ICS 67.240

CCS X04

中华人民共和国国家标准

GB/T XXXX—202×/ ISO 11132:2021

感官分析方法　定量描述评价小组表现评估导则

Sensory analysis—Methodology—Guidelines for the measurement of the performance of a quantitative descriptive sensory panel

（ISO 11132:2021, IDT）

（征求意见稿）

202×- - 发布　　202×- - 实施

国家市场监督管理总局
国家标准化管理委员会　　发布

目　次

前　言

本文件按照 GB/T 1.1-2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草。

本标准等同采用 ISO 11132:2021《感官分析 方法学 定量描述评价小组表现评估导则》。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由全国感官分析标准化技术委员会（SAC/TC566）提出并归口。

本标准起草单位：

本标准主要起草人：

感官分析方法　定量描述感官评价小组表现评估导则

1　范围

本文件给出了对定量描述感官评价小组整体表现和每个成员个人表现进行评估的方法导则。

本文件适用于评价员个人或小组培训效果的验证，以及已建立评价小组的表现评估。

本文件不适用于定性描述性方法评价小组表现评估，如每个评价员没有个人评分记录、所有评价员没有使用相同的属性列表或者是优势属性测量而并非属性强度测量。因此，本文件不适用于使用了共识性感官剖面、自由选择剖面、闪现剖面和动态主导感官属性测试等方法的描述性评价小组表现评估。

本文件中给出的方法是用于监测和评价一个评价小组及其评价员区分产品的能力、同一小组评价员之间的一致性以及小组成员在属性强度评分中的可重复性和再现性。本文件不适用于评价小组之间的比较以及同一评价小组在不同条件下（如不同的时间段）评价的比较。

本文件中给出的方法可由感官分析师/小组长全部使用或选择部分使用，以持续评估小组或评价员个人的表现。列出的方法并不详尽，也能使用其他适当的方法。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

ISO 5492 感官分析 术语（Sensory analysis—Vocabulary）

注：GB/T 10221-2021 感官分析 术语(ISO 5492:2008,IDT)

3　术语和定义

ISO 5492 中界定的以及下列术语和定义适用于本文件。

3.1

一致性 agreement

不同的评价小组或评价员在给定属性得分时表现出相同产品差异的能力。

3.2

评价小组漂移 panel drift

不同的评价小组或评价员在给定属性得分时表现出相同样品差异的能力随着时间的推移，一个评价小组的灵敏度发生变化，或与恒定参照样品相比标度位置发生变化而引起偏差的一种现象。

3.3

表现 performance

一个小组或一个评价员对刺激和刺激属性作出可靠和有效评价的能力。

3.4

验证 validation

证实一个评价小组或评价员能够满足指定表现（3.3）的过程。

3.5

轮次 session

进行产品感官评价的时段。

注：单一轮次可以是一个或多个评价员对一个或多个产品进行评价。对于一个评价员来说，无论是单独评价还是作为小组的一份子参与评价，轮次间由时间间隔开来。

[来源：GB/T 10221-2021, 6.63]

3.6

重复 replicate

在一个实验设计中，一个特定条件的出现。

注 1：这个术语通常意味着该事件是几种同一类型的事件之一，但它可指一个单一的事件。当条件执行两次时，措辞为“两次重复”等。

注 2：要指定条件的多个出现，术语“重复”或“重复轮次”更明确。

注 3：“重复轮次”是指评价员、产品、测试条件和任务相同的轮次。

3.7

评价员偏差 assessor bias

当真实评分值已知时，评价员始终给出高于或低于真实分值的评分，或真实评分值未知时，评价小组始终给出高于或低于真实分值的倾向。

[来源：GB/T 10221-2021, 3.40]

3.8

序列偏差 order bias

由一个产品在一组产品中所处的空间或时间位次而引起的偏差。

注：该术语包括“位置偏差”和“顺序偏差”。

[来源：GB/T 10221-2021, 3.42]

3.9

重复性 repeatabiIity

在相同测试条件下，同一评价员或评价小组对同一测试产品评价结果的一致性。

注：重复性能在一个轮次或几个明显独立的轮次中测量，只要重复评价是在能被认为是相同的测试条件下进行的。如果重复评价是在明显不同的轮次/实验中进行的，那么轮次通常只相隔几天。在这种情况下，短期内的重复性和重现性之间的区别很小，并且与被认为相同或不同的测试条件有关。

[来源：GB/T 10221-2021, 3.45]

3.10

再现性 reproducibility

在不同的测试条件下，或由不同的评价员或评价小组对同一测试样品评价结果的一致性。

注：再现性可通过以下任何一种方法进行测定：

—评价小组（或评价员）的短期再现性，以天为间隔的两个或多个轮次之间进行测量；

—评价小组（或评价员）在中长期再现性，以月为间隔的不同轮次之间进行测量；

—不同评价小组之间在同一实验室或不同实验室之间的再现性。

[来源：GB/T 10221-2021, 3.46]

4 原理

4.1 两种可行的方法

4.1.1 概述

本文件涉及用于评估一个或多个感官属性强度的感官评价小组，以便对产品进行定量描述或剖面分析（见 GB/T 39625-2020）。可采用不同的方法来评估评价小组差异测试的表现。

定量感官小组的表现可通过使用已有的评估方法（称为“持续监控”）从专门为获得表现（称为“专用流程”）而进行小组轮次中的评估。

4.1.2 通过专用程序进行表现评估

专用流程是对评价员个体的评估或其他评估目的首选方法。对于评估的更新，应根据需要定期重复此专用流程。图 1 是流程图。

这种方法通常能在小组的训练阶段结束时使用，以确保小组和评价员个人已经达到了预期的表现水平，并能被视为训练有素的感官评价员或专家评价员（取决于表现标准）。

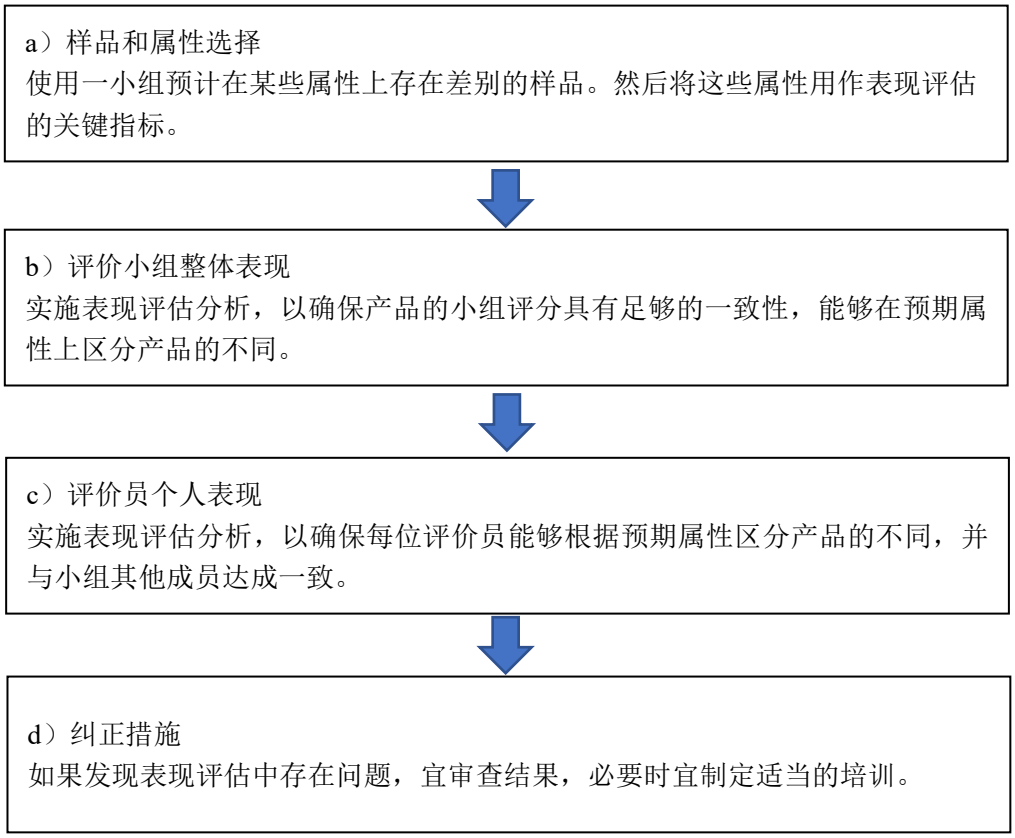


图 1 通过一个专用的程序来进行表现评估的步骤

4.1.3 通过常规产品分析进行持续监控

另一种方法是分析已经收集到的剖面分析数据。为了分析由评价小组生成的持续的剖面分析数据，可使用来自类型、编号等完全不同的产品的剖面分析实验数据。过程如图 1 所示。然而，由于没有预先定义的差别，建议将给定的剖面数据中被整个评价小组用于显著区分产品的属性用作评估感官评价员个人表现的关键依据。没有导致显著差异的属性不能用于一致性的检查，这是由于评价小组间或评价小组内部缺乏一致性可能意味着产品在这些特征上非常相似。

4.2 小组或评价员个体的表现指标

对于一次评价，可确定以下指标：

—评价小组辨别力，以评价小组展现产品之间显著差异的能力来衡量；

—评价员辨别力，以评价员评估产品之间显著差异的能力来衡量；

—评价员的一致性，以评价员的平均产品得分与评价小组的平均产品得分之间的一致性程度来衡量；

—评价小组一致性，以评价员的平均产品得分之间的一致性程度来衡量。

对于重复实验：

—评价员的重复性，对于重复评价，衡量同一产品重复评价之间的同质性。

—评价小组的可重复性，以每个评价员对同一产品的重复评价之间的平均同质性程度来衡量。

4.3 统计分析

本文描述了一种单一的、一致的结果统计分析方法。然而，评价小组表现的一些指标能通过多个指标来评价。例如，误差均方差和误差标准差(SD)（其平方根）都表达了产品评价中的可变性。所使用的措施宜是在应用领域中通常采用的措施。

在使用一个属性的标度时，其他衡量评价员之间一致性的相关标准是评价员和产品之间的相互作用，以及一个评价员的得分与小组均值之间的相关系数。评价员可没有偏差，但以不同的方式使用量表。相关性接近 1，回归斜率接近 1，回归截距接近 0，表明评价员与评价小组的其他成员之间有良好的一致性。

当每个评价员评价少量样品（少于 6 个）时，相关系数宜谨慎解释，因为它能有很高的偶然性（高达 0.7）。

5 条件要求

5.1 实验条件

环境设施宜符合 ISO 8589 的要求。

5.2 评价员资格

评价员宜具有符合 ISO 8586 或更高标准筛选出评价员的资格和经验水平。

6 表现评估专用程序

6.1 样品和属性选择

在每项专用流程中，应向小组中的评价员提供一组样品，该样品需与评价产品时所要评

价的样品相似，并且对于每个相关属性，预期至少在一对样品之间存在统计上的显著差异。

为了确保产品的所有关键属性都经过检验，在测试中应包含足够多样化的属性集。

这些相关的属性被用作衡量评价小组表现的关键标准。样品集应包含重复。每个样品的重复次数应相同。重复可在单个或超过两个或多个轮次中进行评价。评价员、样品和重复次数等取决于产品、被评价的感官属性和设计该流程的目的。例如，可使用 3 个或 4 个样品的 2 或 3 次重复。应注意限制在一轮次实验中的评价次数，以避免感觉疲劳。样品的属性宜与评价小组在评价产品时评价的值的范围相似。

6.2 实验设计

6.2.1 概述

根据要回答的最重要目标，可在专用流程中使用几种类型的实验设计。

6.2.2 随机区组设计

使用随机区组实验设计，其中以评价员来作为“区组”。当一个样品到下一个样品之间没有延滞效应时，这种设计是合适的。否则，应注意采用平衡实验设计（见 6.2.3）。

6.2.3 平衡和随机设计

如果从一个样品到下一个样品之间存在延滞效应，威廉姆斯拉丁方设计是一个适合的实验设计[12]，表 1 展示了 4 个评价员和 4 个样品的威廉姆斯拉丁方平衡设计。

表 1-4 个评价员和 4 个样品的拉丁方设计

评价员	轮次	顺序			
		1	2	3	4
1	1	A	B	C	D
2	1	B	D	A	C
3	1	C	A	D	B
4	1	D	C	B	A
1	2	B	D	A	C
2	2	C	A	D	B
3	2	D	C	B	A
4	2	A	B	C	D

在这个设计中，每个评价员在给定的轮次中将以不同的顺序抽样四种产品，任何特定的产品之后为每个评价员提供不同的产品。例如，在轮次 1 中，评价员 1 在抽样 A 样品之后，将为其提供样品 B，对于评价员 2，样品 A 后为其提供样品 C，评价员 3 在样品 A 后为其提

供样品 D，而对于评价员 4，样品 A 后将不提供其他样品。

对于每个产品评价的重复，推荐为每个评价员以不同的顺序提供样本，以减少顺序效应和延滞效应。

如果评价员人数是 4 的倍数，则可每组 4 名评价员并重复相同的设计。

也可选择一个随机的产品顺序设计，即在每个轮次中每个样品随机出现在每个位置。

这些方法的优点是在评价小组层面上最小化延滞效应，从而在小组层面上更好的估计产品平均值，以进行表现评估。如果产品存在顺序效应，则评价员之间的一致性将受到影响，这是由于为每个评价员提供的产品顺序不同。为了在完全相同的任务上的比较评价员的表现，可对所有的感官评价员提供相同的产品顺序（见 6.2.4）。

6.2.4 相同的顺序设计

为了关注评价员个人的表现，并尽量在最相似条件下的比较评价员的表现，提出了一种替代设计，即所有评价员按照相同的顺序评价产品，见表 2。

表 2 四名评价员和四个样品的相同顺序设计

评价员	轮次	顺序			
		1	2	3	4
1	1	A	B	C	D
2	1	A	B	C	D
3	1	A	B	C	D
4	1	A	B	C	D
1	2	A	B	C	D
2	2	A	B	C	D
3	2	A	B	C	D
4	2	A	B	C	D

值得一提的是，在这种情况下，评价员不是在评价产品本身，而是在评价给定位置的产品（产品和位置效应混淆）。这将导致对产品效应的有偏估计（因顺序效应而有偏），但对于评价员效应和产品*评价员之间的相互作用会导致无偏估计。

6.3 统计分析

表 3 说明了一种对结果制表和汇总的方法。一些计算机软件可能有不同的数据排列要求，如样本按列排列，评价员按行排列。

表 3 评价员对一个属性的评价结果

样品	评价员								均值
	1		...	j		...	n_q		
	分数	均值		分数	均值		分数	均值	
1	Y_{111} Y_{112} ... Y_{11n_r}	$\overline{Y_{11.}}$		Y_{1j1} Y_{1j2} ... Y_{1jn_r}	$\overline{Y_{1j.}Y_{1j.}}$				$\overline{Y_{1..}}$
...									
i	Y_{i11} Y_{i12} ... Y_{i1n_r}	$\overline{Y_{i1.}}$		Y_{ij1} Y_{ij2} ... Y_{ijn_r}	$\overline{Y_{ij.}Y_{ij.}}$				$\overline{Y_{i..}}$
...									
n_p									$\overline{Y_{n_p..}}$
均值	$\overline{Y_{.1.}}$			$\overline{Y_{.j.}}$					$\overline{Y_{...}}$
在此表中假定： n_p=样品数（i=1,2...,n_p）； n_q=评价员人数（j=1,2...n_q）； n_r=每次样品重复次数　(k=1,2..,n_r)。									

除偏差以外，要衡量评价小组整体和评价员个人的表现，可通过方差分析（ANOVA）来分析数据[7]。

本文件中没有显示基本计算的细节，因为这些分析通常是由一套计算机程序来进行。

当在单次中进行重复评价时，每个评价员的数据可采用单因素方差分析进行分析（见表4）。

如果评价员的重复评价是在明显独立的不同轮次中进行的，并且根据研究人员/实验室的标准实验，可使用单因素方差分析模型（即样品效应）或双因素方差分析模型（即轮次和样品效应）（见表 4 和表 5）。

表 4 评价员个人一个属性的单因素方差分析

变异来源	自由度	平方和	均值平方	*F* 比率
样品间 误差	$v_1 = n_p$-1 $v_2 = n_p(n_r$-1)	s_1 s_2	$MS_1=s_1/v_1$ $MS_2 = s_2/v_2$	$F=MS_1/MS_2$
总体	$v_3 = n_p n_r$-1	s_3		
n_p = 样品数 n_r = 每个样品重复次数				

表 5 单次实验中一个属性的评价员个人的单因素方差

变异来源	自由度	平方和	均值平方	*F* 比率
样品间 轮次间 误差	$v_1 = n_p$-1 $v_4 = n_s$-1 $v_5 = n_p(n_r$ - 1)-(n_s-1)	s_1 s_4 s_5	$MS_1=s_1/v_1$ $MS_4 =s_4/v_4$ $MS_5 = s_5/v_5$	$F=MS_1/MS_5$ $F=MS_4/MS_5$
总体	$v_3 = n_p n_r$-1	s_3		
n_p = 样品数 n_r = 每个样品重复次数				

完整数据集采用随机区组方差分析来进行分析（见表 6）。

评价员效应可是固定的或是随机的[8]。对于评价员表现的衡量，通常选择将评价员效应固定，因为关注的重点是特定评价员的表现。但是，为了更好地预测在实际评价条件下的表现，也可选择随机的评价员因素(见表 6、7 和脚注 a)。

当在单个轮次中进行重复评价时，完整数据集可通过双因素方差分析进行分析（见表 6）。

如果在明显独立的轮次中重复评价，并且根据研究人员/实验室的标准实验，可使用双因素方差分析模型（即小组成员和样品效应）或三因素方差分析模型（即小组、轮次和样品效应）（见表 6 和表 7）。

在附录 A 中给出了一个实际应用的例子。

表 6 一个属性的完整数据集（具有重复性）的双因素方差分析

变异来源	自由度	平方和	均值平方	*F* 比率

样品间	$v_6 = n_p$-1	s_6	$MS_6=s_6/v_6$	$F=MS_6/MS_9$[a]
评价员间	$v_7 = n_q$-1	s_7	$MS_7 =s_7/v_7$	$F=MS_7/MS_9$[a]
相互作用	$v_8 = (n_p$-1)(n_q-1)	s_8	$MS_8 = s_8/v_8$	$F=MS_8/MS_9$
误差	$v_9 = n_p n_q(n_r$ -1）	s_9	$MS_9 = s_9/v_9$	
总体	$v_{10} = n_p n_q n_r$-1	s_{10}		

n_p = 样品数

n_q = 评价员人数

n_r = 每个样品重复次数

[a] 考虑到评价员效应是固定的，给出了公式。考虑到评价员随机效应、样品间效应的 F 的分母变为 MS_8 而不是 MS_9。

表 7 对于具有会话效果的一个属性的完整数据集（具有重复性）的三因素方差分析

变异来源	自由度	平方和	均值平方	*F* 比率
样品间	$v_6 = n_p$-1	s_6	$MS_6=s_6/v_6$	$F=MS_6/MS_{12}$[a]
评价员间	$v_7 = n_q$-1	s_7	$MS_7 =s_7/v_7$	$F=MS_7/MS_{12}$[a]
轮次间	$v_{11} = n_s$-1	s_{11}	$MS_{11} = s_{11}/v_{11}$	$F=MS_{11}/MS_{12}$[a]
相互作用	$v_8 = (n_p$- 1)(n_q-1)	s_8	$MS_8 = s_8/v_8$	$F=MS_8/MS_{12}$
误差	$v_{12} = n_p n_q(n_r$ -1)-(n_s-1）	s_{12}	$MS_{12} = s_{12}/v_{12}$	
总体	$v_{10} = n_p n_q n_r$-1	s_{10}		

n_p = 样品数

n_q = 评价员人数

n_r = 每个样品重复次数

[a] 考虑到评价员效应是固定的，给出了公式。考虑到评价员随机效应、样品间效应的 F 的分母变为 MS_8 而不是 MS_{12}。

6.4 评价小组表现—对统计输出的解释

6.4.1 关键属性区分

按照预期显著区分的关键属性的比例应被确定。在完整数据集的方差分析表中，用样本

间α为 0.05 水平来表示每个属性的显著差异（见表 6 和 7）。被显著区分的关键属性比例越高，评价小组的表现就越好。对于未按照预期显著区分的关键属性，小组应对其接受进一步培训。

6.4.2 小组层面的一致性

当任意评价员与小组的其他成员意见不一致时，则小组意见不一致（见 6.5.4）。

如果在方差分析中，样品和评价员之间的相互作用在α水平为 0.05 时显著，则认为小组意见不一致。

评价小组意见的一致程度与相互作用项 s_i 成反比，见如公式（1）所示：

$$s_i=\sqrt{\frac{MS_8-MS_9}{n_r}} \text{ 或 } s_i=\sqrt{\frac{MS_8-MS_{12}}{n_r}} \quad (1)$$

详见表 6 和 7。

导致样本与评价员之显著相互作用的关键属性数量是基于每一个属性的方差分析（见表 6 和表 7）所确定的。导致显著相互作用的关键属性的数量越多，评价小组表现出的一致性就越低。如果相互作用显著，则应在小组成员层面上研究相互作用的性质，并在需要时采取措施。例如，如果一个小组成员不同意小组其他成员评价的产品之间的差异，那么这个小组成员应进行再培训。相互作用的性质通常是由绘制评价员与产品均值之间的关系图来进行研究。另一种选择是使用混合评价模型（MAM）[6][10]。

6.4.3 评价小组的重复性

评价小组的重复性可由评价员个人的重复性估计，与误差项 s_e 成反比，如公式（2）所示：

$$s_e=\sqrt{MS_9} \text{ 或 } s_e=\sqrt{MS_{12}} \quad (2)$$

取决于所选的模型（有或没有轮次影响）。

详见表 6 和表 7。

$$s_R=\sqrt{s_e^2+s_a^2+s_{sess}^2+s_{\mathrm{a\times sess}}^2+s_{\mathrm{prod\times sess}}^2}$$

6.5 评价员个人表现—统计输出的解释

6.5.1 评价员的辨别力

辨别能力是由按照预期被显著区分的关键属性所占比例来衡量。在方差分析表中，用样本间α水平为 0.05 来表示每个属性的显著差异 (见表 4 和 5)。显著区分的关键属性比例越高，评价员的表现就越好。对于未按照预期显著区分的关键属性，评价员需要接受进一步培训。

6.5.2 评价员的重复性

评价员的重复性与评价员的误差项 s_e 成反比，如公式（3）所示:

$$s_e = \sqrt{MS_2} \text{ 或 } s_e = \sqrt{MS_5} \quad (3)$$

取决于所选的模型（有或没有轮次影响）。

详见表 4 和表 5。

6.5.3 评价员的一致性

评价员的一致性与从每个样品中计算出的偏差项的标准差成反比。

(对于评价员 j，样品 i 的偏差项是该样品的评价员均值与评价小组均值之间的差，$\overline{Y_{ij.}}$-$\overline{Y_{i..}}$，详见表 3）。

当评价员的表现缺乏一致性时，从一个评价员得分相对于小组均值的散点图、回归和相关性分析，可看出这种不一致性是随机的，还是由于其采取与小组其他成员使用不同量表的模式。

6.5.4 评价员之间的一致性

当至少有一名评价员与小组的其他成员不一致时，该小组就不是同质的。

可通过以下方式检测到：

—某评价员有明显偏差；

—某评价员的残差明显大于小组整体的；

—评价员得分与小组均值之间的相关系数非常小或为负值；

—评价员分数对于小组均值的回归斜率与 1 差异显著，或截距与 0 在统计上具有显著差异，或两者都有。

评价员之间的一致性与“评价员间”项s_a成反比，如公式（4）所示：

$$s_a = \sqrt{\frac{MS_8 - MS_9}{n_q n_r}} \quad (4)$$

评价员之间的差异应使用“评价员之间”的 F-比率进行显著性实验，并将其与相关自由度的 F 值进行比较。如果该差异是显著的，则有充分的证据表明评价小组一致性存在一个需要解决的问题。缺乏显著性本身不能证明不存在上述问题，因为这个问题可能被较差的重复性（高于预期的误差项，s_e）所掩盖。

6.5.5 使用不同标度的偏差

评价员偏差方差分析结果的显著性可表明评价员以不同的方式使用该标度。

在大多数情况下，“真”值不是已知的，对一个评价员的总体偏差被认为是该评价员的

平均值和该小组的平均值之间的差。评价员 *j* 的偏差由公式（5）给出：

$$\overline{Y_{j.}} - \overline{Y_{..}} \qquad (5)$$

评价员可不同的方式使用量表(见 ISO 4121)。在“通用”标度的使用中，每个属性的强度是根据评价员对特定产品类型的总感官变化的认知情况进行评定的。针对一个或仅针对几个产品类别的小组通常会使用并改进这种方法。在“相对”标度使用中，评价员用于评价强度的参考框架与给定测试中一组产品所显示的感官变化有关。这种方法更可能被针对广泛产品类别的小组所使用。确保标度方法在小组内的一致性可减少标度偏差。

6.6 评价员表现的有关问题

6.6.1 概述

一旦确定了表现的问题，就可列出并制定相应的培训计划。

6.6.2 评价小组

针对那些出现问题的属性，可针对整个小组组织培训课程。

6.6.3 评价员个人

对于评价员个人表现的具体问题，可使用中立或积极的语气一对一地私下讨论问题所在，例如，提供个人水平的数据与小组平均值的比较作为反馈。然后就可进对小组整体进行培训。

6.7 随着时间推移的跟踪表现实验设计

如果计划研究评价小组一段时间内的一致性，一年内每月一次感官评价实验可提供足够的数据。每次感官评价实验的平衡设计应按 6.2.3 所示。随着时间的推移，对表现的审查可用于确定后续方案，如小组漂移或再培训后的表现改进。

7 通过常规产品分析进行持续监测的程序

7.1 属性选择

该流程与通过专用流程进行的表现评估相同（见第 6 条）。由于没有预先定义的差异，建议将小组整体对一个给定的剖面分析数据中被显著区分的属性作为检查评价员个人表现的关键依据。因为评价员内部和评价员之间缺乏一致性可能是由于产品在这些属性上非常相似，所以没有呈现显著差异的属性不能用于检查一致性。

7.2 统计分析

统计分析中除了评价员效应始终被认为是随机效应以外，与通过专用流程评估表现的测

试相似（见第 6 章）。

7.3 一段时间内的表现

如果已经有了几轮次的常规评价的数据，就能对它们进行分析，以显示随着时间的推移而发生的变化。随着时间的推移，对表现的审查能用于确定后续方案，如小组漂移或小组再培训后的表现改进。

7.4 数据随时间变化的统计分析

应使用重复测量方差分析对多个轮次的数据进行全局分析。在实验中，同一评价员可能不会参加所有轮次的感官评价，有必要使用方差分析的一般线性模型选项来获得每个评价员的偏差以及其他参数和方差分量的无偏估计。

对于小组评价实验，能得到估计 a)和 b)：

a) 如果在一系列的轮次中收集了相同的质控样品的数据，则小组的一致性能从每轮次期间进行评估（见表 7）。

b) 内部一致性：当评价员个体发生偏差时，评价员和评价轮次的相互作用可评估它们的稳定性。

对于每个评价员，每个属性能得到三个估计值。

—总体偏差：在重复和（或）轮次中，评价员得分和相应的小组整体平均值之差。

—一致性：与不同轮次之间的偏差项的变化成反比。

—重复性：相同样品的得分之间的差异，通过每个轮次的残差标准差的估计值来确定。

公式：

$$s_R = \sqrt{s_{res}^2 + s_{a\times p}^2 + s_p^2}$$

7.5 完整的统计分析

上述各子条款中描述的统计分析方法分别应用于每个属性，以评估小组和评价员对他们必须回答的属性（问题）的表现。

此外，为了获得数据的整体概况，可使用多维技术，如主成分分析(PCA)、判别分析(DA)和广义普氏分析(GPA)。这些方法能用于表现评估或持续监测。有关这些多变量分析方法的更多详细信息，见参考文献[5]、[9]和[11]。

附录 A

（资料性附录）

应用示例

A.1 数据表

在一个环节中，四名评价员在六个样品的三次重复中为一个属性评分。表 A.1 显示了本例的结果。

注：这是一个具有说明性的例子。通常会有四名以上的评价员参加。

表 A.1 评价员结果数据列表

样品	评价员								均值
	评价员 1		评价员 2		评价员 3		评价员 4		
	分数	均值	分数	均值	分数	均值	分数	均值	
1	8	8.3	5	7.3	6	6	9	8.3	7.5
	8		8		7		8		
	9		9		5		8		
2	6	7	6	5.7	5	5.3	7	6.7	6.17
	8		7		4		7		
	7		4		7		6		
3	4	4.7	5	3.3	4	4	5	5	4.25
	5		2		3		5		
	5		3		5		5		
4	6	5.7	6	5.3	4	3.3	6	5.3	4.92
	6		4		2		5		
	5		6		4		5		
5	4	4	3	3	4	4.3	4	4.3	3.92
	5		2		4		5		
	3		4		5		4		
6	5	5.7	4	4.3	5	5	7	6.3	5.33
	6		2		4		5		
	6		7		6		7		
均值	5.89		4.83		4.67		6		5.35

A.2 本例详细介绍方差分析，见表 A.2、A.4、A.4 和 A.5。

表 A.2 完整数据集的方差分析（评价员=固定效应）

变异来源	自由度	平方和	均值评方	F-ratio
样品间	5	104.90	20.98	16.39[a]
评价员间	3	26.04	8.68	6.79[a]
相互作用	15	16.04	1.07	0.84[b]
残差	48	61.33	1.28	
总体	71	208.31		
[a] 显著差异水平 α= 0.05 [a] 无显著差异水平 α = 0.05				

表 A.3 方差分析-评价员个体

变异来源	自由度	评价员							
		评价员 1		评价员 2		评价员 3		评价员 4	
		MS	*F*	MS	*F*	MS	*F*	MS	*F*
样品间	5	7.42	13.36[a]	7.83	2.66[b]	2.80	2.40[b]	6.13	13.80[a]
残差	12	0.56		2.94		1.17		0.44	
SD 残差，*s*		0.75		1.71		1.08		0.67	
[a] 显著水平 α = 0.05 [b] 无显著差异水平 α = 0.05									

表 A.4 评价员个体偏差和残差（SDs）

评价员	偏差	SD 残差
1	5.89 - 5.35 = +0.54	0.75
2	4.83 - 5.35 = -0.52	1.71
3	4.67 - 5.35 = -0.68	1.08
4	6.00 - 5.35 = +0.65	0.67
注：偏差是评估者的平均值和总体平均值之差，两者都在表 A.1 中给出。		

表 A.5 个体样品偏差项

样品	评价员			
	1	2	3	4
1	0.83	-0.17	-1.50	0.83
2	0.83	-0.50	-0.83	0.50
3	0.42	-0.92	-0.25	0.75
4	0.75	0.42	-1.58	0.42
5	0.08	-0.92	0.42	0.42
6	0.33	-1.00	-0.33	1.00
SD,s	0.31	0.56	0.78	0.24

注：个体偏差是评价员对样品的平均值与该样品的小组平均值之差，两者均在表 A.1 中给出。

A.3 整个小组的整体表现—对统计输出的解释

从表 A.2 可看出，“样品间”效应是显著的（α水平为 0.05），表明评价小组区分产品之间差异的一致性。

从同一张表中也可看出，相互作用在α为 0.05 的水平上并不显著，说明小组成员之间的差异不存在显著的不一致。

显著的“评价员之间”F 比率表明评价员给出了不同的平均分数（在所有产品中）。评价员平均值的变化程度可由评价员标准偏差来描述。

在此示例中，比起固定评价员效应，使用随机评价员效应将得出相同的结论。当产品*评价员相互作用显着时，可能会出现一些解释上的差异。

A.4 评价员个体的表现—对统计输出的解释

A.4.1 概述

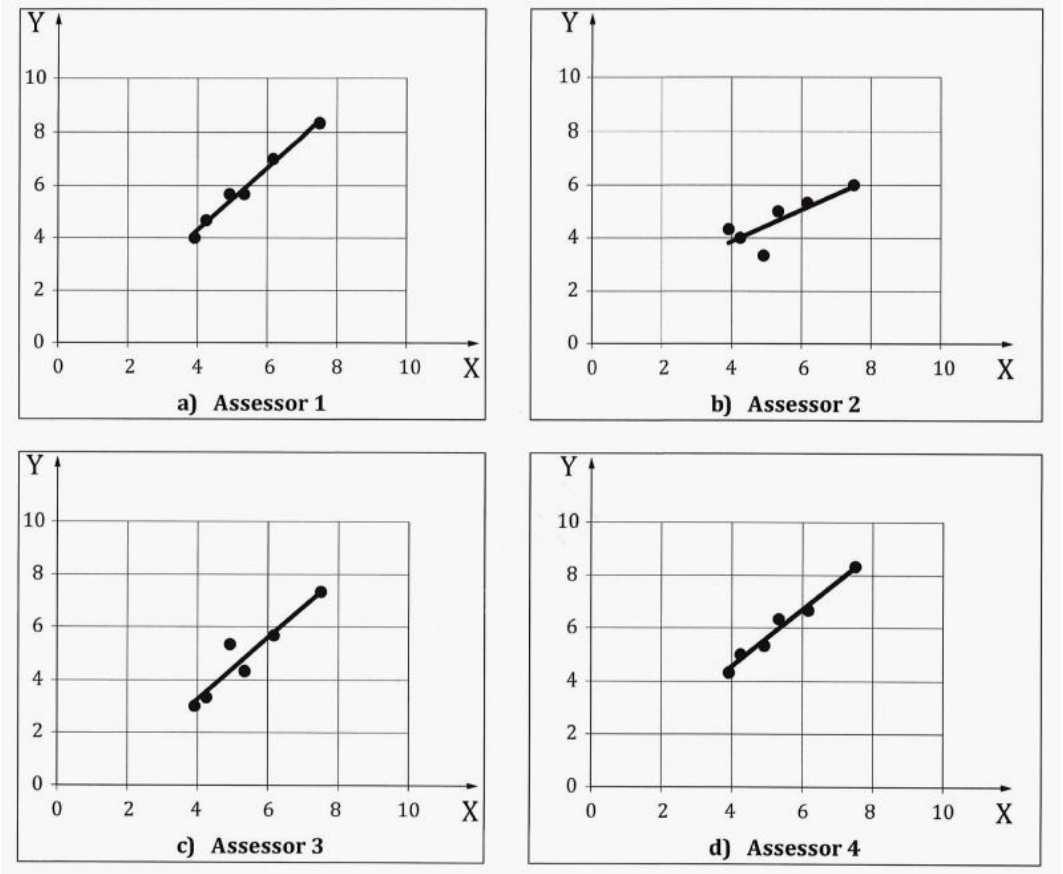
评价员 2 和 3 的残差的标准差最高（见表 A.4），表明在同一样品的重复中，其重复性低于评价员 1 和 4。

评价员 3 的平均值也有很高的负偏差，这表明他倾向于给出比小组的其他成员更低的分数。该评价员的一致性也不如其他成员，从低于小组平均值 1.58 到高于小组平均值 0.42，具有最大的标准差（0.78）。

评价员 4 具有+0.65 的较高正偏差，但由于偏差的标准差仅为 0.24，因此保持了一致性。由于评价员 1 和 4 的一致性很好并且可变性低，他们的分数是值得信赖的，并且评价员 2 和 3 降低了小组均值，因此评价员 4 的“偏差”不必关注。

A.4.2 回归性和相关性统计数据

图 A.1 显示了针对 6 个产品评价员评分与小组均值的关系图。



X 评价小组均值

Y 评价员的评分

图 A.l 评价员 1/2/3/4 与小组平均值

在本例中，评分没有标准值。小组均值作为每个评价员的参考评分。理想的图是一个显示评价员和小组均值之间完全一致的图，数据点靠近一条斜率 b=1.00 截距 a=0.00 的线。相关系数应接近于+1.00。

四名评价员的回归和相关性统计数据见表 A.6。

表 A.6 回归分析和相关性统计数据

参数	评价员			
	1	2	3	4
相关系数	0.99	0.95	0.81	0.99
斜率, *b*	1.18	1.16	0.59	1.07
x 轴-截距, *a*	-0.42	-1.36	1.49	0.29

评价员 4 显然是最好的，相关系数接近于 1，斜率接近于 1，截距最小。

评价员 3 的斜率偏小，表明比其他评价员使用的标度范围更窄。

评价员 2 的截距为负，表示存在负偏差。

A.5 其他表现评估问题

A.5.1 概述

折线图可用于揭示有待进一步研究的问题。

A.5.2 个人评价员

图 A.2~A.4 展示了比较三个小组中评价员个人表现的例子。

图 A.2 展示了除了一个评价员外，所有人的样品区分具有良好的一致性。评价员 10 在样本之间几乎没有分辨能力。其余的评价员对除样品 A 外的所有样品都展示了良好的一致性。

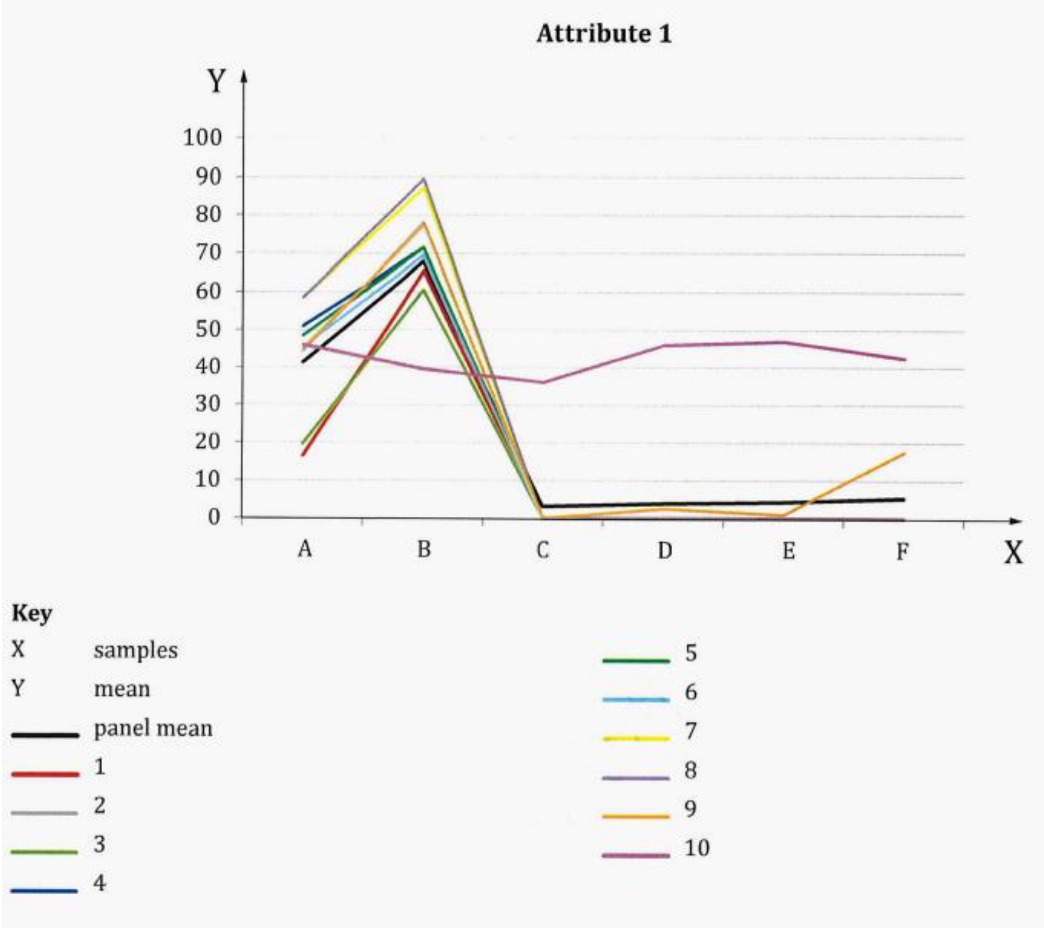


图 A.2 十名评价员在一个属性（属性 1）上对六个样品的小组评分

图 A.3 展示了大多数评价员对样品排序达成一致的情况，但评价员 10 的辨别力较差，对标度的使用范围小。

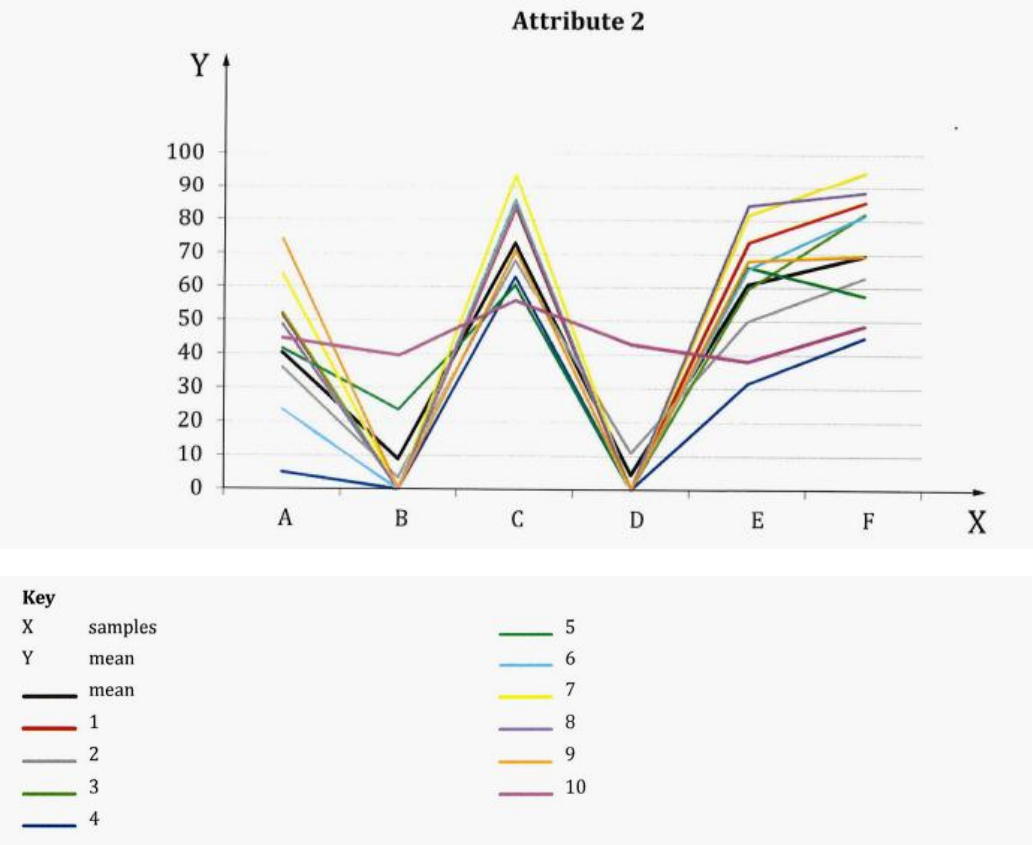


图 A.3 十名评价员在一个属性（属性 2）上对六个样品的小组评分

图 A.4 展示了所有评价员在样品区分和量表使用方面均表现不佳的情况，甚至在样品排序方面也没有表现出一致性，有两名评价员对所有样品的评分都很低。

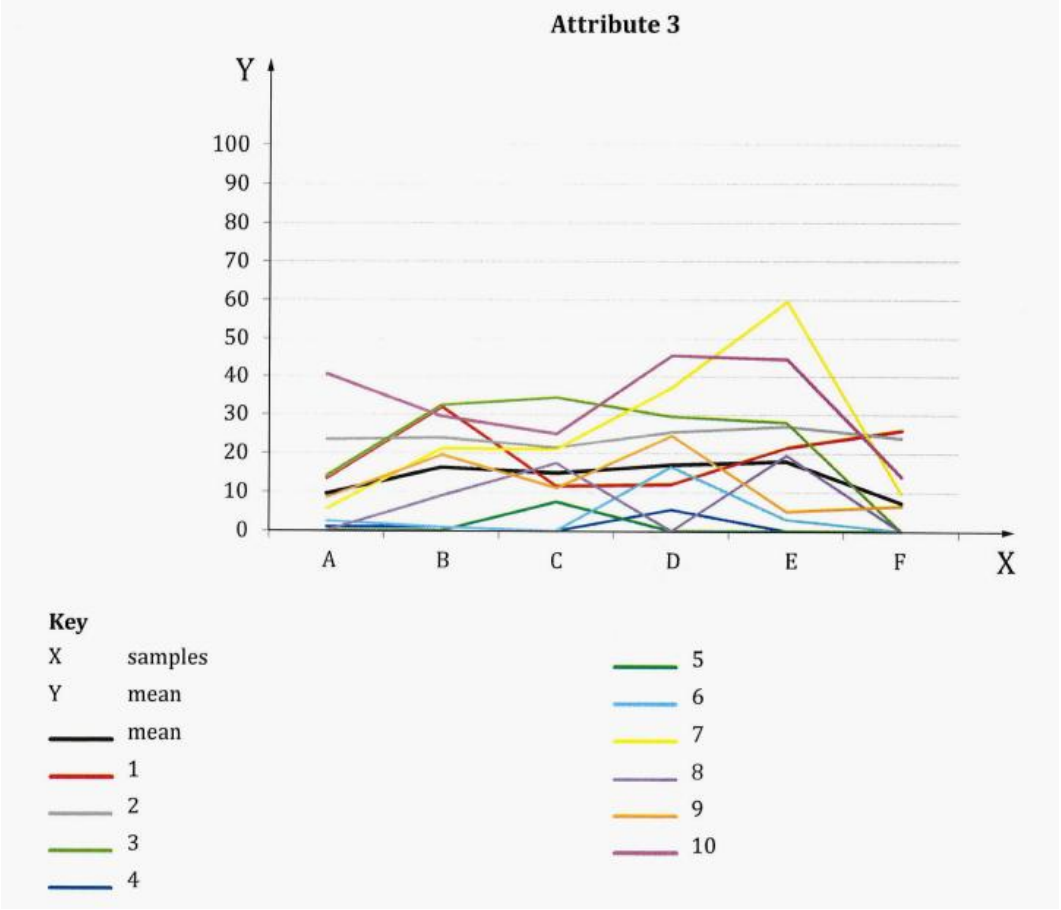


图 A.4 十名评价员在一个属性（属性 3）上对六个样品的小组评分

参 考 文 献

[1] ISO 4121 Sensory analysis Guidelines for the use of quantitative response scale.

[2] ISO 8586 Sensory analysis General guidelines for the selection, training and monitoring of selected assessors and expert sensory assessors.

[3] ISO 8589 Sensory analysis General guidance for the design of test rooms.

[4] ISO 13299 Sensory analysis Methodology General guidance for establishing a sensory profile

[5] Arnold G.M., Williams A.A., The use of generalised Procrustes techniques in sensory analysis. In: Piggott J.R. (ed.) Statistical procedures in food research. *Elsevier Applied Science,* London,1986, pp. 233-254

[6] Brockhoff P.B. Schlich P. Skovgaard I., Taking individual scaling differences into account by analyzing profile data with the Mixed Assessor Model. *Food Quality and Preference*. 2015, 39, pp. 156-166.

[7] Lea P., N/es T., Rodbotten M., Analysis of variance for sensory data. *Chichester: Wiley*, 1997

[8] Lundahl D.S., McDaniel M.R. The panellist effect - fixed or random? *Journal of Sensory Studies*. 1988, 3, pp. 113-121

[9] Naes T., Brockhoff P.B., Tomic 0., Statistics for Sensory and Consumer Science. *John Wiley and Sons*, UK, 2010

[10] Peltier C. Visalli M. Schlich P., Multiplicative decomposition of the scaling effect in the Mixed Assessor Model into a descriptor-specific and an overall coefficient. *Food Quality and Preference*. 2015, 48, pp. 268-273

[11] Varela P., Ares G., Novel Techniques in Sensory Characterization and Consumer Profiling，*CRC Press*, 2014

[12] Williams E.J., Experimental designs balanced for the estimation of residual effects of treatments. *Aust. J. Sci. Res. Ser. A*. 1949, 2 pp. 149-168