国　　际　　电　　信　　联　　盟

ITU-R
国际电联无线电通信部门

ITU-R BT.500-13 建议书
(01/2012)

电视图像质量的主观评价方法

BT 系列
广播业务
(电视)

ITU 国际电信联盟

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或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）

ITU-R的IPR政策述于ITU-R第1号决议的附件1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/en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 http://www.itu.int/publ/R-REC/en）

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
2012年，日内瓦

ITU-R BT.500-13建议书

电视图像质量的主观评价方法

（ITU-R 81/6号课题）

（1974-1978-1982-1986-1990-1992-1994-1995-1998-1998-2000-2002-2009-2012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、等级量表和观看条件。本建议书推荐了双激励损伤量表 (DSIS) 法和双激励连续质量量表 (DSCQS) 法，以及替代评价方法，比如单激励 (SS) 法、激励比较法、单激励连续质量评价 (SSCQE) 法和同时双激励连续评价 (SDSCE) 法。

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，在测量的诸多方面存在着相当程度的一致性；

c) 采用一种标准的方法，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的质量和/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) 数字编码和比特率压缩等新型电视信号处理的引入、使用时间复用分量的新型电视信号的引入，以及增强电视和HDTV等新业务可能的引入，都可能需要改变进行主观评价的方法；

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有可能受到信号链中之前各部分所进行的处理的制约，

建议

1 在实验室实验中，应采用下列各附件中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，也应在运行评价中采用；

2 在不远的将来，尽管存在可替代方法并会开发一些新方法，仍应尽可能采用本建议书附件1的第4和第5节所述的那些方法；

3 鉴于确定主观评价的基础很重要，在所有测试报告中应给出测试配置、测试素材、观察者和所用方法可能最全面的描述；

4 为便于在不同的实验室之间交换信息，应按照本建议书附件2中详述的统计技术处理收集到的数据。

注1 – 附件1给出了关于确定电视系统性能的主观评价方法的资料。

注2 – 附件2给出了关于处理在主观测试过程中收集到的数据所用统计技术的说明。

附件1

评价方法说明

1 引言

主观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受测系统的人的反应。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。第一类评价是确定在最佳条件下系统的性能。这类评价通常称为质量评价。第二类评价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。这类评价通常称为损伤评价。

为开展适宜的主观评价，首先必须对那些与要解决的评价问题的目标和环境最符合的不同选项进行选择。为帮助完成这一任务，除在第2节给出一般特性外，还在第3节提供了一些每种方法要解决的评价问题的资料。第4和第5节则对两种主要的推荐方法做了详细说明。最后，第6节给出了关于正在研究的替代方法的一般性资料。

本附件的用途限于对评价方法进行详细说明。但选择最适宜的方法则由待测系统所针对的业务目标决定。因此，特定应用的完整评价程序在其他建议书中给出。

2 共同的特性

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建议书中给出。

2.1 通用观看条件

说明不同观看条件的不同环境。

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者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近家庭环境。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。

探讨与监视器分辨率和对比度有关的某些方面。

2.1.1 实验室环境

2.1.1.1 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)	未激活显像管屏幕亮度与峰值亮度之比：	≤ 0.02
b)	在全暗的房间内显示时，仅显示黑电平的屏幕亮度与相应仅显示峰白电平的屏幕亮度之比：	≈ 0.01
c)	显示器亮度和对比度：	通过PLUGE建立（见ITU-R BT.814建议书和ITU-R BT.815建议书）
d)	相对于标称值的最大观察角度（该数字适用于阴极射线管显示器，其他显示器适用的数字则正在研究）：	30°
e)	图像监视器后的背景亮度与图像峰值亮度之比：	≈ 0.15
f)	背景：	D_{65}
g)	照明：	低

2.1.2 家庭环境

2.1.2.1 家庭环境中主观评价的通用观看条件

a)	未激活显像管屏幕亮度与峰值亮度之比：	≤ 0.02 （见第2.1.4节）
b)	显示器亮度和对比度：	通过PLUGE建立（见ITU-R BT.818建议书和ITU-R BT.815建议书）
c)	相对于标称值的最大观察角度（该数字适用于阴极射线管显示器，其他显示器适用的数字则正在研究）：	30°
d)	4/3图像宽高比的屏幕尺寸：	该屏幕尺寸应满足优选观看距离(PVD)规则
e)	16/9图像宽高比的屏幕尺寸：	该屏幕尺寸应满足PVD规则

f) 监视器处理： 未经数字处理

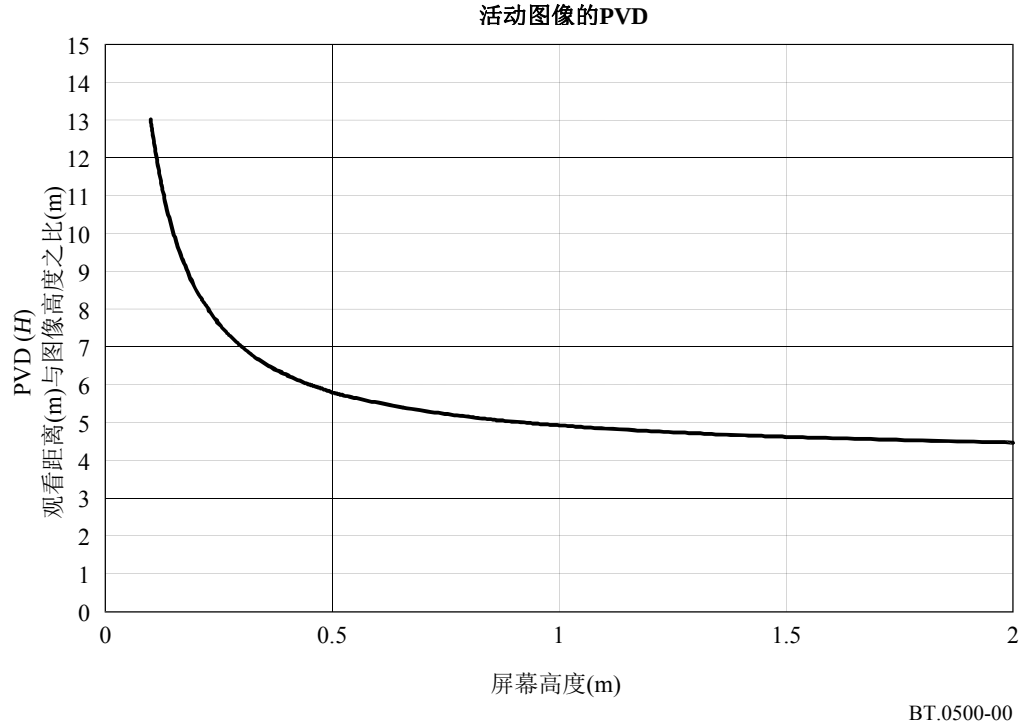
g) 监视器分辨率： 见第2.1.3节

h) 峰值亮度： 200 cd/m^2

i) 屏幕上的环境照度（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

之所以选择这样的观看距离和屏幕尺寸，是为了满足PVD。PVD（随屏幕尺寸而变）在下表和下图中给出。所给数字对标准清晰度电视 (SDTV) 和高清晰度电视 (HDTV) 都有效，因为没发现什么差别。

屏幕对角线 (英寸)		**屏幕高度 (*H*)**	**PVD**
4/3宽高比	**16/9宽高比**	**(m)**	**(*H*)**
12	15	0.18	9
15	18	0.23	8
20	24	0.30	7
29	36	0.45	6
60	73	0.91	5
> 100	> 120	> 1.53	3-4



上表和上图旨在为关于特定应用的建议书所采用的PVD和相关屏幕尺寸提供资料。

2.1.3 监视器分辨率

采用专业阴极射线管 (CRT) 的专业监视器在各自亮度工作范围内通常符合主观评价所需的分辨率标准。

并非所有监视器都能达到200 cd/m^{2}的峰值亮度。

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。

采用消费型CRT的消费型电视机对主观评价而言，根据亮度值的不同，分辨率可能不够。

在这种情况下，强烈建议对最大和最小分辨率（屏幕中心和四角）进行检验和报告。

目前在主观评价执行过程中，对检验监视器分辨率或消费型电视机分辨率来说最实用的系统采用了由电子设备生成的扫描测试图形。

可采用视觉分析来检验分辨率。视觉门限大致定为–12/–20 dB。这套方法的主要缺陷是由荫罩产生的失真加大了视觉评价的难度，但另一方面，存在失真表明视频信号超出了荫罩限定的范围，造成对视频信号的抽样不足。

可建议对CRT的分辨率测试进行进一步研究。

2.1.4 监视器对比度

对比度可能会受到环境照度的强烈影响。

专业监视器的CRT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CRT采用技术措施获得高照度环境下更强的对比度。

要计算给定CRT的对比度，需要得到该CRT的屏幕反射系数K。最好的情况是，屏幕反射系数近似为$K = 6\%$。

在I为200 lux的漫射环境下，$K = 6\%$，3.82 cd/m^2，未激活显像管屏幕区的亮度反射采用下式计算：

$$L_{反射} = \frac{I}{\pi} K$$

代入给定的值，则反射的亮度 (cd/m^2) 接近入射照度 (lux) 的2%。

CRT正面的玻璃被认为不存在镜面反射，正面玻璃对对比度产生的确切影响难以量化，因为这与照明条件有极大关系。

在第2.1.1和第 2.1.2节中，对比度之比CR由下式表示：

$$CR = L_{min} / L_{max}$$

式中：

L_{min}： 在周围照度下未激活显像管屏幕区的亮度(cd/m^2)（对于给定的值，$L_{min} = L_{未激活区} + L_{反射} = 3.82$ cd/m^2）

L_{max}： 在周围照度下显像管白色区的亮度(cd/m^2)（对于给定的值，$L_{max} = L_{white} + L_{反射} = 200 + 3.82$ cd/m^2）。

采用上述各值算出$CR = 0.018$，与第2.1.1.1节的a)和第2.1.2.1节的a)给出的值0.02相当接近。

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。

以数字方式存储的图像和序列是最能再现的源信号，所以它们是优选的类型。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。还有可能会有录像带和计算机磁带格式。

在短期内，35 mm幻灯片扫描器为静止图像提供了一个优选信号源。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。

在信号的形成过程中，早期阶段完成的任何处理所产生的效果都可能影响待测系统的性能，因此常常会需要考虑这种影响是如何产生的。有鉴于此，如果希望检查信号链上分步处理引起的损伤是如何累积的，则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。

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

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其他导则，见附件1的附录1和附录2。

某些参数可能会对大多数图像和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，纳入严格素材是很重要的，因为在分析结果时可能要考虑这种情况，而从非严格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“严格但不过分严格”素材。“但不过分严格”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4个素材。例如，其中一半肯定是严格的，另一半是中等严格的。

一些组织已经开发了测试静止图像和序列。将来有望将其纳入到ITU-R的框架内。ITU-R建议书中提出了一些具体的图像素材，用于各种应用的评价。

关于选择测试素材的其他见解在附件1的附录1和附录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引入的图像具有专长。非专家（“无知”）观察者对测试系统引入的图像不具备专长。无论怎样，不应使观察者直接参与，从而在对所研究的系统的开发中掌握具体面详细的情况。

在测试阶段开始之前，应对观察者进行筛选，使之对Snellen氏E字视力表或Landolt氏C字视力表具有（校正至）正常的视敏度，并采用专门选定的表（例如石原氏色盲检查表）使之具有（校正至）正常的彩色视觉。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感受性和信度，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知识应体现在报告中。

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感受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。但在过渡期间，实验者应纳入尽可能详细的评价人员的特点，以促进对这一因素的进一步研究。建议提供的数据包括：职业类别（例如广播机构雇员、大学生、办公室工作人员等），性别和年龄范围。

2.6 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、顺序及定时。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，但具有可相比较的感受性。对于质量评价的情况，可以规定质量为具体的可感知属性。

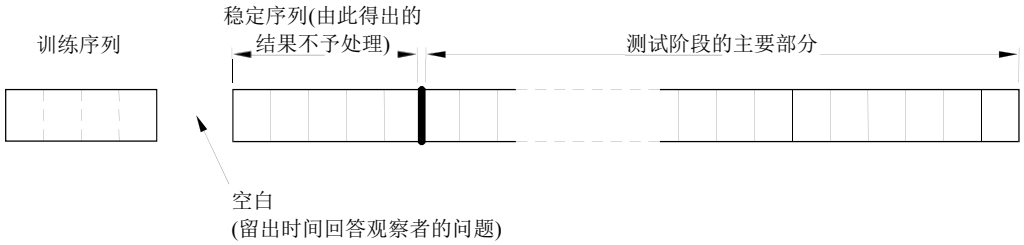
2.7 测试阶段

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”，以稳定观察者的评分。这几个演示中给出的数据不能在测试结果中考虑。如果需要若干测试阶段，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（例如从希腊-拉丁方导出）；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。为检查相干性，有些演示可在不同的测试阶段予以重复。

图1

测试阶段的演示结构



BT.0500-01

2.8 结果的表示

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

对每一测试参数，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的附件2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；
– 图像源和显示监视器的类型（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（更确切地说是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。表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方法的调查结果。

表2

测试方法的选择

评价问题	所用的方法	说明
测量系统相对于某一基准的质量	双激励连续质量量表(DSCQS)法(1)	ITU-R BT.500建议书, 第5节
测量系统的牢靠程度(即降质特性)	双激励损伤量表(DSIS)法(1)	ITU-R BT.500建议书, 第4节
量化系统的质量(如果未提供基准的话)	比率量表法(2)或类别量表(正在研究)	ITU-R BT.1082报告
比较替代系统的质量(如果未提供基准的话)	直接比较法、比率量表法(2)或类别量表(正在研究)	ITU-R BT.1082报告
识别出影响系统出现可感知差别的因素并衡量这些因素的可感知影响	所用方法正在研究	ITU-R BT.1082报告
确定损伤变为可见的那一点	迫选法中的门限估值或调整的方法 (正在研究)	ITU-R BT.1082报告
确定系统是否出现可感知差别	迫选法(正在研究)	ITU-R BT.1082报告
测量立体图像编码的质量	双激励连续质量量表(DSCQS)法(3)	ITU-R BT.500建议书, 第5节
测量两个受损视频序列之间的保真度	同时双激励连续评价(SDSCE)法	ITU-R BT.500建议书, 第6.4节
比较不同的容错工具	同时双激励连续评价(SDSCE)法	ITU-R BT.500建议书, 第6.4节

(1) 对DSCQS和DSIS法开展了关于背景效应的一些研究。研究发现，DSIS法会因背景效应而出现一定程度的偏差。其他细节在附件1的附录3给出。

(2) 有些研究显示，在能获得各种质量的情况下，该方法更为稳定。

(3) 由于在评价立体图像时有可能特别疲劳，一个测试阶段总的持续时间应缩短为不到30 min。

4 双激励损伤量表（DSIS）法（EBU法）

4.1 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5节中的方法。

双激励（EBU）法是一种交替方法，因为在这种方法中，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。背景效应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。

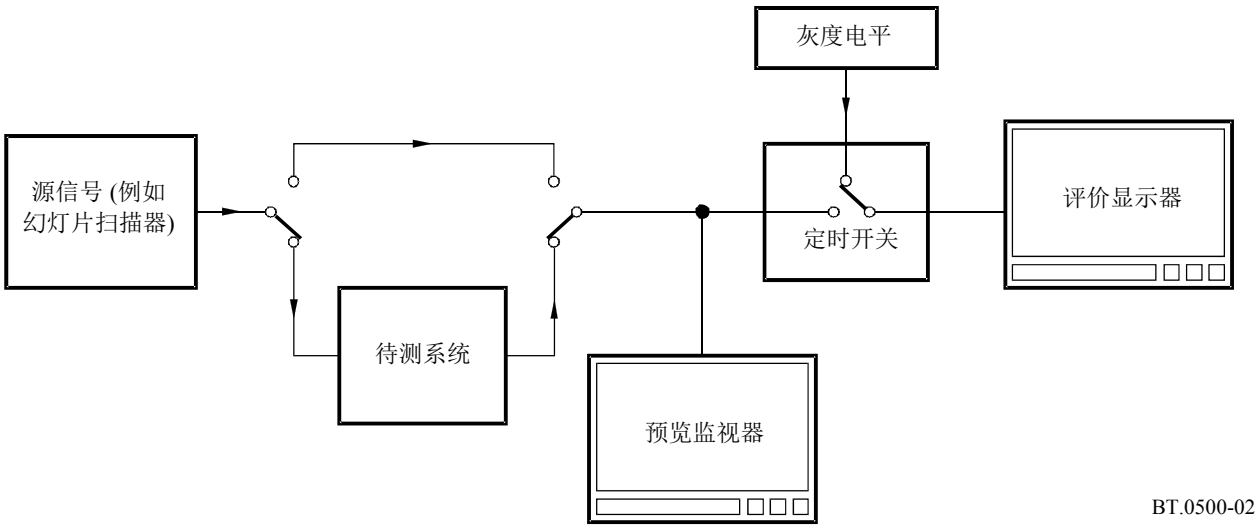
4.2 总体布置

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

测试系统的总体布置应如图2所示。

图2

DSIS法中测试系统的总体布置



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。

4.3 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3a)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3b)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。

4.4 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，有编了号的框或其方式来记录分级。

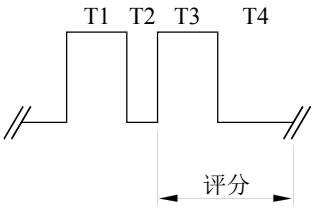
4.5 评价须知

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、测试图像、评分期）。应在图像中显示要评价的损伤的范围和类型，该图像不同于测试中要用的图像，但具有可相比较的感受性。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。

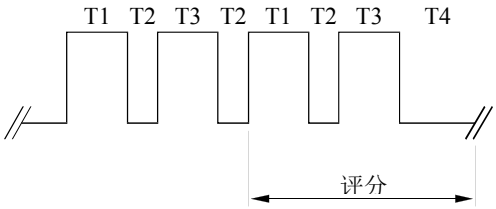
应要求观察者在T1和T3的整个持续时间内观看图像。只允许在T4期间内评分。

图3

测试素材的演示结构



a) 变型I



b) 变型II

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 = 5-11 s 中灰度场

实验显示，将T1和T3延长至10 s以上不会提高评价者确定图像或序列的等级的能力。

BT.0500-03

4.6 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。

关于损伤程度的其他见解在附件1的附录2中给出。

5 双激励连续质量量表（DSCQS）法

5.1 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。

5.2 总体布置

观看条件、源信号、测试素材、观测者以及结果的表示在第2节中做了规定或按照第2节加以选择。测试阶段的说明见第4.6节。

测试系统的总体布置应如图4所示。

5.3 测试素材的演示

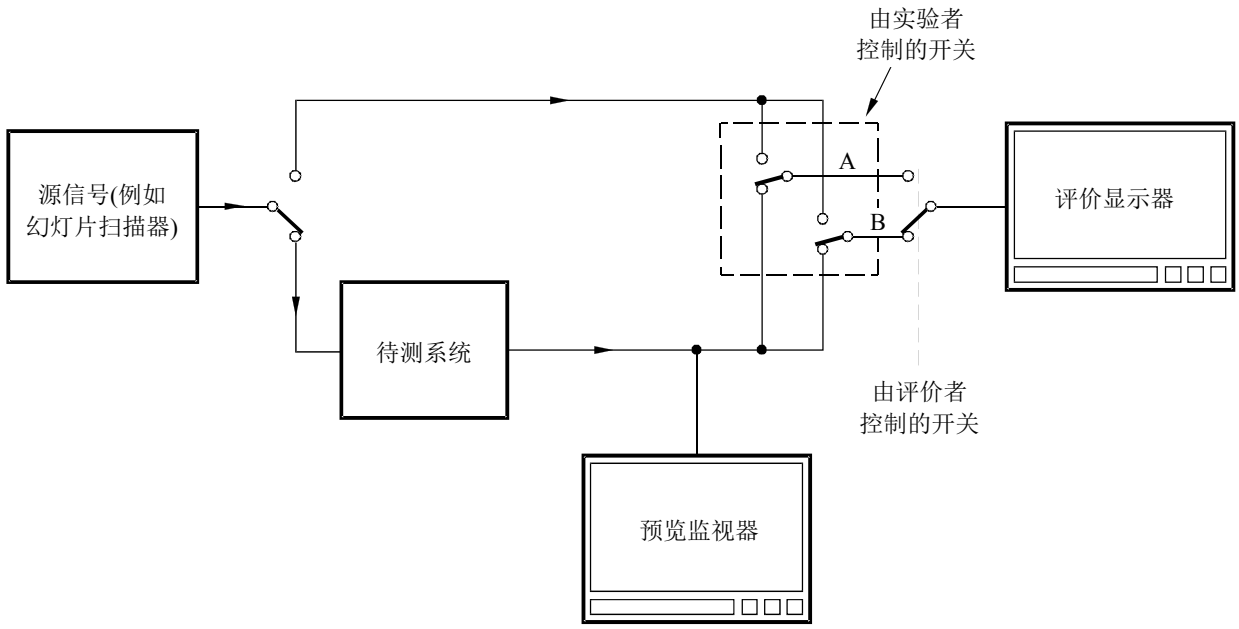
一个测试阶段由多次演示组成。对于只有一位观察者的变型I，每次演示时观察者都可以在信号A和信号B之间自由转换，直到观察者得出与每一信号的质量相关的心理尺度为止。对于同时有几位观察者的变型II，在记录结果之前，条件对要显示一次或多次，每次持续时间相同，以便让观察者得出与这一对条件的质量相关的心理尺度，然后再把条件对显示一次或多次，同时记录结果。重复的次数取决于测试序列的长度。对于静止图像，使用3-4 s的序列并重复5次（在最后2次期间评分）可能是合适的。对于受到时变扰动的活动图像，10 s的序列和2次重复（在第2次重复期间评分）可能是合适的。图5示出了演示的结构。

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4

DSCQS法中测试系统的总体布置



这种方法有下述I和II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。质量范围有限的该变型得出的结果的稳定性被认为尚需进一步研究。

BT.0500-04

5.4 分级量表

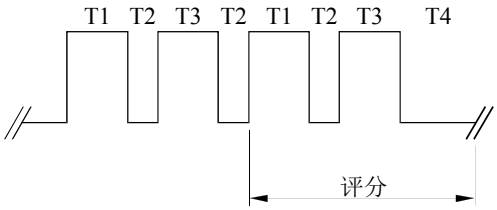
这种方法要求评价每一测试图像的两种版本。每对测试图像中，有一个是无损的，而另一个可能包含损伤，也可能不包含损伤。无损的图像就作为基准，但不告诉观察者哪个是基准图像。在测试系列中，基准图像的位置是以伪随机方式变化的。

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4示出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5

测试素材的演示结构



演示阶段：

T1 = 10 s 测试序列**A**

T2 = 3 s 由200 mV图像电平产生的中灰度场

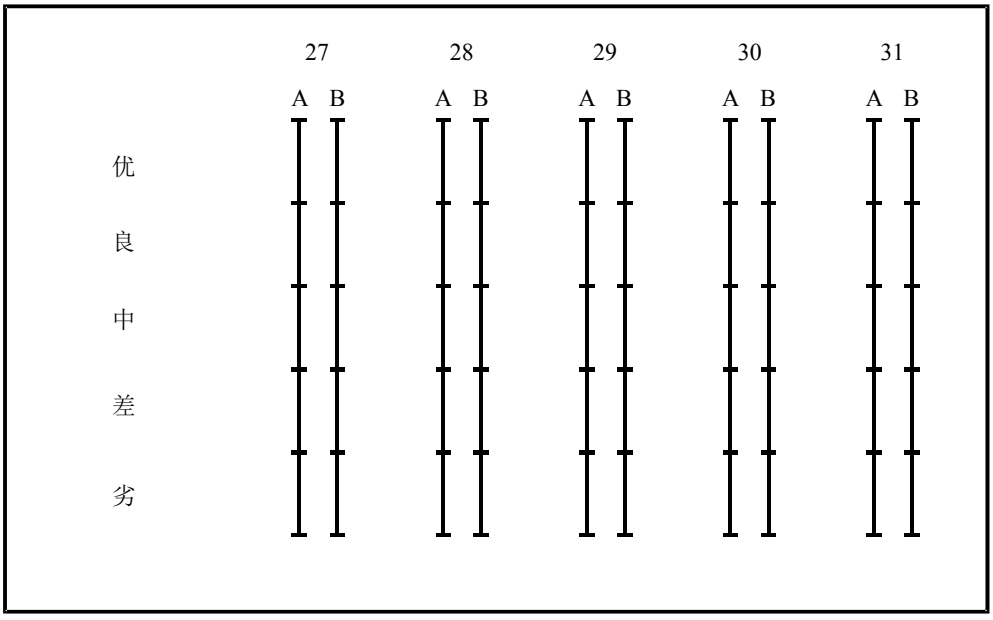
T3 = 10 s 测试序列B

T4 = 5-11 s 中灰度场

BT.0500-05

图6

采用连续标尺的质量评分表的一部分*



* 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。

BT.0500-06

5.5 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（见附件1的附录1）。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。

5.6 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。

6 评价的替代方法

在合适的环境中，应采用单激励法和激励比较法。

6.1 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。

6.1.1 总体布置

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

6.1.2 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）。

6.1.3 测试阶段

测试阶段由一系列评价实验组成。这些评价实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（单激励（SS））和II（多次重复的单激励（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。

   典型的评价实验由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，在监视器上公布一条消息（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；实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。

6.1.4 单激励法的种类

一般而言，在电视评价中采用了三种单激励法。

6.1.4.1 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3给出了ITU-R的量表。在运行监测中，有时也用到半级。在特殊情况下也使用了评价文字的清晰程度、易读性和图像实用性的量表。

表3

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。

6.1.4.2 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法在感受性和稳定性方面具有明显的优势。

6.1.4.3 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”）。

6.1.4.4 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。

6.2 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。

6.2.1 总体布置

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

6.2.2 测试素材的选择

按照与单激励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。

6.2.3 测试阶段

评价实验将使用一个监视器或两个匹配良好的监视器，并且一般像单激励情况那样进行。如果使用一个监视器，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。如果使用两个监视器，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。

6.2.4 激励比较法的种类

在电视评价中采用了三种激励比较法。

6.2.4.1 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4示出了ITU-R的比较量表。

表4

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、条件的等级或主要趋势、各条件之间的“距离”等）。

6.2.4.2 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。

6.2.4.3 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。

6.3 单激励连续质量评价（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，已经开发出了下述新的SSCQE技术并进行了测试。

6.3.1 总体质量的连续评价

6.3.1.1 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 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。

6.3.1.2 测试协议的一般形式

应向被试提供下述格式的测试阶段：

– *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 min；
– *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全部节目段(PS)和质量参数(QP)，但不必含有全部的PS/QP组合；每个测试阶段的长度应在30 min和60 min之间；
– *测试演示 (TP)*：一个测试演示代表某次测试的总体性能。一个测试演示可以划分为若干测试阶段(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，测试演示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。

最简单的测试格式是使用单一的节目段和单一的质量参数。

6.3.1.3 观看参数

观看条件应为ITU-R BT.500建议书、ITU-R BT.1128建议书、ITU-R BT.1129建议书和ITU-R BT.710建议书中目前规定的那些。

6.3.1.4 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5.4节所述的连续质量量表是相互对应的。

6.3.1.5 观察者

应聘用至少15位非专家被试，且具备目前在第2.5节中推荐的条件。

6.3.1.6 观察者须知

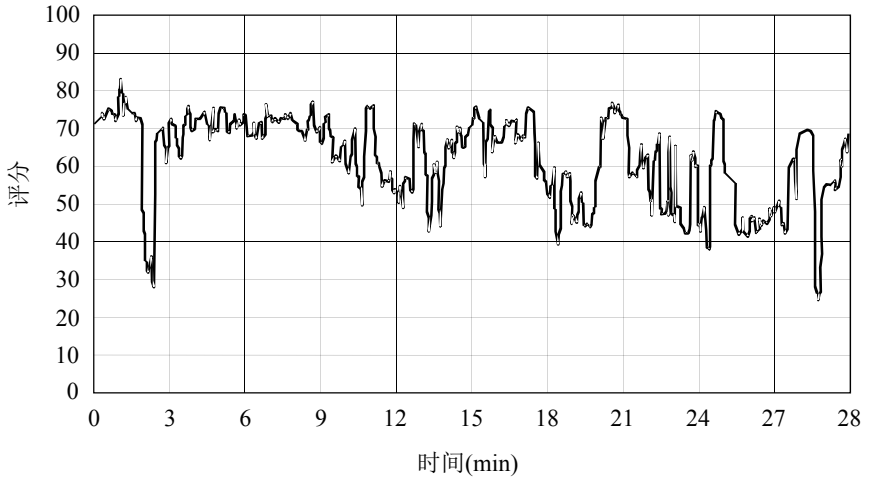
对于服务质量评价（带有伴音）的情况，应告知观察者考虑总体质量，而不只是视频的质量。

6.3.1.7 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7中的示例）。

图7

测试条件：Codex X/节目段：Z



BT.0500-07

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8中的示例）。

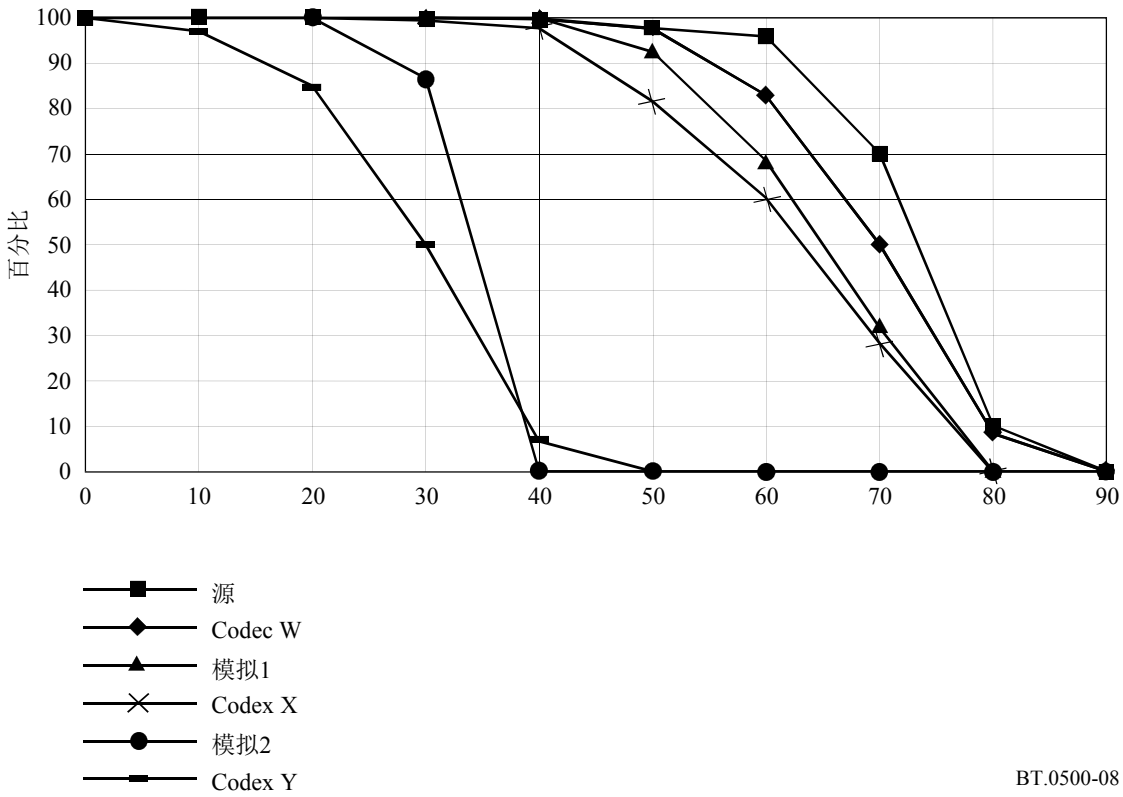
6.3.2 连续质量评价结果的校准和单一质量评分的导出

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，在单激励连续质量评价 (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样本使用原有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价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8

节目段Z的评分序列的平均评分



6.4 同时双激励连续评价（SDSCE）法

ITU-R之所以提出连续评价，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。数字扰动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。

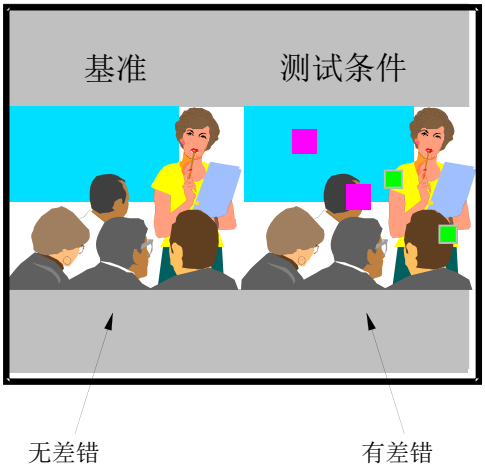
有鉴于此，制定了下述新的SDSCE技术并进行了测试。

6.4.1 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序列格式（SIF）或更短，则这两个序列可以并排在监视器上显示，不然就用两个对齐的监视器（见图9）。

图9

显示格式示例



BT.0500-09

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，要让被试知道那个序列是基准，并请他们在观看序列期间给出评分意见。

6.4.2 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。

6.4.3 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。每个视频段(VS)应按照至少一个测试条件处理。另外，应在测试条件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视频段(VS)/测试条件(TC)组成，中间没有间隔，按随即顺序排列。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视频段(VS)/测试条件(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。所有评分段采用互不重叠的成组的20次连续评分（相当于10 s）获得。

6.4.4 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测试演示（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。

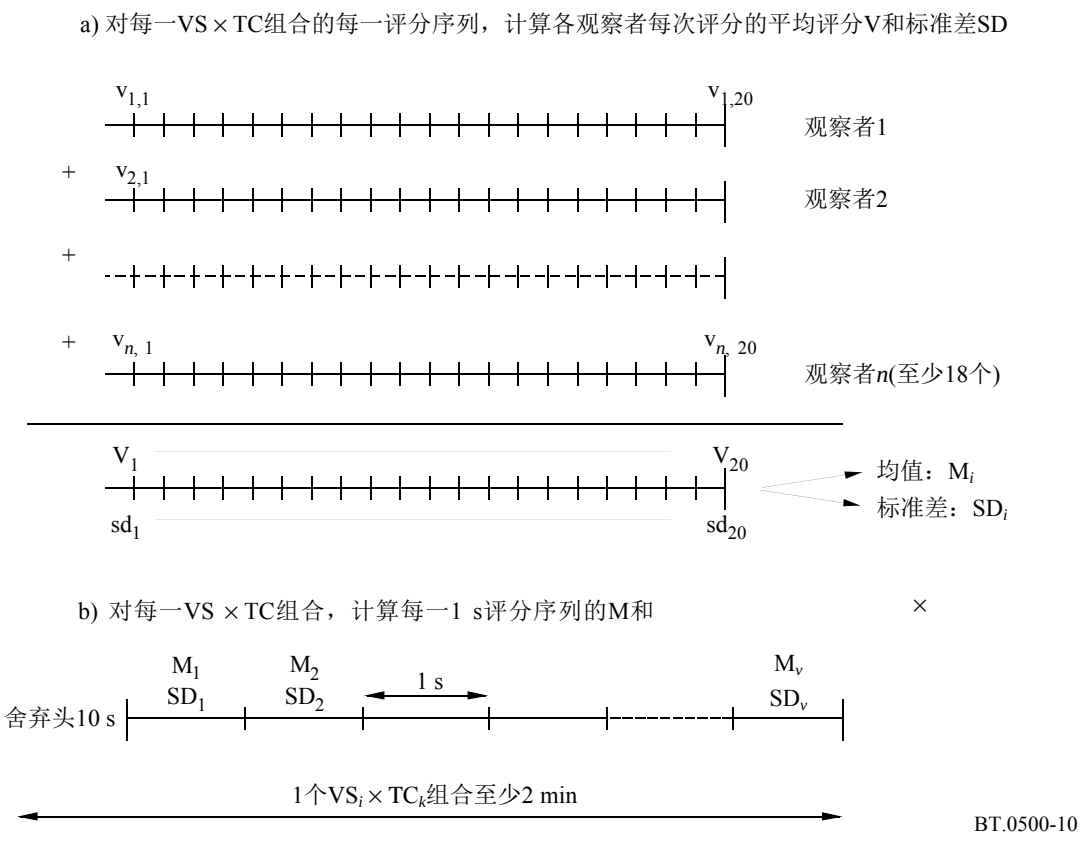
– 计算出每一SOV的均值和标准差，如图10所示。这一步的结果可用一幅时间图表示，见图11。

– 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要弃用。

– 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12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10

数据处理



6.4.5 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，可以采用与附件2第2.3.2节中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11

原始时间图

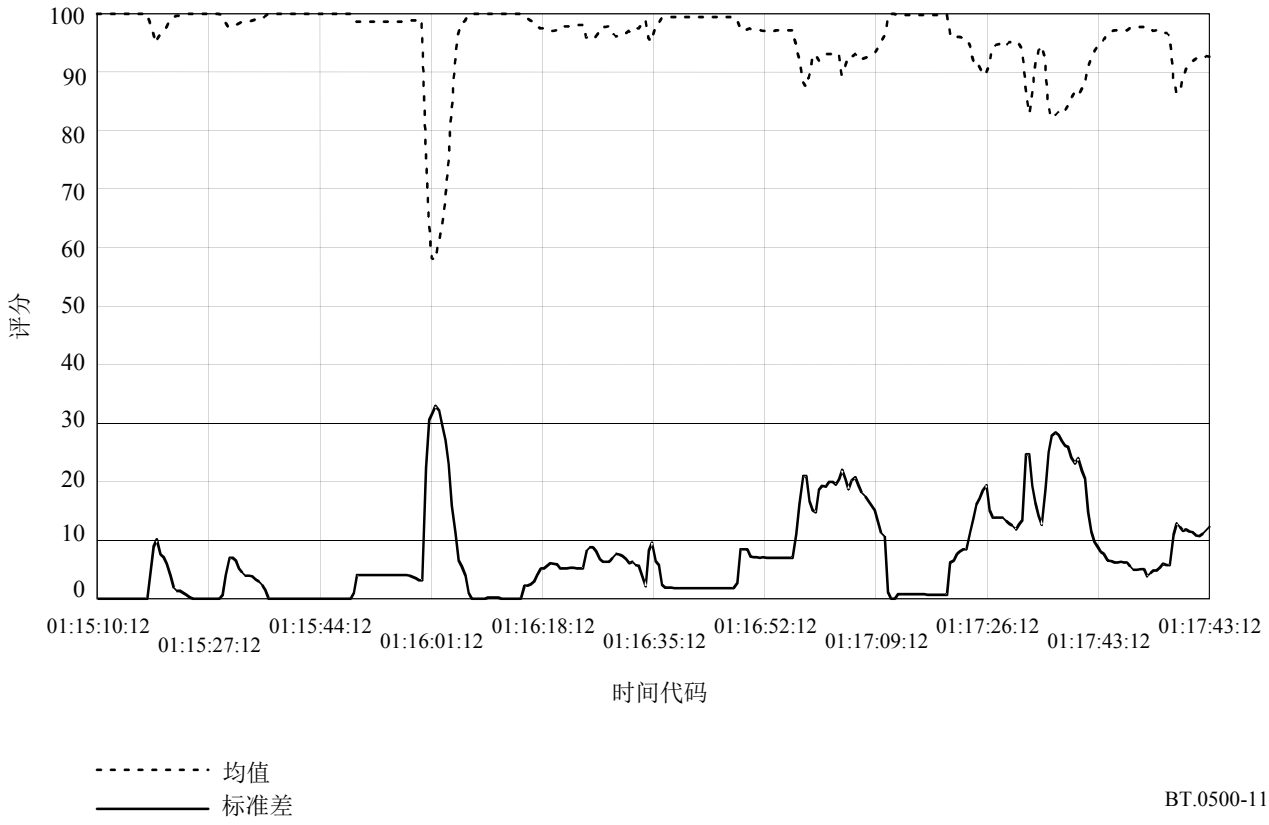
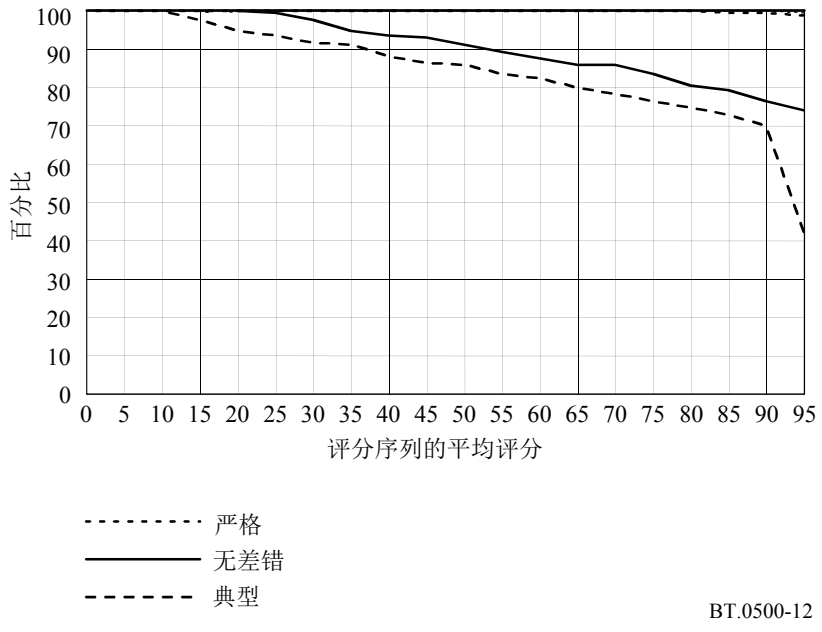


图12

在考虑置信区间的同时从统计分布计算整体讨厌特性



6.5 备注

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，比如要么使用几种方法，要么使用多维方法。

附件1的
附录1

图像内容降质特性

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。

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3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人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14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15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13所示形式的图像内容降质特性。

图13

图像内容降质特性示例的图形表示

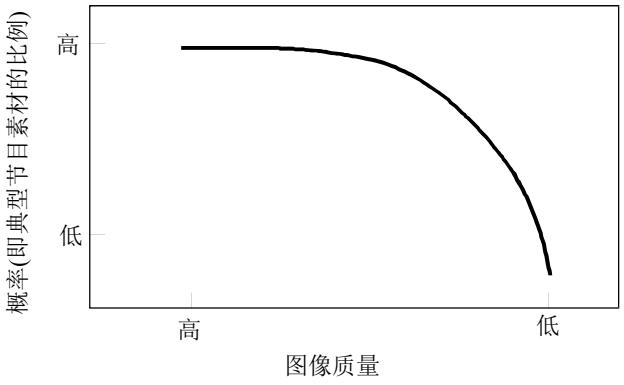


图14

不同临界性水平的素材出现的概率

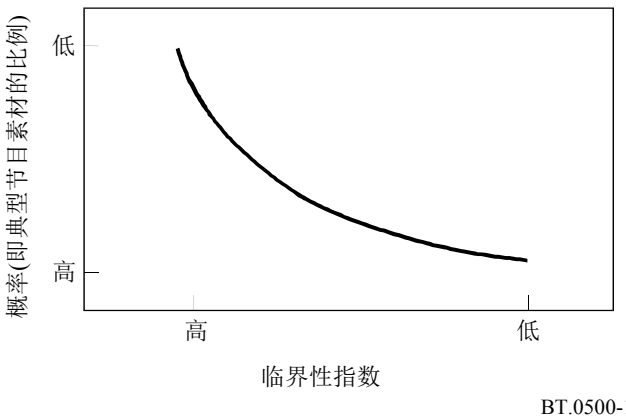
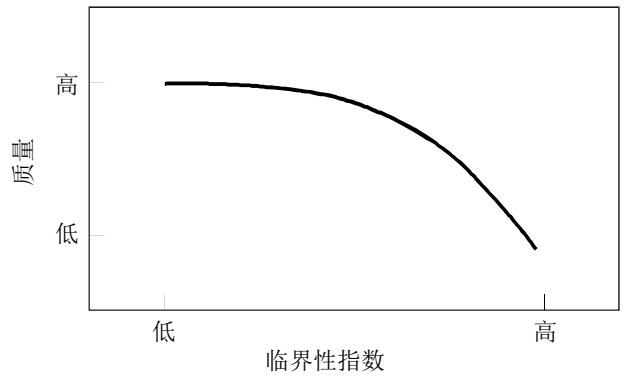


图15

将质量与节目素材的临界性关联起来的函数示例



BT.0500-15

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。

ITU-R BT.1129建议书的附件1介绍了一个节目内容降质特性的例子。

附件1的
附录2

确定节目内容和传输条件的
复合降质特性的方法

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。

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/场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格式的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1的附录1）。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。

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。

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（见附件1的第5节）。此处基准图像是某种适当格式（例如分辨率、帧速率和图像宽高比都适合待测系统的格式）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，对于较低的严格水平，抽样较稀，对于较高的严格水平则抽样较密。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。

附件1的
附录3

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，对于，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。

附件2

结果的分析和表示

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（附件1第4、第5和第6节）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。

2 分析的常用方法

按照附件1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。

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$\bar{u}_k$。

2.2 置信区间的计算

2.2.1 原始（未补偿或未近似）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：

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

其中：

$$\delta_{jkr} = 1{,}96\frac{S_{jkr}}{\sqrt{N}} \tag{2}$$

每一演示的标准差S_{jkr}由下式给出：

$$S_{jkr} = \sqrt{\sum_{i=1}^{N}\frac{(\bar{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \tag{3}$$

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。

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。

2.3 观察者的筛选

2.3.1 用于DSIS、DSCQS和替代方法的筛选，SSCQE法除外

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1）。

注1 – 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：

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \qquad \text{其中} \qquad m_x = \frac{\sum_{i=1}^{N} (u_{ijkr} - \bar{u}_{ijkr})^x}{N} \tag{4}$$

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）。

2.3.2 用于SSCQE法的筛选

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 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：

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

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$2 \le \beta_{2jklr} \le 4$，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 $u_{njklr} \geq \bar{u}_{jklr} + \sqrt{20}\ S_{jklr}$ 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17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，检测仍以本建议书附件2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：

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

同样，每一观察者每一测试配置的平均评分规定如下：

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

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：

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

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：

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

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。

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。

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：

$$p = (\bar{u} - u_{min}) / (u_{max} - u_{min}) \tag{5}$$

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：

$$p = 1/[1 + \exp(D - D_M) G] \tag{6}$$

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：

$$I = (1/p - 1) \tag{7}$$

D_M和G的值可从经过下述变换后的实验数据中导出：

$$I = \exp(D - D_M) G \tag{8}$$

对I采用对数尺度后，可由上式得出一种线性关系：

$$\log_e I = (D - D_M) G \tag{9}$$

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：

$$S = \frac{D_M - D}{\log_e I} = \frac{1}{G} \tag{10}$$

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。

3.2 用非对称函数逼近

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 (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(ms)来衡量失真参数，则关系式（8)必须用下式替代：

$$I = (d / d_M)^{1/G} \tag{11}$$

关系式(6)因此变为：

$$p = 1/\left[1 + (d / d_M)^{1/G}\right] \tag{12}$$

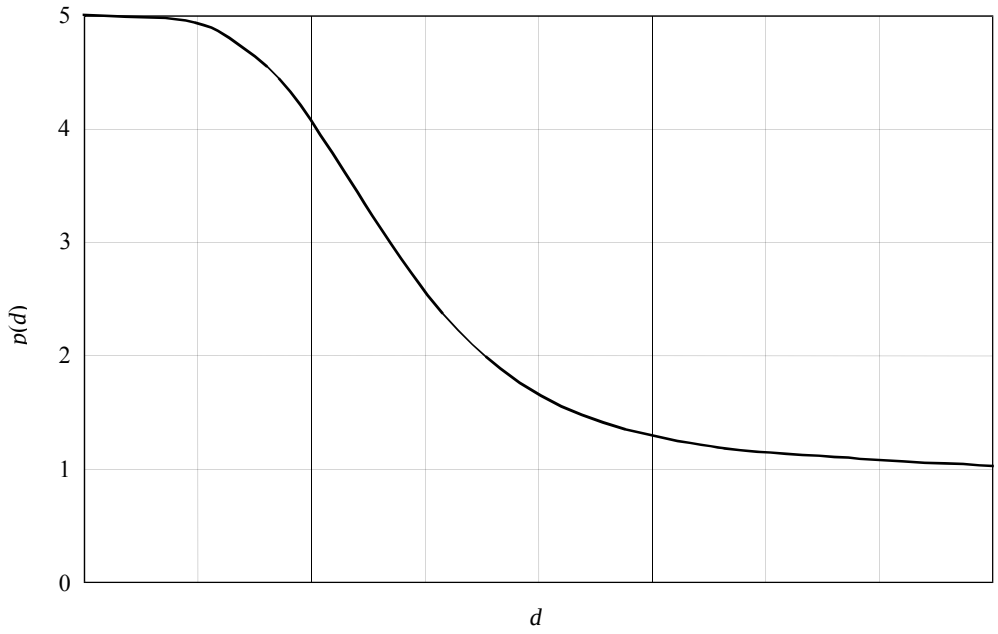
该函数以非对称的方式逼近了逻辑斯谛函数。

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16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16

非对称逼近



BT.0500-16

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5）。

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5

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。

3.4 信度性能的图形合并

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3个系列对参数的估值是分别进行。得到的3个函数则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17）。这样就得到了95%连续置信区域的估值。

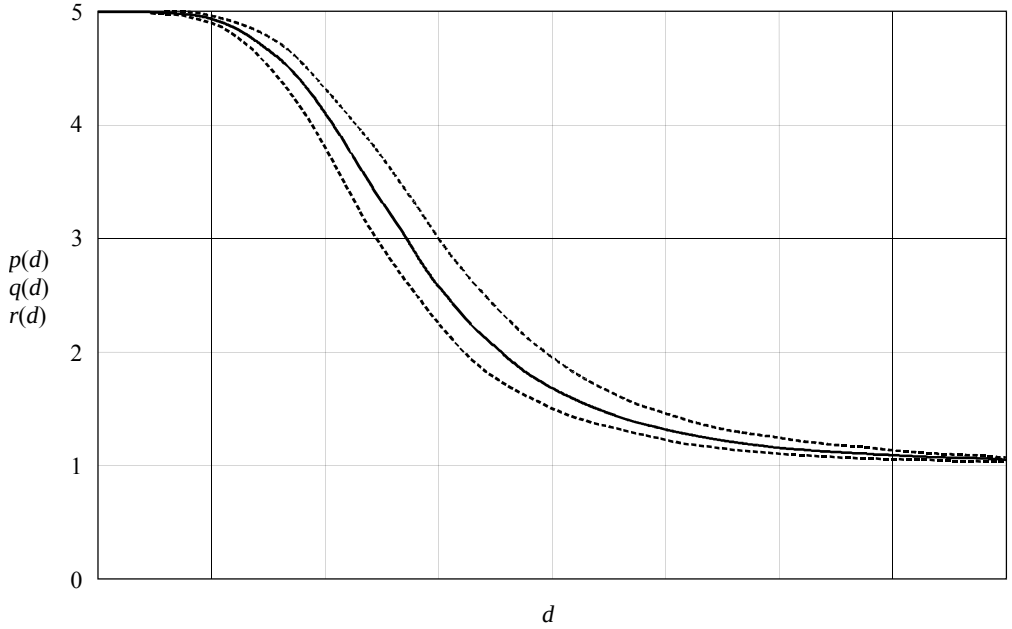
对于4.5的等级（用于该方法的可见性门限），也因此能够从图上直接读出估计的95%置信区间，可用于确定容差范围。

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17

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.0500-17

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程序做了说明。

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3

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，分析的结果。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（通常严格但并不过分严格），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ITU-R BT.500建议书中所述的评价方法。

该格式由表6和表7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6

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↵ 监视器尺寸 = 整数↵ 监视器制造商和型号 = 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*↵ 结果(*j*).文档名(s) = 字符串.DAT↵ .... 结果(*j*).名称 = 字符串↵ 结果(*j*).实验室 = 字符串↵ 结果(*j*).观察者的数目 = *1* ≤ 整数≤ *N*↵ 结果(*j*).训练 = “是”或“否”↵ [结果(*j*).阶段(*i*).观察者] ↵ O(*k*).名 = 字符串↵ O(*k*).姓 = 字符串↵ O(*k*).性别 = “男”或“女”↵ O(*k*).年龄 = 整数↵ O(*k*).职业 = 字符串↵ O(*k*).距离 = 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 测试执行实验室的标识 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，每次测试执行都会得出这样的文档。

表7

结果.DAT原始数据文本文档的格式

文档名.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中名称为结果(*j*).文档名(s)(1)的不同DAT文档中。

(1) 见表6的注释(1)。