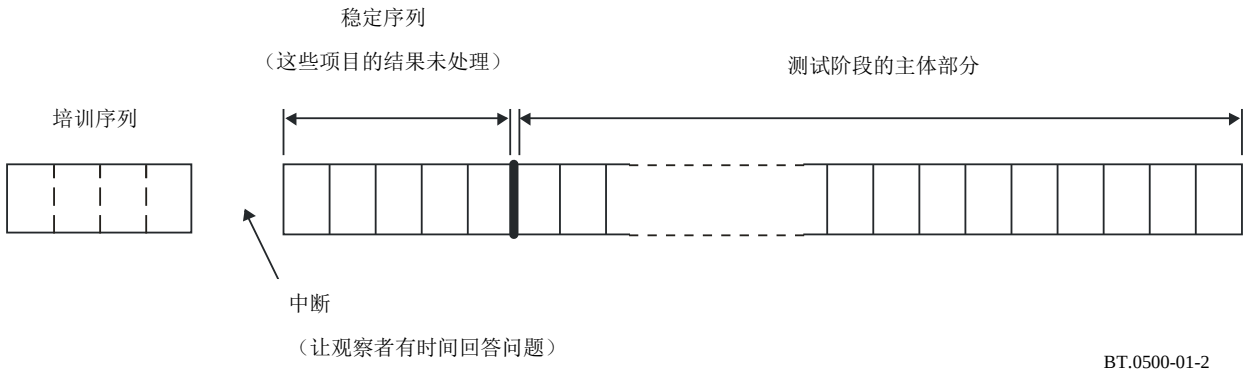
2.6 测试阶段

一个测试阶段应持续半小时以内。第一阶段开始时，应播放5个左右“模拟演示”，以稳定观察者的意见。这几个演示中给出的数据不可在测试结果中考虑。如果需要多个测试阶段，则在后续阶段开始时仅需要3个左右的模拟演示。

演示的播放应采用随机顺序（例如从Graeco-Latin方阵导出）；但测试条件的顺序应加以安排，使得疲倦或适应对分级的影响在不同测试阶段之间得以平衡掉。为检查一致性，有些演示可在不同的测试阶段予以重复。

图1-2

测试阶段的演示结构



2.7 结果的表示

由于结果会在一定范围内变化，从绝对意义上（例如图像或图像序列的质量）分析从大多数评价方法中得出的判断就不合适了。

对每个测试参数，必须给出评价等级的统计分布的均值和95%的置信区间。如果这种评价认为损伤随参数值的变化而变化，则应使用曲线拟合技术。逻辑曲线拟合和对数轴将允许采用直线表达方法，这是优选的表示形式。关于数据处理的其他资料在本建议书第1部分的附件1中给出。

结果必须与下列信息一起给出：

– 测试配置的详情；
– 测试素材的详情；
– 图像源和显示显示器的类型（见注1）；
– 评价者的数目和类型（见注2）；
– 所用的基准系统；
– 实验的总平均分；
– 原始评分和调整后的平均分以及95%的置信区间，如果一位或多位观察者按下述程序被排除在外的话。

注1 – 有某种证据表明，显示器尺寸可能会影响主观评价的结果，因此要求实验者明确报告屏幕尺寸，并指出任何实验中所用显示器的品牌和型号。

注2 – 有证据显示，观看人员（甚至非专家小组人员之间）熟练程度的差异会影响主观观看评价的结果。为便于进一步研究这一因素的影响，要求实验者尽可能详细提供所使用的观看人员的特性。相关因素包括：小组人员的年龄和性别构成，或者小组人员的教育程度或职业类别。

3 测试方法的选择

电视评价中采用了种类繁多的基本测试方法。但在实践中，解决特定的评价问题应采用特定的评价方法。本建议书第3部分提供了各图像制式和应用的图像质量的主观评价导则。

第1部分
附件1

结果的分析和表示

A1-1 引言

在为评价某一电视系统的性能而进行的主观实验期间，会收集大量数据。这些数据以观察者评分表或其电子版本的形式出现，必须用统计技术加以提炼，以便形成图形和/或数字/公式/算法形式的结果，并由此归纳出待测系统的性能。

下面的分析适用于本建议书第2部分附件1、2、3中用于评价电视图像质量的单激励（SS）法、双激励损伤量表（DSIS）法和双激励连续质量量表（DSCQS）法得出的结果，也适用于采用数值量表的其他替代方法。对于第一和第二种情况，使用五级或多级量表进行评分。对于最后一种情况，使用连续评分量表，并将结果（基准图像与实际待测图像之间的评分差值）归一化为0和100之间的整数。

A1-2 分析的常用方法

按照第1部分第2节所述的各方法的原则完成的测试将产生整数值的分布，比如1至5和0至100之间的整数值的分布。由于各观察者的判断之间存在差别，也由于与实验有关的各种条件的影响，比如使用了若干图像或序列，这些分布将会存在一些差异

一次测试由若干演示L组成。每个演示包括若干测试条件J，施加在若干测试序列/测试图像K之一上。在某些情况下，测试序列/测试图像与测试条件的每种组合都可能重复R次。

A1-2.1 平均评分的计算

对结果进行分析的第一步是计算每一演示的平均评分$\bar{u}_{jkr}$，：

$$\bar{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \tag{1}$$

其中：

u_{ijkr}： 观察者i在测试条件j、序列/图像k、重复r次情况下的评分

N： 观察者数目。

同样，可算出每一测试条件和每一测试序列/图像的总平均评分$\bar{u}_j$和$\bar{u}_k$，。

A1-2.2 置信区间的计算

A1-2.2.1 原始（未补偿和/或未近似）数据的处理

在表示某一测试的结果时，所有的平均评分都应有相应的从每一样本的标准差和大小导出的置信区间。

建议采用由下式给出的95%置信区间：

$$\left[\bar{u}_{jkr} - \delta_{jkr},\ \bar{u}_{jkr} + \delta_{jkr}\right] \tag{2}$$

其中：

$$\delta_{jkr} = 1.96 \frac{S_{jkr}}{\sqrt{N}} \tag{3}$$

每一演示的标准差S_{jkr}由下式给出：

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\bar{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \tag{4}$$

在采用95%的概率时，实验平均评分与（对于数目极多的观察者而言的）“真实”平均评分之间的差的绝对值小于95%的置信区间，条件是各个评分的分布满足某些要求。

类似地，可以算出每一测试条件的标准差S_j。但要注意，在测试序列/测试图像数目较少的情况下，相对于参与评价的评价者之间的评价差别而言，所用测试序列之间的差别对标准差的影响更大。

A1-2.2.2 补偿和/或近似数据的处理

对于评价量表的残余损伤/增强效应和边界效应已得到补偿的那些数据，或者以损伤响应形式或损伤加法律形式表示的数据，由于实验质量平均评分与这些失真存在依存关系，置信区间应采用统计变量变换来计算，同时顾及变量值的离中趋势。

如果质量评价的结果表示为损伤响应（即实验曲线），则置信区间的置信下限和上限将是每一实验量值的函数。要计算这些置信限，必须计算标准差并对初始损伤响应的每一实验量值评价其近似值。

A1-2.3 观察者的后筛选

A1-2.3.1 基于峰度的DSIS、DSCQS和除SSCQE方法之外的替代方法的后筛选

首先用β_2测试（通过计算函数的峰态系数，即四阶动差与二阶动差平方的比值）确定测试演示的这种评分分布正常与否。如果β_2在2和4之间，则这一分布被视为正常。对于每次演示，每一观察者的评分u_{ijkr}必须与平均值$\bar{u}_{jkr}$，加上相关标准差S_{jkr}乘以2（若属正常）或乘以$\sqrt{20}$（若属异常），也就是与P_{jkr}相比较，并与相关平均值减去同样的标准差乘以2或乘以$\sqrt{20}$，也就是与Q_{jkr}相比较。每当发现观察者的评分高于P_{jkr}，与每一观察者P_i相关的计数仪就递增。同样，每当发现观察者的评分低于Q_{jkr}，与每一观察者Q_i相关的计数仪就递增。最后，必须计算下面两个比值：$P_i + Q_i$除以每一观察者在整个测试阶段内的总评分次数，以及$P_i - Q_i$除以$P_i + Q_i$得出的绝对值。如果第一个比值大于5%而第二个比值小于30%，则观察者i必须舍弃（见注）。

注 – 对于某次给定实验得出的结果，这一程序的使用应不超过一次。另外，程序的使用应限于观察者人数较少（例如不到20人）且均为非专家的情况。

推荐将这一程序用于EBU法（DSIS）；这一程序也已在DSCQS法和替代方法中得到了顺利应用。

上述过程可用数学方式表示为：

对于每次测试演示，计算均值$\bar{u}_{jkr}$、标准差S_{jkr}和峰态系数β_{2jkr}，其中，β_{2jkr}由下式给出：

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \quad \text{其中} \quad m_x = \frac{\sum_{i=1}^{N}(u_{ijkr} - \bar{u}_{ijkr})^x}{N} \tag{5}$$

对于每一观察者i，找出每一P_i和Q_i，即：

对于j, k, r= 1, 1, 1至J, K, R

若$2 \le \beta_{2jkr} \le 4$，则：

若 $u_{ijkr} \ge \bar{u}_{jkr} + 2\,S_{jkr}$ 则P_i=P_i+ 1

若 $u_{ijkr} \le \bar{u}_{jkr} - 2\,S_{jkr}$ 则Q_i=Q_i+ 1

否则：

若 $u_{ijkr} \ge \bar{u}_{jkr} + \sqrt{20}\,S_{jkr}$ 则P_i=P_i+ 1

若 $u_{ijkr} \le \bar{u}_{jkr} - \sqrt{20}\,S_{jkr}$ 则Q_i=Q_i+ 1

若 $\frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05$ 且 $\left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3$ 则舍弃具有如下参数的观察者i，

包括：

N： 观察者数目

J： 测试条件的数目，包括基准在内

K： 测试图像或序列的数目

R： 重复次数

L： 测试演示的次数（在大多数情况下，演示的次数等于$J \cdot K \cdot R$，不过要注意，有些评价对每一测试条件都采用数目不等的序列）。

A1-2.3.2 基于峰度的用于SSCQE法的后筛选

在采用SSCQE法时，对于具体的观察者筛选而言，应用域不再是一种测试配置（测试条件与测试序列的组合），而是某种测试配置的一个时间窗口（例如10秒（s）的评分段）。筛选分两步，第一步的目标是检测，然后舍弃与平均性能相比评分存在显著偏差的观察者；第二步是检测出并舍弃前后不一致的观察结果，而不考虑系统偏差。

步骤1：局部评分反演的检测

此时首先还是用β_2测试确定每一测试配置的每一时间窗口评分的分布“正常”与否。如果β_2在2和4之间，则这一分布被视为“正常”。然后按照下文的数学表达方式，将此过程应用于每一测试配置的每一时间窗口。

对于每一测试配置的每一时间窗口，采用每一观察者的评分u_{ijkr}计算均值$\bar{u}_{jklr}$、标准差S_{jklr}和系数β_{2jklr}。β_{2jklr}由下式给出：

$$\beta_{2jklr} = \frac{m_4}{(m_2)^2} \quad \text{其中} \quad m_x = \frac{\sum_{n=1}^{N}(u_{njklr} - \bar{u})^x}{N} \tag{6}$$

对于每一观察者i，找出P_i和Q_i，即：

对于j, k, l, r= 1, 1, 1, 1至J, K, L, R

若2 ≤β_{2jklr}≤ 4，则：

若 $u_{njklr} \geq \bar{u}_{jklr} + 2\,S_{jklr}$ 则P_i=P_i+ 1

若 $u_{njklr} \leq \bar{u}_{jklr} + 2\,S_{jklr}$ 则Q_i=Q_i+ 1

否则：

若 $u_{njklr} \geq \bar{u}_{jklr} + \sqrt{20}\,S_{jklr}$ 则P_i=P_i+ 1

若 $u_{njklr} \leq \bar{u}_{jklr} - \sqrt{20}\,S_{jklr}$ 则Q_i=Q_i+ 1

若 $\frac{P_i}{J \cdot K \cdot L \cdot R} > X\%$ 或 $\frac{Q_i}{J \cdot K \cdot L \cdot R} > X\%$ 则舍弃具有如下参数的观察者i，

包括：

N： 观察者数目

J： 在测试条件与测试序列的某种组合内时间窗口的数目

K： 测试条件的数目

L： 测试序列的数目

R： 重复次数。

这一过程可以将得出的评分显著偏离平均评分的观察者舍弃。图1-3显示出了两个例子（显示出重大偏差的两条极值曲线）。但这种舍弃准则无法检测出可能的反演，这是产生偏差的另一个重要原因。因此提出了第二步。

步骤2：局部评分反演的检测

对于步骤2，检测仍以本附件给出的筛选公式为基础，但对应用域做了稍许改动。输入数据集合仍由所有测试配置的所有时间窗口（例如10 s）的评分组成。但这一次，评分是初步的，集中在总均值附近，以便将第一步中已经处理过的偏差效应降至最弱。然后采用通常的过程。

首先用β_2测试确定每一测试配置的每一时间窗口评分的分布“正常”与否。如果β_2在2和4之间，则这一分布被视为“正常”。然后按照下文的数学表达方式，将此过程应用于每一测试配置的每一时间窗口。

过程的第一步是计算每一观察者每一时间窗口的居中评分。每一测试的平均评分 $\bar{u}_{klr}$ 规定如下：

$$\bar{u}_{klr} = \frac{1}{N} \cdot \frac{1}{J} \sum_{n=1}^{N} \sum_{j=1}^{J} u_{njklr} \tag{7}$$

同样，每一观察者每一测试配置的平均评分规定如下：

$$\bar{u}_{nklr} = \frac{1}{J} \sum_{j=1}^{J} u_{njklr} \tag{8}$$

其中u_{njklr}对应着观察者i在时间窗口j、测试条件k、序列l、重复r次情况下的评分。

对于每一观察者，居中评分u^*_{njklr}按下式计算：

$$u^*_{njklr} = u_{njklr} - \bar{u}_{nklr} + \bar{u}_{klr} \tag{9}$$

对于每一测试配置的每一时间窗口，计算均值$\bar{u}^*_{jklr}$、标准差S^*_{jklr}和系数$\beta_2{}^*_{jklr}$。$\beta_2{}^*_{jklr}$由下式给出：

$$\beta_2{}^*_{jklr} = \frac{m_4}{(m_2)^2} \quad \text{其中} \quad m_x = \frac{\sum_{n=1}^{N} (u^*_{njklr})^x}{N} \tag{10}$$

对于每一观察者i，找出 P^*_i和Q^*_i，即：

对于j, k, l, r= 1, 1, 1, 1至J, K, L, R

若$2 \le \beta_2{}^*_{jklr} \le 4$，则：

若 $u^*_{njklr} \ge \bar{u}^*_{jklr} + 2\, S^*_{jklr}$ 则$P^*_i = P^*_i + 1$

若 $u^*_{njklr} \le \bar{u}^*_{jklr} - 2\, S^*_{jklr}$ 则$Q^*_i = Q^*_i + 1$

否则：

若 $u^*_{njklr} \ge \bar{u}^*_{jklr} + \sqrt{20}\, S^*_{jklr}$ 则$P^*_i = P^*_i + 1$

若 $u^*_{njklr} \le \bar{u}^*_{jklr} - \sqrt{20}\, S^*_{jklr}$ 则$Q^*_i = Q^*_i + 1$

若 $\frac{P^*_i + Q^*_i}{J \cdot K \cdot L \cdot R} > Y$ 且 $\left|\frac{P^*_i - Q^*_i}{P^*_i + Q^*_i}\right| < Z$ 则舍弃具有如下参数的观察者i，

包括：

N： 观察者数目

J： 在测试条件与测试序列的某种组合内时间窗口的数目

K： 测试条件的数目

L： 测试序列的数目

R： 重复次数。

根据经验，这一方法适用的参数（X, Y, Z）的推荐值为0.2、0.1、0.3。

A1-2.3.3 基于相关性的后筛选

各观察者必须用稳定的和相关的方法来对各个场景和算法的质量大幅下降做出判断。舍弃准则依据所有观察者对某个给定测试会议的平均分来验证某个观察者分值的一致程度。与在DSQCS方法中一样，在SAMVIQ方法中，可以考虑所有算法（隐含的基准、低锚、经编码的片段）。判定准则基于测试的所有观察者相应的平均分对应的单个分值相关性。该过程比前面几节中描述的相应过程更容易实现。

A1-2.3.3.1 皮尔森相关

质量尺度与观察者分值范围之间的关系被认为是线性的，以便应用皮尔森相关。

主要目的是，如果某个观察者的分值与整个测试会议所有观察者的平均分一致，那么用一种简单的方法来验证。隐含的基准被认为是高质量的锚。如果包括了低的和高的锚，那么它们提高了相关值，观察者之间的相关偏移值反而降低了。

$$r(x,y)=\frac{\sum_{i=1}^{n}x_i y_i-\frac{\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}\right)}} \tag{11}$$

其中：

x_i： 表示三个参数（算法、比特率、场景）所有观察者的平均分

y_i： 对同样的三个参数，某个观察者的单个分值

n： （算法数目）×（场景数目）

i： {编解码器数目、比特率数目、场景数目}。

A1-2.3.3.2 斯皮尔曼等级相关

即使不认为质量尺度与观察者分值范围之间的关系是线性的[4]，也可以应用斯皮尔曼等级相关。

$$r(x,y)=\left[1-\frac{6\times\sum_{i=1}^{n}[R(x_i)-R(y_i)]^2}{n^3-n}\right] \tag{12}$$

其中：

x_i： 对三个参数（算法、比特率、场景）所有观察者的平均分

y_i： 对相同的三个参数，某个观察者的单个分值

n： （算法数目）×（场景数目）

$R(x_i$或$y_i)$： 排列次序

i： {编解码器数目、比特率数目、场景数目}。

A1-2.3.3.3 放弃一名测试观察者的最终舍弃标准

根据以下条件，为了放弃观察者，实施斯皮尔曼等级和皮尔森相关：

IF [均值（r）－标准差（r）]>最大相关门限（MCT）。
舍弃门限＝最大相关门限（MCT）。
ELSE 舍弃门限＝[均值（r）－标准差（r）]。

IF [r（观察者 i）]>舍弃门限。
THEN 不放弃测试的观察者“i”。
ELSE 放弃测试的观察者“i”。

[4] 通常，皮尔森相关结果非常接近斯皮尔曼相关结果。

其中：

r＝ 最小（皮尔森相关，斯皮尔曼等级相关）

均值（r）： 测试的所有观察者相关的平均值

标准差（r）： 测试的所有观察者相关的标准差

最大相关门限（MCT）＝0.85。

最大相关门限值0.85对SAMVIQ和DSCQS方法是有效的，否则，对SS和DSIS方法必须考虑最大相关门限值0.7。

A1-2.4 在具有挑战性的测试条件下计算平均分数和置信区间

很多时候，主观测试需要在具有挑战性的条件下进行。例如，在众包测试中，受试者暴露在比实验室更不受控制的环境中。在多个实验室进行的大规模测试中，实验室间的差异可能会导致收集的评级存在较大差异。A1-2.1至A1-2.3段中介绍的方法通常不太适合这种情况。本节介绍一种先进的数据分析技术，该技术已被证明可以提高恢复的平均分数和置信区间的数据质量。参考Python实现也可以在本附件的后附资料1中找到。

这种方法背后的原理如下。显性地模拟每个受试者的行为是有用的；特别是，受试者的偏见和一致性是影响受试者投票的两个突出的人为因素。通过迭代程序，该技术试图共同评估每个演示的真实质量和每个受试者的偏差和一致性。每种呈现的估计真实质量可解释为“消除偏差的一致性加权平均意见得分”。与A1-2.3.1中段所述的受试者后筛选（保留或拒绝受试者的所有评分（“硬拒绝”））相比，该技术可描述为“软拒绝”。也就是说，对于投票不一致的异常受试者，受试者的投票将具有很小的权重，因此对整体MOS的贡献很小。此方法的一个副作用是对每个测试对象的偏差和一致性的估计。这些信息对于受试者是否适合执行主观测试是有价值的信息，因此可用于为未来测试筛选受试者。例如，如果一个受试者已经证明投票非常不一致，他/她可能会被排除在以后的会话之外。

该方法首先估计所有受试者和重复情况中每个呈现的平均分：

$$\bar{u}_{jk} = \frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} u_{ijkr} \tag{13}$$

其中u_{ijkr}是条件j、序列/图像k、重复r的观察者i的得分，N是观测者的数量，R代表重复次数。

在第二步中，用下式估计每个观测者b_i的偏差：

$$b_i = \frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} u_{ijkr} - \bar{u}_{jk} \tag{14}$$

其中J和K分别是条件的数量和序列的数量。然后，在迭代循环中执行以下步骤。

当前对每个演示的平均分的估计值记录为$\bar{u}_{jk}^{c}$，即

$$\bar{u}_{jk}^{c} = \bar{u}_{jk} \tag{15}$$

接着计算每个观测评分中不能用平均分和观测者偏差解释的残差：

$$e_{ijkr} = u_{ijkr} - \bar{u}_{jk} - b_i \tag{16}$$

然后这些残差用来计算每个观测者的不一致σ_i：

$$\sigma_i = \sqrt{\frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} (u_{ijkr} - \mu_{e_i})^2} \tag{17}$$

其中：

$$\mu_{e_i} = \frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} e_{ijkr} \tag{18}$$

然后，可以通过以下公式获得新的平均分的估算值：

$$\bar{u}_{jk} = \frac{\sum_{i=1}^{N} \sum_{r=1}^{R} \sigma_i^{-2} (u_{ijkr} - b_i)}{\sum_{i=1}^{N} \sum_{r=1}^{R} \sigma_i^{-2}} \tag{19}$$

随后按照公式（12）更新偏差。

如果出现下列情况，循环将终止：

$$\sum_{j=1}^{J} \sum_{k=1}^{K} (\bar{u}_{jk} - \bar{u}_{jk}^{c})^2 \tag{20}$$

终止之后，每次演示的评分标准差为：

$$S_{jk} = \frac{\sigma_j}{\sqrt{N}} \tag{21}$$

其中：

$$\sigma_j = \sqrt{\frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} (e_{ijkr} - \mu_{e_j})^2} \tag{22}$$

且

$$\mu_{e_j} = \frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} e_{ijkr} \tag{23}$$

然后根据公式（2）和（3）计算最终的置信区间。

A1-3 确定平均评分与图像失真主观尺度之间关系的处理方法

如果是为了研究失真的客观尺度与平均评分 $\bar{u}$（ $\bar{u}$的计算按照第A1-2.1节）之间的关系而开展主观测试，可采用下述过程，该过程包括找出$\bar{u}$ 与损伤参数之间的关系。

A1-3.1 用对称逻辑斯谛函数逼近

用一个逻辑斯谛函数逼近这一实验关系是特别值得关注的。

数据$\bar{u}$的处理可采用如下方式：

按下式取连续变量p，将$\bar{u}$的量表值归一化：

$$p = (\bar{u} - u_{min}) / (u_{max} - u_{min}) \tag{24}$$

其中：

u_{min}： u量表上表示最低质量的最低评分

u_{max}： u量表上表示最佳质量的最高评分

p与D之间关系的图形表示说明，该曲线呈现出反称S形，条件是D取值的固有上下限远远超出u快速变化的那段区域。

至此，可以用一个精心选择的逻辑斯谛函数来逼近函数$p = f(D)$，将其表示为下式的普遍关系：

$$p = 1 / [1 + \exp(D - D_M) \cdot G] \tag{25}$$

其中，D_M和G为常量，G可正可负。

由优化逻辑斯谛函数逼近得到的p值用于按下面的关系得出I的推断数值：

$$I = (1/p - 1) \tag{26}$$

D_M和G的值可从经过下述变换后的实验数据中导出：

$$I = \exp(D - D_M) \cdot G \tag{27}$$

对I采用对数尺度后，可由上式得出一种线性关系：

$$\log_e I = (D - D_M) \cdot G \tag{28}$$

采用直线的内推法比较简单，在某些情况下具备一定准确度，足以考虑用直线代表由D作为衡量尺度的效应引起的损伤。

该特性的斜率可表示为：

$$S = \frac{D_M - D}{\log_e I} = \frac{1}{G} \tag{29}$$

由此形成G的优化值。D_M为$I = 1$时的D值。

该直线可用于界定与待测损伤有关的损伤特性。要注意，直线可由逻辑斯谛函数的特征值D_M和G来规定。

A1-3.2 用非对称函数逼近

A1-3.2.1 函数说明

在失真参数D可由一个关联单位，比如S/N（分贝）（dB）来衡量的情况下，用对称逻辑斯谛函数来逼近实验评分与图像失真客观尺度之间的关系相当成功。如果用一个物理单位d，比如时间延迟（毫秒）（ms）来衡量失真参数，则关系式（27）必须用下式替代：

$$I = (d/d_M)^{1/G} \tag{30}$$

关系式(25)因此变为：

$$p = 1/\left[1 + (d/d_M)^{1/G}\right] \tag{31}$$

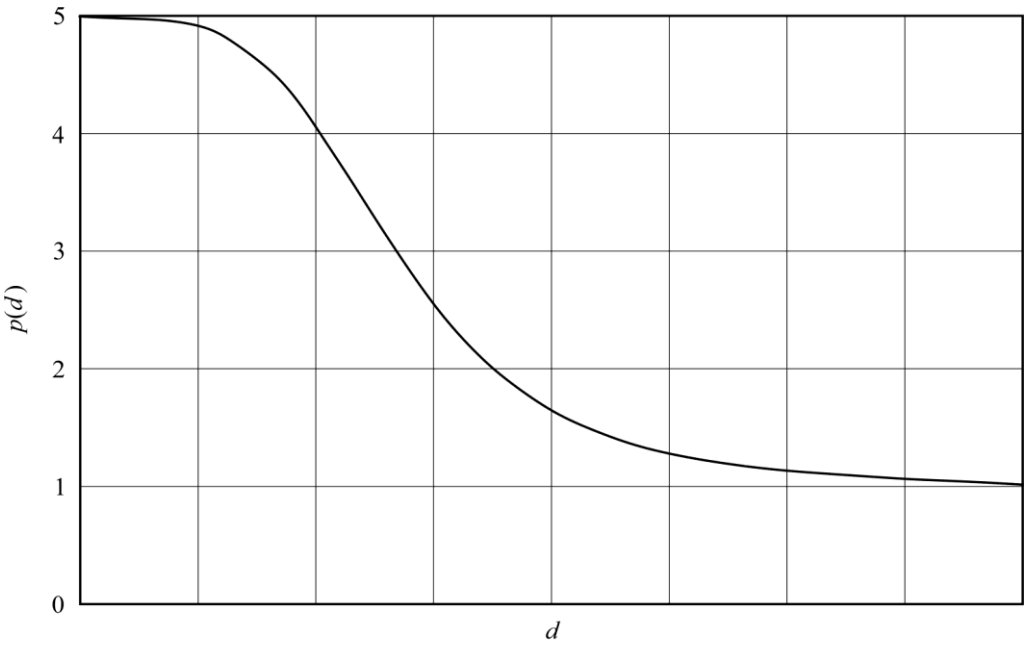
该函数以非对称的方式逼近了逻辑斯谛函数。

A1-3.2.2 参数估值的逼近

函数优化参数的估值可提供实际数据与函数之间的最小残余误差，这一估值用任何回归估值算法都可做到。图 1-3示出了一个采用非对称函数表示实际主观数据的例子。这种表达方式可以让具体的客观尺度与感兴趣的主观数值，比如五级量表上的4.5，相对应。

图1-3

非对称逼近



BT.0500-01-3

A1-3.3 残余损伤/残余增强的校正和量表边界效应的校正

在实践中，使用逻辑斯谛函数有时无法避免实验数据与逼近值之间出现某些差别。这些差异可能是由量表末端效应引起的，也可能是由于测试中同时存在若干种损伤，这都有可能影响统计模型和曲解理论逻辑斯谛函数。

有一种量表边界效应已经确定，表现为观察者倾向于不用判断尺度中的极端值，对于较高的质量评分尤其如此。这可能是由若干因素造成的，包括心理上对做出极端判断的迟疑。另外，在接近量表的边界处采用符合等式(1)的算术方式的判断，可能会出现有偏差的结果，因为在这些范围内评分出现了非高斯分布。

在测试中常常会说明存在残余损伤（即便在基准图像中，平均得评分也只能达到$\bar{u}_0 < u_{max}$)的数值）。

有几种有用的方式可校正评价的原始数据，以得出有效的结论（见表1-3）。

如果实验数据中存在边界效应，则边界效应的校正是数据处理中非常重要的一部分。因此，选择程序时必须特别谨慎。请注意，这些校正程序涉及一些特别的假设，所以在使用中要留心；在表示结果的时候应说明所用的程序。

表1-3

量表边界效应校正方法的比较

边界效应补偿方法	特性		
	残余损伤补偿	残余增强补偿	偏离量表中心
无补偿	否	否	否
线性尺度变换	是	可产生显著误差	否
非线性尺度变换(1)	是	是	否
以损伤加法为基础的方法	是	否	是
积性方法	是	否	是

(1) 采用非线性尺度变换时，必须计算校正后的评分：

$$u_{corr} = C(\bar{u} - u_{mid}) + u_{mid}$$

$$C = \frac{\bar{u} - u_{0\,min}}{u_{0\,max} - u_{0\,min}} \frac{u_{max} - u_{mid}}{u_{0\,max} - u_{mid}} + \frac{u_{0\,max} - \bar{u}}{u_{0\,max} - u_{0\,min}} \frac{u_{min} - u_{mid}}{u_{0\,min} - u_{mid}}$$

其中：

U_{corr}： 校正后的评分

$\bar{u}$： 未经校正的实验评分

u_{min}, u_{max}： 评分量表的边界

u_{mid}： 评分量表的中值

$u_{0\,min}, u_{0\,max}$： 实验评分倾向的下限和上限。

A1-3.4 将信度性能纳入图形

从每一受测损伤的平均等级和相关的95%置信区间可构建3个等级系列：

– 最小等级系列（均值 – 置信区间）；

– 平均等级系列；

– 最大等级系列（均值+置信区间）。

这三个系列对参数的估值分别进行。得到的三个函数之后可以绘在同一幅图中。最大和最小等级系列的两个函数用虚线绘制，平均等级系列的估值用实线绘制。实验量值也绘在这幅图中（见图1-4）。这样就得到了95%连续置信区域的估值。

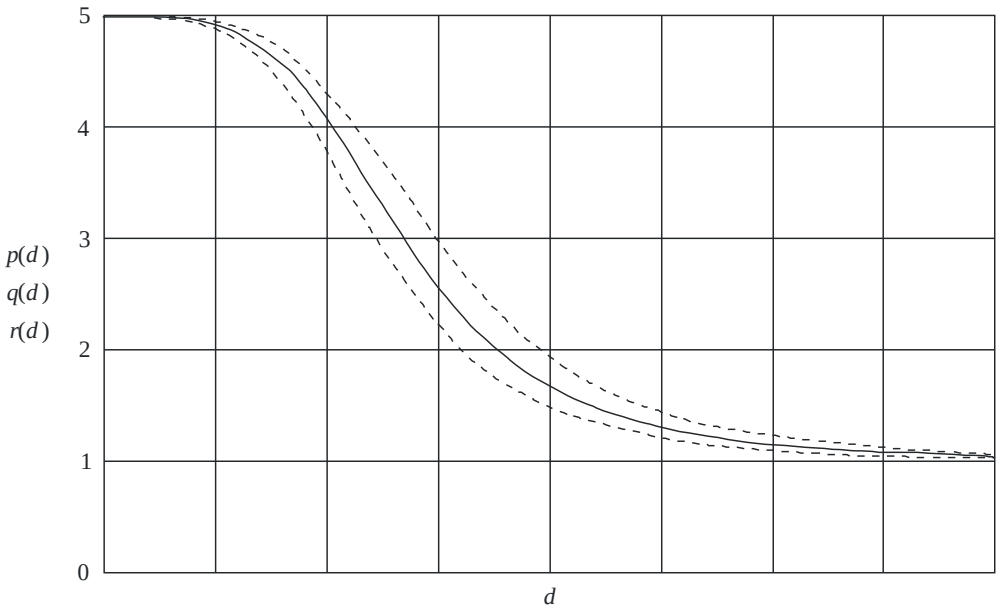
对于4.5的等级（用于该方法的可见性门限），也因此能够从图上直接读出估计的95%置信区间，可用于确定容差范围。

最大和最小曲线之间的空间并非一个95%的区间，而是其平均估值。

至少有95%的实验量值应位于置信区域内；不然就可以断定测试过程中出现了问题或者所选的函数模型并非最佳的。

图1-4

非对称损伤特性的情况



p(d): 平均等级系列
q(d): 最小等级系列
r(d): 最大等级系列
d: 客观损伤尺度

BT.0500-01-4

A1-4 结论

对评价置信区间的程序，也就是评价一组主观评价测试的准确性做了说明。

也可由这一程序得出总体平均质量的估值。总体平均质量不仅与要研究的特定实验有关，也与采用同样方法进行的其他实验有关。

因此，这种质量可用于绘制置信区间性能图，为主观评价提供帮助，并为规划未来的实验提供帮助。

附件1
后附资料1

第A1-2.4节中方法的参考实现

本附件包括A1-2.4节中介绍的数据分析方法的参考Python实现。使用的代码和示例数据也可在SUREAL Python包中公开获得：
https://github.com/Netflix/sureal/tree/master/itur_bt500_demo.

输入数据准备如下。原始投票组织在2D矩阵中，用逗号分隔。每行对应一个演示（测试条件下的源图像）；每栏对应一个受试者。

并非每个受试者都需要对每个演示进行投票。如果受试者*i*没有对演示*jk*进行投票，则在位置（*jk*,*i*）放置一个“nan”（不是数字）。输入数据将放入.csv文件中。以下是来自20个受试者和30个演示的一小部分样本.csv两次重复的投票文件。

```
5.0,nan,5.0,4.0,2.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,5.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
3.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,3.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,4.0,3.0,4.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
4.0,5.0,nan,3.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
4.0,3.0,2.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,4.0,5.0,1.0,4.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,4.0,3.0,5.0
3.0,5.0,4.0,2.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0
5.0,2.0,1.0,3.0,3.0,4.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,2.0,1.0,1.0,3.0,1.0,1.0,1.0,1.0,3.0,1.0,2.0,2.0,1.0,1.0,1.0,2.0,1.0,1.0,2.0
5.0,5.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,3.0,2.0,3.0,4.0,2.0,1.0,2.0,2.0,1.0,2.0,2.0
5.0,2.0,4.0,3.0,4.0,2.0,2.0,2.0,2.0,4.0,3.0,3.0,3.0,5.0,2.0,2.0,2.0,4.0,2.0,2.0
5.0,5.0,5.0,5.0,4.0,3.0,3.0,3.0,3.0,5.0,3.0,4.0,4.0,3.0,2.0,2.0,3.0,3.0,3.0,3.0
5.0,5.0,4.0,3.0,5.0,4.0,4.0,4.0,4.0,5.0,4.0,4.0,5.0,4.0,3.0,3.0,4.0,3.0,3.0,4.0
1.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,3.0
1.0,4.0,1.0,4.0,3.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,4.0
4.0,2.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
2.0,5.0,3.0,2.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
5.0,5.0,5.0,5.0,3.0,3.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0
4.0,5.0,5.0,3.0,5.0,2.0,2.0,3.0,1.0,3.0,3.0,2.0,3.0,5.0,1.0,1.0,2.0,2.0,2.0,2.0
1.0,2.0,2.0,4.0,5.0,1.0,2.0,2.0,1.0,3.0,2.0,2.0,4.0,2.0,3.0,1.0,2.0,2.0,1.0,3.0
4.0,5.0,3.0,5.0,2.0,3.0,2.0,3.0,3.0,4.0,2.0,3.0,4.0,3.0,3.0,1.0,2.0,2.0,2.0,3.0
1.0,5.0,3.0,5.0,4.0,2.0,3.0,3.0,3.0,5.0,3.0,3.0,4.0,2.0,3.0,2.0,3.0,3.0,2.0,3.0
5.0,5.0,5.0,5.0,1.0,4.0,4.0,3.0,3.0,5.0,3.0,4.0,4.0,4.0,4.0,3.0,4.0,3.0,3.0,4.0
5.0,5.0,5.0,5.0,4.0,5.0,4.0,4.0,4.0,5.0,5.0,4.0,4.0,5.0,5.0,5.0,5.0,3.0,4.0,4.0
5.0,1.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0
3.0,4.0,4.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
4.0,1.0,3.0,5.0,3.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0
3.0,3.0,1.0,3.0,1.0,1.0,2.0,3.0,1.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0
5.0,3.0,2.0,2.0,5.0,3.0,1.0,3.0,1.0,4.0,3.0,4.0,3.0,4.0,3.0,3.0,3.0,2.0,1.0,2.0
,
5.0,nan,5.0,4.0,2.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,5.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
3.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,3.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,4.0,3.0,4.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
4.0,5.0,nan,3.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
4.0,3.0,2.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,4.0,5.0,1.0,4.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,4.0,3.0,5.0
3.0,5.0,4.0,2.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0
5.0,2.0,1.0,3.0,3.0,4.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,2.0,1.0,1.0,3.0,1.0,1.0,1.0,1.0,3.0,1.0,2.0,2.0,1.0,1.0,1.0,2.0,1.0,1.0,2.0
```

```
5.0,5.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,3.0,2.0,3.0,4.0,2.0,1.0,2.0,2.0,1.0,2.0,2.0
5.0,2.0,4.0,3.0,4.0,2.0,2.0,2.0,2.0,4.0,3.0,3.0,3.0,5.0,2.0,2.0,2.0,4.0,2.0,2.0
5.0,5.0,5.0,5.0,4.0,3.0,3.0,3.0,3.0,5.0,3.0,4.0,4.0,3.0,2.0,2.0,3.0,3.0,3.0,3.0
5.0,5.0,4.0,3.0,5.0,4.0,4.0,4.0,4.0,5.0,4.0,4.0,5.0,4.0,3.0,3.0,4.0,3.0,3.0,4.0
1.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,3.0
1.0,4.0,1.0,4.0,3.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,4.0
4.0,2.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
2.0,5.0,3.0,2.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
5.0,5.0,5.0,5.0,3.0,3.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0
4.0,5.0,5.0,3.0,5.0,2.0,2.0,3.0,1.0,3.0,3.0,2.0,3.0,5.0,1.0,1.0,2.0,2.0,2.0,2.0
1.0,2.0,2.0,4.0,5.0,1.0,2.0,2.0,1.0,3.0,2.0,2.0,4.0,2.0,3.0,1.0,2.0,2.0,1.0,3.0
4.0,5.0,3.0,5.0,2.0,3.0,2.0,3.0,3.0,4.0,2.0,3.0,4.0,3.0,3.0,1.0,2.0,2.0,2.0,3.0
1.0,5.0,3.0,5.0,4.0,2.0,3.0,3.0,3.0,5.0,3.0,3.0,4.0,2.0,3.0,2.0,3.0,3.0,2.0,3.0
5.0,5.0,5.0,5.0,1.0,4.0,4.0,3.0,3.0,5.0,3.0,4.0,4.0,4.0,4.0,3.0,4.0,3.0,3.0,4.0
5.0,5.0,5.0,5.0,4.0,5.0,4.0,4.0,4.0,5.0,5.0,4.0,4.0,5.0,5.0,5.0,5.0,3.0,4.0,4.0
5.0,1.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0
3.0,4.0,4.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
4.0,1.0,3.0,5.0,3.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0
3.0,3.0,1.0,3.0,1.0,1.0,2.0,3.0,1.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0
5.0,3.0,2.0,2.0,5.0,3.0,1.0,3.0,1.0,4.0,3.0,4.0,3.0,4.0,3.0,3.0,3.0,2.0,1.0,2.0
```

实现该方法的Python代码位于文件*demo_bt500.py*中。

demo_bt500.py：

```
import argparse
import csv
import sys
import pprint

import numpy as np
from scipy import linalg


def read_csv_into_3darray(csv_filepath):
    """
    Read data from CSV file.

    The data should be organized in a 2D matrix, separated by comma. Each row
    correspond to a PVS; each column corresponds to a subject. If a vote is
    missing, a 'nan' is put in place.

    If some subjects evaluated a PVS multiple times, another 2D matrix of the
    same size [num_PVS, num_subjects] can be added under the first one. A row
    with a single comma (,) should be placed before the repetition matrix.
    Where the repeated vote is not available, a 'nan' is put in place.

    :param csv_filepath: filepath to the CSV file.
    :return: the numpy array in 3D [num_PVS, num_subjects, num_repetitions].
    """

    data = []
    data3dlist = []
    with open(csv_filepath, 'rt') as datafile:
        datareader = csv.reader(datafile, delimiter=',')

        for row in datareader:
            if row != ["", ""]:
                data.append(np.array(row, dtype=np.float64))
```

```
            else:
                data3dlist.append(data)
                data = []
        data3dlist.append(data)

    data3d = np.zeros([len(data3dlist[0]), len(data3dlist[0][0]), len(data3dlist)])

    for r_idx, r_mat in enumerate(data3dlist):
        data3d[:, :, r_idx] = r_mat

    return data3d


def weighed_nanmean_2d(a, wts, axis):
    """
    Compute the weighted arithmetic mean along the specified axis, ignoring
    NaNs. It is similar to numpy's nanmean function, but with a weight.

    :param a: 1D array.
    :param wts: 1D array carrying the weights.
    :param axis: either 0 or 1, specifying the dimension along which the means
    are computed.
    :return: 1D array containing the mean values.
    """

    assert len(a.shape) == 2
    assert axis in [0, 1]
    d0, d1 = a.shape
    if axis == 0:
        return np.divide(
            np.nansum(np.multiply(a, np.tile(wts, (d1, 1)).T), axis=0),
            np.nansum(np.multiply(~np.isnan(a), np.tile(wts, (d1, 1)).T), axis=0)
        )
    elif axis == 1:
        return np.divide(
            np.nansum(np.multiply(a, np.tile(wts, (d0, 1))), axis=1),
            np.nansum(np.multiply(~np.isnan(a), np.tile(wts, (d0, 1))), axis=1),
        )
    else:
        assert False


def one_or_nan(x):
    """
    Construct a "mask" array with the same dimension as x, with element NaN
    where x has NaN at the same location; and element 1 otherwise.

    :param x: array_like
    :return: an array with the same dimension as x
    """
    y = np.ones(x.shape)
    y[np.isnan(x)] = float('nan')
    return y


def get_sos_j(sig_j, u_jkir):
    """
    Compute SOS (standard deviation of score) for presentation jk
    :param sig_j:
    :param u_jkir:
    :return: array containing the SOS for presentation jk
    """
    den = np.nansum(
        stack_3rd_dimension_along_axis(one_or_nan(u_jkir) / np.tile(sig_j ** 2,
(u_jkir.shape[1], 1)).T[:, :, None],
                                       axis=1),
        axis=1)
    s_jk_std = 1.0 / np.sqrt(np.maximum(0., den))
    return s_jk_std
```

```
def stack_3rd_dimension_along_axis(u_jkir, axis):
    """
        Take the 3D input matrix, slice it along the 3rd axis and stack the resulting 2D
matrices
        along the selected matrix while maintaining the correct order.
        :param u_jkir: 3D array of the shape [JK, I, R]
        :param axis: 0 or 1
        :return: 2D array containing the values
            - if axis=0, the new shape is [R*JK, I]
            - if axis = 1, the new shape is [JK, R*I]
    """

    assert len(u_jkir.shape) == 3
    JK, I, R = u_jkir.shape

    if axis == 0:
        u = np.zeros([R * JK, I])

        for r in range(R):
            u[r * JK:(r + 1) * JK, :] = u_jkir[:, :, r]

    elif axis == 1:
        u = np.zeros([JK, R * I])

        for r in range(R):
            u[:, r * I:(r + 1) * I] = u_jkir[:, :, r]

    else:
        NotImplementedError

    return u



def run_alternating_projection(u_jkir):
    """
    Run Alternating Projection (AP) algorithm.

    :param u_jkir: 3D numpy array containing raw votes. The first dimension
    corresponds to the presentation (jk); the second dimension corresponds to the
    subjects (i); the third dimension correspons to the repetitions (r).
    If a vote is missing, the element is NaN.

    :return: dictionary containing results keyed by 'mos_j', 'sos_j', 'bias_i'
    and 'inconsistency_i'.
    """
    JK, I, R = u_jkir.shape

    # video by video, estimate MOS by averaging over subjects
    u_jk = np.nanmean(stack_3rd_dimension_along_axis(u_jkir, axis=1), axis=1)  # mean
marginalized over i

    # subject by subject, estimate subject bias by comparing with MOS
    b_jir = u_jkir - np.tile(u_jk, (I, 1)).T[:, :, None]
    b_i = np.nanmean(stack_3rd_dimension_along_axis(b_jir, axis=0), axis=0)  # mean
marginalized over j

    MAX_ITR = 1000
    DELTA_THR = 1e-8
    EPSILON = 1e-8

    itr = 0
    while True:

        u_jk_prev = u_jk

        # subject by subject, estimate subject inconsistency by averaging the
        # residue over stimuli
        e_jkir = u_jkir - np.tile(u_jk, (I, 1)).T[:, :, None] - np.tile(b_i, (JK, 1))[:,
```

```
:, None]
        sig_i = np.nanstd(stack_3rd_dimension_along_axis(e_jkir, axis=0), axis=0)
        sig_j = np.nanstd(stack_3rd_dimension_along_axis(e_jkir, axis=1), axis=1)

        # video by video, estimate MOS by averaging over subjects, inversely
        # weighted by residue variance
        w_i = 1.0 / (sig_i ** 2 + EPSILON)
        # mean marginalized over i:
        u_jk = weighed_nanmean_2d(
            stack_3rd_dimension_along_axis(u_jkir - np.tile(b_i, (JK, 1))[:, :, None],
axis=1),
            wts=np.tile(w_i, R),  # same weights for the repeated observations
            axis=1)

        # subject by subject, estimate subject bias by comparing with MOS,
        # inversely weighted by residue variance
        b_jir = u_jkir - np.tile(u_jk, (I, 1)).T[:, :, None]
        # mean marginalized over j:
        b_i = np.nanmean(stack_3rd_dimension_along_axis(b_jir, axis=0), axis=0)

        itr += 1

        delta_u_jk = linalg.norm(u_jk_prev - u_jk)

        msg = 'Iteration {itr:4d}: change {delta_u_jk}, u_jk {u_jk}, ' \
              'b_i {b_i}, sig_i {sig_i}'.format(
            itr=itr, delta_u_jk=delta_u_jk, u_jk=np.mean(u_jk),
            b_i=np.mean(b_i), sig_i=np.mean(sig_i))

        sys.stdout.write(msg + '\r')
        sys.stdout.flush()

        if delta_u_jk < DELTA_THR:
            break

        if itr >= MAX_ITR:
            break

    u_jk_std = get_sos_j(sig_j, u_jkir)
    sys.stdout.write("\n")

    mean_b_i = np.mean(b_i)
    b_i -= mean_b_i
    u_jk += mean_b_i

    return {
        'mos_j': list(u_jk),
        'sos_j': list(u_jk_std),
        'bias_i': list(b_i),
        'inconsistency_i': list(sig_i),
    }


if __name__ == "__main__":
    parser = argparse.ArgumentParser()

    parser.add_argument(
        "--input-csv", dest="input_csv", nargs=1, type=str,
        help="Filepath to input CSV file. The data should be organized in a 2D "
             "matrix, separated by comma. The rows correspond to PVSs; the "
             "columns correspond to subjects. If a vote is missing, input 'nan'"
             " instead.", required=True)

    args = parser.parse_args()
    input_csv = args.input_csv[0]

    o_jir = read_csv_into_3darray(input_csv)

    ret = run_alternating_projection(o_jir)
```

```
pprint.pprint(ret)
```

第1部分
附件2

数据文档互换通用格式说明

数据文档互换通用格式的目的是促进参与国际协作主观评价活动的各实验室之间的数据交换。

任何主观评价都是按照5个相互关联的连续阶段开展的：测试准备，测试执行，数据处理，结果的表示和分析。在大型国际活动中，通常的情况是将工作分配给参与活动的不同实验室：

– 在其他参与方的协助下，其中一个实验室负责组织测试，包括确定要评价的质量参数，要使用的测试素材（当前临界但并不过界），测试框架（例如方法、观看距离、各阶段的布置、测试项目演示的顺序），以及测试环境（例如观看条件、介绍性说明）。

– 请自愿参与的实验室提供采用适当技术处理的测试素材（仿真或借助硬件），这些技术对待评质量参数而言具有代表性。

– 另有一方负责剪辑测试磁带。

– 由不同的自愿参与的实验室用经过初步剪辑的磁带进行测试。这一测试可以是盲测。在这种情况下，实验室通过收集评价者的评分来完成测试，而不一定让评价者了解待评质量参数。

– 一般会要求另一参与方协调最终的原始数据的收集，用于结果的处理和编辑，这也可以采用盲测的方式进行。

– 最后，用一种文字/表格或图形表示法来分析结果，并公布最后报告。

提出的格式能够用于收集按照测试定义阶段规定的程序得出的结果。

该格式符合本建议书第1部分和第2部分中所述的评价方法。

该格式由表1-4和表1-5所示结构的文本文档组成。其句法由标签和字段组成，还包括一组有限的保留符号（例如“**[**”“**]**”“”“↵”和“**=**”）。

在容量方面（例如参与实验室、观察者、测试序列和质量参数的数目，评分量表边界，或评分设备的类型）没有固有的限制。

表1-4

用于识别结果的文本文档的格式

识别文档的格式和句法	备注
[测试框架]↵ 类型=“*DSCQS*”或“*DSIS I*”，“*DSIS II*”等↵ 阶段的数目=*1* ≤整数≤ *x*↵ 量表下限=整数↵ 量表上限=整数↵ 显示器尺寸=整数↵ 显示器制造商和型号=字符串↵	[段落标识符] 所用的ITU-R BT.500建议书方法的标识 测试中分配的阶段的数目（(1) 量表的定义（见具体的方法要求，若有的话） 显示器对角线长度（英寸）
[结果] ↵ 结果的数目=*1* ≤整数≤ *y*↵ Result(*j*).Filename(s) =*character string*.DAT ↵ .... Result(*j*).Name =字符串↵ Result(*j*).Laboratory =字符串↵ Result(*j*).Number of observers =*1* ≤整数≤ *N*↵ Result(*j*).Training =“是”或“否”↵ [Result(*j*).Session(*i*).Observers] ↵ O(*k*).First Name =字符串↵ O(*k*).Last Name =字符串↵ O(*k*).Sex =*“F”*或*“M”*↵ O(*k*).Age =*integer*↵ O(*k*). Occupation =字符串↵ O(*k*).Distance =整数↵	[段落标识符] 要考虑的结果文档的数目(1) 完整.DAT（见表7）文档名，包括路径 惯用的结果文档名称 测试实施实验室的标识 观察者的总数 表明训练期间收集的评分是否含在所附的DAT文档中 [段落标识符] 观察者标识 选用 选用 主要的社会-经济群体（例如工人、学生） 以显示器高度表示的观看距离(3 *H*，4 *H*，6 *H*)

(1) 阶段：一个测试可分为若干不同的阶段，以适应最大测试时长要求。不同的阶段可由相同的观察者参加，也可由不同的观察者参加，其间要求他们评价不同的测试项目。将不同测试阶段收集到的结果合并，可得出一套完整的测试结果（演示次数× 每次演示的评分次数）。结果可附在不同的.DAT文档中，每次测试实施都会得出这样的文档。

表1-5

Results.DAT原始数据文本文档的格式

filename.DAT文档的格式和句法	备注
整数整数整数.....↵ 整数整数整数.....↵ 整数整数整数.....↵	DAT原始数据文档由以空格分开的评分值组成。每一观察者应占一行。 原始数据按输入的顺序存放。 数据可以分放在表6中名称为Result(j).Filename(s)(1)的不同DAT文档中。

(1) 见表1-4的注释(1)。

第1部分
附件3
（资料性）

图像内容降质特性

A3-1 引言

某一系统在投入使用之后，可能会处理范围广泛的节目素材，其中一些如果不降低质量，就不适用。在考虑系统的适用性时，必须既了解对于系统来说较严格的节目素材的保护，又了解此时预计出现的质量的降低。其实对正在考虑的系统而言，需要了解的是某种图像内容降质特性。

某些系统的性能可能不是随着素材越来越严格而均匀降低的，这种降质特性对此类系统尤为重要。例如，某些数字和自适应系统对于很大范围的节目素材都能维持较高的质量，但超出这个范围，性能就降低了。

A3-2 降质特性的导出

从概念上讲，图像内容降质特性确定了在系统达到特定质量水平的较长的时间内可能出现的节目素材的比例。图1-5对此做了图示。

图像内容降质特性可用四个步骤导出：

– 步骤1：确定能排定若干图像序列的等级顺序的“临界性”算法尺度，这些序列经过相关系统或系统类别后产生了失真，算法尺度确定的等级顺序相当于观察者完成任务后得到的顺序。这一临界性尺度可能涉及视觉建模的若干方面。

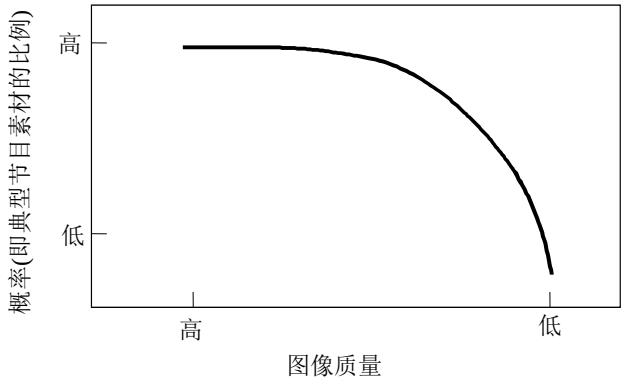
– 步骤2：将临界性尺度用于从典型电视节目中抽取的大量样本，导出可用于估算节目素材出现概率的某种分布，这些节目素材体现了所考虑的系统或系统类别不同水平的临界性。图1-6给出了这种分布的一个示例。

– 步骤3：采用经验方法导出在节目素材的临界性不断提高的情况下系统维持其质量的能力。在实践中，这就要求对系统能得到的质量进行主观评价，采用的节目素材是为抽取步骤2确定的临界性范围的样本而选定的。由此得出一个函数，将系统能得到的质量与节目素材的临界性关联起来。图1-7示出了这种函数的一个例子。

– 步骤4：综合步骤2和步骤3得出的信息，以便导出具有图1-5所示形式的图像内容降质特性。

图1-5

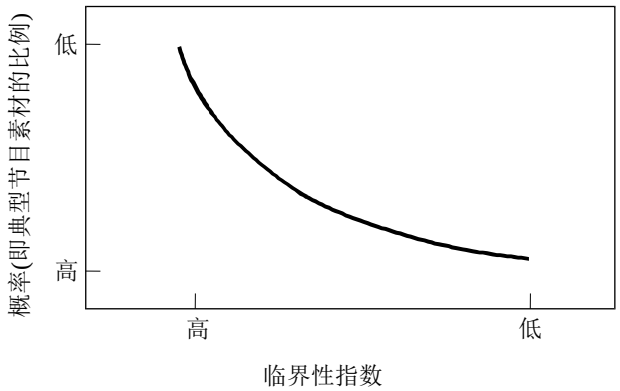
可能的图像内容降质特性示例的图形表示



BT.0500-01-5

图1-6

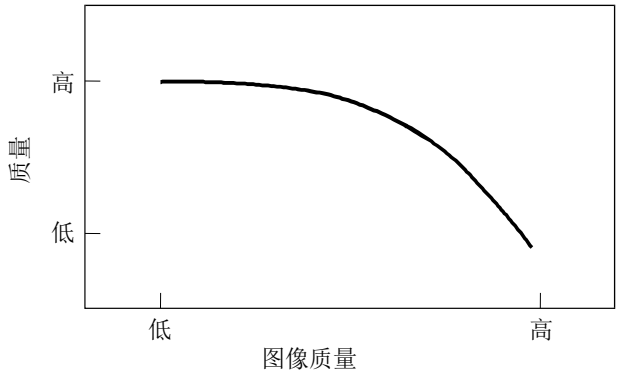
不同临界性水平的素材出现的概率



BT.0500-01-6

图1-7

将质量与节目素材的临界性关联起来的可能函数示例



BT.0500-01-7

A3-3 降质特性的使用

降质特性是研究系统适用性的一个重要工具，可用于全面了解在可能遇到的节目素材范围内有望达到的性能。降质特性可用于以下3个方面：

– 在系统设计阶段可用于优化参数（例如图像源的分辨率、比特率、带宽），使之更符合服务需求；

– 用于研究单一系统的适用性（例如预测运行期间降质的出现与否和严重程度）；

– 用于评价替代系统的相对适用性（例如比较降质特性以确定哪个系统更适用）。应注意，虽然类型相似的替代系统可能使用相同的临界性指数，但类型不同的系统却有可能使用不同的临界性指数。不过，降质特性表示的只是实践中所见的不同质量等级的概率，即便是从系统特定的不同临界性指数导出的特性，也可以直接加以比较。

本建议书所述的方法尽管提供了衡量某个系统图像内容降质特性的手段，但仍无法全面预测电视观众对系统的可接受性。为了获得这一信息，可能有必要让一些观察者观看由所研究的系统编码的节目，并考察其评论。

第3部分附件1介绍了一个数字电视图像内容降质特性的例子。

第1部分
附件4
（资料性）

确定节目内容和传输条件的复合降质特性的方法

A4-1 引言

复合降质特性以既明确考虑节目内容又明确考虑传输条件的方式将感知的图像质量与实践中的出现概率联系起来。

原则上讲，这种特性可从观察次数、测试次数和接收点足够多的主观研究中导出，以形成一个代表可能的节目内容和传输条件流行程度的样本。但在实践中，这类实验可能不太实用。

本附录描述一种更易实现的用于确定复合降质特性的替代程序。这种方法分为3个阶段：

– 节目内容分析，
– 传输频道分析，
– 导出复合降质特性。

A4-2 节目内容分析

这一阶段包括两个操作。首先导出适于衡量节目内容的尺度，然后估计该尺度的各数值在实践中出现的概率。

节目内容尺度是个统计指标，用于捕捉节目内容的各个方面，这些方面强调的是待测系统以能感知的方式忠实再现节目素材的能力。显然，这种尺度若以适当的感知模型为基础将是有益的。但是没有这种模型，某种尺度若能在某一方面捕捉到视频帧/域内部或之间存在的空间多样性的程度，也就足够了，条件是这一尺度与感知的图像质量之间存在大致的单调关系。对于采用完全不同的图像显示方式的系统（或系统类别），可能有必要采用不同尺度。

一旦选定了合适的尺度，就有必要估计这一统计指标各数值出现的概率。要做到这一点，可采用下述两种方式中的一种：

– 采用经验程序，对分辨率、帧速率和图像宽高比都适合待测系统的演播室制式的10 s节目段，随机抽取约200段进行分析。对这一样本的分析可得出统计指标各数值的相对出现频次，作为实践中出现概率的估计值；或者
– 采用理论方法，用一个理论模型来估计出现概率。应注意，尽管优先选用经验方法，但在一些特定情况下（例如随着新的制作技术的出现，关于节目内容的信息不足）可能有必要采用理论方法。

上述分析可形成内容统计指标各数值的一个概率分布（也见附件3）。将这一概率分布与传输条件分析的结果相结合，为替代程序的最后阶段做好准备。

A4-3 传输频道分析

这一阶段也包括两个操作。首先导出适于衡量传输频道性能的尺度，然后估计该尺度的各数值在实践中出现的概率。

传输频道尺度是个统计指标，用于捕捉频道性能的各个方面，这些方面影响的是待测系统以能感知的方式忠实再现源素材的能力。显然，这种尺度若以适当的感知模型为基础将是有益的。但是没有这种模型，某种尺度若能在某一方面捕捉到频道产生的制约，也就足够了，条件是这一尺度与感知的图像质量之间存在大致的单调关系。对于采用完全不同的图像编码方式的系统（或系统类别），可能有必要采用不同尺度。

一旦选定了合适的尺度，就有必要估计这一统计指标各数值出现的概率。要做到这一点，可采用下述两种方式中的一种：

– 采用经验程序，在约200个随机选定的时刻和接收点衡量频道性能。对这一样本的分析可得出统计指标各数值的相对出现频次，作为实践中出现概率的估计值；或者

– 采用理论方法，用一个理论模型来估计出现概率。应注意，尽管优先选用经验方法，但在一些特定情况下（例如随着新的传输技术的出现，关于频道性能的信息不足）可能有必要采用理论方法。

上述分析可形成频道统计指标各数值的一个概率分布。将这一概率分布与节目内容分析的结果相结合，为替代程序的最后阶段做好准备。

A4-4 导出复合降质特性

这一步包括一次主观实验，其中节目内容和传输条件按照头两步确定的概率联合变化。

所用的基本方法是双激励连续质量程序，具体到活动序列而言推荐采用10 s的形式（见第2部分附件2）。此处基准图像是某种适当制式（例如分辨率、帧速率和图像宽高比都适合待测系统的制式）的具有演播室质量的图像。对比而言，在测试过程中显示的图像与待测系统在选定的频道条件下要收到的图像相同。

按照这一方法的头两步确定的概率来选择测试素材和频道条件。在按照内容统计指标对测试素材的每一段进行分析以确定其支配值之后，由测试素材的各段组成一个选择库。然后从这个库中对素材进行抽样，使得样本覆盖统计指标的所有可能取值，对于较低的临界水平，抽样较稀，对于较高的临界水平则抽样较密。频道统计指标的可能取值以类似方式选取。然后将这两种来源独立的影响随机组合在一起，形成已知概率的内容与频道条件的某种组合。

这些研究结果将感知的图像质量与实践中的出现概率联系起来，可用于研究某个系统的适用性或用于从适用性强弱的角度比较各系统。

第1部分
附件5
（资料性）

背景效应

在某一图像的主观评分受到损伤的出现顺序和严重程度的影响时，就产生了背景效应。例如，如果在一连串轻微受损的图像之后显示一个严重受损的图像，观察者对这一图像的评分可能无意中会比通常情况下的评分低。

不同国家的4个实验室共同对与评价图像质量的3种方法（DSCQS法、DSIS法的变型II和一种比较方法）产生的结果相关的背景效应进行了调研。测试素材用MPEG（ML@MP）编码形成，同时降低了水平分辨率。对每一测试系列应用4个基本测试条件（B1、B2、B3、B4）和6个背景测试条件，其中一个测试系列说明弱背景损伤，另一个说明强损伤。对两个测试系列均采用了上述3种方法。背景效应表明以弱损伤为主的测试的结果与以强损伤为主的测试的结果之间的差别。用基本测试条件B2和B3确定背景效应。

实验室共同研究的结果表明DSCQS法不存在背景效应。对于DSIS法和比较方法，背景效应明显，而DSIS法的变型II则存在最强的背景效应。结果显示，以弱损伤为主的图像可引起较低的评分，而以强损伤为主的图像则可引起较高的评分。

调查结果表明，ITU-R推荐的DSCQS法是将主观图像质量评价中的背景效应降质最弱的较好方法。

ITU-R BT.1082报告给出了关于上述调研的更多资料。

第1部分
附件6
（资料性）

空间和时间信息测量

以下给出的空间和时间信息测量法为完整测试片段上的各个帧单独赋值。在时间序列值中，该结果通常将在某种程度上有所变化。以下给出的感知信息测量法用最大函数（片段的最大值）消除了这种可变性。可以对可变性本身开展有益的研究，例如，逐帧形式的空间－时间信息图。在测试片段上使用信息分发也允许用场景剪辑对场景进行更好的评估。

空间感知信息（SI）：是一种通常用于表示图像空间细节数量的测量法。它通常高于空间上更复杂的场景。它并不意味着是一种信息熵测量法，也与通信理论中定义的信息无关。空间感知信息*SI*基于Sobel滤波器。在时间*n*（F_n），各个视频帧（亮度平面）首先用Sobel滤波器[Sobel（F_n）]进行滤波。然后，计算各个经Sobel滤波器滤除后的帧中像素的标准差（std_{space}）。为视频片段中各个帧重复该操作，产生场景空间信息的时间序列。选择时间序列（$\max_{time}$）中的最大值，以表示场景的空间信息内容。该过程可以用方程式的形式来表示：

$$SI = max_{time} \{std_{space} [\text{Sobel}(F_n)]\}$$

时间感知信息（TI）：是一种通常用于表示视频片段时间变化次数的测量法。它通常高于高速运动的片段。它并不意味着是一种信息熵测量法，也与通信理论中定义的信息无关。

时间信息测量法TI，当作所有i和j的$M_n(i, j)$空间（std_{space}）上的标准差最大时间值（max_{time}）来计算。

$$TI = max_{time} \{std_{space} [M_n(i, j)]\}$$

其中，$M_n(i, j)$指的是帧中相同位置上各像素之间的差异，但属于两个随后的帧，就是说：

$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$$

其中，$F_n(i, j)$是时间上第n帧第i行和第j列处的像素。

注 – 对包含场景剪辑的场景，可以给定两个值：在一个值中，时间信息测量法中包括了场景剪辑，在另一个值中，测量法不包括场景剪辑。

第1部分
附件7
（资料性）

术语和定义

Algorithm（算法）	一项或多项图像处理操作
AVI	音频视频交错
CCD	电荷耦合器件
CI	置信区间
CIF	通用中间制式（H.261建议书中为视频电话定义：352行×288像素）
CRT	阴极射线管
DSCQS	双激励连续质量量表法
DSIS	双激励损伤量表法
LCD	液晶显示器
MOS	平均评分法
SC	激励比较法
PDP	等离子显示板
PS	节目片段
QCIF	四分之一CIF制式（H.261建议书中为视频电话定义：176行×144像素）
SAMVIQ	多媒体视频质量的主观评价
Sequence（序列）	经综合处理或未经处理的场景

Scene（场景）	视听内容
S/N	信噪比
SI	空间信息
SIF	标准中间制式[ISO 11172（MPEG-1）中定义：352行×288像素×25帧/秒和352行×240像素×30帧/秒]
SP	同时呈现
SQCIF	子QCIF
SS	单激励法
SSCQE	单激励连续质量评估法
std	标准差
TI	时间信息
TP	测试演示
TS	测试阶段
VTR	磁带录像机

第2部分

主观图像评价方法描述

1 引言

本部分详细介绍了进行主观图像质量评价所需的各个图像评价的方法。在一些情况下，这与第1部分第2节提供的通用评价特性有所差异。

为确保主观图像质量评价的结果能够被其他实验室正确解读，重要的是要使流程的详细说明可用，以及将所使用的方法的任何变型与其他想要复制该评价流程的实验室可能要求的所有需要的额外信息一同记录在案。

2 建议的图像评价方法

附件1 双激励损伤量表（DSIS）

附件2 双激励连续质量量表（DSCQS）

附件3 单激励（SS）法

附件4 激励比较法

附件5 单激励连续质量评估（SSCQE）

附件6 同时双激励连续评估（SDSCE）

附件7 多媒体视频质量的主观评价（SAMVIQ）

附件8 用于评估视频素材质量的专家观看协议（EVP）

3 说明

ITU-R BT.1082报告对多维标度法和多元法等其他技术做了说明，这些技术还有待进一步研究。

至今所描述的所有方法都有其优势和限制，目前不可能从中明确推荐一种。因此，研究人员仍可自行选择最适合其所处环境的方法。

各种各样方法的限制表明，单单特别重视某一种方法是不明智的。因此，考虑更“完备的”方法可能比较合适，比如要么使用多种方法，要么使用多维方法。

第2部分
附件1

双激励损伤量表（DSIS）法
（EBU法）

A1-1 总体说明

典型的评价要么会要求评价一个新系统的损伤，要么会要求评价传输路径对损伤的影响。对于测试组织者来说，第一步包括选择足够的测试素材，以便要进行的评价富有意义，并确定应使用的测试条件。如果参数变化的影响受到关注，则有必要按照大致相等的为数不多的几个步长，选择覆盖损伤等级范围的一组参数值。而对参数值不是如此变化的新系统进行评价时，要么需要加上附加的但主观上类似的损伤，要么应使用另一种方法，如第2部分附件2中的方法。

双激励损伤量表（DSIS）法（EBU法）是一种循环方法，在这种方法中，评价者首先看到无损伤的基准图像，然后又看到受损伤的同一图像。随后要求评价者根据第一幅图像来评价第二幅。在持续半小时以内的测试阶段里，向评价者以随机的顺序演示一系列带有随机损伤的图像或序列，涵盖所有必要组合。无损伤的图像包含在这些待评图像或序列中。在一系列测试阶段结束时，计算每一测试条件和测试图像的平均评分。

该方法使用损伤量表，相对于较大的损伤而言，通常可以发现这种量表对较小的损伤可得出更为稳定的结果。虽然该方法有时用于有限的损伤范围，但它更适合用于整个的损伤范围。

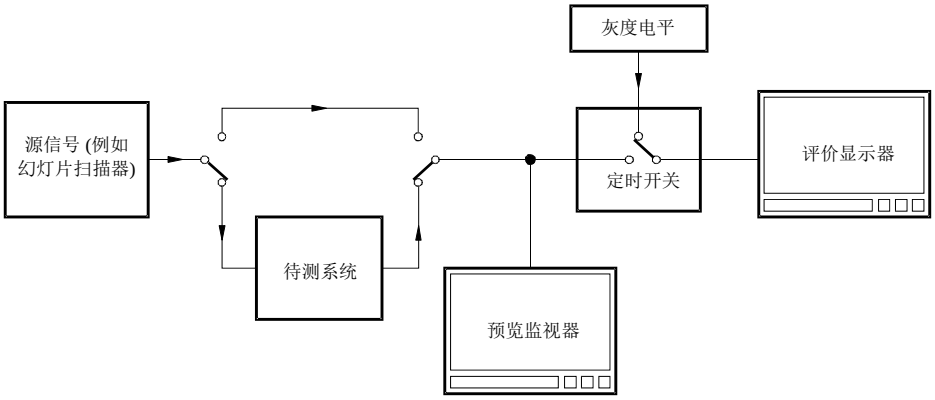
A1-2 总体布置

观看条件、源信号、测试素材、观察者以及结果的表示在第1部分第2节中做了规定或按照第1部分第2节加以选择。

测试系统的总体布置应如图2-1所示。

图2-1

DSIS法中测试系统的总体布置



BT.0500-02-1

评价者观看的是一台评价显示器，其信号来自一个定时开关。与定时开关相连的信号通路可直接连至源信号，也可通过待测系统间接连至源信号。评价者会看到一系列图像或序列，它们是成对排列的，每对中的第一个是直达的自源信号，第二个是经过待测系统的相同图像。

A1-3 测试素材的演示

一个测试阶段由多次演示组成。演示的结构有下述I和II两种变型。

变型I： 基准图像或序列以及测试图像或序列只演示一次，如图2-2(a)所示。

变型II： 基准图像或序列以及测试图像或序列演示两次，如图2-2(b)所示。

变型II： 比变型I费时，在需要鉴别的损伤非常小或待测的是活动序列时可以使用。

A1-4 分级量表

应采用五级损伤量表：

5 不可察觉

4 可察觉，但不讨厌

3 稍微讨厌

2 讨厌

1 很讨厌。

评价者应使用一种给出非常明确的量表的表格，有编了号的框或以其他方式来记录它们的分级。

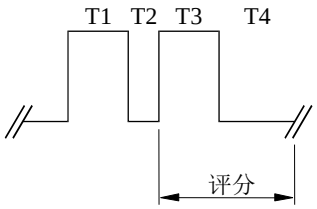
A1-5 评价须知

在每一测试阶段开始时，应向观察者解释评价类型、分级量表、顺序及定时（基准图像、灰度、测试图像、评分期）。在待评价图像中显示要评价的损伤的范围和类型，该图像应不同于测试中要用的图像，但具有相似的灵敏度。不能暗示看到的最低质量必须对应于最低的主观等级。应要求观察者根据图像给出的总体印象来做出其判断，并把这种判断用规定主观尺度的措词来表示。

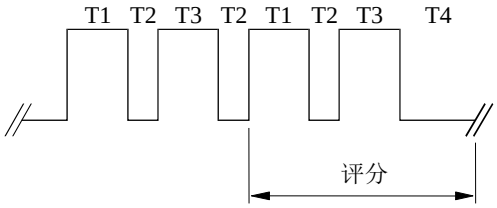
应要求观察者在T1和T3的整个持续时间内观看图像。只允许在T4期间内评分。

图2-2

测试素材的演示结构



A)变型I



B) 变型II

BT.0500-02-2

演示阶段

T1 =	10 s	基准图像
T2 =	3 s	由200 mV图像电平产生的中灰度场
T3 =	10 s	测试条件
T4 =	5至11 s	中灰度场

经验显示，将T1和T3期间延长超过10 s并不会提高评价者对图像序列进行评级的能力。

A1-6 测试阶段

图像和损伤的演示应以伪随机顺序进行，每一测试阶段最好采用不同的序列。在任何情况下，同一测试图像或序列，不管损伤程度是否相同，绝不应连续演示两次。

在选择损伤范围时，应使得大多数观察者用到所有等级；应以总平均分（实验中所有判断的平均值）接近3为目标。

一个测试阶段应大致不超过半小时，包括解释和准备时间；测试序列可从表示损伤范围的几幅图像开始；对这几幅图像的判断在最后结果中不予考虑。

关于损伤程度选择的其他见解在第1部分附件2中给出。

第2部分
附件2

双激励连续质量量表（DSCQS）法

A2-1 总体说明

一次典型的评价可能需要评价一个新系统的质量，或需要评价传输路径对质量的影响。在无法提供可展示各种质量的测试激励和测试条件的情况下，双激励法被认为特别有用。

该方法是一种交替方法，因为在这种方法中，要求评价者观看一对图像，每一个都来自同一信号源，只不过一个经过要检查的流程，另一个是直达的信号源。要求评价者评价二者的质量。

在持续半小时以内的各测试阶段里，向评价者以随机的顺序演示一系列带有随机损伤的图像对（每对中两幅图像的顺序是随机的），涵盖所有必要的组合。在所有测试阶段结束时，计算每一测试条件和测试图像的平均评分。

A2-2 总体布置

观看条件、源信号、测试素材、观察者以及对测试的介绍在第1部分第2节中做了规定或按照第1部分第2节加以选择。测试阶段的描述见第2部分附件1的第A1-6节。

测试系统的总体布置应如图2-3所示。

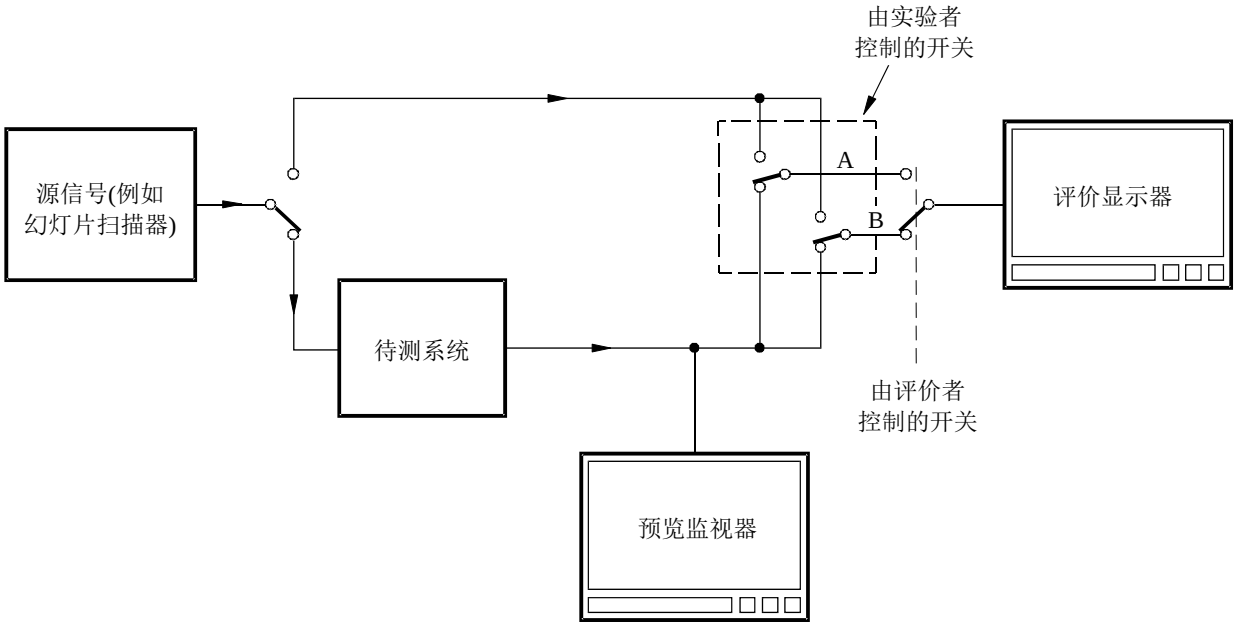
A2-3 测试素材的演示

一个测试阶段由多次演示组成。对于只有一位观察者的变型I，每次演示时观察者都可以在信号A和信号B之间自由转换，直到观察者得出与每一信号的质量相关的心理尺度为止。对于同时有几位观察者的变型II，在记录结果之前，条件对要显示一次或多次，每次持续时间相同，以便让观察者得出与这一对条件的质量相关的心理尺度，然后再把条件对显示一次或多次，同时记录结果。重复的次数取决于测试序列的长度。对于静止图像，使用3-4 s的序列并重复5次（在最后2次期间评分）可能是合适的。对于受到时变扰动的活动图像，10 s的序列和2次重复（在第2次重复期间评分）可能是合适的。图2-4显示出了演示的结构。

如果现实情况把可用序列的长度限制在不到10 s，则可以把这些比较短的序列组合成段，将显示时间扩展到10 s。为了把连接点处的不连续性降至最低，由连续的序列组成的段在时间上可能是逆向的（有时称为“回文式”显示）。必须多加小心，确保作为逆向的段显示的测试条件能体现因果过程，即测试条件是逆向显示的源信号通过待测系统而得到的。

图2-3

DSCQS法中测试系统的总体布置



这种方法有下述两种变型。

变型I 评价者一般是单独的，评价者可以在A和B两种条件之间切换，直到他对每一种条件都认为得出了满意的评分为止。A线路和B线路都提供了直达基准图像，或通过待测系统提供了图像。但哪条线路得到哪个图像则在一个测试条件和下一个测试条件之间是随机变化的，它们由实验者注明，但不公布。

变型II 来自A线路和B线路的图像连续显示给评价者，供评价者给出对每一图像的评分。对于每次演示，A线路和B线路都像上述变型I那样得到图像。

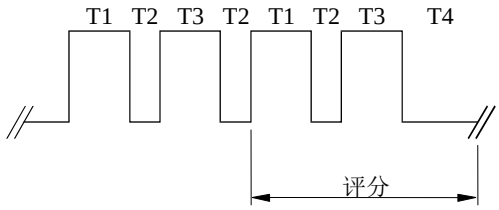
A2-4 分级量表

这种方法要求评价每一测试图像的两种版本。每对测试图像中，有一个是无损的，而另一个可能包含损伤，也可能不包含损伤。无损的图像就作为基准纳入，但不告诉观察者哪个是基准图像。在测试系列中，基准图像的位置是以伪随机方式变化的。

只要求观察者在垂直标尺上标出记号来评价每次演示的总体图像质量。垂直标尺是成对打印的，涵盖了每个测试图像的两次演示。为了防止量化误差，标尺提供了连续的评分系统，但分成了相等的5段，与ITU-R的五级质量量表相对应。对5个等级进行分类所用的相关术语与平常所用的一样；不过此处是将其当做一般性的指导，在分数表中按对排布的10个标尺的每一行第一个标尺的左侧标出。图A2-5显示了典型评分表的一部分。为了防止在标尺的划分与测试结果之间可能出现的混淆，标尺用蓝色打印，结果用黑色记录。

图2-4

测试素材的演示结构



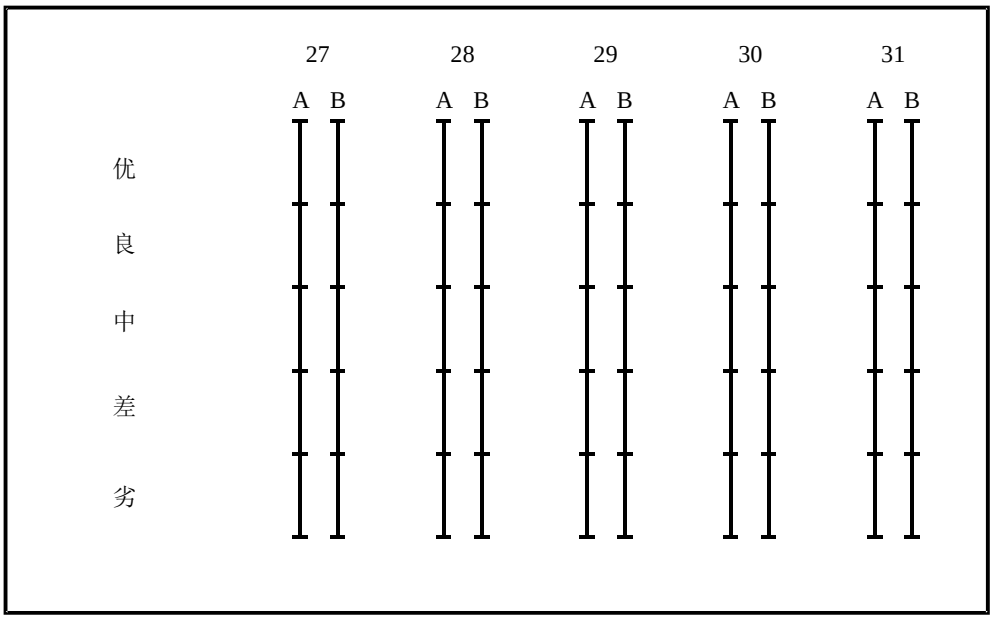
BT.0500-02-4

演示阶段

T1 =	10 s	基准图像
T2 =	3 s	由200 mV图像电平产生的中灰度场
T3 =	10 s	测试条件
T4 =	5至11 s	中灰度场

图2-5

采用连续标尺的质量评分表的一部分*



BT.0500-02-5

* 在采用DSCQS法的测试阶段内规划测试项目的布置时，实验者最好应进行检验，确信实验中未产生系统差错。不过完成这种置信检验的方法还有待研究。

A2-5 结果的分析

将每一测试条件的评价对（基准和测试）从评分表上的度量长度转换为归一化的0至100范围内的评分。然后计算基准条件与测试条件之间存在的评价差别。其他程序在第1部分附件2中给出。

经验显示，从不同测试序列中获得的评分取决于所用测试素材的临界性。对不同的测试序列分别显示结果，可更全面地了解编解码器的性能，而将结果表示为评价中所用的所有测试序列的一个综合平均分则无法做到这一点。

如果将单个测试序列的结果在横轴上按照测试序列临界性的高低顺序排列，就有可能给出待测系统图像内容降质特性的概约图形说明。不过这种表达形式只是说明了编解码器的性能，并未表明具有给定临界性的序列出现的可能性（见第1部分附件2）。在能够获得系统性能的这种更完整的说明之前，需要对测试序列的临界性和具有给定临界性的序列出现的概率开展进一步研究。

A2-6 结果的分析

在使用这种DSCQS法时，将DSCQS数值与其他测试协议所用的形容词（例如DSIS法中的不可察觉，可察觉但不讨厌……）形成关联，从而得出关于待测条件的质量的结论，会有一定风险，甚至出现差错。

要注意，用DSCQS法得出的结果不应看做绝对评分，而应看做基准条件与测试条件之间的评分差值。因此，将评分与某个说明质量的术语联系起来是不对的，即便是与DSCQS协议本身所用的术语（例如优，良，中……）联系起来也是不对的。

在评价开始之前决定可接受标准，这在任何测试程序中都很重要。在采用DSCQS法时这一点极为重要，因为缺乏经验的使用者对于由这种方法产生的质量量表值有误解的趋势。

第2部分
附件3

单激励（SS）法

在单激励法中，显示单一的图像或一个图像序列，并为评价者提供一份整个演示的索引。测试素材可以只包含测试序列，也可以既包含测试序列，又包含其相应的基准序列。对于后一种情况，基准序列作为一个单独的激励显示，并像其他测试激励那样进行评分。

A3-1 总体布置

观看条件、源信号、条件的范围和锚定、观察者、对评价的介绍以及结果的表示在第1部分第2节中做了规定或按照第1部分第2节加以选择。

A3-2 测试素材的选择

对实验室测试而言，测试图像的内容应按照第1部分第2.3节所述加以选择。

一旦选定了内容，就要准备测试图像，以反映正在考虑的设计选项或者某一（或某些）因素的范围。在考察两个或多个因素时，可以以两种方法来准备图像。第一种，每个图像只代表每一因素的一个等级。在另一种方法中，每个图像代表要考察的每一因素的一个等级，但在几个图像之间，每一因素的每一等级都与所有其他因素的每一等级同时存在。两种方法都能将结果明确地划归具体因素。后一种方法还可以检测不同因素之间的相互作用（即非加性效应）。

A3-3 测试阶段

测试阶段由一系列评价试验组成。这些评价试验应以随机顺序给出，每一观察者最好采用不同的随机顺序。在采用单一随机顺序的序列时，演示结构有I（SS）和II（SSMR）两种变型，分别如下：

a) 在测试阶段，测试图像或序列只演示一次；第一阶段开始时，应播放几个“模拟演示”（见第1部分第2.7节的说明）；实验通常要确保同一图像不会以同样的损伤程度连续演示两次。

典型的评价试验由3种显示组成：一个是中灰度适应场，一个是激励场，还有一个是中灰度后期曝光场。这些显示的持续时间随着观察者的任务、素材和要考虑的意见或因素而变化，但分别为3、10和10 s并不罕见。观察者指数要么在激励场显示期间收集，要么在后期曝光场显示期间收集。

b) 将测试阶段分成3个演示，测试图像或序列演示3次。每个演示都只包含所有待测图像或序列一次；每一演示开始时，在显示器上公布一条消息（例如“演示1”）；第一个演示用于稳定观察者的意见；从这次演示中得出的数据在测试结果中不予考虑；对图像或序列的评分是对从第二个和第三个演示中得出的数据进行平均得到的；试验通常要确保每一演示中图像或序列的随机顺序采用下述限定：

- 某一给定图像或序列的所在位置与其他演示中的位置不同；
- 某一给定图像或序列的所在位置不能正好在其他演示中同一图像或序列的位置之前。

典型的评价实验由2种显示组成：一个是激励场，另一个是中灰度后期曝光场。这些显示的持续时间随着观察者的任务、素材和要考虑的意见或因素而变化，但建议分别为10和5 s。观察者指数只能在后期曝光场的显示期间收集。

变型II（SSMR）引入了完成一个测试阶段所需的明确的额外时间（45 s与23 s，对每一待测图像或序列而言）；尽管如此，它还是降低了一个测试阶段内变型I的结果对图像或序列的顺序的强烈依赖。

另外，实验结果显示，变型II在评分范围内可以形成约20%的跨度。

A3-4 SS的种类

一般而言，在电视评价中采用了三种SS法。

A3-4.1 形容词分类判断法

在形容词分类判断中，观察者将图像或图像序列划归一组类别中的某一类别，这组类别通常按语义来规定。类别可以表明关于是否检测到某种属性的判断（例如用于确定损伤门限）。评价图像质量和图像损伤的类别量表使用最为频繁，表2-1给出了ITU-R的量表。在运行显示中，有时也用到半级。在特殊情况下也使用了评价文字的易读性、阅读费力度和图像实用性的量表。

表2-1

ITU-R质量和损伤量表

五级量表	
质量	**损伤**
5 优	5 不可察觉
4 良	4 可察觉，但不讨厌
3 中	3 稍微讨厌
2 差	2 讨厌
1 劣	1 很讨厌

对于每个条件，由这种方法可得出量表各类别之间的判断分布。对响应进行分析的方式取决于判断（检测等）和想要获取的信息（检测门限、条件的等级或主要趋势、各条件之间的心理“距离”）。有许多分析方法可以使用。

A3-4.2 数值分类判断法

对采用11级数值分类量表的单激励程序（SSNCS）进行了研究，并与图形和比率量表做了比较。ITU-R BT.1082报告对这项研究做了说明。研究表明，在无法得到基准的情况下，SSNCS法在灵敏性和稳定性方面具有明显的优势。

A3-4.3 非分类判断法

在非分类判断中，观察者为显示的每一图像或图像序列指定一个数值。这种方法有两种形式。

连续量表是分类法的一种变型。在连续量表中，观察者在连接两个语义标号（例如表3中分类量表的两端）的直线上为每一图像或图像序列指定一个点。这种量表有可能在中间点上包括另外的标号作为基准。将距量表某一端的距离作为每一条件的指标。

在数值量表中，评价者为每一图像或图像序列指定一个数字，该数字反映了在某一规定的尺度（例如图像锐度）方面得出的图像或图像序列的判断等级。所用数字的范围有可能受限制（例如0-100），也有可能不受限制。有时，指定的数字从“绝对”意义上说明判断等级（不像某些形式的幅度估值那样直接提及其他图像或图像序列的等级）。在其他情况下，数字用于说明相对于之前所用“标准”的判断等级（例如幅度估值、分段法和比率估值）。

由两种形式都可得出每一条件的某种数值分布。所用的分析方法取决于判断的类别和所需的信息（例如等级、主要趋势、心理“距离”）。

A3-4.4 性能法

正常观看的某些方面可以用由外部控制的任务（寻找目标信息、阅读文字、辨别目标等）的性能表示。然后可以将某种性能尺度，例如完成这种任务的准确度和速度，作为衡量图像或图像序列的一个指标。

由性能法可得出每一条件的准确度或速度评分的分布。分析集中在确立具有集中趋势（或离中趋势）的各条件之间的关系上，并常常使用方差分析或类似技术。

第2部分
附件4

激励比较法

在激励比较法中，显示两个图像或图像序列，由观察者给出一个指标，表示两个演示之间关系。

A4-1 总体布置

观看条件、源信号、条件的范围和锚定、观察者、对评价的介绍以及结果的表示在第1部分第2节中做了规定或按照第1部分第2节加以选择。

A4-2 测试素材的选择

按照与SS法相同的方式产生所用的图像或图像序列。形成的图像或图像序列则加以组合，形成评价实验中所用的图像对。

A4-3 测试阶段

评价实验将使用一个显示器或两个匹配良好的显示器，并且一般像单激励情况那样进行。如果使用一个显示器，尝试将包括一个额外的激励场，持续时间与第一个相同。在这种情况下，比较好的做法的是确保在各次尝试中，一对中的两个组成部分在第一个位置和第二个位置上出现的频度相同。如果使用两个显示器，则激励场要同时显示。

判断是比较所有可能的条件对，与此同时激励比较法对各条件之间的关系进行更为全面的评价。但如果这样做需要的观察量过大，则有可能在评价者之间分配观察量，或者使用从所有可能的对中抽出的一些样本。

A4-4 激励比较法的种类

在电视评价中采用了三种激励比较法。

A4-4.1 形容词分类判断法

在形容词分类判断中，观察者将某一对中各组成部分的关系划归一组类别中的某一类别，这组类别通常按语义来规定。这些类别可以表明可察觉的差别存在与否（例如“相同”“不同”），或者表明可察觉差别的存在与否和方向（例如“小”“相同”“大”），或者表明对程度和方向的判断。表2-2示出了ITU-R的比较量表。

表2-2

比较量表

–3	甚差
–2	较差
–1	稍差
0	相同
+1	稍好
+2	较好
+3	甚好

对于每个条件对，由这种方法可得出量表各类别之间的判断分布。对响应进行分析的方式取决于判断（例如差别）和想要获取的信息（刚能看出差别、条件的等级、各条件之间的“距离”等）。

A4-4.2 非分类判断法

在非分类判断中，观察者用一个数值表明一个评价对中各组成部分的关系。这种方法有两种形式：

– 在连续量表中，观察者在连接两个标号（例如“相同”-“不同”或表4中分类量表的两端）的直线上为每一关系指定一个点。这种量表有可能在中间点上包括另外的基准标号。将距直线某一端的距离作为每一条件对的值。

– 在另一种方式中，评价者为每一关系指定一个数字，该数字反映了在某一规定的尺度（例如质量差别）方面得出的这一关系的判断等级。所用数字的范围有可能受限制，也有可能不受限制。指定的数字从“绝对”意义上或者用“标准”对中的术语对关系加以说明。

由两种形式都可得出每一对条件的某种数值分布。所用的分析方法取决于判断的类别和所需的信息。

A4-4.3 性能法

在某些情况下，性能尺度可从激励比较程序中导出。在迫选法中，准备条件对时，让其中一个组成部分含有特定级别的某种属性（例如损伤），而另一个含有其他级别的该属性或不含该属性。请观察者决定哪个组成部分的该属性级别更高/更低，或决定哪个组成部分包含该属性；将性能的准确度和速度作为衡量条件对中各组成部分关系的指标。

第2部分
附件5

单激励连续质量评估（SSCQE）

数字电视压缩的引入将对随场景和内容变化的图像质量产生损伤。即便在很短的数字编码视频片段内，质量也会随场景内容的不同而有很大变化，并且损伤存在的时间有可能非常短。常规的ITU-R方法本身不足以评价这种素材。另外，实验室测试中的双激励法没有再现单激励家庭观看条件。因此，曾认为有益的做法是连续衡量数字编码视频的主观质量，其中被试观看素材一次，没有基准源信号。

有鉴于此，已经开发出了单激励连续质量评价（SSCQE）技术并进行了测试。

A5-1 记录设备和设备配置

应使用连接至计算机的电子记录手持设备来记录被试得出的质量评价。这种设备应具备如下特性：

– 不带弹簧复位的滑块机构，
– 10厘米（cm）的直线移动范围，
– 位置固定或能安装在桌面上，
– 每秒记录两个样本。

A5-2 测试协议的一般形式

应向被试提供下述制式的测试阶段：

– *节目段*（PS）：一个节目段对应着按某一待评质量参数（QP)（例如比特率）处理的一种节目类型（例如体育、新闻、戏剧）；每个节目段应持续至少5分钟；
– *测试阶段*（TS）：一个测试阶段是由PS/QP的一种或多种不同组合构成的一个序列，其中没有间隔且按随即顺序排列。每个测试阶段至少有一次含有PS和QP，但不必含有全部的PS/QP组合；每个TS的长度应在30分钟和60分钟之间；
– *测试演示*（TP）：一个TP代表某次测试的总体性能。一个TP可以划分为若干TS，以便符合最大时间长度要求和评价所有PS/QP对的质量。如果PS/QP对的数目有限，TP可由相同的测试阶段重复构成，以便在足够长的时间段内进行测试。

对于服务质量评价，应引入伴音。在这种情况下，应认为在进行测试之前对伴音素材的选择与对视频素材的选择具有同等的重要性。

最简单的测试制式是使用单一的PS和单一的QP。

A5-3 观看参数

观看条件应为第1部分目前规定的那些或第3部分给出的应用特定条件。

A5-4 分级量表

在测试须知中，应让被试了解手持设备滑块机构的移动范围与第1部分第5.4节所述的连续质量量表是相互对应的。

A5-5 观察者

应聘用至少15位非专家被试，且具备目前在第1部分第2.5节中推荐的条件。

A5-6 观察者须知

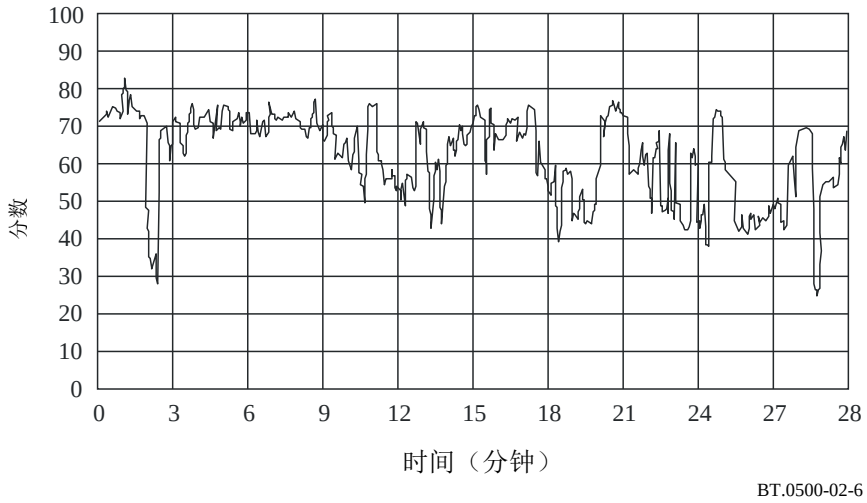
对于服务质量评价（带有伴音）的情况，应告知观察者考虑总体质量，而不只是视频的质量。

A5-7 数据的表示、结果的处理和表示

应将所有测试阶段的数据合并。这样就能得到单一一幅图，表示随时间而变的平均质量评分$q(t)$，作为所有观察者针对每一节目段、质量参数或每一完整测试阶段的质量分级的平均值（见图2-6中的示例）。

图2-6

测试条件：Codex X/节目段：Z



无论如何，只有在计算某一节目段的平均值时，不同观察者反应时间上的差异才有可能影响评价结果。正在开展研究，以评价不同观察者的反应时间对得出的质量分级的影响。

这一数据库可以转换为质量等级q出现概率$P(q)$的直方图（见图2-7中的示例）。

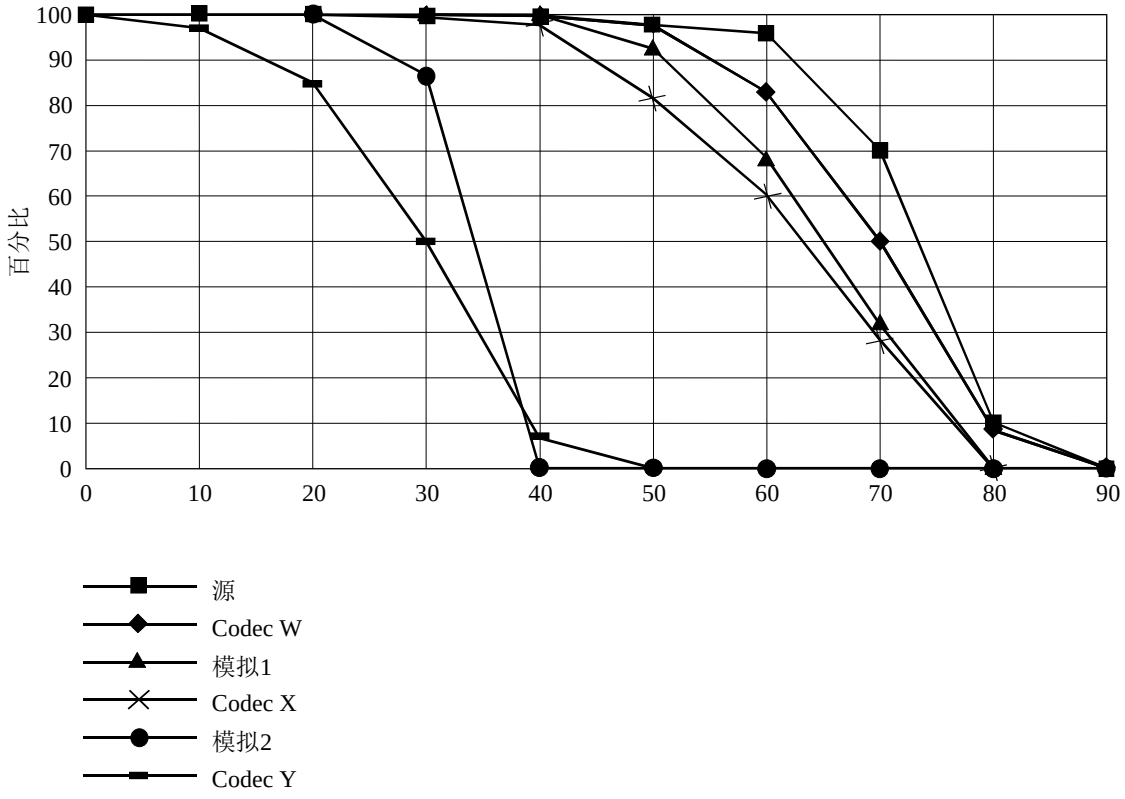
A5-8 连续质量评价结果的校准和单一质量评分的导出

尽管有人指出，较长时间的数字编码视频单一评分DSCQS测试阶段存在记忆上的偏差，但最近已经证实，这种影响对长度为10 s的视频片段的DSCQS评价影响不大。因此，在SSCQE过程中有可能出现第二阶段，以便根据从直方图数据中抽取的有代表性的10 s样本使用现有DSCQS法校准质量直方图。目前正对该第二阶段展开研究。

过去所用的常规ITU-R方法能够产生电视序列的单一质量评分。已经进行了一些实验，考察了已编码视频序列的连续评估与同样段落的总体单一质量评分之间的关系。已经确定，如果序列的最后大约10-15 s出现显著损伤，则人的记忆效应会扭曲质量评分。但也已经发现，人的这种记忆效应可用递减的指数加权函数来模拟。因此，在SSCQE法中有可能出现第三阶段，用于处理这些连续质量评价，以便获得一个等效的单一质量尺度。目前正对此进行研究。

图2-7

节目段Z的评分序列的平均评分



BT.0500-02-7

第2部分
附件6

同时双激励连续评估（SDSCE）法

ITU-R之所以提出连续评估，是由于原先的方法对数字压缩方案的视频质量测量存在某些不足。原先那些标准化方法的主要缺陷是由于在显示的数字图像中出现了与环境有关的扰动。在原先的协议中，待评视频序列的观看时长一般限制在10 s，观察者要对现实服务中出现的情况得出有代表性的判断，这段时间显然不够。数字伪像在很大程度上取决于源图像的空间和时间内容。这种情况在压缩方案中存在，但也与数字传输系统的容错性能有关。采用原先的标准化方法很难选出有代表性的视频序列，或者说至少很难评价其代表性。为此，ITU-R引入了SSCQE法，这种方法能够衡量较长序列的视频质量，衡量视频内容的代表性，以及衡量差错统计值。为了让再现的观看条件尽可能接近实际情况，在SSCQE中未采用基准。

在需要评价保真度时，必须引入基准条件。SDSCE是以SSCQE为基础制定的，但在向被试显示图像的方式上以及在评分量表上有稍许变化。提出这种方法是供活动图像专家组（MPEG）评价甚低比特率情况下的抗错性，但对于必须评价受到时变降质影响的视觉信息保真度的那些情况，这种方法也适用。

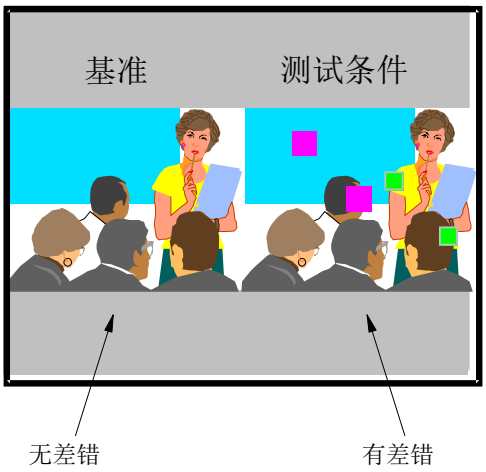
有鉴于此，制定了下述新的SDSCE技术并进行了测试。

A6-1 测试程序

被试小组同时观看两个序列：一个是基准序列，另一个是测试条件。如果这两个序列采用标准图像制式（SIF）或更短，则这两个序列可以并排在同一个显示器上显示，不然就应采用两个对齐的显示器（见图2-8）。

图2-8

显示制式示例



请被试检查两个序列之间的差别，并通过移动手持评分设备上的滑块来判断视频信息的保真度。如果保真度理想，则滑块应放在量表范围的顶部（代码为100）；如果保真度全无，则滑块应移动到量表的底部（代码为0）。

在整个观看期间，要让被试知道哪个序列是基准，并请他们在观看序列期间给出评分意见。

A6-2 不同的阶段

*训练阶段*是这种测试方法的一个关键部分，因为被试可能会误解其任务。应提供书面须知，确保所有被试获得完全一样的信息。须知中应解释被试将要观看的是什么，要评价的是什么（例如质量差别），以及如何表达其评分意见。被试提出的任何问题都应得到解答，以尽可能避免因测试管理员而产生的评分偏差。

在了解须知后，应运行一个*示范阶段*。这种方式可让被试熟悉评分程序和损伤种类。

最后运行一个*模拟测试*，显示若干有代表性的条件。这些序列与测试中所用的序列应有所不同，应一个接一个地显示，中间没有间隔。

在*模拟测试*结束之后，实验者应主要检查在测试条件等同于基准序列的情况下，评价结果是否接近一百（即看不出差别）；如果情况相反，被试声称看出了某些差别，则实验者应再次进行解释和模拟测试。

A6-3 测试协议的特性

下述定义适用于对测试协议的说明：

– 视频段（VS）：一个视频段对应着一个视频序列。

– 测试条件（TC）：一个测试条件要么是一个具体的视频过程，要么是一个传输条件，也可以是二者。每个VS应按照至少一个测试条件处理。另外，应在TC清单中加入基准序列，以便能够对基准/基准对进行评价。

– 阶段（S）：一个阶段由一系列不同的成对VS/TC组成，中间没有间隔，按随机顺序排列。每一阶段至少有一次含有全部VS和TC，但不必含有全部的VS/TC组合。

– 测试演示（TP）：一个测试演示由一系列涵盖所有VS/TC组合的阶段组成。必须由同样数目的观察者（但不一定是同样的观察者）对VS/TC的所有组合进行评分。

– 评分期：请每位观察者在一测试阶段内连续评分。

– 评分段（SOV）：用于评分的10 s的段。所有SOV采用互不重叠的成组的20次连续评分（相当于10 s）获得。

A6-4 数据处理

一旦测试完成，就会得到一个（或多个）数据文档，纳入了不同阶段（S）的所有评分，这些不同阶段代表了TP的打分总次数。通过验证每一VS/TC对都已得到处理且每一对都分配了相同次数的评分，就完成了数据有效性的第一次校验。

在按照这一协议完成的测试中收集到的数据可用3种不同的方式处理：

– 每一单独VS的统计分析；

– 每一单独TC的统计分析；

– 所有VS/TC对的总体统计分析。

每种情况都需要进行多步骤分析：

– 根据对观察结果的累计算出每次评分的均值和标准差。

– 计算出每一SOV的均值和标准差，如图2-9所示。这一步的结果可用一幅时间图表示，见图2-10。

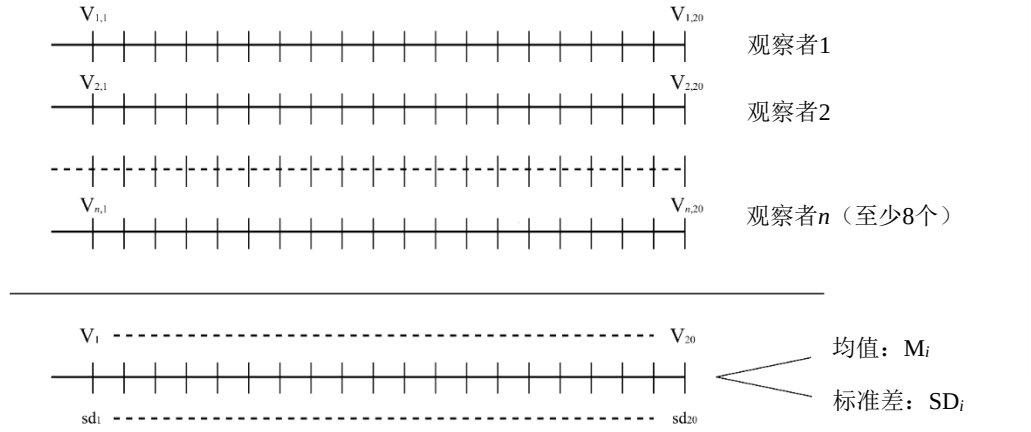
– 分析前一步算出的均值（即与每一SOV相对应）的统计分布及其出现频次。为了避免由前一个VS × TC组合产生的近因效应，每一VS × TC样本的头10个SOV要舍弃。

– 根据对出现频次的累计算出总体讨厌特性。这一计算要考虑置信区间，如图2-11所示。总体讨厌特性因示出了每一评分段的均值与其累积出现频次之间的关系而与这一累积统计分布函数形成对应。

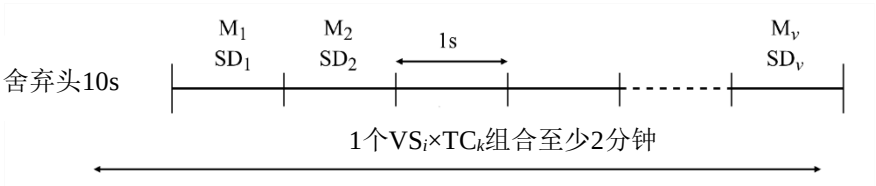
图2-9

数据处理

a) 计算平均分、V和标准差、每次观察者为每个VS X TC组合的每个评分序列评分的SD。



b) 计算每个VS X TC组合的每个1秒评分序列的M和SD。



BT.0500-02-9

A6-5 被试信度

通过检验被试在显示基准/基准对时的表现就可以定性评价被试信度。在这种情况下，预计被试将给出特别接近100的评价结果。可由此证明他们了解自己要承担的任务，不会随意打分。

另外，对于SSCQE法，可以采用与第1部分附件1第A1-2.3.2节所述程序接近的程序来检查被试信度。

在SDSCE程序中，评分的信度取决于下面两个参数：

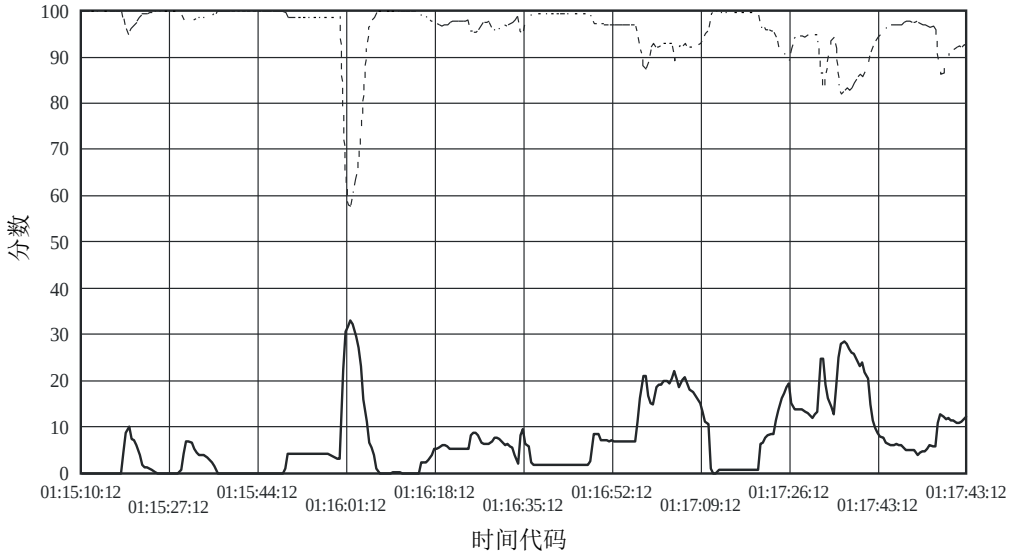
系统偏差：在测试期间，有的观察者可能过于乐观或过于悲观，甚或误解了评分程序（例如评分量表的含义）。这样就可能导致某一系列评分与平均系列之间或多或少存在系统偏差，甚至完全超出平均范围。

局部反演：在其他一些为人熟知的程序中，观察者有时可能没有特别留心观看和跟踪所显示的序列的质量。在这种情况下，总体评分曲线相对而言尚处在平均范围内，但仍可观察到局部反演。

这两种不合意的结果（反常行为和反演）是可以避免。参与者接受训练固然重要，但采用某种工具检测并在必要时舍弃前后不一致的观察结果也应该是可能的。本建议书对一种拟议中的可进行这种筛选的二步程序做了说明。

图2-10

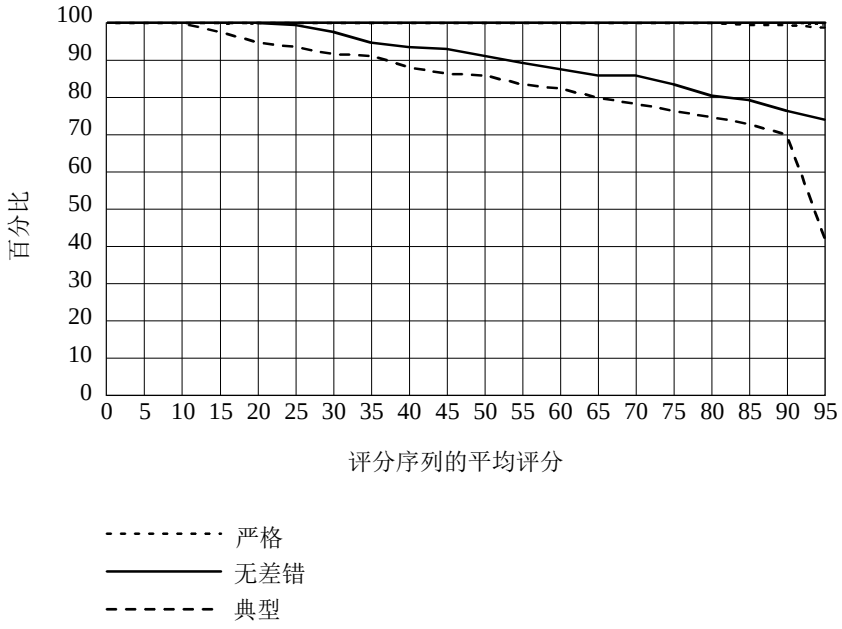
原始时间图



BT.0500-02-10

图2-11

在考虑置信区间的同时从统计分布计算整体讨厌特性



BT.0500-02-11

第2部分
附件7

多媒体视频质量（SAMVIQ）的主观评价

A7-1 引言

SAMVIQ质量评价方法使用连续质量尺度，以提供对视频片段内在质量的测量。各个观察者在从0到100评级的连续尺度上移动一个滑条，该连续尺度用5个线性排列的质量项目来注释（很好、好、一般、差、很差）。

在SAMVIQ方法中，观察者准许使用一个片段的若干个版本。当所有版本都经观察者评定后，可对之后的片段内容进行评估。

不同版本可由观察者通过计算机图形接口随机选择。根据需要，观察者可以停止、评审并修改某个片段各个版本的评分。该方法包括一个显性基准（即未经处理的）片段，以及相同片段的若干个版本，这些版本包括经处理的和未经处理的（即隐含基准）片段。片段的各个版本都单独显示，并使用一个类似于在DSCQS方法中使用的连续质量尺度来评价。因此，该方法在功能上与利用随机访问的单激励方法十分类似，但只要观察者想要观测，他就可以观测显性基准，这使得该方法类似于使用一个基准的方法。

SAMVIQ质量评估方法使用连续质量尺度，以提供对视频片段内在质量的测量。各个观察者在从0到100评级的连续尺度上移动一个滑条，该连续尺度用五个线性排列的质量项目来注释（很好、好、一般、差、很差）。

逐个场景地进行质量评估（见图2-12），包括显性基准、隐含基准和各种各样的算法。

为更好地理解这一方法，定义了以下特定词汇：

场景：视听内容

序列：综合处理过或未经处理的场景

算法：一种或多种图像处理方法。

A7-2 显性、隐含的基准与算法

评估方法通常包括质量锚，以稳定结果。在SAMVIQ方法中，出于以下原因，考虑了两个高质量锚。已经完成的一些测试表明，可以使用显性基准来最大限度地缩小分值的标准差，而不使用隐含的基准或不使用基准。尤其是对多媒体数字信号编解码器性能的评估，最好使用显性基准来获得最可靠的结果。为了评估基准的内在质量，也可加上隐含基准，而不是显性参考，原因是陈述是匿名的，并且是经过处理的片段。显性名称“基准”会对大约30%的观察者产生影响。这些观察者对显性基准可能给出最高分（100分），而该分值总的说来有别于隐含基准对应的分值。值得注意的是，当没有可用的基准时，测试仍有可能进行，但标准的偏差会显著增大。

SAMVIQ方法适用于多媒体内容，原因是它可能结合图像处理的不同特点，例如多媒体数字信号编解码器类型、图像制式、比特率、时间更新、图像缩放等。算法这个名称总结了这些特点的其中一个特点或其组合。

A7-3 测试条件

在场景期间，临界点的变化是受到限制的，原因是在提供一个综合分值的其它方法（如单激励方法）隐含使用的相同规则后选择同样的内容。最大的片段观测期为10秒或15秒，对获得稳定的和可靠的质量分值而言，这已足够。应使用专用的解码器－播放器或其产品的屏幕拷贝，以保持适当的显示性能。

A7-4 测试机构

a) 如图2-12所示，逐个场景地进行测试。

b) 对当前场景，可能以任何次序来播放任何片段，并为其打分。每个片段都可以多次播放和打分。

c) 从一个场景到另一个场景，对片段的访问是随机的，防止观察者试图根据已排好的次序、以完全相同的方式来做出判定。实际上，在一个测试中，算法的次序仍保持相同，以便简化对结果的分析和陈述。只有来自相同按钮的相应访问是随机的。

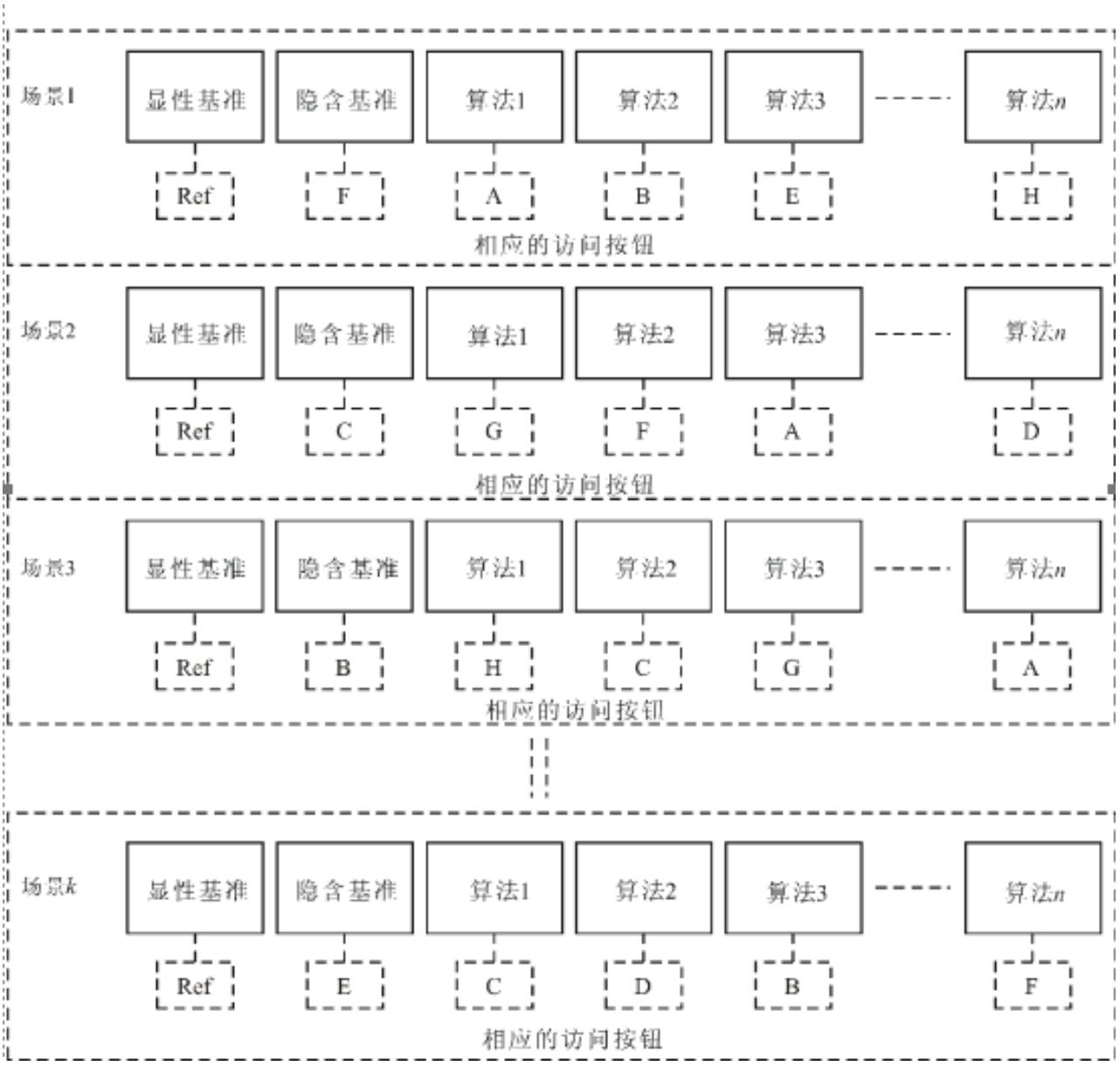
d) 对第一次观测，当前的片段必须在打分之前全部播放过；否则，可能立即打分和停止。

e) 为测试下一个场景，必须为当前场景的所有片段打分。

f) 为完成测试，必须为所有场景的所有片段打分。

图2-12

SAMVIQ方法的测试机构举例



BT.0500-02-12

SAMVIQ方法通过软件来实现。除了图2-12中所示的访问按钮，“播放”“停止”“下一个场景”和“上一个场景”按钮都是必需的，以便允许观察者管理不同场景的表述（例如，参见第A7-6节）。当观察者已给出一个分值，那么应在该场景对应的访问按钮下方显示出来。当一个片段的所有不同版本都已经过评级时，仍允许观察者对分值进行比较，并且如有必要，可以对分值进行修改。不必评估当前的整个片段，原因是，在第一遍观测中，已经突出了大的差别。

A7-5 数据表述与分析

A7-5.1 摘要信息

为了复制测试或比较不同测试的结果，需要提供有关测试环境的精确数据。因此，如表2-3所示，建议报告有关测试环境的信息。

表2-3

测试摘要信息

方法名称	
显示技术	
显示器的参考名称	
最大亮度等级（cd/m²）	
黑色亮度等级（cd/m²）	
黑色等级设置：PLUGE（前面所述可察觉的黑色等级距离门限＝8）。 否则表示门限值	
背景亮度等级（cd/m²）	
亮度（lux）	
观测距离： 不受限制的：在显示器之前 受限制的：nH	
显示器尺寸（对角线，以英寸表示）	
宽/高显示比	
显示制式（行与列的数目）	
图像输入制式（行与列的数目）	
图像输出制式(1)（行与列的数目）	
白色色温：D65，否则 白色彩色坐标（x，y）	
有效观察者数目	

(1) 当处理输入图像时，例如在显示器上重新调节输入图像时，需要该信息。

显示特性可能影响测试结果。其他信息，例如亮度响应（百万分之一的保真度）和基色，应对平板显示器提出要求。

视频片段的特性对设计测试或解释其结果而言是重要的。如第1部分附件1所述，建议报告空间－时间特性。该信息应在视频素材库中收集适于多媒体应用中视频质量主观评估的测试片段时加以考虑。

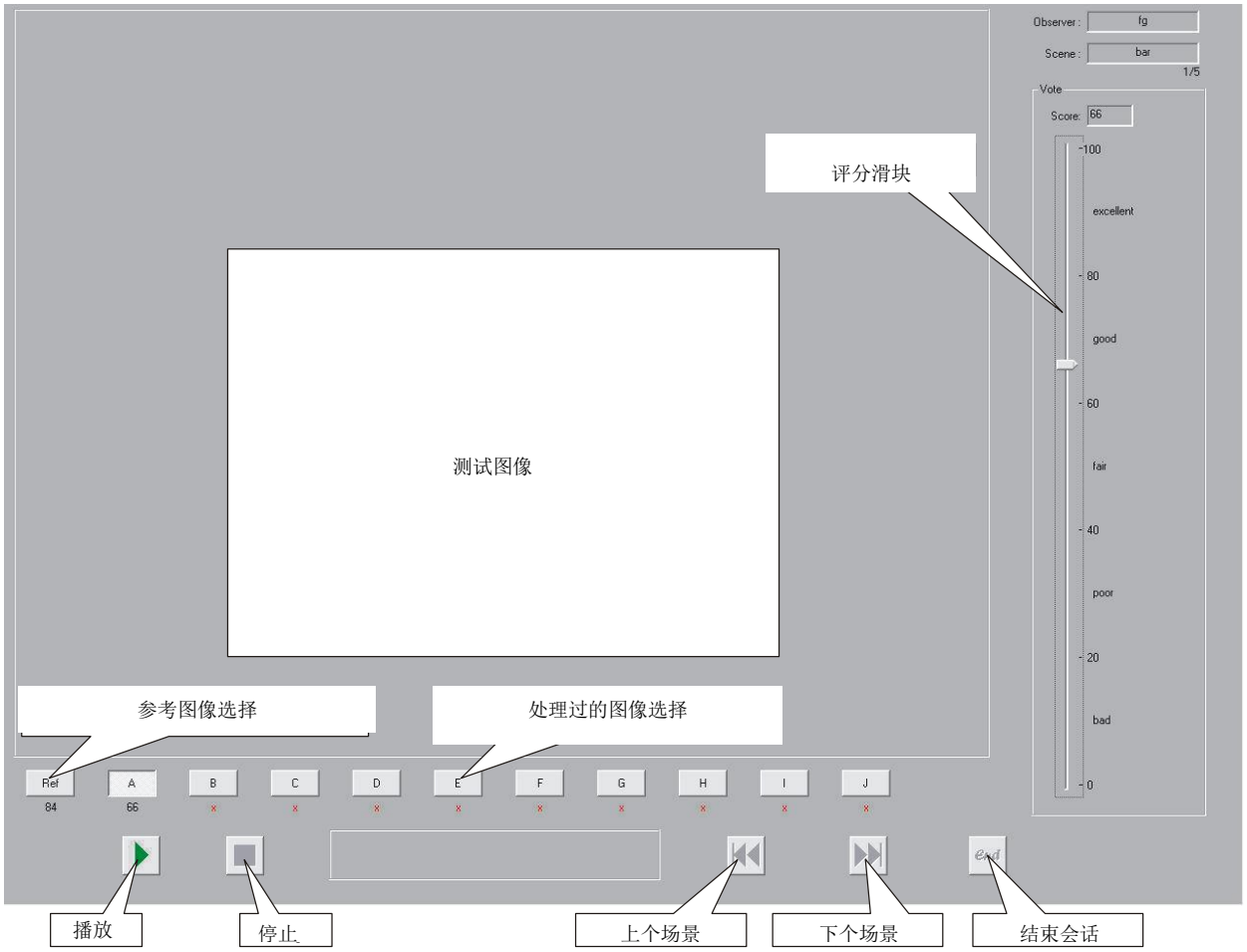
A7-5.2 分析方法

分析方法是第1部分附件1中所述的那些方法。

A7-5.3 观察者筛选

A1-2节描述对SAMVIQ的筛选。

A7-6 SAMVIQ界面示例（资料性）



BT.0500-02-12a

第2部分
附件8

用于评估视频素材质量的专家观看协议（EVP）

本附件阐述利用专家观看协议主观评价移动图像视频质量的方法，其中包括更少数量观看者的参与。这些观察者都是从相关视频处理领域的专家中挑选出的。

A8-1 实验室设置

A8-1.1 显示器的选择与设置

应使用具有专业应用（例如，广播播音室或广播车）典型特性的平板显示器；显示器对角线的尺寸范围在22’（最低）至40’（建议值）之间，但在评估HDTV或更高清晰度的图像系统时可将尺寸放大至50’或更高。

允许使用处于工作状态的显示器观看区域的一个缩小的部分；在此情况下，处于工作状态的显示器部分的周边应设置为中灰色。在此使用条件下，不允许将显示器分辨率设置为其自身分辨率之外的其它分辨率。

显示器应能够合理设置并使用专业的轻型测量仪表测量亮度和颜色。测试中显示器的测量应遵守相关建议书规定的参数。

A8-1.2 观看的距离

专家观看座位距离的选择应依据屏幕的分辨率和屏幕工作部分的高度而定，其依据为第1部分第2.1.3.2节设计的观看距离，或按关键观看条件设定的更短观看距离。

A8-1.3 观看的条件

专家观看协议（EVP）不一定要在测试实验室进行，但重要的是测试位置不会受到可听和/或可视干扰（例如，亦可使用安静的办公室或会议室）。

应当消除直接或通过反射落在屏幕上的光线；其它周边光线应较暗，保持在可填写得分表（如果使用）的最低亮度。

在显示器前就座的专家数量可能因屏幕尺寸大小而变化，以确保所有观看者看到同样的图像和激励呈现。

A8-2 观看者

参加EVP试验的观看者应是相关研究领域的专家。

鉴于观看者是在有资格的人士中选拔，因此没有必要进行视觉灵敏度或色盲筛查。

不同观看者的最低数量为九位。

为达到最低数量的观看者，同一试验既可在同一地点重复进行，也可在一个以上的地点实施。专家观看会不同地点的得分可合并在一起进行统计处理。

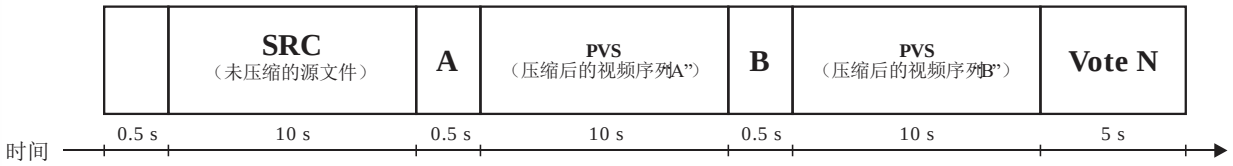
A8-3 基本测试单元

针对每组需要评估的编码条件，向专家呈现的素材应组织在一起，形成一个基本测试单元（BTC），见图2-13。

BTC中需要考虑的源参考序列（SRC）和经处理的视频序列（PVS）剪辑应总与相同的视频序列相关，这样专家们才有可能确定受测压缩算法提供的视频质量是否有改进。

图2-13

专家观看协议中基本测试单元的时间轴



BT.0500-02-13

BTC应按如下方式组织：

- 屏幕设置为中灰0.5秒（亮度表的中间值）；
- 未压缩基准视频剪辑呈现10秒；
- 在中灰背景下显示消息“A”（第一个要评估的视频）0.5秒；
- 呈现受损版本视频剪辑10秒；
- 在中灰背景下显示消息“B”（第二个要评估的视频）0.5秒；
- 呈现受损版本视频剪辑10秒；
- 显示请观看者表述其意见的消息5秒。

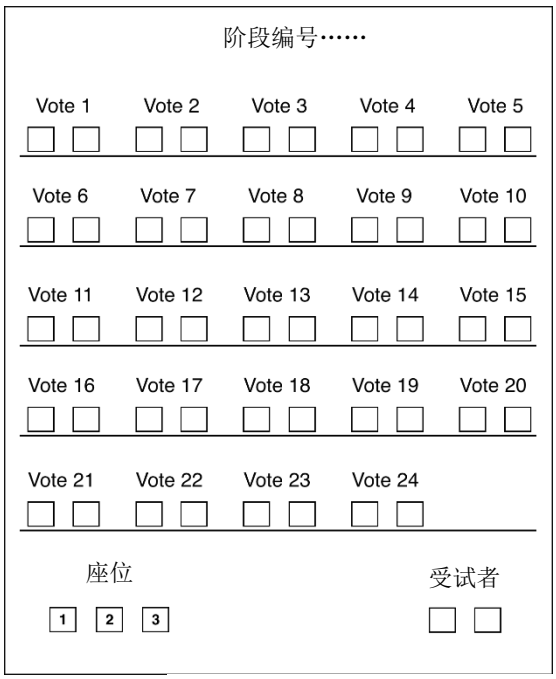
“Vote”消息后应紧跟一个数字，为得分表同步提供帮助。

A8-4 得分表和评级

如图2-13所示，视频剪辑呈现的安排应将未受损基准（SRC）置于最前端，接下来再安排两个受损视频序列（PVS）。对各BTC而言，PVS的呈现顺序应随机变化且观看者不应了解呈现的顺序。

图2-14

24-BTC专家观看测试的得分表示例



BT.0500-02-14

评分使用11级数字标尺，得分从10（难以察觉的损伤）至0（十分令人厌恶的损伤）。表2-4提供了有关11级数字标尺含义的指导。

表2-4

11级数字标尺的含义

得分	受损程度	
10	难以察觉	
9	可轻微察觉	某些位置
8		所有位置
7	可察觉	某些位置
6		所有位置
5	明显察觉	某些位置
4		所有位置
3	令人厌恶	某些位置
2		所有位置
1	特别令人厌恶	某些位置
0		所有位置

针对每个BTC，请观看者填写一份由两个框（标为“A”“B”）组成的问卷调查表，在两个框内填写11级数字标尺中选出的一个得分。

图2-14提供了一个由24个BTC构成的测试得分表示例。

对各个BTC而言，观看者将填写标有字母**A**（对第一个视频剪辑进行评级）和标有字母**B**（对第二个视频剪辑进行评级）的框。

呈现原始无损的视频剪辑使专家们能够轻松评估任何损坏。

11级数字标尺的含义应在下文“培训部分”中详细解释。

A8-5 测试设计和测试的设立

测试设计者应将BTC呈现的顺序设置为随机，使相同的视频剪辑和相同的受损剪辑不会连续出现两次。

任何观看测试均应以“稳定阶段”开始，其中应当包含“最好”“最恶劣”和两个“中等质量”的BTC。这将使观看者在测试之始便对质量范围有一个直接印象。

如果观看测试超过20分钟，则测试设计者应将其分为两个（或多个）独立的观看测试分场，每场的时间不超过20分钟。在此情况下，应在各观看测试前提供“稳定阶段”。

A8-6 培训

即便预计此程序将在专家的参与下使用，仍最好在每次试验前组织一场简短（5-6 BTC）的培训。

培训课题中使用的视频素材可与实际工作中的相同，但呈现的顺序应有所差别。

应当培训观看者如何使用11级标尺，请他们仔细观看屏幕上出现“A”和“B”后立即呈现的视频剪辑，并查看他们是否能看出其与第一个呈现的视频剪辑（SRC）的差别。

A8-7 数据收集和处理

每次测试结束后应收集得分并登入电子表格计算平均值。

最好对观看者进行“后筛选”（使用线性皮尔森相关）。

“相关性”函数的应用应考虑到各受试者与平均意见得分（MOS）间的关系；可就观看者是“可以接受”还是“应被拒绝”定义一个门限值（ITU-T P.913建议书提议将0.75作为“拒绝”的门限值）。

A8-8 使用专家观看协议成果的条件

当时间和资源不允许运行一个正式的主观评估试验时，可使用专家观看协议（EVP）。

专家观看协议比正式主观评估需要的时间更少，且可在没有外部音视频干扰的“非正式”环境中进行。

唯一强制条件涉及上文各段所述周围环境的亮度和观看条件（显示器、观察角度和观看的距离）。

A8-9 使用EVP结果的限制

尽管EVP展示出仅通过九位观看者就能提供可接受结果，但EVP试验提供的MOS不能被视作正式主观评估试验成果的替代品。

使用EVP获取的MOS数据可用于获取受损水平的初步指标。

使用EVP获取的MOS数据可用于对评估中的视频处理机制做初步排名。

在方便或必要时，EVP试验可在多地并行实施，假设观看条件、观看距离和测试的设计完全相同。

如果在不同地点运行的同一EVP试验中观看专家的数量大于等于15，则可通过加工原始主观数据获得MOS、标准差和置信间隔数据，这有助于更精确地对各种测试案例做出排名。在最后一个案例中，可实施更精准的推导统计分析，例如T-Student测试。

第2部分
附件8
的后附资料1
（资料性）

在有大量专家评估员出现的情况下
专家观看协议的应用及其表现

该参考性附录提供了两种针对编码HD和UHD视频剪辑的不同的主观评估专家观看协议测试，该测试在第117届MPEG会议上进行，采纳了附件8的条款以快速和可靠地对两种不同的源编码方法进行评级。

由于很多专家都出席了第117届MPEG会议，参与两种专家观看协议测试的评估员数量远远超过了ITU-R BT.2095建议书中建议的9位；有30位专家参加了HD EVP测试，32位专家参加了UHD EVP测试。

专家评估员的广泛参与使MOS数据分析成为了可能，以便在对编码的视频剪辑进行评级时验证使用附件8条款的固有信度等级。

在评估中，四组观看者（即：9、12、15和18）被选中对比通过9位专家得出的MOS值和通过12、15和18位专家得出的MOS值。

目标是对比9位专家得出的评级（以与专家观看协议一致）与12、15和18位专家得出的评级（类似于正式主观评估测试）。

图2-15（针对UHD的测试）和图2-16（针对HD的测试）展示了四种情况的评级结果非常相近。

如果我们通过将18位专家得出的结果视作一种“基础真值”，我们可以根据通过18位专家得出的MOS值（连续的红线）绘制出图2-15和图2-16的图表，对测试点进行评级。

图表中的其他线展示了通过9位专家（点状红线）、12位专家（蓝色虚线）和15位专家（连续的绿线）得出的结果。

观察图2-15和2-16绘制的结果，值得注意的是：

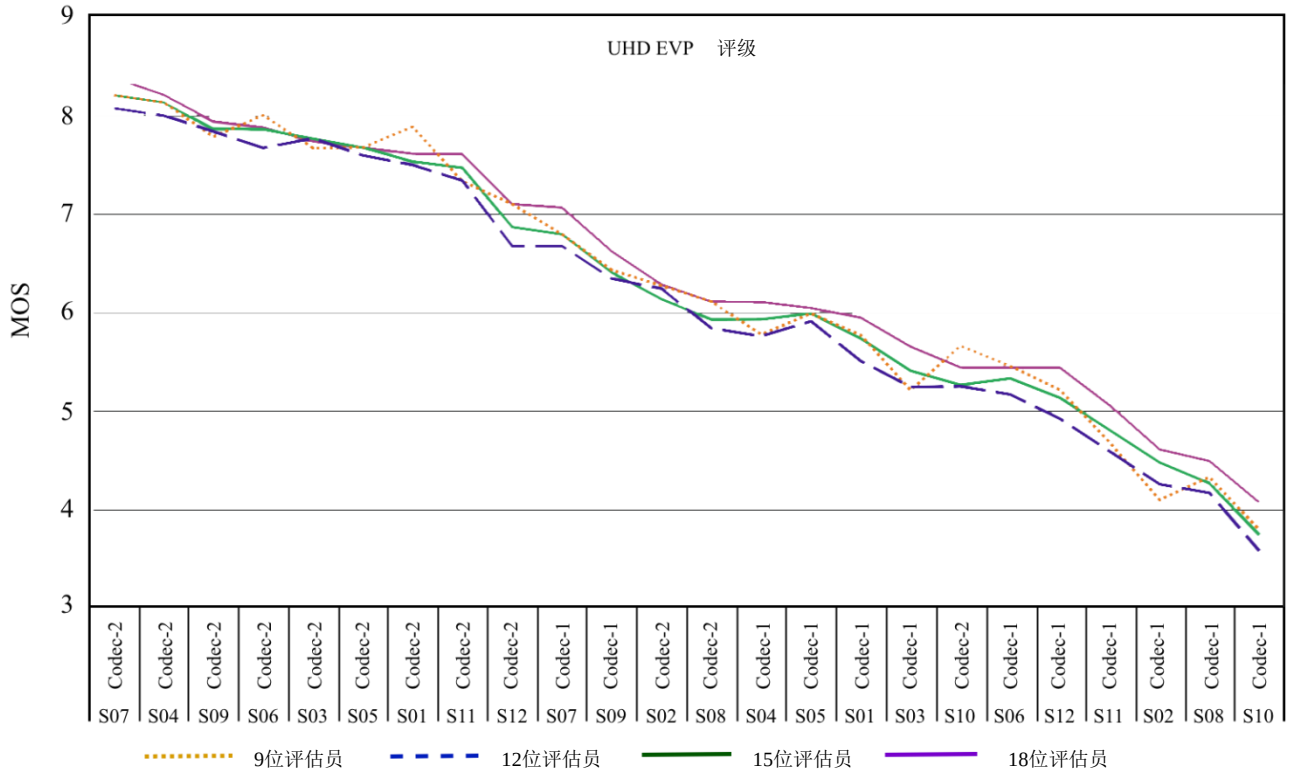
– 15位和18位专家的线从呈现了从高品质到低品质MOS值的均匀斜率；

– 9位和12位专家的线对18位的对比呈现出了一些“倒置”，即使评分的变化在其扩展中相当有限。

总而言之，此处展示的EVP协议实验描述了EVP协议的良好表现，验证了附件8中提到的内容，即：在更多的观看者可以参加测试且正式主观评估完成后，即使EVP协议不能作为正式主观实验的完全替代，其仍是一种稳定的评估程序，且结果非常接近所获得的结果。

图2-15

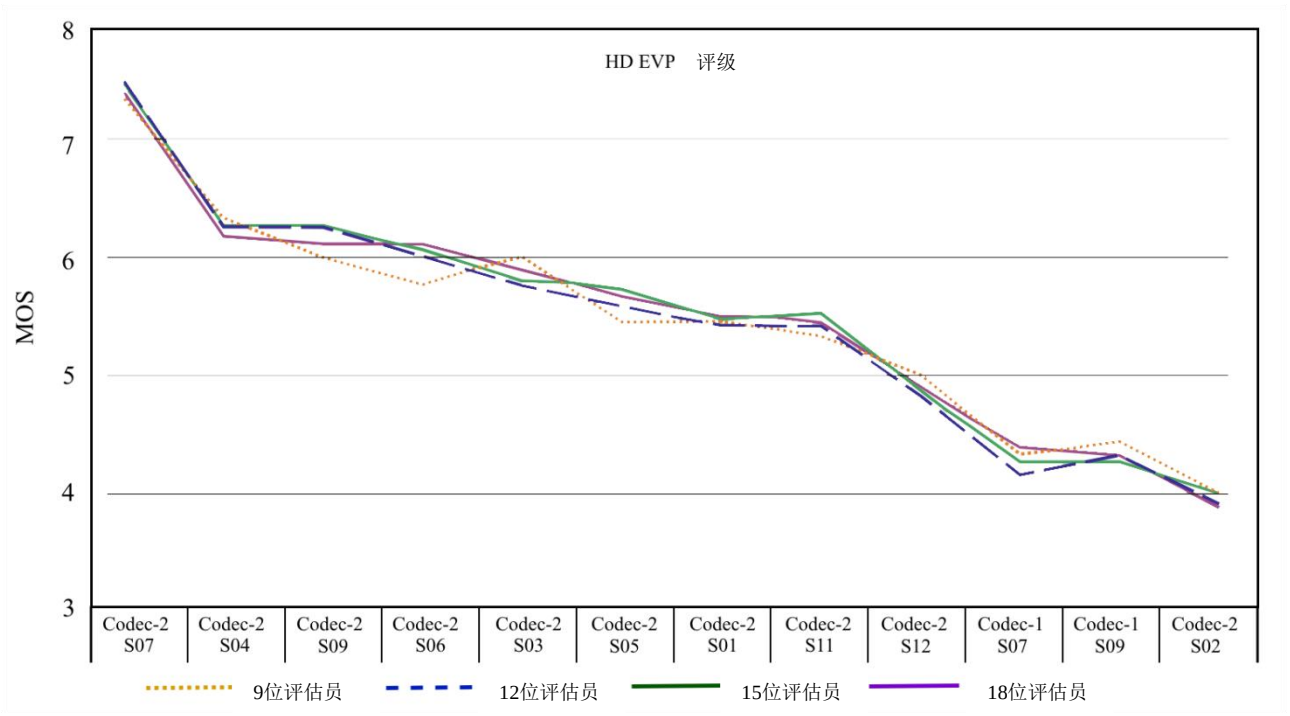
以评估员数量为函数的UHD实验评级



BT.0500-02-15

图2-16

以评估员数量为函数的HD实验评级



BT.0500-02-16

第3部分

应用特定的图像质量主观评价方法

应对主观评价测试设计进行应用特定考虑。第3部分为各图像制式和应用中的图像质量主观评价提供了导则：

附件1 标准清晰度数字电视（SDTV）系统的主观评价

附件2 高清电视的图像质量主观评价

附件3 图文电视和类似文本业务中的字母数字和图形图像的质量主观评价

附件4 多节目业务图像质量主观评价

附件5 剧场大屏幕数字图像的数字显示系统图像质量的专家观看评价

附件6 多媒体应用中的视频质量的主观评价

附件7 立体三维电视（3DTV）系统的主观评价

第3部分
附件1

标准清晰度数字电视（SDTV）系统的主观评价

A1-1 引言

旨在与本建议书第1和第2部分共同使用的本附件提供了有关本建议书为提供达到（或接近）传统电视系统质量水平的数字系统进行主观评价给出的通用方法的应用的详细信息。此处给出的程序性详情，与相关的背景信息一道，适合于用于传输根据ITU-R BT.601建议书关于馈给和分布式应用生成的材料以及在发射应用中使用的材料的编解码器（或系统）的测试。

对于分布式应用，可以用观察者主观判断来表达质量规范。理论上，可以根据这些规范来对此类编解码器进行主观评价。然而，为馈送应用设计的编解码器的质量理论上无法根据主观性能参数制定，因为其输出并非用于立即观看，而是用于为进一步传输进行演播室后处理、存储和/或者编码。因为为各类后处理操作定义这一性能存在难度，更受青睐的做法是为一连串设备规定性能，包括后处理功能在内（该功能被认为是实际馈送应用的代表）。这一设备串通常可由一个编解码器、其后的一个演播室后处理函数（或在基本馈送质量评价中可为其他编解码器），和再之后的信号到达观察者前的另一个编解码器构成。采用这一用于馈送编解码器的规范的策略意味着，本建议书中给出的测量程序亦可以被用于对其进行评价。

在存在大量经验的主观评价领域，可推荐测试条件和方法。但是，必须记住，在规定质量或损伤目标时，现有方法无法给出绝对的主观评分，而是在某种程度上受参考和/或锚定条件的选择影响的结果。同一方法可用于固定和可变字长编解码器，以及域内和帧间编解码器，虽然测试图像序列的选择可能会受影响。

最完全可靠的评价高质量编解码器等级顺序的方法是在同一时间，在完全相同的条件下，对所有候选系统进行评价。单独进行的、涉及质量细微差异的测试应被用作指导，而非无可争辩的优越性的证据。

被确定为发生在编码器和解码器之间的传输链路上的误码率的函数的损伤或许是一个有用的主观度量。目前，关于真实传输误差统计的实验知识不足，无法为解释错误聚类或丛发的模型推荐参数。在这一信息可用之前，可使用泊松分布（Poisson-distributed）错误。

A1-2 观看条件

第1部分第2节给出了主观评价的通用观看条件。以下段落给出了数字系统主观评价的具体观看条件。

A1-2.1 实验室环境

实验室环境旨在提供临界条件，以对系统进行检查。以下表3-1提供了在实验室环境中进行主观评价的具体观看环境。

表3-1

在实验室环境中进行数字系统主观评价的具体观看环境

条件	项目	值
a	观看距离与图像高度之比	4 H和 6 H (1)
b	峰值亮度	70 cd/m^2
c	符合规格的背景部分所对的观看角度	≥43° H × 57° W
d	显示器	高质量屏幕 尺寸≥ 20" (50 cm) (2)

(1) 数字标清系统的评价的设计观看距离（DVD）为6 H，但评价者距离在4 H也是可以接受的，条件是结果单独给出。

(2) 因为有证据表明显示器尺寸可能影响主观评价，实验者被要求明确地报告在任何实验中使用的显示器的屏幕尺寸、品牌和型号。

A1-2.2 家庭环境

此环境旨在为数字电视链的消费者一侧提供质量评价的手段。表3-2给出了在家庭环境中进行数字标清电视（SDTV）主观评价的具体观看条件。

表3-2

在家庭环境中进行数字系统主观评价的具体观看环境

条件	项目	值
a	观看距离与图像高度之比	6 *H*
b	4/3长宽比的屏幕尺寸	25"至29"(1)
c	16/9长宽比的屏幕尺寸	32"至36"(1)
d	显示器标准	SDTV
e	峰值亮度	200 cd/m2
f	屏幕上的环境照度 (落在屏幕上的来自环境的入射光应在屏幕上垂直测量)	200 Lux

(1) 该屏幕尺寸满足 PVD = 6 *H*的较佳观看距离(PVD)的规定。

A1-3 评价方法

A1-3.1 基本图像质量的评估

当对编解码器进行分布式应用评价时，该质量指的是在单一通过一个编解码器对之后解码的图像。对于馈送编解码器，可以在多个编解码器串联之后对基本质量进行评价，从而模拟典型的馈送应用。

当待评价的质量范围不大的时候（一般是电视编解码器），要使用的测试方法为本建议书描述的双激励连续质量量表法的变型II。原始源序列被用作基准条件。进一步考虑演示序列的持续时间。在最近对4:2:2分量视频的编解码器进行的测试中，我们认为对本建议书提供的演示进行调整是有利的。合成图像可被用作附加参考，以据此判断编解码器在较低质量等级的性能。

建议在评价中使用至少六个图像序列，加上一个在试验开始时候用于培训目的的附加序列。在考虑比特率减缩应用环境下，序列范围应介于适度临界到临界之间。

在整个附件中，采用在电视比特率减缩的环境下具备临界性的图像序列测试数字编解码器的重要性得到强调。因此，对于特定的比特率缩减任务，询问特定的图像序列的临界性如何，或者一个序列是否比另一个序列的临界点高都是合理的。一个简单但并非特别有帮助的回答是，对于不同编解码器来说，“临界性”并不相同。举例来说，对于域内编解码器来说，包含许多细节的静止图像很可能就达到临界，而对于能够利用帧到帧相似处的帧间编解码器来说，同样情况下一点难度都没有。一些采用移动纹理和复杂运动的序列对于所有类别的编解码器来说都达到临界，因此这些序列类型对于生成或识别来说最为有用。复杂运动可采取对于观察者来说可预测的，但对于编程算法来说并非如此的运动形式，例如曲折周期运动。

对可能的图像临界性的数据测量的检查，例如通过关联法、波谱法、条件熵法等进行的检查，揭示了一个简单但有用的基于域内/帧间自适应熵测量的度量。该方法被用于“校准”被建议用于34、45和140 Mbit/s的编解码器ITU-R试验的图像序列，并已被证明在使用的序列选择方面有用。在图像序列上进行此类测量，最容易实现的方法是将它们传输到图像处理计算机中，并通过软件对其进行分析。

在对这些技术的访问不可用的情况下，以下提供了对于如何选择临界素材的一般导则。

a) 固定字长域内编解码器

在静止图像上评价这些编解码器是可能并且有效的，但建议使用移动序列，因为编码噪声过程更容易观察，对于电视应用来说更加实际。如果在编解码器的计算机模拟中使用了静止图像，应在整个评价序列上进行处理，从而，比如说保存任何源噪音的时间方面。选择的场景应尽可能多地包含以下细节：静止和移动纹理区域（一些带有彩色纹理）；静止和移动对象，具有鲜明的高对比度边缘（一些为彩色）；静止普通中灰区域。整体之中至少一个序列应只展示可察觉噪音源，至少一个序列应是合成的（即由计算机生成的），这样不存在摄影机缺陷，例如扫描孔径和迟延。

b) 固定字长帧间编解码器

选择的测试场景均应包含动作，以及尽可能多地以下细节：移动纹理区域（一些为彩色）；具有鲜明的高对比度边缘、沿垂直于这些边缘的方向和不同方向移动的对象（一些为彩色）。整体之中至少一个序列应只展示可察觉噪音源，至少一个序列应是合成的。

c) 可变字长字域内编解码器

建议使用移动图像序列素材对这些编解码器进行测试，原因同固定字长编解码器。应注意，由于其可变的字长编码和相关的缓冲存储器，这些编解码器可在整个图像中动态地分配编码位容量。因此，举例来说，如果半个图像由不需要许多码位来编码的无特色的天空构成，容量被保留给图像另一部分，这样，即便处于临界点，也可以高质量地复制。由此得出的重要结论是，如果图像序列对于此类编解码器来说要达到临界，屏幕的每个部分的内容都应该详细呈现。屏幕应被移动和静态纹理、尽可能多的色彩差异，以及具有鲜明的高对比度边缘的物体填满。整体之中至少一个序列应只展示可察觉噪音源，至少一个序列应是合成的。

d) 可变字长帧间编解码器

这是最复杂的一类编解码器，这类编解码器需要要求最高的素材来对其施加压力。不仅每个场景都应被细节填满（同域内可变字长编解码器的情况一样），而且细节也应展示动作。此外，由于许多编解码器采用运动补偿方法，整个系列中的动作应是复杂动作。复杂动作的例子有：采用同时缩放和平移摄影机的场景；有纹理的或充满细节的背景幕布在风中飘动的场景；包含在三维世界中旋转的对象的场景；包含在屏幕上加速的充满细节的对象的场景。所有场景都应包含大量不同速度、质地和高对比度边缘以及不同颜色的内容的对象的大量活动。整体之中至少一个序列应只展示可察觉噪音源，至少一个序列应具有来自天然静止图像的复杂的计算机生成摄像机运动（从而使其没有噪音和摄像机延迟），以及至少一个序列应完全由计算机生成。

A1-3.2 顺流处理之后的图像质量评估

本评价旨在允许对用于馈送应用的编解码器在特定后处理（例如背景调色、慢动作、电子变焦）方面的适应性做出判断。对此类评价的设备的最少布置是单次通过受测编解码器之后进行相关后处理，然后由观看者观看。但是，可能更具代表性的做法是，馈送应用在后处理之后使用更多编解码器。

要使用的测试方法是双激励连续质量量表方法的变型II。但是，在此，基准条件为经过与解码图像相同的后处理的源程序。如果纳入低质量基准被认为是有利的，那么该基准也要经过相同的后处理。

后处理评价所需的测试序列应遵守与用于其他数字应用的序列完全相同的临界标准。但是，色度键前景序列中实现这可能有困难，因为它们通常有大面积的无特征蓝色背景。

因为可能必须用多个后处理来对编解码器进行评测的实际限制，使用的测试图像序列的数量可以最少为三个，加上用于示范目的的附加序列。序列性质取决于被研究的后处理任务，但在电视比特率减缩的背景下，对于研究的的程序，应介于适度临界和临界之间。对慢动作评价，显示率为信源率的十分之一可能是合适的。

A1-3.3 对降质特性的评估

在对由于传输或发射信道缺陷造成的编解码器图像损伤进行的主观评估中，应选择至少五个（但最好更多）误码率或选择的传输/发射条件，采用近似对数性间隔，对导致编解码器损伤的范围（从“不可察觉”到“非常讨厌”）进行充分采样。

可能要求以导致罕见的、被认为可能不会在一个10 s的测试序列周期内发生的可见瞬态的传输误码率进行编解码器评价。这里建议的演示时间显然不适合这样的测试。

如果编解码器输出在非常低的误码率条件下进行录制（导致10 s周期内产生少量可见瞬态），用于稍后编辑成主观评价演示，须谨慎确保使用的录像是在较长时间跨度内观看的编解码器输出是典型的。

由于在传输误码率范围内探索编解码器性能的需要，实际限制表明，三个测试图像序列加上一个附加示范序列很可能足矣。序列的持续时间应为10 s左右，但应注意，测试观看者可能更偏好15-30 s的持续长度。在电视比特率缩减的背景下，范围应介于适度临界到临界之间。

由于测试覆盖整个损伤范围，双激励损伤量表的方法是合适的，应予以采用。

A1- 3.4 图像内容降质特性

第1部分附件1给出了图像降质特性的一般概念。为将这一概念应用至标清数字电视系统，应采用以下流程。

A1- 3.4.1 临界性的定义

应对被称为“临界性”的某个度量进行定义。该度量表示受测数字电视系统的特性，用客观测量来度量。作为数字电视系统的一个示例，MPEG-2 MP@ML被使用，应用ITU-R BT.1210建议书描述的基于临界性的熵的固定量化器方法。

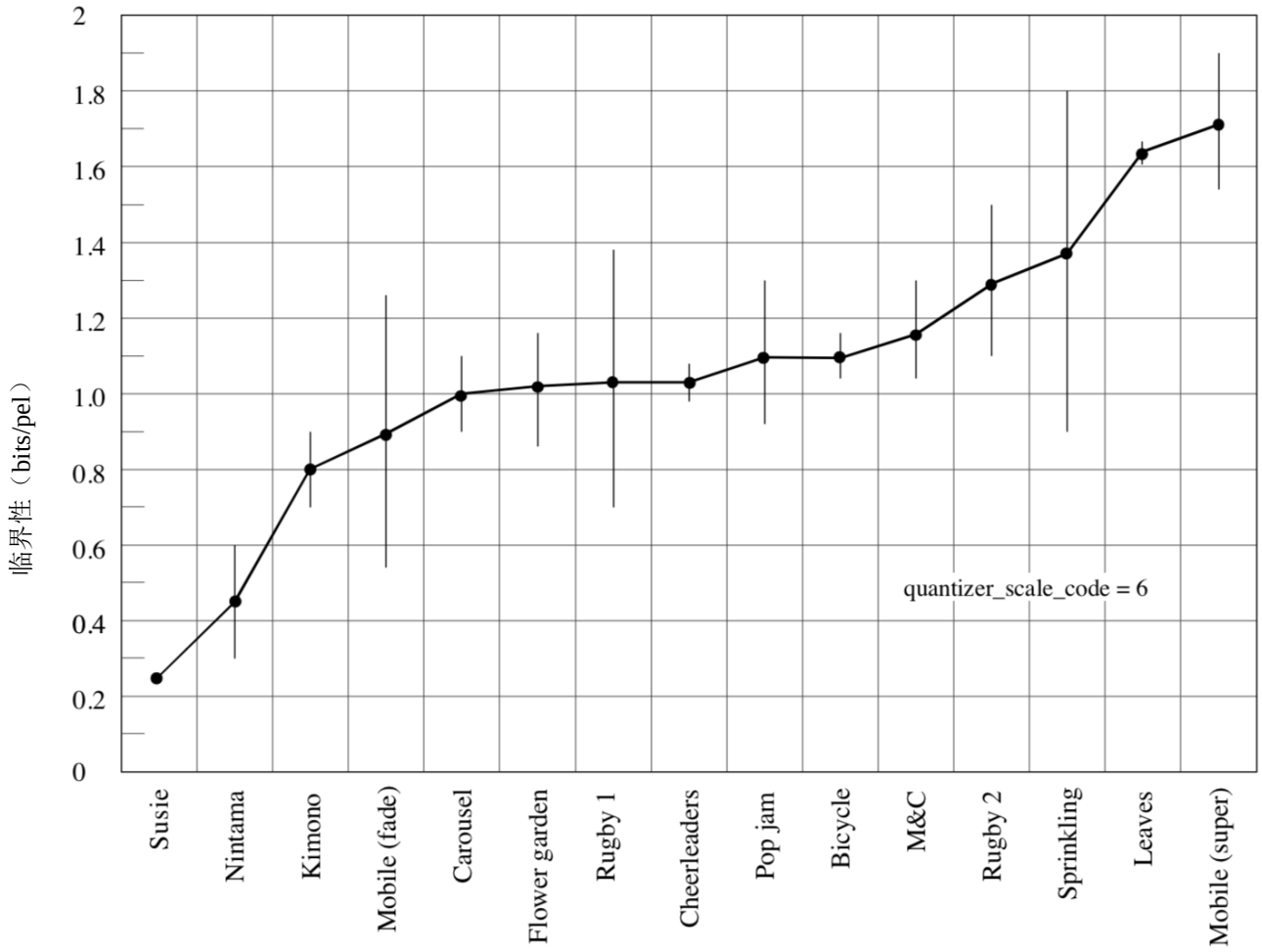
A1- 3.4.2 图像内容降质特性的推导过程

– 步骤1： 测定主观评价中使用的测试序列的临界性

测定以下步骤3中描述的主观评价中使用的测试序列的临界性。图3-1显示了示例系统中每个序列的均差和标准差。大部分序列的临界值介于0.8-1.4比特/像素（bits/pixel）。一些序列的标准差很大，因为序列中的图像内容差异很大。

图3-1

测试序列临界性的均差和标准差



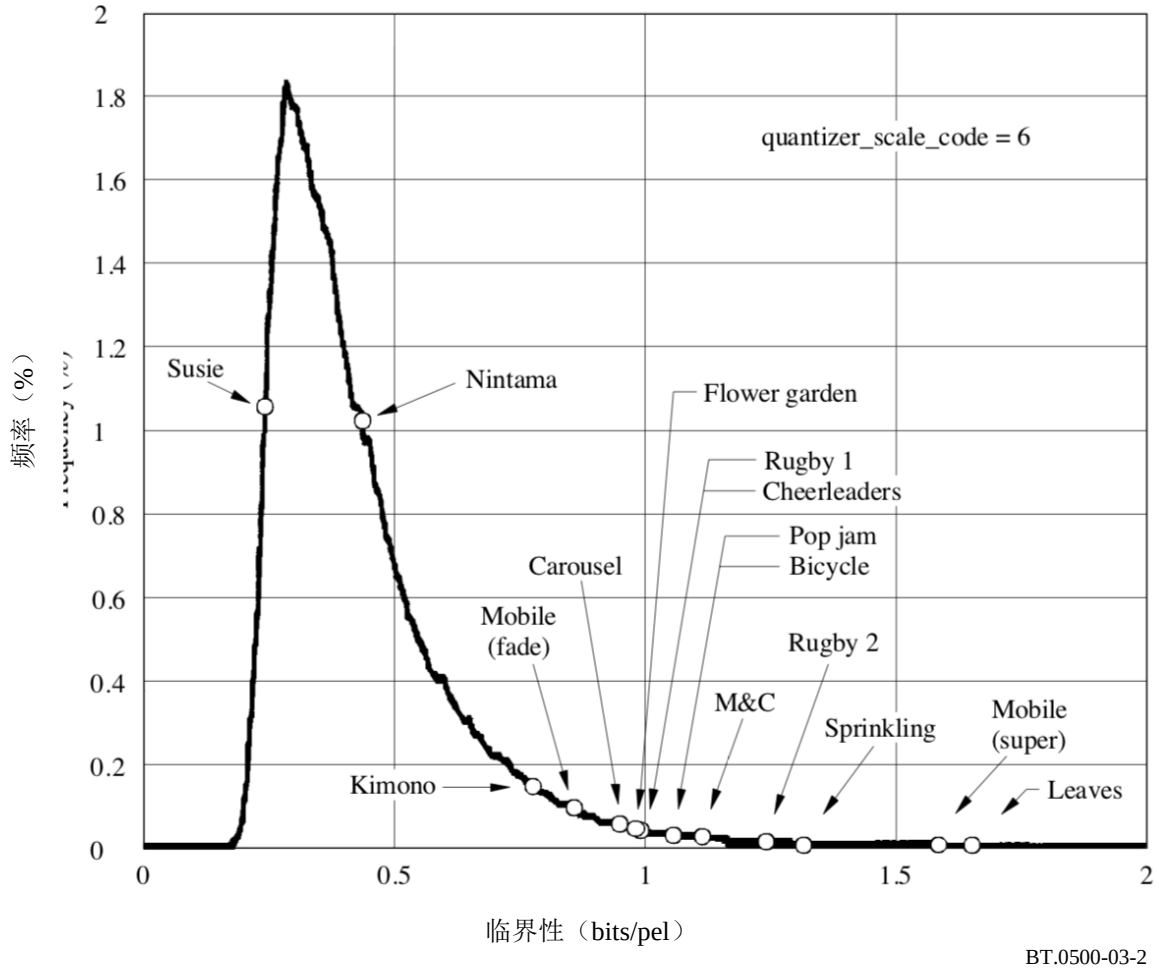
BT.0500-03-1

– 步骤2： 测量广播节目的长时间临界性分布

在足够长的时间周期内（例如一周）测量广播电视节目的临界性分布。图3-2显示了一周测量的分布的示例，总共130 h的NTSC广播信号，这些被转换为用于测量的分量*Y*/*C*信号。电视节目的临界性的发生频率按照每5×10^{-3} bits/pixel计算。该图表还显示了用于主观评价的测试序列的临界性。

图3-2

广播节目的临界分布和测试序列的临界分布



BT.0500-03-2

– 步骤3: 对受测系统的图像质量进行主观评价，得出临界性与主观图像质量之间的关系

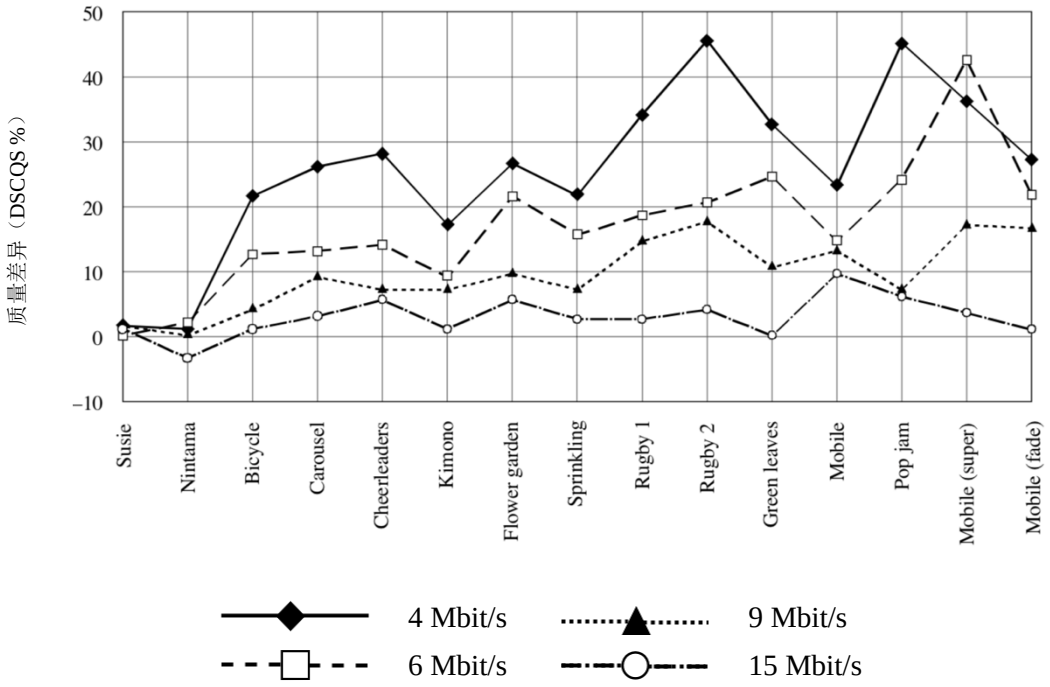
采用DSCQS方法评价数字电视系统图像质量。结合步骤1中获得的主观评价结果和临界性结合，可得出临界性和评价测试得分之间的关系。图3-3显示了示例系统在4、6、9和15 Mbit/s比特率的图像质量。图中的质量差异（DSCQS %）代表从基准降质的程度，初始为4:2:2分量序列。图3-4显示了临界性和质量差异之间的关系。本示例中假设了临界性和图像质量之间的线性关系，使用最小二乘法得出回归线。图中显示了每个比特率上的回归线。一般来说，可应用非线性关系，取决于评价结果。

– 步骤4: 通过结合步骤3（临界性与质量之间的关系）和步骤2（临界性与发生频率之间的关系）的结果来得出图像内容的降质特性（质量与发生频率之间的关系）。

通过结合步骤2和步骤3得到的结果，得出图像内容降质特性，即数字编码的电视节目图像质量的分布。广播电视节目中的图像降质被转换成累积发生频率。图3-5显示了示例系统的图像内容降质特性。

图3-3

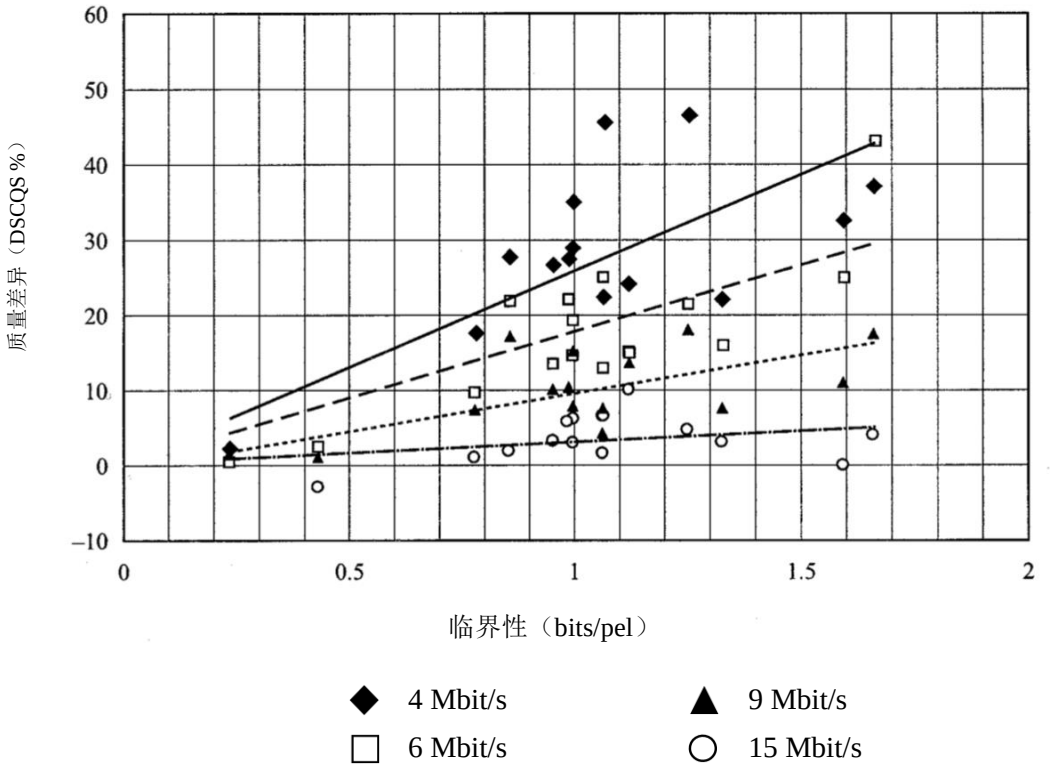
主观评价结果（在6H的MP@ML）



BT.0500-03-3

图3-4

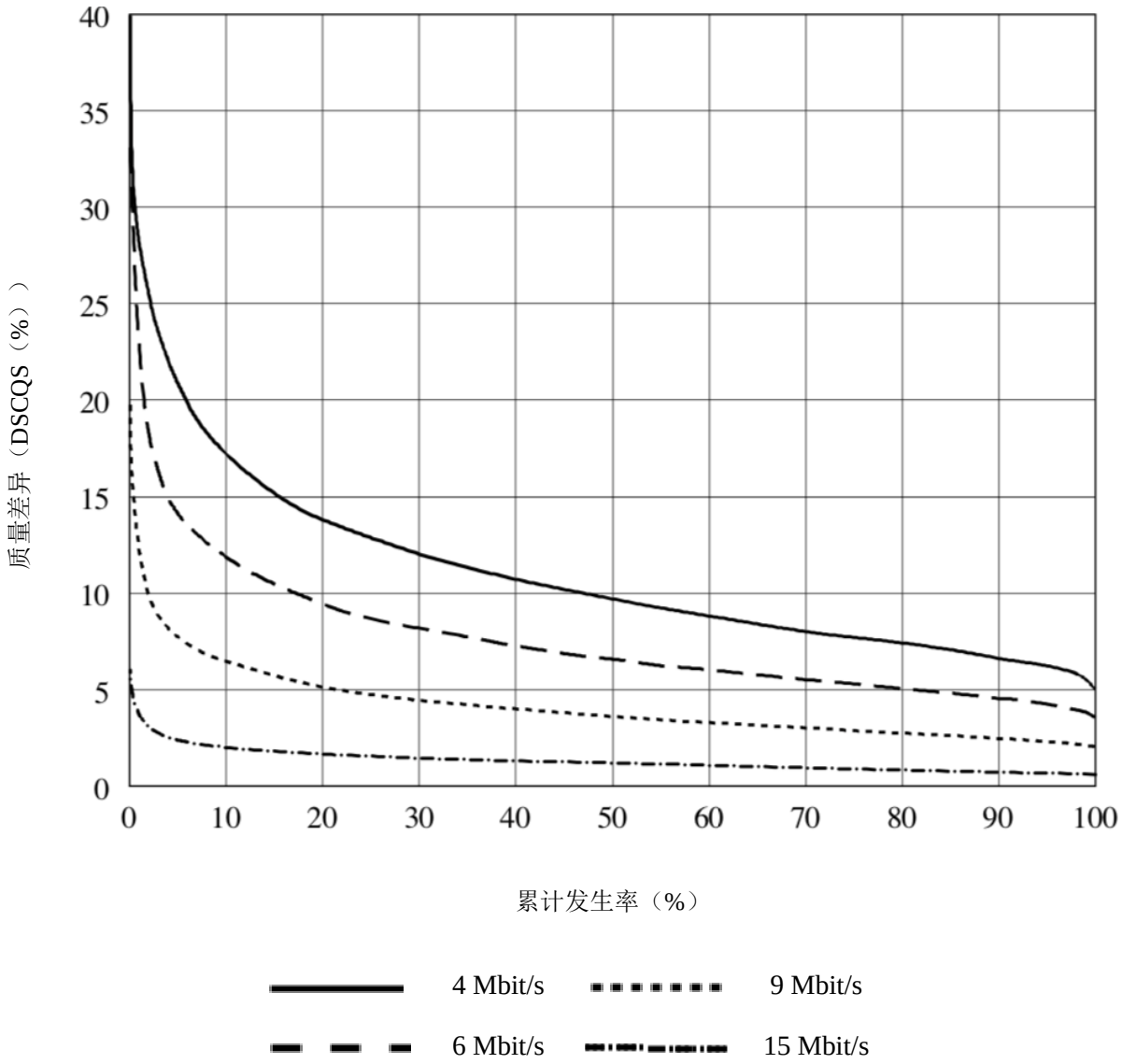
临界性和评价分数之间的关系（在6H的MP@ML）



BT.0500-03-4

图3-5

图像降质累积发生频率（在6H的MP@ML）



BT.0500-03-5

A1-4 应用说明

在不要求对编解码器绝对质量或损伤进行判断，仅需要评分排名的情况下，或者需要对从双激励结果中获得的评分排名进行确认的情况下，应使用成对激励比较的方法。

正如本建议书中所描述，该方法提供了灵敏性比较和确定系统对之间的关系程度的方法。拓展此方法，为超过两个系统的质量或损坏进行评分是可能的。在这一方法中，从观察者对所有可能的图像序列对的评分中得出总体评分排名。

该分析的复杂性在于，比如，观察者可以对图像A的评分高于图像B，图像B高于图像C，但图像C的评分又高于图像A。这被称为“三元转移关系”。

这一方法的一个问题是，要求的演示数量随着测试图像序列和编解码器数量的平方而增加，可能变得不切实际。

如果使用广播信道来提供多个节目流或可扩展或分层编码方案，也许有必要修改评估方法，将以下情况纳入考虑：

– 可接受业务的标准在源代码方面可能并不透明；相反，可以是系统能力，在一个给定比特率划分中，提供传统业务的可行替代。相应地，作为质量测试的基准，使用典型接收条件下的传统系统提供的素材而非未压缩的数字形式的素材可能是合适的。进一步来说，使用遴选出的代表当前和未来节目内容的范围的测试素材（见第1部分附件3）也许是合适的。在测试中，观看条件应为第1部分和本附件第A1-2节给出的观看条件，而通用测试方法应为双激励连续质量量表法（第2部分附件2）；以及

– 系统能否在全信道负载和传输受损的情况下保持单个节目流的完整性是一个问题。相应地，在降质测试中，确保全信道负载和使用遴选出的代表可能的接收条件范围（见第1部分附件4）的降质电平是合适的。在测试中，观看条件应为第1部分和本附件第A1-2节给出的观看条件，而通用测试方法应为双激励损伤量表法（见第2部分附件1）。

注1 – 当模拟和数字系统在同一个情境中测试时，重要的是选择一套可以反映模拟和数字系统的平衡难度的测试素材。在这一情况下，为辅助分析，采用多维标度法可能有用。

第3部分
附件2

高清电视系统(HDTV)的图像质量主观评价

A2-1 观看环境

除非在以下表3-3中列出，否则观看环境应按照第1部分第2节所述。

表3-3

进行HDTV图像质量主观评价的观看环境

条件	项目	值
a	观看距离与图像高度之比	3
b	屏幕的峰值亮度 (cd/m^2)(1)	150-250
c	不活跃屏幕与峰值亮度之比(2)	≤ 0.02
d	在全黑房间中仅显示黑色电平时的屏幕亮度与对应的峰值之比(3)	约为 0.01
e	图像显示器背景亮度与图像的峰值亮度之比	约为 0.15
f	来自其他来源的照度(4)	低
g	背景色度	D65

表3-3（完）

进行HDTV图像质量主观评价的观看环境

条件	项目	值
h	符合上述规格的背景部分所构成的角度(5)。.为所有观察者保留这一角度。	53°高度x 83°宽度
i	观察者安排	距离显示器中心水平方向 ± 30°以内。垂直限制正在研究中
j	显示屏尺寸(6)	1.4 m (55 in)

(1) 与100%幅值的视频信号对应的屏幕峰值亮度。

(2) 该项目可受室内照明以及显示器对比度范围的影响。

(3) 黑色电平对应0%幅值的视频信号。

(4) 室内照明应调节为可满足条件c和e。

(5) 建议最小为28°高度 x 48°宽度

(6) 如果规定尺寸显示器不可用，则应使用值≥ 76.2 cm (30 in)的显示器。见第1部分注3。

A2-2 评价方法

发射系统传送的HDTV图像的总体质量主观评价应采用双激励连续质量量表法（第2部分附件2）进行，以HDTV演播室质量图像为基准。

HDTV发射系统降质特性评价应采用双激励缺损量表法（第2部分附件1），以HDTV演播室图像或未缺损发射图像作为基准。

当实践中可能遇到的性能超出节目内容范围和传输条件成问题时，应考虑第1部分附件4的复合降质特性。

使用这些方法时，应谨慎区分显示器制式和基本系统制式（例如任何上转换）的影响。如果觉得可用且适当的话，可使用不同显示器进行辅助评价，从而将不同显示器制式考虑在内。

一些HDTV发射系统可能包含嵌入的传统电视制式（向后兼容）。因此，有必要从图像质量的角度评估嵌入HDTV发射中的传统电视图像的适当性。对于这些系统，应执行第3部分附件1给出的观看条件和评价方法。

采用比特率缩减方案的数字HDTV发射系统应执行第3部分附件1描述的基本概念和流程。

A2-3 测试素材

ITU-R BT.2245报告列出了大量静止图像和移动序列。这些应被优先用作HDTV质量测试的通用测试素材。

第3部分
附件3

图文电视和类似文本业务中的字母数字和图形图像的图像质量主观评价

引言

有些系统处理的是图形和字母数字图像，并通过适当的数字编码传输这些图形和字母数字图像。字母数字和图形图像具有一个与传统电视图像不同的特性，在对他们进行主观评价时的思维过程亦可能有所差别。

本建议书建议了评估当前电视节目中包含的图像的主观质量的方法。需要对用于通过电视信道传输，并使用数字代码来描述字母数字图像和图形图像的若干新业务的字母数字图像和图形图像的质量进行研究。一些传输参数会对显示的图像的质量造成影响：在图文电视的alpha-mosaic编码情况下的页面解析（每个页面的行数和每行的字符数）；DRCS（动态可重定义字元集，见ITU-R BT.653建议书）情况下的字元分辨率（像素数和每单元行数）；在广播听力描记法、传真或图文电视情况下的编码和图像分辨率。进一步来说，可能影响编码的传输误差的影响也应被考虑在内。因此，对这些参数的质量进行测量和确定客观−主观之间的关系是必要的。

研究显示，对这些图像进行质量评价需要不同的方面。这些图像可能包含与传统电视图像不同的特性。诸如像素制式、字元分辨率、间隔、色彩和布局等参数对各类质量属性造成影响：易读性、质量、舒适性、令人讨厌度、阅读费力度、疲劳度和审美考虑。在此考虑三个主要方面：观看条件、评价方法和评价环境。

考虑到建立字母数字和图形图像的质量主观评价基础的重要性，应在所有测试报告中提供对测试配置、测试素材、观察者和方法的尽可能最完整的描述。

A3-1 观看条件

第1部分定义了对应室内低亮度水平的电视图像的观看条件。字母数字和图形图像很可能也在正常照明条件下观看。因此，为研究建议了观看条件的补充集合：500 lux照明、屏幕最大亮度从70至200 cd/m^2，屏幕对比度从30至50，背景亮度（从房间墙壁）与最大屏幕亮度之比的值为1/4。屏幕高度的四到八倍的观看距离也应被考虑在内。

A3-2 评价方法

在印刷排版方面已经做了大量研究。大部分研究使用“性能测量”，例如检测或识别阈值、识别率、阅读速度等。传统上被用于评价电视图像质量的“性能测量”很少使用。人们认为通过电视信道传输的新系统应具备良好性能（例如，超过95%的字母良好辨识率百分比）。因此，本建议书给出的质量和损伤量表可被有效使用，虽然还需要进行研究，以建立这些量表与易读性联系起来的方式。已尝试与语音质量评价方式（ITU-T）进行比较，并建议了“阅读费力度”五级量表，用于进一步研究。

另一个方法将使用表3-4给出的两个不同五级量表进行主观评价结果的比较。

表3-4

易读性和阅读费力度量表

易读性质量量表	阅读费力度量表
易读性优	阅读不费力
易读性良	需要注意力，但不太费力
易读性中	阅读比较费力
易读性差	阅读费力
易读性劣	阅读很费力

研究发现，重要的是要使每个等级量表措辞十分明确。阅读费力度量表分数的平均值通常高于易读性量表的平均值，观察者在阅读费力度量表中给出的分数范围更广。

另一个使用第2部分第A3-4.1节描述的质量量表的实验是评价由具备不同线路标准和带宽的电视系统传输的文字稿的总体质量和总体易读性。对于每个意见都提供两个模型，一个复杂性和准确性更高，但两个模型都引发了“损伤量表”添加的概念，描述了有限的水平和垂直定义的综合效果。亦从正确识别出的字符的比例来衡量易读性。但是，在这方面，当质量低的时候易读性仍然很高；很明显，通常前一个标准有用性较低。

另一个研究利用定宽字符和可变宽度字符对印刷文本素材的性能和主观方法进行了比较。显示主观方法灵敏度更高。采用阴极射线管显示器重复了同样类型的研究，这次仅使用主观方法。这些主观方法的使用产生了处理固定和可变矩阵的视觉最优大小的结果。

A3-3 评价环境

业务评价的新方法考虑正在研究的业务的用户活动可被准确定义的情况。评价并不是根据传统的演示图像并简单地要求观看者做出标准主观评价的方式来进行，而是观看者像使用正在研究的业务那样使用演示的图像，所有的评价都是在这一环境中做出的。

业务使用模拟并不妨碍对于传统主观测量的使用。但是，它打造了一个更适合正在研究的业务的主观评估环境。它还使观察者性能的客观测量以及尤其适合正在研究的业务和参数的新的主观测量的开发成为可能。最后，它为将实验室进行的评估推广到在服务条件下进行的评估打造了更加安全的基础。

第3部分
附件4

多节目业务图像质量主观评价[5]

引言

对在一个多节目业务中使用恒定比特率（CBR）来压缩和编码的各节目质量的主观评价，应使用在第3部分附件1或2详述的主观程序以及本附件第A4-2节详述的流程。

对在一个多节目业务中使用可变比特率（VBR），通过使用统计多路复用或联合编码的方式来压缩和编码的各节目的质量的主观评价，应使用第3部分附件1或2详述的主观程序以及本附件第A4-3节描述的流程。

A4-1 总体评价详情

– 对基于主题的信道的质量进行评价应使用与通常在这些信道上传输的内容和临界性相似的测试素材。

– 为了对在一段时间内“瞬时”质量不同的节目编排的总体感知质量进行评价，应使用第A4-2和第A4-3节描述的流程。

– 根据包含在DSCQS法的描述中的评论，对与低质量素材进行比较的多节目业务，应对涉及低质量基准的系统的结果进行缩放和进一步研究。

A4-2 恒定比特率多节目业务的主观图像评价流程

每个SDTV和HDTV节目的主观图像质量评价可以使用第3部分附件1（SDTV）或附件2（HDTV）中描述的方法独立进行。对系统基本质量的评价，应使用在第2部分附件2中描述的通用测试方法DSCQS。对存在传输损伤的节目的评价，应使用第2部分附件1中描述的通用测试方法DSIS。

A4-3 可变比特率多节目业务的主观图像评价流程

对于VBR编码的SDTV和HDTV节目的主观图像质量评价，可使用DSCQS方法执行。还必须注意对测试素材的选择，因为图像质量可取决于所有多路复用的节目的图像质量。

---

5 包括“统计多路复用”或“统计复用（Stat-Mux）”业务。

第3部分
附件5

剧场大屏幕数字图像[6]的数字显示系统图像质量的专家观看

A5-1 引言

在过去几年中，专家观看经常被用来对通用视频流程的性能进行快速检验。

本附件描述了使用有限数量专家评价者进行的、可确保不同实验室获得的结果的一致性的一个专家观看测试方法。

A5-2 为什么新方法以“专家观看”为基础

指出应用所建议的方法所带来的好处是有益的。

首先，正式的主观评价测试通常要求使用至少15名作为“非专家”选出的观察者。需要对新观察者进行长时间的测试和连续的研究。这一观察者数量是达到必要灵敏度所需要的，以便可以自信地对被测系统进行区分和评分，或被自信地判定为等效。

其次，传统测试使用非专家观察者，有可能无法揭示在长时间曝光下可能变得明显（即便对非专家来说也是如此）的差异。

再次，传统评价通常确立质量（或质量差异）度量，但并不直接识别导致这些措施产生的伪像或物理表现。

在此建议的方法尝试解决所有这三个问题。

A5-3 专家受试者定义

在本附件中，“专家观看者”是知道用于进行评价的素材、知道“要看什么”和可能有或没有被深入告知用于处理待测视频素材的算法详情的人。在任何情况下，“专家观看者”具备质量调查方面的长期经验，在测试所涉及的特定领域从事专业工作。例如，在组织对LSDI素材的“专家观看”测试阶段时，应选择电影制作或后期制作，或者高质量视频内容制作方面的专家（例如摄影导演、色彩修正师等）；这一选择必须考虑到对LSDI图像质量和压缩伪像做出独特主观判断的能力。

A5-4 评估员的选择

专家观看测试是基于评价者意见的评价阶段，在该阶段观看者就视觉质量和/或损伤可见性提供判断。

基本专家组由五到六名受试者构成。这一小数字使得汇聚评价者更加容易，并能更快做出决定。

根据实验需要，使用超过一组基本专家组构成更大的专家集合（例如来自不同实验室的专家）是可以接受的。

---

6 大屏幕数字图像（LSDI）是适用于戏剧、比赛、体育活动、音乐会、文化活动等节目的一个数字图像系统家族，从采集到以高清质量在配备适当设备的电影院、礼堂和其他场所的大屏幕呈现。

在测试自己开发的技术时，专家可能倾向于在打分中存在偏见，这是公认的。因此，应避免纳入直接参与受测系统开发的人员。

应对所有评价者进行视力检查，以确定其视力是否正常或矫正至正常（斯内伦试验，Snellen test），以及具备正常色觉（Ishihara Test）。

A5-5 测试素材

选择出的测试素材应是预见受测系统在真实环境中使用时的制作水准范围和困难程度的样本。选择应优先考虑具有挑战性但又不过分极端的素材。理想情况下，应使用5-7个测试序列。

选择素材的方法也可能因针对受测系统所设计的应用而不同。

在这方面，此处未对测试素材的选择规则做出规定，留待测试设计者根据上述考虑做出决定。

A5-6 观看条件

须在测试报告中完整地描述的观看条件须遵守表3-5的规定，并在测试中保持恒定。

表3-5

观看条件概述

观看条件	设置	
	最小值	最大值
屏幕尺寸（米）	6	16
观看距离(1)	1.5 H	2 H
投影仪亮度（中心屏幕，白色峰值）	34cd/m^2	48cd/m^2
屏幕亮度（关掉投影仪）		<1/1 000投影仪亮度

(1) 当观看距离更接近1.5 H时应使用“蝴蝶型”演示。如果使用并排演示，观看距离应更接近2 H的值。

A5-7 方法

A5-7.1 评估阶段

每个评估阶段（定义为一个观察者组的测试轮的集合）由两个阶段（即，阶段I和阶段II）构成。

A5-7.1.1 阶段I

阶段I是在受控环境中进行的正式主观测试（见第A5-6节）。该测试可产生有效、灵敏和可重复的测试结果。在测试中，专家使用下文描述的评分量表，独自对显示的素材评分。不允许小组成员之间讨论他们正在观看的素材，或者对演示进行控制。在此阶段，专家们不应知晓受测的编码方案，或者受测素材的演示顺序。受测素材应随机化，以避免在测试中产生任何偏见。

A5-7.1.1.1 素材的演示

演示方法结合了同时双激励连续评估（SDSCE）法（第2部分附件6）和双激励连续质量量表（DSCQS）法（第2部分附件2）的元素。也可被称为同时双激励（SDS）法，供参考。

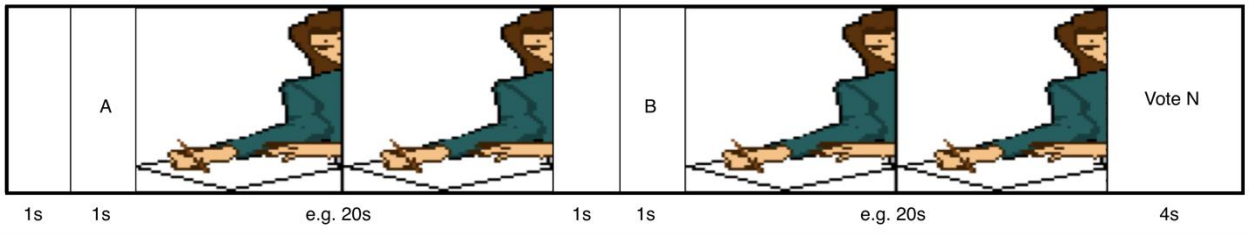
采用SDSCE方法，每个试验包含来自两个图像素材的分屏演示。在大多数情况下，其中一个图像来源将是基准（即源图像），而另一个是测试图像；在其他情况下，两个图像均取自基准图像。基准应为透明演示（即，不经过除对源记录介质来说是隐式的压缩之外的压缩）的源素材。测试素材应为经过受测系统之一加工处理的源素材。比特率和/或质量水平应遵守测试设计的规定。与SDSCE方法不同，观察者将不知晓图像对中的两个图像所代表的条件。

应使用无镜像的传统分屏或通过蝴蝶技术（屏幕右侧的图像被水平翻转）来进行分屏演示。由于将使用全幅图像，一次只能展示每个图像的一半。在每个演示中，显示器的每一侧都将显示图像同一侧。

使用DSCQS方法，图像对被连续展示两次，一次用于熟悉和细看，一次用于确认和评分。每个序列持续时间为15-30 s。可以在每个片段的开头对每个序列标记，以对评价者提供协助（见图3-6所示的非镜像分屏示例）。

图3-6

非镜像分屏示例

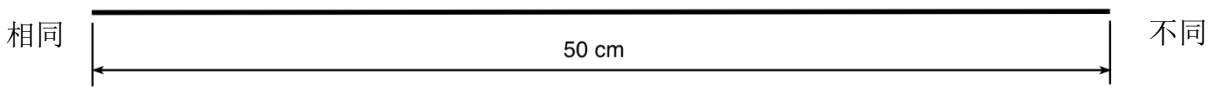


BT.0500-03-6

A5-7.1.1.2 判定量表

LSDI应用可接受性的标准为，测试（即，压缩的）图像无法与基准图像区别开。可使用多种常用的评分方法来对受测系统进行评价。所提出的方法是建议的激励比较量表（第2部分附件4）。具体的示例量表为第2部分附件4第A4-4.2节描述的非分类（连续）SAME-DIFFERENT量表：

图3-7



BT.0500-03-7

A5-7.1.1.3 判定阶段

可能包含超过一轮（取决测试条件数量）的本阶段应包含两个试验类型：测试试验和核对试验。在测试试验中，显示器半边显示基准，而另外半边显示测试。在核对试验中，两个半边都显示基准。核对试验的目的是对判断偏差进行评估。

对于每个受测系统，需要对每个测试序列进行以下测试试验：

表3-6

左显示面板	右显示面板
左半基准	左半测试
右半基准	右半测试
左半测试	左半基准
右半测试	右半基准

上述案例每个最好重复至少两次。对每个系统，需要对每个测试序列进行以下核对试验：

表3-7

左面板	右面板
左半基准	左半基准
右半基准	右半基准

上述案例还是每个最好重复至少两次。

测试阶段应分成几轮进行，每轮不超过一小时，每轮中间休息15分钟。编解码器和测试序列组合产生的测试和核对试验应通过伪随机分配在各轮中分布。对这个过程施加一些限制要更加复杂，但是有益。例如，如果有四轮，可以将一个给定编解码器和测试序列的四个测试试验中的每一个指派至其中一轮中随机确定的位置。这一做法的好处是确保每个系统的测试试验在整个测试轮中分布。

A5-7.1.1.4 测试分数的处理

对一个给定测试试验，测试分数为0-100量表上从“相同（SAME）”端点到观察者所做记号之间的距离。结果将根据平均意见评分（MOS）进行分析，而MOS将被用于建立受测系统的等级排序。取决于每个系统的观测数量（观察者×测试序列×重复），数据可能要经过方差分析（ANOVA）[7]。核对试验的性能可被用于得出基线“概率”判断差异。

A5-7.1.2 阶段II

阶段II的主要目标之一是细化阶段I结果的相对排名，其精度和信度可能会被有限数量的观察者和/或判断试验降低。进一步并且重要的目标是为了得到以下观察结果：图像是基于哪些特性被认为是不同的，以及阶段I中的判断是基于哪些特性做出的。

这一部分涉及专家小组对所显示素材的复审。在这一部分，专家被允许讨论所示素材，如有必要可按需多次重复部分或所有素材，以进行复审和/或者示范，从而达成一致判断和对所见的描述。如果专家观看者要求，允许使用包括模态（例如慢动作和静止帧）在内的“特技播放”。这些技术需要与测试经理进行一些交涉，以及测试经理的干预。

[7] 在感兴趣的最低阶条件下，总共有10-20个观察值就足以应用推断统计处理，如ANOVA。

A5-7.1.2.1 对受测素材分组

为妥善实施阶段II的测试，有必要对受测素材按照内容分组，获得所谓的基本专家观看集（BES），即从相同源序列获得的所有编码序列都必须分组，之后按照从阶段I导出的排名来排序。

测试素材将被按照从最低MOS值到最高MOS值的顺序排列。BES的数量与用于测试的序列数一样多。

A5-7.1.2.2 基本专家观看测试子阶段

基本专家观看（BEV）测试子阶段是一个讨论阶段。在这一阶段中，专家检查BES中包含的所有素材；一项任务是确定或修改从阶段I正式测试中得出的排名顺序。因此，必须对差异的相关可见性进行确定或修正。

A5-7.1.2.3 阶段II规划

在阶段II，所有BEV都必须实施。专家被告知演示顺序为阶段I的排名结果。专家将不知晓拥护的系统和排名之间的关系。

阶段II将以集体工作的形式，在评价者之间形成一致意见。

在开始阶段II之前，将指示评价者（可能通过书面文本形式）执行以下任务：

– 观看每个BEV中的素材。
– 讨论每个BEV中的素材的排名；如果小组对排名有异议，确定一个新排名顺序。
– 对每个案例进行评价，提供关于所见的差异的性质的详细评论（如有）。
– 记录排名、评论和观察结果。

从各组收集所有评论，并检查不符之处是测试经理的责任。在测试进行期间，从各小组获得的阶段I和II的结果将保密，防止对后续小组造成影响。在可能的条件下，测试经理被授权对不符进行鉴别，并通过对有争议的排名进一步测试来支持结论。最后一步的目的是确保整体共识。

A5-8 报告

测试的最后报告由测试经理负责。

在这份报告中将提供以下信息：

– 阶段I的结果（包括MOS表格，以及数据分析的结果（如适用））。
– 在阶段II收集的专家意见。
– 对于排名的任何重新评估的意见。
– 关于观看条件、输入信号特性、信号处理、投影仪特性、投影仪设置、色度、观看者选择和测试条件的所有相关信息
– 对于显示设备性能的完整表征（平均故障间隔时间等）。
– 概述和结论。

第3部分
附件6

多媒体应用中的图像质量的主观评价

A6-1 引言

许多国家已着手部署数字广播系统，它将允许提供包括视频、音频、静态图像、文本和图表等在内的多媒体和数据广播应用。

需要标准的主观评估方法来规定性能要求，并验证为各项应用而考虑的技术解决方案的适宜性。主观方法是必要的，原因是它们提供了测量法，允许业界更直接地预测最终用户的反应。

广播系统需要交付明显不同于当前在用的多媒体应用：信息通过固定与/或移动接收机访问；帧速率可以是固定的，或者是可变的；可能的图像尺寸变化范围很大（即从 SQCIF 到 HDTV）；典型地，视频与嵌入的音频、文本与/或语音相关；视频可以通过先进的视频编解码器来处理；并且理想的观测距离很大程度上取决于应用。

在第 2 部分中规定的主观评估方法应在这一新的背景下应用。此外，可以采用新的方法完成对多媒体系统的调查，以满足用户对多媒体领域特性的要求。

本附件描述了多媒体应用的视频质量的非交互式主观评估。这些方法可用于不同目的，包括但不限于：算法的选择、对视听系统性能的评定，以及在视听连接期间对视频质量等级进行评估。

A6-2 共性

A6-2.1 观看条件

表 3-8 列出了建议的观看条件。所用的显示器尺寸和类型应符合正在调查的应用。由于多媒体应用中使用了若干种显示技术，因此，所有有关评估中所用显示器的相关信息（例如制造商、型号和规范），都应予以报告。

当使用基于个人计算机（PC）的系统来展示序列时，还应报告系统的特性（例如视频显示卡）。

表 3-9 显示了一个有关正在测试的多媒体系统配置数据记录的例子。

如果通过使用特定的解码器-播放器组合来获取测试图像，那么这些图像必须独立于特有的外观，以便获得匿名的显示器。有必要确保质量评估不受原始环境知识的影响。

当测试中评估的系统使用降低的图像制式时，例如CIF、SIF或QCIF等，应在显示屏的一个窗口上显示片段。屏幕上背景的颜色应为50%的灰色。

表3-8

用在多媒体质量评价中的、建议的观看条件

参数	设置
观看距离(1)	限制的：1-8 H 非限制的：取决于观察者的喜好
屏幕峰值亮度	70-250 cd/m^2
非活动屏幕亮度与峰值亮度之比	≤ 0.05
当在完全黑暗的屋内仅显示黑色电平时，屏幕亮度与相应的白色电平峰值之比	≤ 0.1
图像显示器背景亮度与图形亮度峰值之比(2)	≤ 0.2
背景色度(3)	D_{65}
屋内背景亮度(2)	≤ 20 lux

(1) 观看距离通常取决于应用。

(2) 该值表示允许最大可察觉失真的设置，对某些应用，允许更高值或者取决于应用。

(3) 对PC显示器，背景色度应尽可能接近显示器的“白点”色度。

表3-9

测试中的多媒体系统的配置

参数	规范
显示器类型	
显示器尺寸	
视频显示卡	
制造商	
型号	
图像信息	

A6-2.2 源信号

源信号直接提供基准图像以及测试中的系统的输入。源片段的质量应尽可能高。作为一个指导原则，视频信号应使用 YUV（4:2:2、4:4:4 制式）或 RGB（24 或 32 位）记录于多媒体文件中。当实验者有兴趣对来自不同实验室的结果进行比较时，需要使用一组公共的源序列，以消除更大的变化源。

A6-2.3 测试素材的选择

测试场景的数目和类型对解释主观评估的结果而言是至关重要的。某些过程可能导致大多数序列相同程度的损伤。在这种情况下，用少量片段（例如两个）获得的结果应提供一个有意义的评价。不过，新的系统常常具有一定的影响，这很大程度上取决于场景或片段内容。在这种情况下，应选定测试场景的数目和类型，以便为标准的节目编排提供合理的概

括。此外，应为测试中的系统选定“临界但不过界”的素材。“不过界”这个短语暗示，场景可以仍是标准电视节目编排内容可想象的组成部分。有关场景复杂度的一个有用提示可由其空间和时间感知特性来提供。在第1部分附件6中，对空间和时间感知特性的测量有更详细的陈述。

A6-2.4 条件和锚定的范围

由于大多数评估方法对范围变化和观测条件分布是敏感的，因此判断阶段应包括变化因素的全部范围。不过，这可能与更加严格的范围近似，通过提出某些可能成为尺度极限的条件。这些可以作为例子而陈述，并确定为最大的极限（直接锚定），或分布于整个阶段中，并且不被确定为最大的极限（间接锚定）。可能的话，应使用大的质量范围。

A6-2.5 观察者

筛选后的观察者数目应至少为15。他们应当不是专家，在某种意义上，他们的日常工作与图像质量没有直接利害关系，并且他们不是经验丰富的评估者。在阶段开始前，应使用斯内伦（Snellen）或朗多（Landolt）环形视力表，对观察者进行（校正）标准视觉灵敏度筛选，并使用特别选择的视力表（如Ishihara），进行标准颜色视觉筛选。

需要的评估者数目依采用的测试程序的灵敏度和信度而定，并取决于所追求效果的期望大小。

实验者应尽可能详细地包括其评估小组成员的特点，以利于对该因素做进一步研究。提供的建议数据可以包括：职业类别（例如广播机构职员、大学学生、办公室工作人员）、性别和年龄范围。

A6-2.7 实验设计

实验者接下来要选择实验的设计方法，以便实现特定的成本和精度目标。最好是在实验中至少包括两份复制品（即相同条件下的重复试验）。重复使计算个体的信度变得可能，而且如果必要，从某些对象中放弃不可靠的结果。此外，重复确保测试中的学习效果在某种程度上能够得以平衡。通过在各次测试阶段开始之时纳入一些“虚拟演示”，可以在处理学习效果过程中获得进一步的改进。这些条件应是有代表性的演示，这些演示将在之后阶段予以显示。在对测试结果进行统计分析过程中，不考虑初步的陈述。

阶段由一系列演示构成，不应超过半个小时。

当测试多个场景或算法时，场景或算法的陈述次序应是随机的。可能要对随机的次序进行修改，以便确保相同场景或相同算法不会出现在紧邻的时间段中（即连续地出现）。

A6-3 评价方法

利用第2部分描述的方法，可以对多媒体系统的视频性能进行检测。多媒体视频质量的主观评价（SAMVIQ）方法利用了多媒体领域的特性，并可用于多媒体系统的性能评价。

第3部分
附件7

立体三维电视（3DTV）系统的主观评价

A7-1 评价（感知）的角度

立体3DTV通过重现视觉场景下感知物体相对深度的条件，充分利用了人类双眼视觉系统的特性。当前立体成像的主要要求是能用两部水平排列的摄像机捕获同一场景的至少两张视图。场景中描绘的物体图像在左右视图中将有不同的相对位置。两个视图中相对位置的差异通常被称为图像像差（或视差），一般用像素、物理距离（例如，毫米）或相对测量值（例如，与屏幕宽度的百分比）来表示。应将图像像差与角度（视网膜）像差区分开来。事实上，相同的图像像差信息，会因观看距离不同，而产生不同的角度（视网膜）像差。深度感知的程度与方向基于立体图像产生的视网膜像差的程度和方向。

一般而言，适用于单视场电视图像的评估要素，如分辨率、彩色再现、运动表现、整体质量、清晰度等亦可适用于立体电视系统。此外，立体电视系统还有许多特有的评估要素，这可包括深度分辨率（深度方向上的空间分辨率）、深度运动（沿深度方向的运动或移动是否顺利再现）以及空间畸变等要素。对于后一种要素，木偶剧效应（puppet theatre effect）（即感知的物体异乎寻常地大或小）和纸板效应（cardboard effect）（即感知的物体是立体的，但却异乎寻常地扁）是两个众所周知的例子。

我们可以确定立体系统所提供的共同影响体验质量的三个基本感知维度：图像质量、深度质量和视觉舒适度。一些研究人员认为，亦可从更普遍的概念（如自然度和临场感）角度衡量立体成像技术的心理影响。

A7-1.1 主要感知角度

图像质量是指系统所提供图像的感知质量。这是视频系统性能的一个主要决定因素。图像质量主要受技术参数以及编码和/或传输过程等产生的误差的影响。

深度质量是指系统提供增强的深度感的能力。即使是在标准2D图像中，单眼视觉线索（如线性透视、模糊、梯度等）也可传达一定的深度感。但立体3D图像亦包含像差信息，提供额外的深度信息，因此比2D的深度感更强。

视觉（不）舒适度是指对观看立体图像相关（不）舒适度的主观感觉。拍摄不当或显示不当的立体图像可造成严重的不舒适感。

A7-1.2 额外感知角度

自然度是指感知立体图像是现实的如实再现（即感知现实主义）。立体图像可能会出现各种不同的畸变，使其看起来不太自然。例如，有时立体物体看起来异乎寻常地大或小（木偶剧效应），或看起来异乎寻常地扁（纸板效应）。

临场感是指仿佛身临另一个地方或环境的主观体验。

本建议书提供了关于评估上述三个主要角度（图像质量、深度质量和视觉舒适度）的方法和程序的信息。自然度和临场感的评估方法未包含在本建议书中，但计划稍后会将这些方法纳入建议书。

A7-2 主观方法

本建议书列出了许多图像质量评估方法。所有方法都是进行一系列的评判试验，向一组观看者播放用被研究系统（如使用不同参数的算法；使用不同比特率的编码技术；不同的传输情形等）处理过的视频序列集。在每次试验中，观看者被要求用规定的量表评估视频序列的特性（如图像质量）。这些方法各不相同，主要是演示模式（即向观看者播放视频序列的方式）和观看者用来对这些序列进行评分的量表不同。

测试图像是根据第A7-4节所述项目选择的双目立体图像。评价者对下列三项进行评估：

– 图像质量：对测试图像进行处理的系统以及用于显示拟评估图像的显示器对立体3D图像分辨率的影响；

– 深度质量：对测试图像进行处理的系统以及用于显示拟评估图像的显示器对立体3D图像深度感知的影响；

– 视觉舒适度：对测试图像进行处理的系统以及用于显示拟评估图像的显示器对立体3D图像观看舒适度的影响；

本附件包括本建议书中的六种方法，这些方法已在过去二十年中成功用于解决与立体成像技术的图像质量、深度质量和视觉舒适度相关的各种研究问题。这些方法是：

– 单激励（SS）法；

– 双激励损伤量表（DSIS）法；

– 双激励连续质量量表（DSCQS）法；

– 激励比较（SC）法；

– 单激励连续质量评估（SSCQE）法；

– 同时双激励连续评估（SDSCE）法。

在使用这些方法时酌情对其进行了细微修改，如对视觉舒适度使用了不同的量表。与图像质量、深度质量和视觉舒适度的评价方法相关的演示模式和量表分别见表3-10、表3-11和表3-12。

本节中将对每种方法进行简要描述。所有方法中共同的方法要素在随后几节中介绍。

表3-10

图像质量的主观评价方法

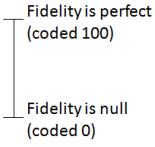
展示模式	序列持续时间	二进制量表	离散量表	连续量表
附件1第6.1节描述的单激励（SS）法	~10 s		5 优 4 良 3 中 2 差 1 劣	
附件1第4节描述的双激励损伤量表（DSIS）法			5 不可察觉 4 可察觉，但不讨厌 3 稍微讨厌 2 讨厌 1 很讨厌	
附件1第5节描述的双激励连续质量量表（DSCQS）法	~10 s			
附件1第6.2节描述的激励比较（SC）法	~10 s	A对B	−3 甚差 −2 差 −1 稍差 0 相同 1 稍好 2 较好 3 甚好	
附件1第6.3节描述的单激励连续质量评估（SSCQE）法	~3-5 min			
附件1第6.4节描述的同时双激励连续评估（SDSCE）法				

表3-11

深度质量的主观评价方法

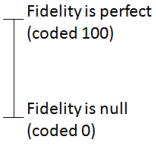
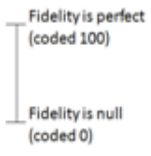
展示模式	序列持续时间	二进制量表	离散量表	连续量表
附件1第6.1节描述的单激励（SS）法	~10 s		5 优 4 良 3 中 2 差 1 劣	
附件1第4节描述的双激励损伤量表（DSIS）法			5 不可察觉 4 可察觉，但不讨厌 3 稍微讨厌 2 讨厌 1 很讨厌	
附件1第5节描述的双激励连续质量量表（DSCQS）法	~10 s			
附件1第6.2节描述的激励比较（SC）法	~10 s	A对B	−3 甚差 −2 差 −1 稍差 0 相同 1 稍好 2 较好 3 甚好	
附件1第6.3节描述的单激励连续质量评估（SSCQE）法	~3-5 min			
附件1第6.4节描述的同时双激励连续评估（SDSCE）法				

表3-12

视觉舒适度的主观评价方法

展示模式	序列持续时间	二进制量表	离散量表	连续量表
附件1第6.1节描述的单激励（SS）法	~10 s		5 非常舒适 4 舒适 3 有些不舒适 2 不舒适 1 非常不舒适	
附件1第4节描述的双激励损伤量表（DSIS）法			5 不可察觉 4 可察觉，但不讨厌 3 稍微讨厌 2 讨厌 1 很讨厌	
附件1第5节描述的双激励连续质量量表（DSCQS）法	~10 s			
附件1第6.2节描述的激励比较（SC）法	~10 s	A对B	−3 甚差 −2 差 −1 稍差 0 相同 1 稍好 2 较好 3 甚好	
附件1第6.3节描述的单激励连续质量评估（SSCQE）法	~3-5 min			
附件1第6.4节描述的同时双激励连续评估（SDSCE）法				

A7-3 一般观看条件

观看条件（包括屏幕亮度、对比度、背景亮度、观看距离等）应与第1部分第2.1节中所述用于2D的那些观看条件一致。这种一致性方法的理由有两方面。其一，在实际中用户以与2D相同的显示器和观看条件观看3DTV。其二，3DTV视频技术性能的进展通常需要与标准HDTV视频技术的进展联系起来（即对比）衡量。

第1部分第2.1节规定了两种可能的观看距离选择标准。这里将选择设计观看距离（DVD）。对数字系统而言，DVD是指两个邻近像素对应观看者眼睛形成1弧分角度的距离。

应注意，两个邻近像素对应观看者眼睛成1弧分的角度，那么在设计观看距离，系统可描绘出的最小角度（视网膜）像差（即系统的深度分辨率）等于1弧分（或60弧秒）。研究表明，近97%的人能够区分等于或小于140弧秒的水平像差，至少80%能察觉30弧秒的水平像差。因此，大多数观看者应该能在设计观看距离分辨目前3D视频系统可描绘出的最小像差。

A7-4 测试素材

应按研究中要解决的实验问题选择测试素材。一般情况下，测试序列在内容（体育、戏剧、电影等）及其时空特性上应在被研究业务所传送的节目中具有代表性。

此外，选定的立体测试序列内容通常亦应可以舒适地观看。立体图像的视觉舒适度主要取决于图像包含的像差（视差）和观看条件。因此，应注意确保这种像差不超过下一节中列出的限值，除非研究的目的是专门衡量视觉舒适度。此外，在可能的情况下，应测量和报告测试序列像差分布的统计数据：平均值、标准偏差和范围（最小/最大）。

在选择易于观看的立体3D图像的测试图像时，可将视差（即左右眼图像之间的不一致性）以及视差的分布和变化列为应考虑的项目。易于观看的立体3D图像与视差（左右眼图像之间的不一致性）以及视差的分布和变化之间的关系见随后几个分节的内容。

A7-4.1 基准视频素材的使用

研究人员可能希望将基准序列（如有的话）纳入测试序列集。基准序列通常是测试序列未经任何处理的版本（即原始的源序列）。对于立体研究，主要的基准是原始的、未经处理的立体序列。但实验计划可能还包含基准序列的单视场版（即原始源序列的单视图版）；例如，在视觉舒适度研究中，将单视场参考序列的视觉舒适度作为基准可能是有帮助的。单视场参考序列应以3D模式演示（如使用与实际的立体序列相同的3D软件设置对左右两眼演示左视图）。在实验计划中纳入基准序列有两个重要优势。其一，提供了衡量被研究算法或技术实现的透明度（亦称保真度）的机会[8]。其二，纳入基准序列就相当于提供了一个高质量的参照，可以帮助稳定观察者的评分[9]。

---

[8] 透明度（保真度）是描述编解码器或系统相对于理想传输系统的性能而没有任何退化的概念。很容易看出，通过比较分配给参考序列的评级与分配给使用所研究算法或技术处理的序列的评级，可以测量透明度。

[9] 认识到，通过使用低质量的锚，也可以提高跨空间（即跨不同实验室）和时间（即在同一实验室在不同时间）评分的稳定性。然而，国际电联确实有立即的计划生产/定义用于立体成像技术评估的标准化低质量锚。

A7-4.2 视觉舒适度限值

像差/视差过大会导致视觉不舒适，原因可能是这会恶化调节与聚焦功能之间的冲突。因此已提出建议，如要尽可能减小调节与聚焦之间的冲突，那么立体图像的像差应足够小，以便感知的物体深度在“舒适区”范围内。已提出了若干方法来界定这些限值。一种方法是用屏幕视差的程度来规定舒适观看的限值，以屏幕水平尺寸的百分比表示。并建议对于交叉/负像差，这一限值为1%，对于非交叉/正视差，为2%（总计约为3%）。而另一种方法是，舒适区根据眼睛的景深界定。对于电视广播典型的观看条件，研究人员已假定景深介于0.2D（屈光度）与±0.3D（屈光度）之间。对于从3.1H的设计观看距离观看1920×1080（ITU-R BT.709建议书）HDTV图像分辨率的系统，这些值约对应屏幕视差的±2%和±3%。最后，第三种方法从视网膜像差角度规定舒适度限值，并设定对于正负像差，这些限值均为视角的±1°。

值得一提的是，这些不同的方法往往会得到相同的舒适度限值。回想在设计观看距离，两个邻近像素对应观看者眼睛成1弧分的角度。因此，60个像素对应1°的视角。这样我们就可以很容易地从视网膜像差角度规定舒适度限值（对一般观看者）。例如，对于1920×1080（ITU-R BT.709建议书）HDTV图像分辨率的系统，1%（~19.2个像素）约对应20弧分，2%约对应40弧分，3%约对应60弧分（或1°）。

应注意，尽管在设计观看距离，两个邻近像素总是对应1弧分的角度，二者之间的物理分隔距离（如以毫米为单位）随显示器的增大（像素的数目不变，但屏幕的物理尺寸增加）而增加。因此，对较大的显示器，较高的限值（如±3%）会导致对应点之间的物理距离（两个视图的视差，以毫米为单位）超过一般观看者的瞳距（~63-65毫米）。这可能会增加不舒适感。

A7-4.3 左右图像之间的差异

在立体3D系统中，通过将左右图像分别呈现给左右两眼而形成双目3D图像。如两个图像之间产生差异，可导致身心应激，在一些情况下可能无法观看3D图像。例如，在拍摄和播放立体3DTV节目时，左右图像之间可能会出现几何畸变，如大小不一致、垂直移位、旋转误差等。测试图像最好不要出现这些几何畸变。欲了解进一步的信息，可参见ITU-R BT.2160-2报告附件4的第3.2.1节。

在选择易于观看的立体3D图像的测试图像时应考虑的关于左右图像之间差异的项目如下：

– 几何差异，包括大小、垂直位移和偏转；
– 亮度差异，包括黑白度；
– 串扰。

A7-4.4 视差的范围、分布和变化

视差分布与立体图像的视觉舒适度相关。

立体图像的视差分布在场景切换帧期间是不连续的。极端视差或视差突变会造成视觉不适感，因此小心控制测试图像的视差甚为重要。欲了解进一步的信息，见ITU-R BT.2160-2报告附件4第3.2.2节。

一般情况下，鉴于使用立体测试序列的研究可能会引起某种程度的视觉不适感，因此推荐在可能的情况下使用像差不超过舒适度限值的测试素材，虽然偶尔超过这些限值的情况可能是允许的。

A7-5 实验装置

实验装置（视频服务器、显示器等）应能够播放全分辨率HD测试序列，如使用HDMI帧封装制式。这可使可开展的研究的范围具有更大的灵活性。

至今尚未规范用于3DTV评估的参考显示器。因此，大多数研究人员预计将使用目前的消费级3DTV显示器。鉴于这类显示器的特性可能因制造商不同亦各不相同，大力提倡研究人员描述研究中使用的显示器的设置信息。

A7-6 观察者

A7-6.1 样本容量

一般情况下，推荐使用至少30名观看者。但人们认识到，既然3D研究的样本容量考量与2D研究不同，那么实际的数量将取决于研究的具体目标。

A7-6.2 视力筛选

应使用目前的临床视力测试方法从视敏度、色盲和立体视觉方面对观察者进行筛选（如对视敏度使用斯内伦（Snellen）视力表，针对色觉使用Ishihara（假同色图或等效方法，针对立体视觉使用Rando立体图或等效方法）。注意，Randot、Stereo Fly或Frisby测试等立体视测试通常测量从约20弧秒至400弧秒的视网膜像差。提倡研究人员描述参与研究的观察者立体视力方面的相关统计数据。如需对参与者的立体视力进行更详细的分析，研究人员可使用本附件附录1中所示的测试素材。

A7-7 观察者须知

应针对研究角度（如深度质量、舒适度等）编写须知。值得一提的是，3D研究的道德准则比2D图像质量评估通常使用的准则更为严格，因为参与者可能会出现视觉不适。一般情况下，3D研究中在向参与者说明研究动机以及暴露于研究中使用的激励要素可能产生的任何负面影响时要更为谨慎。

A7-8 测试阶段持续时间

如观看素材被认为是舒适的，那么测试阶段的持续时间可能与2D研究一样长（即加休息时间~20-40分钟）。如已知素材视差过大，因此存在已知的不舒适的可能性，那么应对持续时间加以限制。

A7-9 响应的可变性

主观评价实验中观看者提供的评分通常是多变的。观看者之间的差异可能只反映了参考人群的特性，因此可通过增加样本容量解决这个问题。

但是，其中部分可变性可能源自实验过程中个体观看者响应模式的变化。这些变化意味着评价标准的改变，由于任务练习的增加以及熟悉伪影特性等情况，这种变化是有可能发生的。为了尽可能降低这种可变性的负面效应，研究人员应提供适当的培训程序（任务、退化程度等），使用多种随机化方式（即以不同的随机顺序向不同的观看者演示测试序列），并重复试验（通过这种方式亦可衡量响应模式可能的变化）。

A7-10 观看者的舍弃标准

针对第A7-2节所述的方法，观看者的舍弃标准（“观察者的筛选”）见第1部分。

A7-11 统计分析

3D成像系统研究的统计分析与2D成像系统相同。

附件7
附录1

视力测试的测试素材

A7-1 视力测试

表3-13列出了视力测试的测试用图。这12种测试是根据人类视觉系统的层次结构从低到高选定的。下文描述了八种主要的视力测试（VT），其他四种是临床测试。观察者必须有正常的立体视觉，这意味着他们必须通过精细立体视VT-04和动态立体视VT 07测试。其余六种测试则针对更细微的表征。测试图应放在显示器屏幕高度的三倍处观看。

表3-13

视力测试的立体测试素材

编号	项目	测试目的	内容
1	同时视	同时感知两眼分视图像且合成位置正确的能力	呈现给一只眼的图像是一个笼子，呈现给另一只眼的图像是一头狮子
2	双眼融像	将左右眼分视图像合二为一的能力	呈现给一只眼的图像有两个点，呈现给另一只眼的图像有三个点，其中一个点是共同的
3	粗略立体视	将有视差的两眼分视图像合成为一个有粗略深度的图像的能力	呈现给两眼的图像是一只翅膀展开的蜻蜓的立体像对
4	精细立体视	将有视差的两眼分视图像合成为一个有精细深度的图像的能力	提供九个测试用菱形图形，每个菱形中有四个圆圈，其中一个略有视差
5	交叉融像限值	将有交叉像差的两眼分视图像合二为一的能力	演示长条图形的立体像对，交叉视差变化率为10'/s
6	非交叉融像限值	将有非交叉视差的两眼分视图像合二为一的能力	演示长条图形的立体像对，非交叉视差变化率为11'/s
7	动态立体视	在移动随机点立体图像中感知深度的能力	动态随机点立体图
8	双目锐度	双目锐度，包括可能影响良好立体视觉的单眼锐度不平衡	各种方向和大小的E字符
9	水平斜视	患者无法克服的眼位水平偏斜	垂直线和水平线
10	垂直斜视	患者无法克服的眼位垂直偏斜	垂直线和水平线
11	不等像	两眼视像的形状和大小不一样的情形	左眼用图包括“[o”字符，右眼用图包括“o]”字符，其中字符“o”的位置是共同的
12	旋转隐斜	融合作用遭到阻断时一只眼绕垂直轴的偏斜趋势	左眼用图是钟表的表盘，右眼用图是指向六点钟的钟表指针

注 1 – 这些素材采用1125/60/I制式（见ITU-R BT.709建议书）。

注 2 – 这些素材可从图像信息与电视工程师学会（ITE）获得，地址3-5-8 Shibakoen, Minato-ku, Tokyo 105-0011, Japan，电话：81-3-3432-4675，电子邮件：ite@ite.or.jp。

下面用并排摆放进行交叉自由融像的左右缩略图加以说明。

1) VT-01：同时视（狮子测试）

测试同时感知两眼分视图像且合成位置正确的能力。呈现给一只眼的图像是一个笼子，呈现给另一只眼的图像是一头狮子，狮子位置变化速率为12′/s。每个图像的大小固定在10°，以便观察者可以在旁黄斑区捕获图像。视力正常的观察者可以在演示阶段内的某个时间看到狮子在笼子里。

图3-8

VT-01测试图

右眼用图

左眼用图

BT.0500-03-8

2) VT-02：双眼融像（worth 4点测试）

测试将左右眼分视图像合二为一的能力。呈现给一只眼的图像有两个点，呈现给另一只眼的图像有三个点，其中一个点是共同的。视力正常的观察者可以看到四个点。

图3-9

VT-02测试图

右眼用图

左眼用图

BT.0500-03-9

3) VT-03：粗略立体视（蜻蜓测试）

测试将有视差的两眼分视图像合成为一个有粗略深度的图像的能力。呈现给两眼的图像是一只翅膀展开的蜻蜓的立体像对。视力正常的观察者可看到蜻蜓的翅膀在显示屏前面。

图3-10

VT-03测试图

右眼用图

左眼用图

BT.0500-03-10

4) VT-04：精细立体视（圆圈测试）

测试将有视差的两眼分视图像合成为一个有精细深度的图像的能力。提供九个测试用菱形图形，每个菱形中有四个圆圈，其中只有一个略有视差。视力正常的观察者可看到具有视差的圆圈在显示屏的前面。表3-14列出了测试编号、正确答案和3 *H*处的立体角度。

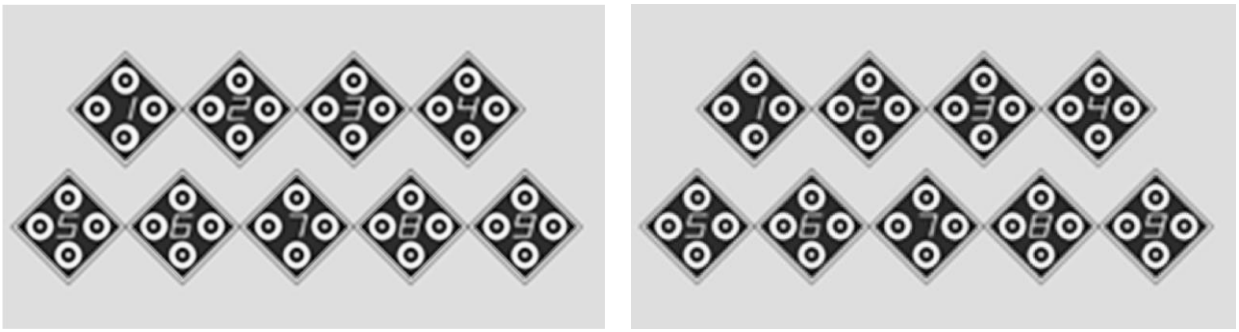
表3-14

正确答案和视差

测试编号	正确答案	3 *H*处的立体角度（"）
1	下	480
2	左	420
3	下	360
4	上	300
5	上	240
6	左	180
7	右	120
8	左	60
9	–	0

图3-11

VT-04测试图



右眼用图　　左眼用图

BT.0500-03-11

5) VT-05:交叉融像限值（长条图形测试）

测试将有交叉像差的两眼分视图像合二为一的能力。演示长条图形的立体像对，视差变化率为10′/s。可以测量上升和下降系列的融像限值。指示观察者在上升系列中一看到两个图像立刻报告融像间断，在下降系列中一看到的分视图像合为一个单一图像时报告其融像恢复。

图3-12

VT-05测试图

右眼用图　　左眼用图

BT.0500-03-12

6) VT-06：非交叉融像限值（长条图形测试）

测试将有非交叉视差的两眼分视图像合二为一的能力。演示的图像与上述交叉测试中的相同，但左右眼用图对调。

图3-13

VT-06测试图

右眼用图　　左眼用图

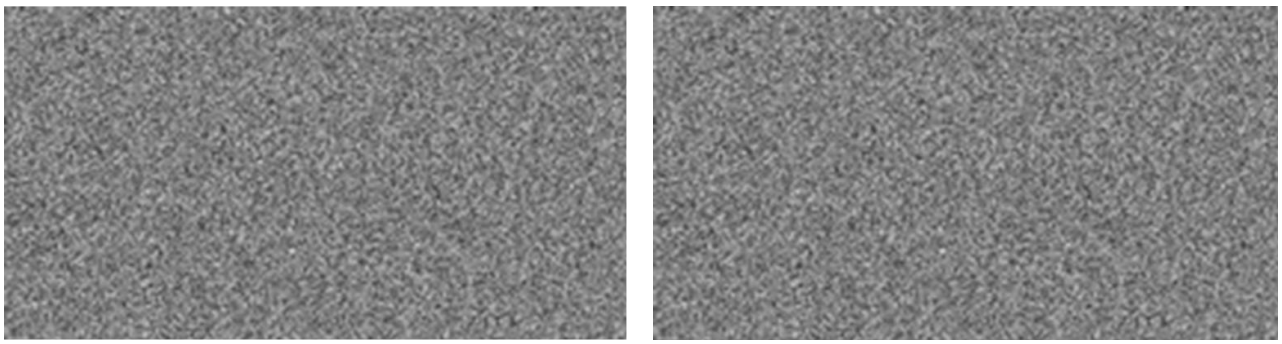
BT.0500-03-13

7) VT-07：动态立体视（动态随机点立体图测试）

测试在移动随机点立体图像中感知深度的能力。视力正常的观察者可在动态随机点立体图中看到一个矩形的形状和正弦深度运动。

图3-14

VT-07测试图



右眼用图　　左眼用图

BT.0500-03-14

8) VT-08：双目锐度（锐度测试）

测试双目锐度及双眼融像，包括可能影响良好立体视觉的单眼锐度不平衡。图像由四列五行各种方向和大小的E字符组成。中间两列，双眼皆看看到，左边两列仅左眼可以看到，右边两列仅右眼可以看到。视力正常的观察者可以正确说出E字符的方向。字符的大小对应3 *H*处的锐度：约1.0、0.5、0.33、0.25和0.125。

图3-15

VT-08测试图

右眼用图　　左眼用图

BT.0500-03-15

9 & 10)　VT-09：水平斜视（水平马氏杆（maddox）测试）和VT-10：垂直斜视（垂直马氏杆测试）

这些图用于衡量眼位的水平和垂直偏斜。双眼视轴假定了一个相对于彼此、与生理条件要求不同的位置。演示的图像包括一条垂直线和一条水平线。视力正常的观察者可以看到两线的交叉点约在二者的中心处。标记旁数字的单位是棱镜度，其中PD（瞳距）=65毫米，观看距离为3.02 H。

图3-16

VT-09测试图

BT.0500-03-16

图3-17

VT-10测试图

BT.0500-03-17

11) VT-11：不等像（“[ ]”字符测试）

两眼视像的形状和大小不一样的情形。左眼用图包括“[o”字符，右眼用图包括“o]”字符，其中字符“o”的位置是共同的。视力正常的观察者可以看到相同大小相同高度的“[”和“]”。

图3-18

VT-11测试图

BT.0500-03-18

12) VT-12：旋转隐斜（钟表测试）

只有在一眼被遮盖或融合作用遭到阻断时才会表现出来的绕垂直轴的眼位偏斜。左眼用图是钟表的表盘，右眼用图是指向六点钟的钟表指针。视力正常的观察者可以看到钟表恰好显示六点钟。

图3-19

VT-12测试图

BT.0500-03-19