国　　际　　电　　信　　联　　盟

ITU-R

国际电联无线电通信部门

ITU-R BS.2132-0 建议书
(10/2019)

在无给定参考的情况下使用多重刺激对音频系统听觉差异进行主观质量评估的方法

BS系列
广播业务(声音)

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/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
2020年，日内瓦

ITU-R BS.2132-0 建议书

在无给定参考的情况下使用多重刺激对音频系统听觉差异进行主观质量评估的方法

（2019年）

范围

本建议书描述了一个在无给定参考情况下使用多重刺激对音频系统听觉差异进行主观质量评估的方法。该方法映射了ITU-R BS.1534建议书规定的带有隐藏参考和锚点的多刺激测试（MUSHRA）方法的很多方面，但与ITU-R BS.1534建议书不同的是，本建议书拓展了对于系统的评估，从而将已知隐藏参考和锚点不可用的情况包含在内。

关键词

聆听测试、音频质量、高级音响系统、主观评估、感知评估

国际电联无线电通信全会，

考虑到

a) ITU-R和国际电联电信标准化部门（ITU-T）建议书中存在很多用于对音频、视频和语音系统的主观质量进行评估的主观测试方法；

b) 使用标准化的主观测试方法对测试数据的交换、兼容性和纠正评估十分重要；

c) 寻求使用标准化的测试方法对高级音响系统进行评估；

d) 在一些应用中，没有可用的或合适的参考信号，这样就不能相对于已知信号对音响系统的主观质量进行评估；相反，必须在没有参考的情况下进行评估；

e) 节目制作过程需要使用技术系统来制作音频信号和表达创作意图，在这种情况下，可能出现目标参考信号或系统行为不可用的情况；

f) ITU-R BS.2051建议书所描述的高级音响系统的引进，为生产中的创造性表达提供了新的工具，需要新的主观评估方法，包括将感知属性与总体感知音频质量关联起来的方法，

建议

当合适的参考信号或系统参考不可用时，应将本建议书附件1中给出的测试和评估程序用于对音频系统听觉差异的主观评估。

附件1

1 引言

主观听力测试是一个测量音频系统感知质量的可靠方法。在将系统与已知的未受损参考进行比较时，存在经过很好描述和证明的对广播环境中的音频质量进行评估的方法。ITU-R BS.1116建议书 – 包括多声道音频系统在内的音频系统中细小损伤的主观评估方法 – 旨在从一个给定的参考信号评价存在细小损伤的高质量的音频系统，而ITU-R BS.1534建议书 – 音频系统中级质量水平的主观评估方法 – 旨在在中级水平上对音频系统质量进行评估，适用于广播应用，但显然不同于参考信号。应注意，这两种方法的制定在很大程度上是由对低比特率音频编码系统的效果进行评估的需要而推动的。

在一些应用中，没有可用或合适的参考信号，因此系统的主观质量评估无法根据对参考的保真度来执行。ITU-R BS.1284建议书 – 主观评估声音质量的一般方法 – 仅描述了专门用于高质量音频范围或不给出音频质量的绝对评分的方法。

本建议书描述了在参考不可用的情况下，使用多重刺激对应用中的音频系统的听觉差异进行主观质量评估的方法。

本方法使用ITU-R BS.1534建议书中采用的多重刺激呈现方法作为对受测音响系统进行比较的基础。评价者被要求从以下方面为每个受测系统评分：

1 总体主观音响质量。

2 属性评分（预先定义的一组所选属性）。

使用ITU-R BS.1534建议书中定义的连续质量量表进行总体主观音响质量评分。

由于在ITU-R BS.1116、ITU-R BS.1284、ITU-T P.835和ITU-T P.806建议书中描述的属性评分阶段是可选的，所以为每个实验预先选择了许多与音响质量相关属性（最好是从现有的和经过验证的词汇表中选择）。评估者在100点线性量表上对这些属性评分。

通过对这两个数据类型进行统计分析，可推导出：

– 每个音响系统的相对总体主观音质。

– 每个音响系统选择属性的感知特性（可选）。

– 不同感知特性对于受测系统感知质量的相对权重（可选）。

2 术语

总体主观质量 – 捕捉被评估的音响质量的所有方面的单一特性，即ITU-R BS.1284建议书规定的“基本音频质量”。这里使用“总体主观质量”一词以避免与ITU-R BS.1116建议书定义的“基本音频质量”产生潜在混淆。

“基本音频质量”和“总体主观质量”之间的主要区别是质量评估过程的不同，而非这两个全局质量术语归纳的不同感知属性。对于“基本音频质量”，评估是通过对两个或更多刺激之间互相比较来进行的，其中的一个刺激被定义为参考（例如，判断音频项目的压缩版本与未压缩的原始版本之间的量化差异）。相比之下，“总体主观质量”为仅与内部参考进行比较的量化判断，即，测听者在无给定外部参考的情况下的预期（例如，不同的双声道再现）。

控制变量 – 在实验中被控制的变量，虑及结构化和控制的实验设计。也被称为自变量，因为变量的值不受其他实验变量的影响。

反应变量 – 评价者向这些变量提供响应，在一个给定量表上对感知到的刺激评分。也被称为因变量，因为变量的值受到其他实验变量（即自变量/控制变量）的影响。

条件 – 在评价中使用的一组控制变量的值。

试验 – 评估过程中的一个环节，在该环节中，被评价的系统（或其子集）在一个给定条件下被呈现，评价者根据响应变量给出评分。

复现 – 重复的测试条件，在该条件下，每个评价者在控制变量值相同的情况下对相同的响应变量打分。

描述性 – 以客观和非评判的方式进行描述。

属性 – 可使用评分量表评估的可感知质量的规定特性。感知的总体主观质量可考虑多个属性的评价。

节目项 – 在评价中与其他控制变量结合使用的一段音频材料。

刺激 – 在一组控制变量值下，通过一个系统对节目项的单独呈现。

词汇表 – 一组描述性感知属性，有明确的属性名称、定义和评分量表。

3 一般做法

很多不同种类的研究策略被用于在科学关注范围内收集可靠信息。在对音频系统质量的主观评价中，应使用正式的实验方法，以确保结果及其解读的牢靠性。从主观实验中收集牢靠数据要求对实验条件进行控制和操纵，这样，当数据以控制的方式被呈现给评价者，实验将产生高质量数据。需要对实验进行精心设计和规划，以确保最大程度减少可造成实验的无用噪音增加的不可控制因素。例如，如果实验中的所有条件以固定的相同序列呈现给所有评价者，会产生无法从数据及其解读中移除的演示序列偏差效应。就这点而言，更好的做法是确保条件或者以随机的顺序呈现给每一位评价者，或者使用平衡设计来使任何潜在序列偏差效应最小化。第5节详细介绍了建议的测试程序。

为确保这类实验中的最佳数据质量，将以下细节纳入考虑非常重要。它们是本建议书的组成部分。

应招募有经验的评价者，因为他们通常会生成高质量数据。按照ITU-R BS.1534建议书的描述筛选有经验的评价者。为了能够对受测系统的性能进行有意义的评测，同样重要的是，使用关键节目项和选择能够最好地区分受测系统并有助于体验的感知质量的感知属性。

对于实验设计，实验者需要精心规划实验的持续时间。纳入足够数量和范围的关键测试项目可生成对被研究的音响系统的性能更具概括性的观点。比较多个音响系统亦是可取的。这些目标很常见，但除了带来使评价者负担过重的风险之外，还带来了时间和成本损失。本建议书附件1的参考性附录1中纳入了有助于资源规划（包括对测试时长的估计）的方法。

为了使实验能够在晚些时候或者在另一地点被如实继续或复现，测试报告中应不仅包括结果，也应包括全部实验详情。ITU-R BS.1116和ITU-R BS.1534建议书阐述了报告指导原则。

4 实验参数

本节中定义了关键实验参数，从而使对照实验的结构化设计成为可能。这些参数被分为两个主要类别，即控制的实验变量和响应变量。

4.1 控制的实验变量

控制的实验变量（或者称自变量）被用于定义实验中受到控制的参数。这些实验虑及可产生全面统计分析的结构化和受控的实验设计。通常，控制变量是为诸如被评估的系统、节目材料、评价者和复现等参数定义的。对于每个控制变量，等级数量由实验者定义。例如，在一个测试中包含10个不同的节目项相当于有10个节目项变量等级。等级的数目随后被用于设计实验和后续的数据分析。

4.1.1 被评估的系统

在这些实验中，实验员关注的是对被评估的技术或系统的感知质量的研究。

根据米勒定律（Miller, G.A.,1956年），被评价的系统的数量应在5-9个的范围内，以最大程度减少评价者评分的误差。如果待评价的音响系统的数量超过9个，可参见5.1小节的进一步指导。

如果可以，实验员应将一个或者更多质量容易理解的系统纳入在内，使受测系统的结果可以在情境中被考虑。

4.1.2 节目素材

对测试素材的选择应遵循ITU-R BS.1116和ITU-R BS.1534建议书概述的流程。虽然没有普遍适用的节目素材可用来评价所有条件下的所有系统，但必须明确寻找重要节目素材。对优质材料的搜索通常耗费时间；但是，除非为每个系统找到关键素材，否则实验将无法揭示系统之间的差别，并且将是无结果的。

4.1.3 评价者

建议应该使用有经验的评价者，以保证所采集测试数据的质量。这些评价者应具备以批判性的方式收听声音的经验。这类评价者将比无经验的测听者更快地给出一个可靠的结果。同样重要的是要注意，大多数无经验的评价者在频繁收听之后趋于变得对各种类型的伪像更加敏感。一个有经验的评价者因其完成一个聆听测试的能力而被选定。根据对评估的重复，将按照评价者在一个测试中的可靠性和辨识技巧来决定这个能力是否合格并被量化，如下所定义：

1 辨识：对感觉到测试项之间差别能力的一个度量。

2 可靠性：对重复对相同测试项进行评分的接近程度的一个度量。

应该仅仅将对任何给定测试被划分为有经验的评价者的那些评价者计入到最终的数据分析中（见ITU-R BS.1116建议书）。这些是基于至少一个由每个评价者进行的重复评分，并且允许在一个实验内对评价者经验进行合格确认和量化。这些方法应该作为在一个先导实验内对评价者的预筛选，或者更可取的是，既作为预筛选又作为后筛选（使用主测试的响应）。一个先导实验通常是与一个主实验相关的小规模测试，包含要在主实验中评估的一个代表性测试样本集。出于对测听者专业知识进行评价的目的，先导实验应该包含测试刺激的一个相关子集，代表在实际主实验期间要被评估的整个刺激和伪像范围。

名义上，大约要招纳20名（多多益善）有经验的评价者。

4.1.4 重复

评估收集的数据质量和评价者绩效的一个方法是要求每名评价者对一个条件做出重复判断。建议重复至少两个样本，以在不过分增加实验规模的条件下检查评价者绩效（见上一节）。

4.1.5 额外的控制实验变量

可能出现实验要求额外的控制实验变量的情况。这是非常正常并且可以接受的，可以采用与第4.1节定义的类似的结构化的方式来增加变量。实验者应知晓，增加控制变量的数量将增加实验规模和时长。

4.1.6 响应变量

对于每个条件，评价者被要求使用响应变量来给出他们的评价。使用两个不同类型的响应变量，包含相关维度的描述如下：

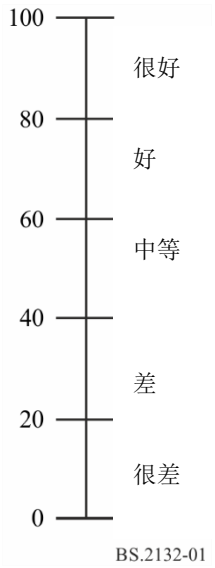
– 总体主观质量（每系统）。

– 属性评分（可选，每系统）。

4.1.7 总体主观音响质量

图1

连续质量量表

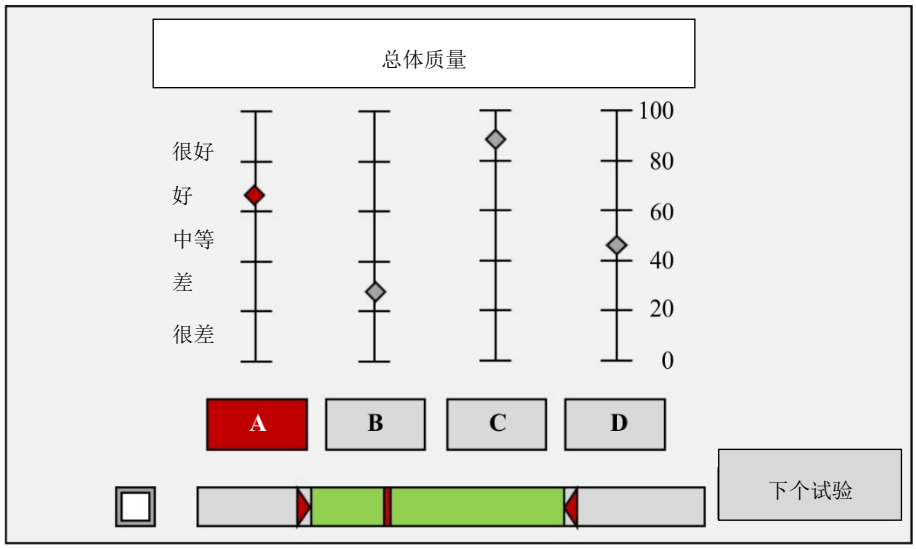


首先要求评价者使用连续质量量表（CQS）[1]为总体主观音响质量评分。评价者被要求评估每个呈现的总体主观音响质量，并在CQS上做出评分。CQS包含一个100点直线量表（通常>10厘米）。它被分成5个相等的间隔，附有如图1中所示的形容词。在一次试验中由多个系统呈现一个共同的节目项，每个系统都有自己的评分量表，如图2所示。

[1] 此量表亦被用于评价图像质量（ITU-R BT.500建议书 – 电视图像质量的主观评价方法和ITU-R BS.1534建议书 – 音频系统中级质量水平的主观评价方法）。

图2

总体主观音响质量测试图形用户界面示例



4.1.8 属性评分

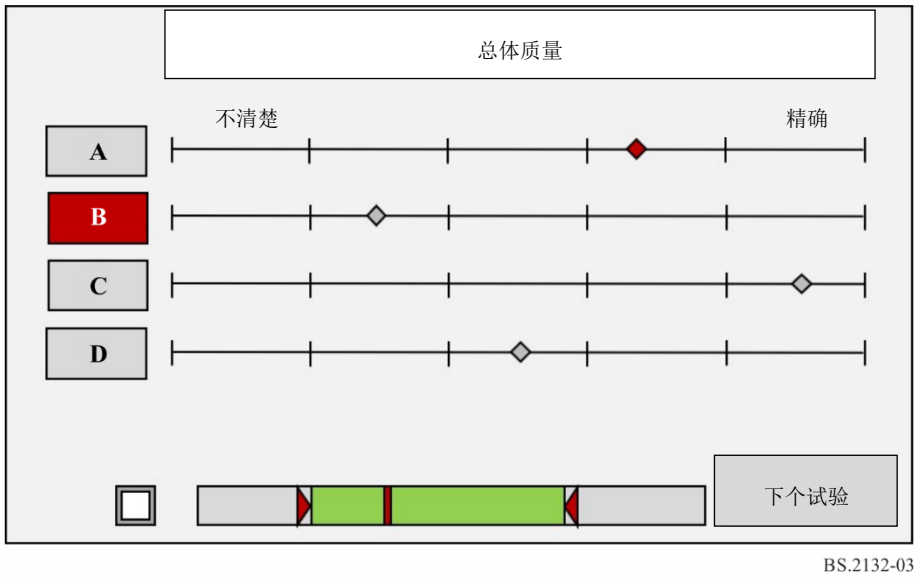
在总体主观音响质量评分之后，评价者被额外要求针对该系统–节目项组合的每个描述属性评分。仍然是在一次试验中用多个系统呈现一个节目项，每个系统都有自己的评分量表。

属性量表通常为100点连续量表。

图3展示了属性评分界面的一个示例。本示例界面展示的属性及其定义为“定位准确性”。

图3

属性评分的图形用户界面示例示例（定位准确性）



5 测试协议

5.1 实验设计

应对实验进行精心设计，以确保生成高质量的数据，同时最大程度减少随机或不受控制的影响的来源。设计还应有助于评估实验规模和时长，并为数据分析提供架构。设计包含两个重要方面，即处理设计和刺激划分设计，描述如下。

5.1.1 处理设计

处理设计规定了在实验中将使用哪些控制变量和排除（评价者变量）。

对于中等规模的实验，建议采用评估所有可能的控制变量等级组合的完全析因设计。对于完全析因实验，条件的数量通过将每个自变量等级数量相乘得到。

5.1.2 刺激划分设计

刺激划分设计定义了条件是如何呈现给每名评价者的。

建议采用“被试内”设计，通过该设计，所有条件都被演示给每名评价者。演示的顺序是受到控制的 – 通常通过随机化 – 以最大程度减少系统性偏差效应的影响。完全平衡的演示顺序是可取的。

5.1.3 细分大型实验

在某些情况下，采用完全析因被试内设计，实验的规模可能变得过大，效率低下。这些情况可能在需要评价多个音响系统或每个评价者的总体测试时长长得不合理的时候发生。

在这些情况下，可考虑更加高级的实验设计。

本节仅描述了可被考虑来适应这些情况的方法。不过，有兴趣的实验者应查阅有关实验设计的文献以获得最佳指导。

以下提供了处理较大实验的两种不同解决方案作为示例。

5.1.3.1 裂区设计

第4.1.1小节建议被评估系统的最大数量限制在5-9个的范围内。在有很多音响系统需要被评估的情况下，可考虑裂区设计。例如，在需要评估14个音响系统的时候，总体评估可由两组试验构成，每组七个系统。为控制裂区呈现相关的任何偏差效应，对于每个试验的音频系统的划分应是随机的。但是，这种随机化不应影响为呈现给每名评价者或总体实验中呈现的条件的总数量。见5.1.1小节。

当使用裂区设计时，重要的是将区块因子纳入作为分析的一部分。理想情况下，区块不应成为统计显著因子。

5.1.3.2 被试间设计

如果每个评价者需要评价的条件的数量很多，可能导致不合理的大数量倾听进程，尤其是在每个评价者的总测试时长可能超过四小时的时候。

一个应对这一问题的方法是采用被试间（或组间）设计。这是一个通用方法，这种方法向不同评价者或评价者组呈现不同条件。减少向每个评价者（或评价者组）呈现的条件的数量的一个简单方法是为每个评价者（或评价者组）分配节目项的不同子集。必须小心谨慎地使用该方法，以确保呈现的条件的总体数量在每个评价者（或评价者组）之间是平衡的。

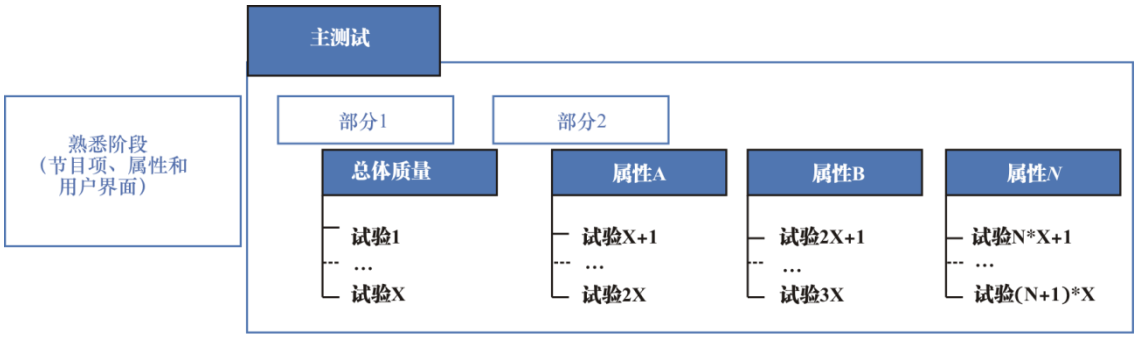
采用这一设计时，重要的是测试将分组因子作为分析的一部分。理想情况下，分组不应成为统计显著因子。

5.2 测试管理

测试应分成两个阶段实施，加上一个熟悉阶段，之后是正式的测试，属性在这一阶段被使用。不同元素被呈现给评价者的顺序见图4。图中假设需要评估N个属性，使用M个节目项。总共有*x*个试验用于总体主观音响质量评分，共有*y*个试验用于属性评分。

图4

总体测试结构的总体流程图，包括熟悉阶段和主测试

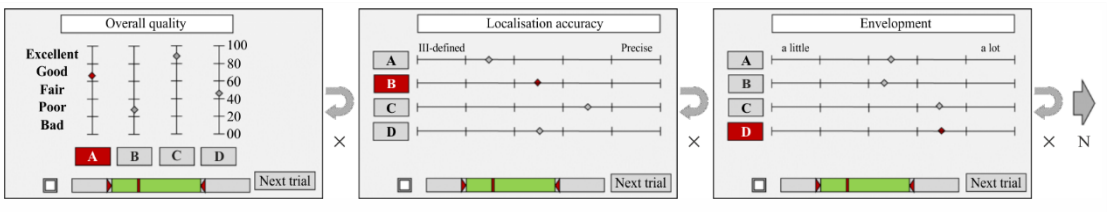


BS.2132–04

为确保高质量数据的收集，评价者应熟悉测试协议、用户界面、节目项，以及感知属性。每名评价者还必须被允许倾听测试刺激，对属性进行检查和试用用户界面。熟悉阶段可包括对包含仅用于熟悉目的的测试子集的刺激的自由盲听。图5显示了上述描述的首选流程。在一个试验中，评价者使用单独界面来评估每个属性的刺激。

图5

使用单独界面对单个刺激的评估过程



BS.2132-05

5.2.1 可选流程

图6描述了总体测试的可选流程。可为评价者呈现单个属性的单独界面，或者多个属性的合并界面。图7显示了图形用户界面（GUI）的示例。在该示例中，评价者使用每个属性的单独界面来对每个刺激进行评估。图8显示了一个GUI示例。在该示例中，评价者在合并界面中评估多个属性。

图6

总体测试结构的可选流程图，包括熟悉阶段和主测试

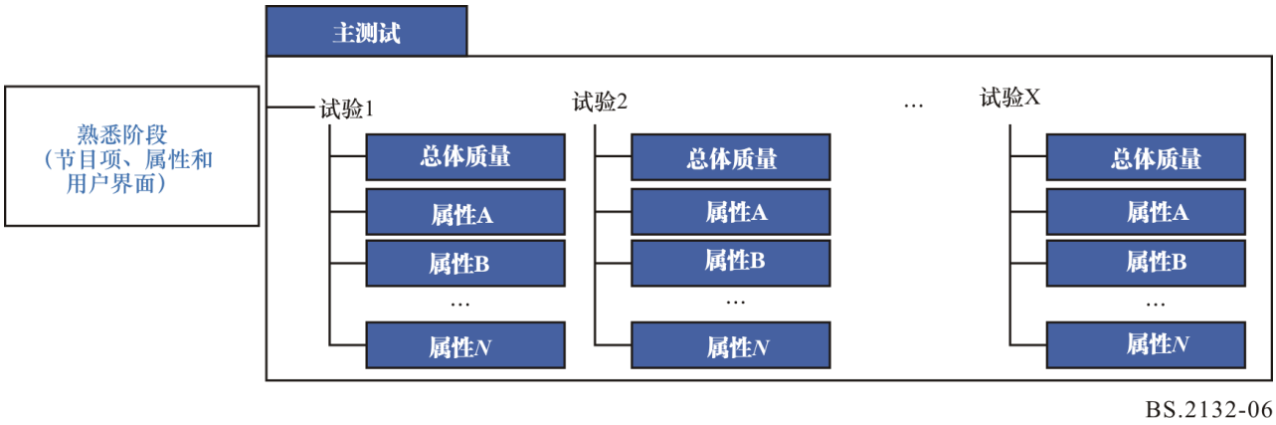


图7

使用单独界面对相同刺激的每个属性进行评估的过程

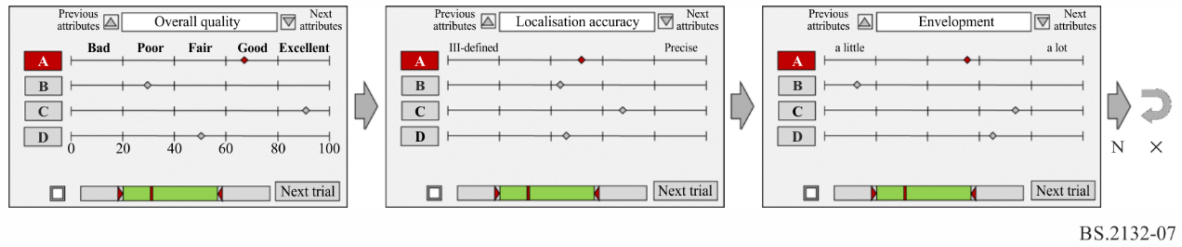
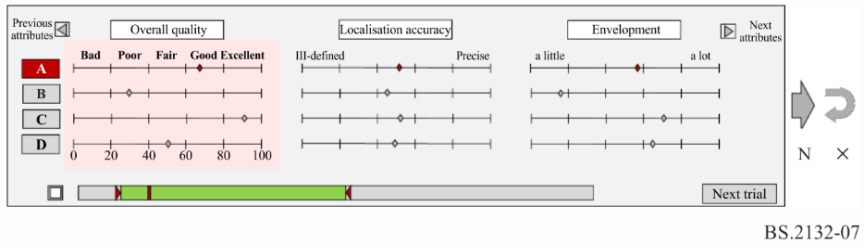


图8

使用合并界面对相同刺激的的多个属性进行评估的过程



5.3 评价者指令

为确保评价者完全知晓要执行的任务，应在实验之前为他们提供书面和口头的指令。指令应使评价者知晓将要执行的任务而不存在过度偏差，并向评价者介绍测试协议，响应变量（总体主观音响质量、属性定义）和如何在测试GUI使用这些变量。附录2提供了这些须知的示例。

6 测试环境

6.1 测听环境

应在符合ITU-R BS.1116建议书要求的测听室中进行测试。

6.2 再现设备

耳机或扬声器都可以用在测试中。但是，在一个测试进程中不允许二者都使用。所有评价者必须使用相同类型的换能器。应使用ITU-R BS.1116建议书规定的参考监控扬声器或耳机。

扬声器的设置和要求以及评价者的测听位置最好按照ITU-R BS.1116建议书的要求配置。

扬声器布局最好从ITU-R BS.2051建议书定义的布局中选择。

6.3 校准

为确保可重复和可再现的结果，慎重地设置测试使用的设备和使用的测试材料时十分重要。

项目的相对响度

不同项目的相对响度不得以任何方式与ITU-R BS.1770建议书规定的响度测量相关。短音频素材应被调整至需要的响度水平。应保留大声（强音）项和安静（弱音）项之间的差异（例如，可以是15分贝）。每个项目的相对响度都必须主观调整，最好是由一组技巧娴熟的评价者进行。为在再现电平中保持这种差异，充分测试不同项目十分重要。

刺激的相对响度

响度的细小差异往往会导致对于更大声的刺激的偏好。这些差异应在同一项目的不同刺激之间移除。应使用ITU-R BS.1770建议书规定的客观（而非主观）响度测量。如果无法使用客观度量，在素材被纳入测试之前，每个素材的响度必须由一组技巧娴熟的评价者主观调整。

项目的同步

相同节目项的不同处理版本的刺激之间的瞬时切换不应产生可感知的时间切换。参考ITU-R BS.1534建议书以了解关于刺激呈现的详情。

6.3.1 再现系统校准

对于使用扬声器进行的测试，最好使用ITU-R BS.775和ITU-R BS.2051建议书定义的扬声器布局之一。或者，应使用适用建议书中定义的表示法来描述用于测试的布局。

再现系统的电平应按照ITU-R BS.1116建议书的要求调整。

超低音音箱和低音管理系统的校准详情在本建议书范围之外。低音管理结果应为单个扬声器和超低音音箱组合的频率响应应为平坦的频率响应（在ITU-R BS.1116建议书规定的范围之内）。

从之前的测试序列已经注意到，单个测听者可能倾向于不同的绝对测听电平。这不是一个首选的选项，但并不总有可能防止测听者要求这种程度的灵活性。目前尚不清楚这是否会影响某些正在接受评估的伪像的可听性。因此，如果测听者确实调整了系统的增益，那么在测试结果中应注意到这一事实。

按照ITU-R BS.1116建议书的规定，耳机的立体声系统信道之间的延时差异不得超过20 μs，扬声器不得超过100 μs。

在系统带有附带图像的情况下，在测系统中参考监控耳机或扬声器总体延时不应超过ITU-R BS.775建议书中设定的限值

注 – 高级音响系统声音参数的测量可能比早期多声道音频系统的情况复杂很多。必须慎重选择测量麦克风及其测量时的朝向。见ITU-R BS.2419报告。ITU-R BS.1116建议书规定了电声要求和操作室响应特性的进一步信息。

7 统计分析

对评价者评分数据进行统计分析，以提供关于在测系统的主观质量的见解。计算平均评分，以提供性能指征，数据方差的估算被用于指示观察到的系统性能差异的可靠性。

在收集数据时，每个评价者提供在测系统的属性评分。使用不同节目项来测试这些系统。评价者使用预先定义的量表，对感知属性列表上的每个系统评分。对于每个节目项，每个评价者对完全相同的一套属性量表上的每个属性进行评分。评价者亦对每个系统节目项组合的总体主观质量进行评分。

对于每个节目项，评价者必须提供对每个系统的属性评分，以及总体主观音响质量评分。

ITU-R BS.1534建议书提供了总体主观音响质量数据和每个描述性属性数据的统计分析详情。

此外，使用本建议书获得的感官数据的分析产生与使用更经典的量化描述分析获得的感官数据分析类似的见解。此类分析包括对每个感官属性进行的方差分析（ANOVA），以及多变量分析（例如使用主成分分析，PCA）。

8 参考文献

[1] ITU-R BS.2051建议书 – 用于节目制作的高级音响系统

[2] ITU-R BS.775建议书 – 伴同或不伴同图像的多声道立体声音响系统

[3] ITU-R BS.645建议书 – 国际声音节目连接中使用的测试信号和计量方法

[4] ITU-R BS.1116建议书 – 音频系统中细小损伤的主观评价方法

[5] ITU-R BS.1534建议书 – 音频系统中级质量水平的主观评价方法

[6] ITU-R BS.1770建议书 – 测量音频节目响度和音频电平实际峰值的算法

[7] ITU-R BS.1864建议书 – 数字电视节目国际交换中响度的操作方法

[8] Miller, G.A. (1956), The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review. 63 (2): 81–9

[9] ITU-R BT.500-13建议书 – 电视图像质量的主观评估方法

[10] ITU-T P.835建议书 – 用于评价包括噪声抑制算法在内的语音通信系统的主观测试方法

[11] ITU-T P.806建议书 – 使用多重评分量表的主观质量测试方法

[12] ITU-R BS.2419报告 – 有关高级音响系统的校准水平和平衡的麦克风定向效应

附件1的
附录1
（资料性）

估算实验的测试时长的电子表格（Excel）工具

以下提供的Excel工具用于估算一个主观音频评估实验的时长，以便进行资源规划。表1显示了一个实验的示例。浅棕色栏为输入栏，浅绿色栏为输出栏。

Copy of
R15-WP6C-190715-T

表1

提供的excel表格的屏幕快照，以估算主观音频评估实验的测试时长

Full Factorial Design Calculator

v4

Input fields

Result fields

Controlled experimental variables

(Independent variables)

Variable	Label	Description	No. of levels	calculation no of levels	Degrees of Freedom (DOF)
1	SYSYEM	No. of systems under test	7	7	6
2	PROGRAMME	No. of progamme items	3	3	2
3	REPLICATE	No. of presentations or repetitions	1	1	0
4	ASSESSOR	No. of assessors	20	20	19
Total			31		27

Response variables

(Dependent variables)

Variable	Label	Description	No. of levels
Total		Overall subjective Quality	1
Descriptive		Envelopment, source width, etc.	6

Test parameters

Parameter	No.	Units	Comments
No. of controlled experimental conditions (total)	21		Excluding assessors
No. of test conditions (per replication)	21		Excluding assessors
Total no. of independent variables	420		Including assessors
No. of blocks	1		1 = within-subjects design >1 = between-subjects design
No. of PROGRAMME items per block	3		Must be integer ≥2
No. of ratings per condition	20		
Total no. of ratings per assessor	147		
Estimated average rating time per condition	20	s	
Estimated total duration per assessor	0,8	hrs	
Session duration	2	hrs	max. < 2 hrs incl. breaks
Estimates no. of sessions per assessor	1	Sessions	
Total no. of sessions	20	Sessions	
No. of data points per response variable	420		
Total no. of data points in experiment	2940		

附件1的
附录2
（资料性）

评价者须知示例

A2.1 通用须知

各位测听者：

欢迎参加测试。在测试中，你将听到并通过一系列节目项来评价不同的音响系统。测试采用“无给定参考的多重刺激”的方式进行。

你总共有两个小时的时间完成测试，包括熟悉阶段和之后的主要测试阶段。在熟悉阶段，你将会了解节目项、用户界面，以及用于评估每个音响系统的属性。在熟悉阶段之后，将进行由两个部分构成的测试。

第1部分包括对音响系统的总体主观质量评分。

总体主观质量 – 捕捉在测音响质量所有方面的单一属性。

第2部分包括对每个音响系统的以下几个属性评分：

– 场景深度
– 环绕感
– 吞噬感
– 定位准确性
– 亮度
– 失真

以下提供了这些属性的定义，在测试开始之前将对定义进行解释。

在进行测试的每个阶段，请注意倾听节目项，对每个项目进行探索和评估，没有时间限制。

如果您对测试协议有任何疑问，或者需要进一步澄清，请提出（最好在熟悉阶段或之后）。

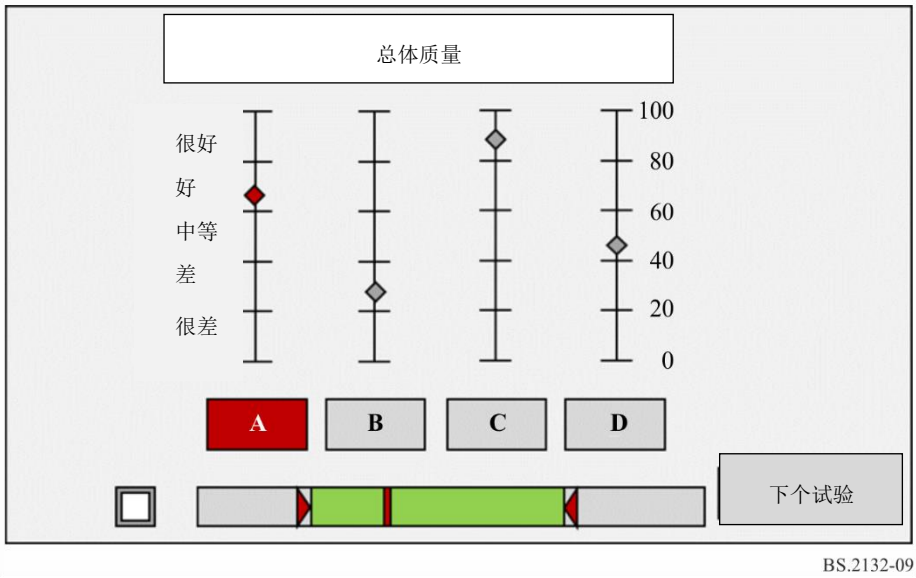
你可以每20-30分钟休息放松一下。

A2.1.1 全局质量评分

你需要根据总体主观质量，采用如图9所示的0-100点连续质量量表，对每个音响样本的全局质量进行评估。请认真倾听每个样本，可根据需要多次倾听，在你准备好的时候进行评分。一旦对所有样本打分完毕，点击“下一个”开始下一轮试验。

图9

总体主观质量评分测试用户界面



A2.1.2 描述性属性

对于每个试验，你需要基于各属性中的某一方面评估每个系统的音响质量（见表2）。

表2

属性标签和定义

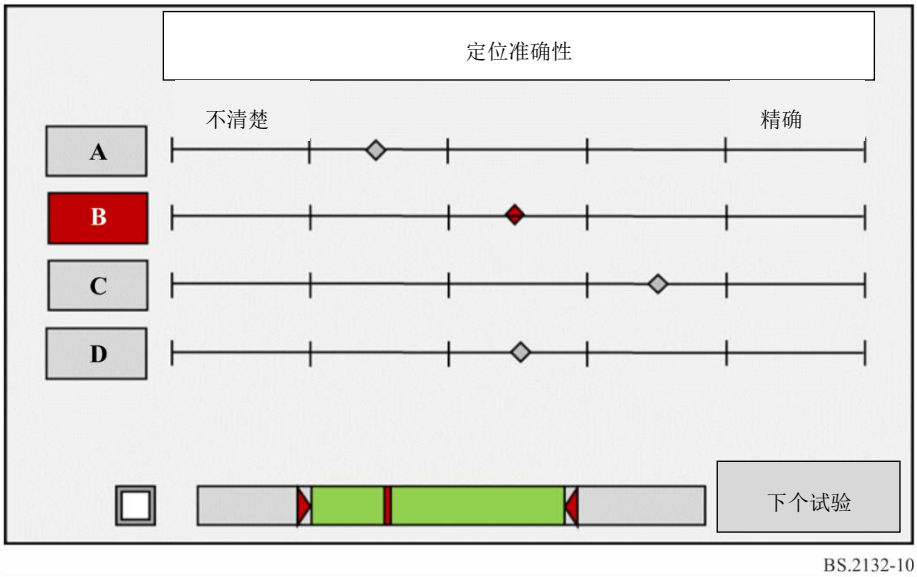
属性	定义	示例	低层标签	高层标签
场景深度	声音图像深度的感知。虑及场景的总体深度和单个音响的相对距离		平	深
环绕感	你是否被再现的音响环绕？它是否给你一种环绕或包围你的空间感？被声音包围的感觉	声音程度	轻微	大量
吞噬感	垂直方向上可感知的声源范围。一种被吞噬感完全环绕或覆盖的感觉	声音在你上下出现的程度和声音垂直环绕的程度	轻微	大量
定位准确性	各个乐器和声音在空间声音图像中放置和分离情况如何？每个声源在房间里的定位有多精确？如果个别声源在无意中扩散或扩大，则精确度低。在空间的声音影像中，个别的乐器和声音能否被清晰地放置和分离？每个声源在房间里的定位的精准程度如何？		不清楚	精确

表2（结束）

属性	定义	示例	低层标签	高层标签
亮度	三倍或高频扩展： – 轻微：如同你隔着门听音乐，压抑、模糊或沉闷。 – 大量:乐器的空间亮度、纯度和清晰度。高频率的清晰度，不尖锐或刺耳，也不失真		轻微	大量
失真	给声音增添了伪像的额外的和不需要的声音	例如，可能以会产生“尖锐”“嘶哑”或“损坏”的声音或一时的嗡嗡声的暂时或音色的伪像的形式出现	轻微	大量

图10

描述性属性测试用户界面



请在你准备好的时候开始进行熟悉过程。

感谢您的配合。

附件1的
附录3
（资料性）

在无给定参考的情况下对音频系统音响质量进行主观评估和描述性剖析的使用案例示例

ITU-R BS.2051建议书中对高级音响系统的介绍，为制作中的创造性表达提供了工具。对这些系统的评估包括在已知隐藏参考和锚点不可用条件下的评估。本建议书中描述的方法适用于这些情况。

此外，其他的案例可能存在。在这些用例中，实验者可以受益于使用这种方法来表征他们的系统或信号的主观质量。

适用这一方法的使用案例的示例包括以下主观质量评估：

– 生产者意图没有参考可用或适用的高级声音系统生产渲染器的行为。
– 由单一产品渲染器在不同的扬声器布局上再现高级音响编排。
– 先进音响系统内容家庭影院再现系统，没有提供已知最佳先验质量目标的系统。
– 上混或下混算法。
– 用于空间音频录制和生产的麦克风阵列。
– 用于空间音频生产的混响处理器。
– 多波段动态处理系统和无线电分配设置。
– 声音修复技术。