国 际 电 信 联 盟

ITU-R

国际电联无线电通信部门

# ITU-R BT.2095-1 建议书

(06/2017)

# 利用专家观看协议 (EVP) 主观评估视频质量

BT 系列

广播业务

(电视)


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 |

**注**：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
2018年，日内瓦

# ITU-R BT.2095-1 建议书

# 利用专家观看协议（EVP）主观评估视频质量

（2016-2017年）

## 范围

本建议书阐述利用专家观看协议主观评估移动图像视频质量的方法，其中包括从相关视频处理领域挑选出的若干专家组成的观看小组的参与。

## 关键词

电视、视频质量、主观评估、专家观看

国际电联无线电通信全会，

*考虑到*

*a)* 用于数字电视应用的源代码技术无论是效率还是视觉表现均在不断改善；

*b)* 视频编码技术的不断演进意味着人们对技术和视觉表现评估方法的需求与日俱增；

*c)* 新视频源编码技术的压缩效率和视觉表现需要新且更加高效的视觉评估和评级方法；

*d)* 现行ITU-R建议书阐述的评估方法对时间和人力资源要求都很高，且通常并未考虑到显示器技术演进以及最终用户的感受；

*e)* 与建立在使用非专家观看者基础之上的方法相比，近来专家观看协议中的新方法在时间和总成本方面均展现出更高的效率和更好的表现；

*f)* 如果专家观看协议的成果不能视作正式主观评估协议成果的替代品，那么可将其作为受测系统宝贵的初步性能指标；

*g)* 平板显示器领域技术的不断演进大幅改变了专家通常使用的观看条件；

*h)* ISO/IEC在评估新视频源编码技术方面已成功地使用了基于专家观看的新协议，

*建议*

**1** 评估新数字视频编码技术时，应考虑使用专家观看协议，见附件1；

**2** 专家观看协议的实施应使用附件1所述专业平板显示器和实验室设置。

注1 – 附件2（资料性）展示了使用*考虑h)*提及的专家观看协议进行的主观实验的结果。

# 附件1

# 用于评估视频材料质量的专家观看协议

## 1 实验室设置

### 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。

### 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ITU-R BT.2022建议书设计的观看距离，或按关键观看条件设定的更短观看距离。

### 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刺激呈现。

## 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。

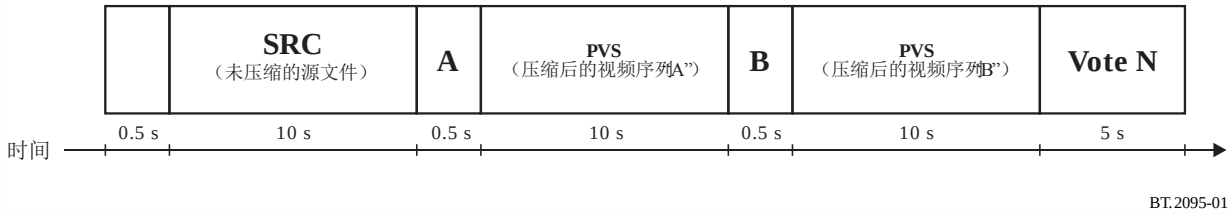
## 3 基本测试单元

针对每组需要评估的编码条件，向专家呈现的材料应组织在一起，形成一个基本测试单元（BTC），见图1。

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1

**专家观看协议中基本测试单元的时间轴**




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。

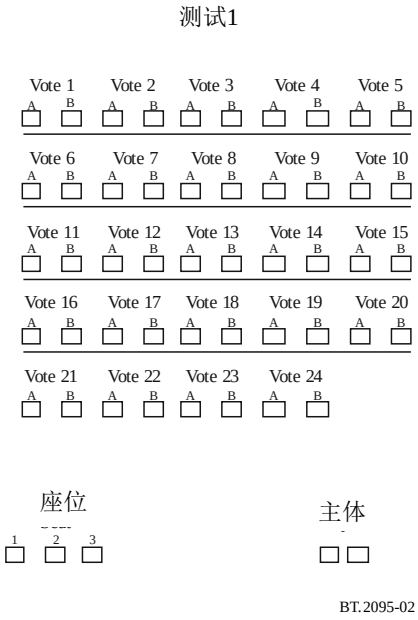
“Vote”消息后应紧跟一个数字，为得分表同步提供帮助。

### 3.1 得分表和评级

如图1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

**24-BTC专家观看测试的得分表示例**




评分使用11级数字标尺，得分从10（难以察觉的损害）至0（十分令人厌恶的损害）。表1提供了有关11级数字标尺含义的指导。

表1

**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。

### 3.2 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”。

### 3.3 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材料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。

## 4 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。

最好使用线性皮尔森相关性，对观看者进行“后筛选”。

“相关性”函数的应用应考虑到各主体与平均得分（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）。

## 5 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）。

## 6 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-学生测试。

# 附件2
# （资料性）

# 在有大量专家评估员出现的情况下<br>专家观看协议的应用及其表现

该资料性附件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ITU-R BT.2095建议书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ITU-R BT.2095建议书条款的固有可靠性等级。

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3（针对UHD的测试）和图4（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3和图4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3和4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，此处展示的专家观看协议实验描述了专家观看协议的良好表现，验证了ITU-R BT.2095建议书中提到的内容，即：在更多的观看者可以参加测试且正式主观评估完成后，即使专家观看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3

**以评估员数量为函数的UHD实验评级**

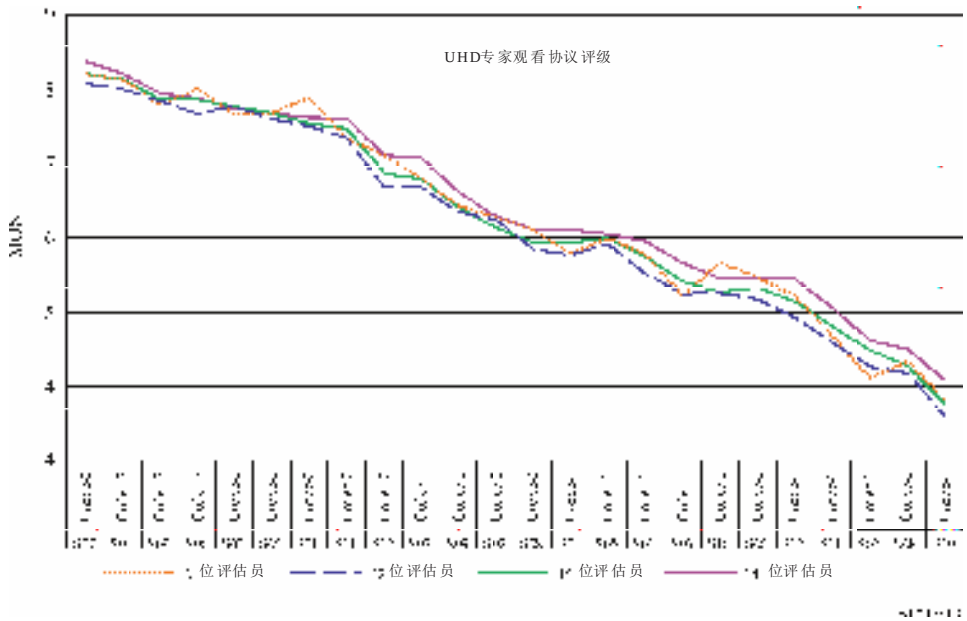


图4

**以评估员数量为函数的HD实验评级**