ICS 67.240
CCS X 04

# 中华人民共和国国家标准

GB/T ××××—202×/ISO 11132:2021

# 感官分析方法　定量描述感官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，Sensory analysis—Methodology—Guidelines for the measurement of the performance of a quantitative descriptive sensory panel，IDT）**

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。

本文件等同采用ISO 11132:2021《感官分析　方法学　定量描述感官评价小组表现评估导则》。

本文件做了下列最小限度的编辑性改动：

a) 为与现有标准协调，将标准名称改为《感官分析方法　定量描述感官评价小组表现评估导则》；

b) 增加了表5中关于“$n_s$”的说明，同时将“$MS_4 = s_1/v_1$”更改为“$MS_4 = s_4/v_4$”，纠正原文错误；

c) 删除了6.4.3和7.4中的2个公式，纠正原文错误；

d) 更正了图A.1中b)和c)的顺序，纠正原文错误。

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/TC 566)提出并归口。

本文件起草单位：中国标准化研究院、内蒙古蒙牛乳业(集团)股份有限公司、内蒙古伊利实业集团股份有限公司、黑龙江飞鹤乳业有限公司、深圳雾芯科技有限公司、中国烟草总公司郑州烟草研究院、完美(广东)日用品有限公司、中山市食品学会、广东博然堂生物科技有限公司、上海百雀羚生物科技有限公司、安利(上海)科技发展有限公司、云南贝泰妮生物科技集团股份有限公司、福建富邦食品有限公司、上海家化联合股份有限公司、广东北工商绿色护肤品研究院有限公司、上海上美化妆品股份有限公司、拉芳家化股份有限公司、宜品乳业(青岛)集团有限公司、哈尔滨敷尔佳科技股份有限公司、华熙生物科技股份有限公司、中国合格评定国家认可中心、东北电力大学、浙江工商大学、广州中妆美业化妆品有限公司、上海康识食品科技有限公司、东阿阿胶股份有限公司、水羊集团股份有限公司、美出莱(杭州)化妆品有限责任公司、上海上水和肌生物科技有限公司、浙江养生堂天然药物研究所有限公司、无限极(中国)有限公司、广州悦荟化妆品有限公司、霸王(广州)有限公司、上海永熙信息科技有限公司、蜜雪冰城股份有限公司、广东碧茜生物科技有限公司、元气森林(北京)食品科技集团有限公司、海南京润珍珠生物技术股份有限公司、浙江百姿化妆品股份有限公司、佳格食品(中国)有限公司、中国绿色食品有限公司、河南腾云实验室科技创新有限公司、北京君翌科技有限公司。

本文件主要起草人：汪厚银、钟葵、赵镭、史波林、李洪亮、苏玉芳、温烃、刘桂荣、安志丛、乔学义、李懿霖、费雅君、项雅科、唐正红、康东方、侯姣靓、姜兴涛、罗霞、张逸、周梅、赵菲菲、王飞飞、骆主胜、徐波、赵毅、曹海磊、牟善波、周滢、朱丹晔、张立国、雷翠婷、黄泽婷、张顶武、路会丽、杨宇阁、孙阳恩、戴跃锋、韩婕珺、施威、黄志明、宋义运、刘晶晶、田师一、吴薇、陈亚非、胡家逢、陈正鹤、王绪瑶、张璐、高飞、胡力、周朔、王旭辉、李思、唐飞、李伟、张默。

# 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评价员个体表现进行评估的导则。

本文件适用于对评价员个体或小组整体培训效果的验证和已建立评价小组表现的管理，以及评价小组及评价员个体对不同产品的辨别能力、同一小组内不同评价员之间的一致性以及评价员在属性强度评分中重复性的表现监测和评估。

本文件不适用于下列情况：

——使用共识性剖面（CP）、自由选择剖面（FCP）、自选特性排序剖面（FP）和动态主导属性测试（TDS）等描述性方法的小组表现评估，包括不记录评价员个体评分，没有对所有评价员通用的单一属性列表，或者评价员只指出主导属性而并不测量特性强度等；

——评价小组再现性的表现评估，包括评价小组之间的比较以及同一评价小组在不同条件下（如不同时间）进行的多次评价的比较。

本文件列出的方法并非全覆盖。评价小组组长（或感官分析师）全部使用或部分使用本文件中给出的方法持续评估感官小组或评价员个体的表现，或使用其他适合的方法。

## 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 界定的以及下列术语和定义适用于本文件。

ISO 和 IEC 维护的用于标准化的术语数据库网址如下：

——ISO 在线浏览平台：https://www.iso.org/obp；

——IEC 电工百科：https://www.electropedia.org/。

3.1

**一致性　agreement**

给同一组产品的某一属性评分时，不同评价小组或评价员表现出相同产品差异的能力。

3.2

**评价小组漂移　panel drift**

随着时间推移，因评价员敏感性发生变化或开始易受偏差的影响，导致评价小组对某个恒定参比

样品特定属性强度的评分在标度上的位置发生偏移的现象。

3.3

**表现　performance**

一个小组或一个评价员对刺激及其相应属性进行可靠且有效评价的能力。

3.4

**验证　validation**

证实一个评价小组或评价员能满足特定表现(3.3)要求的过程。

3.5

**轮次　session**

进行产品感官评价的时段。

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

[来源：GB/T 10221—2021，6.63]

3.6

**重复　replicate**

试验设计中某特定条件出现的次数。

注1：该术语通常指某条件会出现多次，但也可能只出现一次。当某一条件出现两次时，表述为“两次重复”，以此类推。

注2：为了明确表述某一条件的多次出现，术语“复制(replication)”或“复制轮次(replicate session)”会更明确。

注3：“复制轮次”是指评价员、产品、测试条件和任务都相同的轮次(3.5)。

3.7

**评价员偏差　assessor bias**

一种评价员始终给出高于或低于已知真值或小组平均值的评分的倾向。

[来源：GB/T 10221—2021，3.40]

3.8

**序列偏差　order bias**

由一个样品在一组测试样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　repeatability**

在相同测试条件下，同一评价员或评价小组对同一测试样品评价结果的一致性(3.1)。

注：重复性测量是在一个或几个明显独立的轮次(3.5)中进行，且重复(3.6)评价的测试条件是相同的。有时重复评价是在明显不同的轮次/试验中进行，而且这些轮次只相隔几天。这种情况下，短期内的重复性和再现性(3.10)之间的区别很小，主要与测试条件相同或不同有关。

[来源：GB/T 10221—2021，3.45，有修改]

3.10

**再现性　reproducibility**

在不同测试条件下，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：再现性通过以下方法测定：

a)评价小组(或评价员)的短期再现性，以天为间隔的两轮或多轮之间感官评价结果的一致性；

b)评价小组(或评价员)的中长期再现性，以月为间隔的不同轮次之间感官评价结果的一致性；

c)不同评价小组间的再现性，不同评价小组在同一实验室或不同实验室获得的感官评价结果的一致性。

[来源：GB/T 10221—2021，3.46]

## 4 原则

### 4.1 两种可行的方式

#### 4.1.1 概述

本文件涉及对产品一个或多个感官属性强度进行评价以建立产品定量描述或定量剖面的感官评价小组的表现评估(参见 ISO 13299)。差别检验评价小组的表现评估选用其他适宜的方法。

定量感官评价小组的表现评估,可采用为表现评估而专门开展的多轮次小组测试(称为“专用程序”),或使用已有的评估方法(称为“持续监测”)。

#### 4.1.2 通过专用程序进行表现评估

专用程序是评价员个人资格和其他目的验证的首选方法。对于资格的重新验证,宜根据需要定期重复该专用程序,具体程序见图 1。

此法通常在小组培训阶段结束时使用,以确保小组和评价员个体的表现达到了预期水平,同时能基于表现评估指标认定其是优选感官评价员或专家感官评价员。

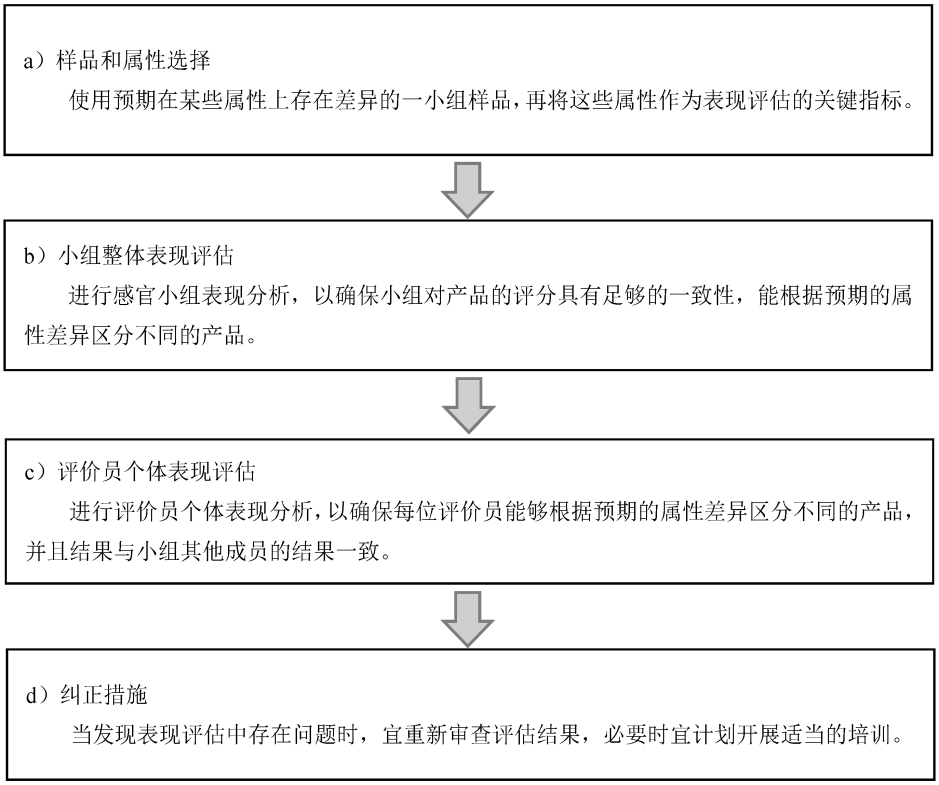



**图 1 采用专用程序进行表现评估的步骤**

#### 4.1.3 通过常规产品剖面进行持续监测

另一种方法是监测已收集的产品剖面数据。为了评估小组持续产出的剖面数据,适合使用不同剖面试验(如使用不同类型产品、不同数量的产品等)产出的数据。程序步骤与图 1 相同。然而,由于没有预设的产品属性差异,建议将小组在一个给定的剖面分析中整体能显著区分产品间差异的属性作为

考察评价员个体表现的关键测量依据。产品间不存在显著差异的属性不能用于一致性的考察,因为评价员自身或评价员之间在这些属性上缺乏一致性可能意味着产品在这些特性上非常相似。

在这种情况下,在给定的一段时间内,有必要选用一组在这些属性上产品间的差异大于评价小组实际能识别出的差异的样品进行考察。

### 4.2 感官小组或评价员个体表现的指标

对于一次评价,能确定的指标如下:

——评价小组的辨别力,以评价小组表现出的区分产品间存在显著差异的能力来衡量;

——评价员个体的辨别力,以评价员表现出的区分产品间存在显著差异的能力来衡量;

——评价员的一致性,以该评价员评价产品的评分均值与评价小组的评分均值之间的一致性程度来衡量;

——评价小组的一致性,以(小组内)每个评价员评价产品的评分均值的一致性程度来衡量。

对于重复评价,能确定的指标如下:

——评价员的重复性,以该评价员对同一产品重复评价结果的一致性程度来衡量;

——评价小组的重复性,以(小组内)每个评价员对同一产品重复评价结果之间的平均一致性程度来衡量。

### 4.3 统计分析

本文件描述了一种单一和常用的结果统计分析方法。然而,评价小组某些表现指标可通过不止一种方式来评价。例如,误差均方和误差标准差(误差均方的平方根)均能衡量产品评价结果的变异性。宜采用实际应用中通常使用的衡量指标。

在使用标度评价某一属性时,评估评价员间一致性的衡量指标还包括分析评价员与产品的交互效应以及评价员分值与小组均值的相关系数。评价员自身可能没有偏差,但不同的评价员可能以不同的方式使用标度。相关系数接近 1、回归斜率接近 1 以及回归截距接近 0,均表明评价员与评价小组的其他成员之间具有良好的一致性。

当评价员评价的样品数量较少(少于 6 个)时,宜谨慎解释相关系数,因为获得一个较高的相关系数(最高达 0.7)可能具有偶然性。

## 5 前提要求

### 5.1 试验条件

感官评价设施宜遵循 ISO 8589 的要求。

### 5.2 评价员资格

评价小组中评价员宜符合或高于 ISO 8586 中优选/初级评价员的资格和经验的要求。

## 6 采用专用程序的表现评估

### 6.1 样品和属性选择

每次使用专用程序时,宜向小组提供一组与待评价产品相似的样品,且针对每个相关属性,预期至少在一对样品之间存在统计上的显著差异。

为确保产品所有关键属性都被评价,测试中宜包含足够多样化的属性集。

这些相关属性被作为衡量评价小组表现的关键指标。样品组宜包含重复,且每个样品重复次数宜相同。重复可在一个、2 个或多个轮次中进行。评价员数量、样品数量和重复的次数都取决于产品、被评价的感官属性和试验的目的,如 3 个或 4 个样品可采用 2 次或 3 次重复。宜控制每轮测试中的评价次数,以避免感官疲劳。样品属性的量值范围宜与评价小组评价产品时所涉及的属性量值范围相近。

### 6.2 试验设计

#### 6.2.1 概述

基于要实现的最重要的目标,在专用程序中使用多种类型的试验设计。

#### 6.2.2 随机化区组设计

采用随机化区组试验设计,设计中将评价员作为“区组”。这种设计适合于相邻的样品之间不存在延滞效应的情况。如果存在延滞效应,则宜考虑采用平衡试验设计(见 6.2.3)。

#### 6.2.3 平衡和随机设计

若预期相邻的样品之间存在延滞效应,则适合采用威廉姆斯拉丁方设计[5],表 1 展示了 4 个评价员和 4 个样品的威廉姆斯拉丁方设计。

**表 1 4 个评价员和 4 个样品的威廉姆斯拉丁方设计**

| 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 |

在该设计中,每个评价员在一个给定轮次中以不同的顺序评价 4 种产品。对于每个评价员,测试任一特定产品后,接着都会测试一个不同的产品。例如,在轮次 1 中,完成了 A 样品测试后,评价员 1 接着测试样品 B,评价员 2 接着测试样品 C,评价员 3 接着测试样品 D,而对于评价员 4,则在评价完 A 样品后就完成了测试。

产品重复评价时,建议为每个评价员提供不同的产品顺序,以减少序列效应和延滞效应。

如果评价员人数是 4 的倍数,则可对每组 4 名评价员重复相同的设计。

也可选择一个随机的产品顺序设计,如在每个轮次中每个样品随机出现在每个位置。

以上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地估计产品均值,以进行表现评估。然而,如果确实存在产品顺序效应,则评价员之间的一致性将受到影响,因为每个评价员不会按照相同的产品顺序评价产品。为了在完全相同的任务上比较评价员,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 相同顺序设计

为了关注评价员个体表现,并尽可能在最相似的条件下比较评价员的表现,还可采用一种替代试验设计,即所有评价员按照相同的顺序评价产品,见表 2。

**表 2 4 名评价员和 4 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 |

需要注意的是,该设计下,评价员不是在评价产品本身,而是评价给定位置的产品(产品和位置效应被混淆)。这将导致对产品效应的估算存在序列偏差(因序列效应而产生的偏差),但对于评价员效应和产品×评价员交互效应的估算不会产生偏差。

## 6.3 统计分析

表 3 列出了一种结果整理与总结的方法。有些计算机软件可能有不同的数据排列要求,如样品按列排列,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 | | … | $j$ | | … | $n_q$ | | |
| | 分值 | 均值 | | 分值 | 均值 | | 分值 | 均值 | |
| 1 | $Y_{111}$<br>$Y_{112}$<br>…<br>$Y_{11n_r}$ | $\overline{Y}_{11.}$ | | $Y_{1j1}$<br>$Y_{1j2}$<br>…<br>$Y_{1jn_r}$ | $\overline{Y}_{1j.}$ | | $Y_{1n_q1}$<br>$Y_{1n_q2}$<br>…<br>$Y_{1n_qn_r}$ | $\overline{Y}_{1n_q.}$ | $\overline{Y}_{1..}$ |
| … | | | | | | | | | |

**表 3　评价员对一个属性的评价结果（续）**

| 样品 | 评价员 | | | | | | | | 均值 |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | | … | $j$ | | … | $n_q$ | | |
| | 分值 | 均值 | | 分值 | 均值 | | 分值 | 均值 | |
| $i$ | $Y_{i11}$<br>$Y_{i12}$<br>…<br>$Y_{i1n_r}$ | $\overline{Y}_{i1.}$ | | $Y_{ij1}$<br>$Y_{ij2}$<br>…<br>$Y_{ijn_r}$ | $\overline{Y}_{ij.}$ | | $Y_{1n_q1}$<br>$Y_{1n_q2}$<br>…<br>$Y_{in_qn_r}$ | $\overline{Y}_{in_q.}$ | $\overline{Y}_{i..}$ |
| … | | | | | | | | | |
| $n_p$ | | | | | | | | | $\overline{Y}_{n_p..}$ |
| 均值 | $\overline{Y}_{.1.}$ | | | $\overline{Y}_{.j.}$ | | | $\overline{Y}_{.n_q.}$ | | $\overline{Y}_{...}$ |
| 在此表中假定：<br>$n_p$=样品数($i$=1,2...$n_p$)；<br>$n_q$=评价员人数($j$=1,2...$n_q$)；<br>$n_r$=每个样品重复次数($k$=1,2...$n_r$)。 | | | | | | | | | |

除偏差以外，对评价小组整体和评价员个体的表现评估需要采用方差分析(ANOVA)来分析数据[6]。

鉴于统计分析通常依赖计算机软件进行，本文件并未展示基础计算的具体细节。

当评估单个轮次中的重复性时，采用单因素方差分析(ANOVA)处理每位评价员的数据(见表 4)。

若评估独立的多轮次间的重复性，根据研究人员/实验室的标准规程，可使用单因素方差分析(即样品因素)或双因素方差分析(即轮次和样品因素)(见表 4 和表 5)。

**表 4　评价员个体评价单个属性的方差分析**

| 变异来源 | 自由度 | 平方和 | 均方 | $F$值 |
|---|---|---|---|---|
| 样品间误差 | $v_1=n_p-1$<br>$v_2=n_p(n_r-1)$ | $S_1$<br>$S_2$ | $MS_1=s_1/v_1$<br>$MS_2=s_2/v_2$ | $F=MS_1/MS_2$ |
| 总体 | $v_3=n_pn_r-1$ | $S_3$ | | |
| $n_p$=样品数；<br>$n_r$=每个样品重复次数。 | | | | |

**表 5　考虑轮次效应的评价员个体评价单个属性的方差分析**

| 变异来源 | 自由度 | 平方和 | 均方 | $F$值 |
|---|---|---|---|---|
| 样品间<br>轮次间<br>误差 | $v_1=n_p-1$<br>$v_4=n_s-1$<br>$v_5=n_p(n_r-1)-(n_s-1)$ | $S_1$<br>$S_4$<br>$S_5$ | $MS_1=s_1/v_1$<br>$MS_4=s_4/v_4$<br>$MS_5=s_5/v_5$ | $F=MS_1/MS_5$<br>$F=MS_4/MS_5$ |
| 总计 | $v_3=n_pn_r-1$ | $S_3$ | | |
| $n_p$=样品数；<br>$n_s$=轮次数；<br>$n_r$=每个样品重复次数。 | | | | |

完整数据集采用随机化区组方差分析进行分析(见表 6)。

评价员效应可被认为是固定或是随机的[7]。为了评估评价员表现，通常将评价员效应固定，因为关注的重点是感官评价中具体评价员的表现。但是，为了更好地预测在实际评价条件下的表现，也可将评价员效应设定为随机效应（见表 6 和表 7 的脚注）。

当评估单个轮次中的重复性时，采用双因素方差分析处理完整数据集（见表 6）。

若评估独立多轮次间的重复性，根据研究人员/实验室的标准规程，可使用双因素方差分析（小组成员和样品因素）或三因素方差分析（小组、轮次和样品因素）（见表 6 和表 7）。

附录 A 提供了一个实际应用示例。

**表 6　单属性完整数据集（带重复的）双因素方差分析**

| 变异来源 | 自由度 | 平方和 | 均方 | $F$ 值 |
|---|---|---|---|---|
| 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] |
| 交互效应 | $v_8=(n_p-1)(n_q-1)$ | $S_8$ | $MS_8=s_8/v_8$ | $F=MS_8/MS_9$ |
| 误差 | $v_9=n_pn_q(n_r-1)$ | $S_9$ | $MS_9=s_9/v_9$ | |
| 总计 | $v_{10}=n_pn_qn_r-1$ | $S_{10}$ | | |
| $n_p$=样品数；<br>$n_q$=评价员人数；<br>$n_r$=每个样品重复次数。 | | | | |
| [a] 公式中评价员效应是固定的。若评价员效应是随机的，则样品间效应 $F$ 值的分母由 $MS_9$ 替换为 $MS_8$。 | | | | |

**表 7　考虑轮次效应的单属性完整数据集（带重复的）三因素方差分析**

| 变异来源 | 自由度 | 平方和 | 均方 | $F$ 值 |
|---|---|---|---|---|
| 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] |
| 交互效应 | $v_8=(n_p-1)(n_q-1)$ | $s_8$ | $MS_8=s_8/v_8$ | $F=MS_8/MS_{12}$ |
| 误差 | $v_{12}=n_pn_q(n_r-1)-(n_s-1)$ | $s_{12}$ | $MS_{12}=s_{12}/v_{12}$ | |
| 总计 | $v_{10}=n_pn_qn_r-1$ | $s_{10}$ | | |
| $n_p$=样品数；<br>$n_q$=评价员人数；<br>$n_r$=每个样品重复次数；<br>$n_s$=轮次数。 | | | | |
| [a] 给定公式中评价员效应是固定的。若评价员效应是随机的，则样品间效应 $F$ 值的分母由 $MS_{12}$ 替换为 $MS_8$。 | | | | |

## 6.4　评价小组整体表现——统计结果的解释

### 6.4.1　关键属性辨别

宜确定预期能被显著区分的关键属性的比例。在完整数据集的方差分析表中，用 $\alpha$ 为 0.05 的水平来表示不同样品间在每个属性的显著差异（见表 6 和表 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，评价小组宜接受进一步培训。

### 6.4.2 评价小组的一致性

若评价小组内任一评价员与其他成员存在不一致时，则表明小组未达到一致性（见 6.5.4）。

如果在方差分析中，样本与评价员之间的交互效应在 $\alpha$ 水平为 0.05 时显著，则表明小组未达成一致。

评价小组的一致程度与交互效应项 $s_i$ 呈负相关，见公式（1）：

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

详见表 6 和表 7。

通过方差分析（见表 6 和表 7）确定样品与评价员之间存在显著交互效应的关键属性数量。能被显著区分的关键属性的比例越高，小组表现的一致性就越低。如果交互效应显著时，宜在小组成员层面上研究交互效应的性质，并在需要时采取措施。例如，如果某一评价员对产品间差异的评价结果与小组其余成员的不一致，则该评价员宜重新进行培训。

通常，通过绘制评价员的产品均值图表来研究交互效应的性质。另一种方法是采用混合评价员模型（MAM）[8,9]。

### 6.4.3 评价小组的重复性

评价小组的重复性通过评价员个体的重复性估计，与误差项 $s_e$ 呈负相关，见公式（2）：

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

取决于所选的模型（无论是否有轮次效应）。

详见表 6 和表 7。

## 6.5 评价员个体表现——统计结果的解释

### 6.5.1 评价员个体的辨别力

评价员个体的辨别力通过辨别出的预期有显著差异的关键属性所占比例来衡量。方差分析表中，不同样品在每个属性上存在差异的显著水平设为 0.05（见表 4 和表 5）。被辨别出的预期关键属性比例越高，表明评价员表现越好。当评价员不能对预期有显著差异的关键属性进行辨别时，宜接受进一步培训。

### 6.5.2 评价员个体的重复性

评价员的重复性与评价员的误差项 $s_e$ 呈负相关，见公式（3）：

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

取决于所选的模型（无论是否存在轮次效应）。

详见表 4 和表 5。

### 6.5.3 评价员个体的一致性

评价员个体的一致性与从每个样品中计算出的偏差项标准差（SD）呈负相关。

（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值，即 $\overline{Y}_{ij.}-\overline{Y}_{i..}$，见表 3）。

当评价员的表现缺乏一致性时，绘制该评价员分值和小组均值的散点图，结合回归和相关性分析，说明这种不一致性结果是随机的还是由于该评价员与小组其他成员使用标度方式的不同所导致的。

#### 6.5.4 评价员之间的一致性

当至少有一名评价员与评价小组的其他成员不一致时，该小组就不具有均匀性。

可通过下列方式检测：

——某评价员有显著偏差；

——某评价员的残差项显著大于小组整体的残差项；

——评价员分值与小组均值之间的相关系数非常小或为负值；

——评价员分值和小组均值的回归曲线的斜率与 1 差异显著，或截距与坐标原点差异显著，或两者兼具。

评价员之间的一致性与“评价员间”的误差项 $s_a$ 呈负相关，见公式(4)：

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

评价员之间的不一致宜用“评价员之间”的 $F$ 值进行显著性检验，并将其与自由度查表所得的 $F$ 值进行比较。若检验结果显著，则充分表明评价小组存在一致性问题，需要采取解决措施。若结果不显著，也不能保证小组一致性没问题，因为小组的一致性问题可能会被较差的重复性(高于预期的误差项 $s_e$)所掩盖。

#### 6.5.5 标度使用方式的偏差

评价员偏差的方差分析结果显著表明评价员可能在标度使用方式上存在差异。

大多数情况下，“真”值未知，评价员个体的整体偏差被认为是该评价员的均值与小组均值的差值。评价员 $j$ 的偏差由公式(5)给出：

$$\overline{Y}_{j\cdot}-\overline{Y}_{\cdots} \qquad \cdots\cdots(5)$$

评价员可采用不同的方式使用标度(参见 ISO 4121)。使用“通用”标度时，产品属性的评分与评价员自身从测试产品类型中体验到的整体感官差异有关。只评价一种或少数几种产品品类的评价小组通常采用此方法。使用“相对”标度时，评价员进行强度评价的参考模式与测试中给出的样品集所表现出的感官差异大小有关。这种方式更适合对宽泛品类产品进行评价的小组。同一个小组内，采用一致的标度使用方式对减小标度偏差十分重要。

### 6.6 评价小组及评价员表现的相关问题

#### 6.6.1 概述

一旦发现评价小组及评价员个体表现中的问题，将问题列出并制定相应的培训计划。

#### 6.6.2 评价小组

对于出现问题的属性，组织整个小组开展针对性的培训。

#### 6.6.3 评价员个体

对于评价员个体表现中存在的具体问题，可使用中立或积极的语气，一对一地私下讨论问题所在，如可提供评价员个人与小组平均水平的对比数据，然后再开展整个小组的培训计划。

### 6.7 长期表现追踪的试验设计

如果计划研究一段时间内评价小组表现的一致性，一年内持续每月一次的感官评价试验能提供足够的数据。宜按 6.2.3 所示，对每次感官评价试验进行平衡设计。持续跟进一段时间后，根据表现评估

结果再确定后续方案，如评价小组漂移或再培训后的能力提升。

## 7 通过常规产品剖面分析进行持续监测的程序

### 7.1 属性选择

属性选择与采用专用程序进行表现评估中的一致（见第 6 章）。然而，由于没有预设差异，建议将小组在一个既定剖面分析中整体能显著区分产品间差异的属性作为考察评价员个体表现的关键测量依据。产品间没有显著差异的属性不能用于一致性的考察，因为评价员自身或评价员之间在这些属性上缺乏一致性可能意味着产品间在这些特性上非常相似。

### 7.2 统计分析

统计分析中，除了评价员效应宜始终被设定为是随机因素外，其余的与采用专用程序进行表现评估的测试相似（见第 6 章）。

### 7.3 长期表现追踪

如果已有了几个轮次的常规评价数据，即可对数据进行分析以展示持续跟进的这段时间内的变化。根据表现评估结果再确定后续方案，如评价小组漂移或小组再培训后的表现能力提升。

### 7.4 长期表现追踪的统计分析

宜使用重复测量的方差分析（ANOVA）对多轮次的数据进行整体分析。实际测试中，同一评价员可能不会参加所有轮次的感官评价，通常需要使用一般线性模型的单变量方差分析来获得每个评价员的偏差、其他参数和方差分子的无偏估计。

对于小组，能得到下列 2 个估计。

a) 小组一致性的估计：如果在多个轮次中有收集同一对照样的数据，小组一致性可通过轮次效应进行估计（见表 7）。

b) 内部一致性的估计：当评价员个体发生偏差时，评价员和轮次之间的交互效应可评估其稳定性。

对于评价员个体，每个属性能得到下列 3 个估计。

a) 总体偏差：多次重复和/或轮次中，评价员分值和相应的小组整体均值之差的平均值。

b) 一致性：与不同轮次之间偏差项的变化呈负相关。

c) 重复性：通过收集每个轮次的残差标准差的估计值来确定相同样品得分之间的差异。

### 7.5 完整剖面数据的统计分析

6.4 和 6.5 中描述的统计分析方法分别应用于每个属性来评估小组和评价员在需作答属性（问题）时的表现。

此外，为了获得数据的整体概况，可使用多维分析技术，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验证或持续监测。这些多维分析技术的更多详细信息见参考文献[10]～[12]。

# 附 录 A
（资料性）
应用示例

## A.1 数据表

在一个轮次中，4 名评价员给样品的一个属性评分，6 个样品重复 3 次。结果如表 A.1 所示。

注：这里仅是一个展示示例，通常会有 4 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 | | |
| | 分值 | 均值 | 分值 | 均值 | 分值 | 均值 | 分值 | 均值 | |
| 1 | 8 | 8.3 | 5 | 7.3 | 6 | 6.0 | 9 | 8.3 | 7.50 |
| | 8 | | 8 | | 7 | | 8 | | |
| | 9 | | 9 | | 5 | | 8 | | |
| 2 | 6 | 7.0 | 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 | 4.0 | 5 | 5.0 | 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 | 4.0 | 3 | 3.0 | 4 | 4.3 | 4 | 4.3 | 3.92 |
| | 5 | | 2 | | 4 | | 5 | | |
| | 3 | | 4 | | 5 | | 4 | | |
| 6 | 5 | 5.7 | 4 | 4.3 | 5 | 5.0 | 7 | 6.3 | 5.33 |
| | 6 | | 2 | | 4 | | 5 | | |
| | 6 | | 7 | | 6 | | 7 | | |
| 均值 | 5.89 | | 4.83 | | 4.67 | | 6.00 | | 5.35 |

## A.2 统计分析

本示例详细的方差分析结果见表 A.2、表 A.3、表 A.4 和表 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 | 均方 | *F*值 |
|---|---|---|---|---|
| 样品间 | 5 | 104.90 | 20.98 | 16.39[a] |
| 评价员间 | 3 | 26.04 | 8.68 | 6.79[a] |
| 交互效应 | 15 | 16.04 | 1.07 | 0.84[b] |
| 残差 | 48 | 61.33 | 1.28 | |
| 总计 | 71 | 208.31 | | |
| [a]显著性水平=0.05时显著。<br>[b]显著性水平=0.05时不显著。 | | | | |

**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 | |
| | | 均方 | *F*值 | 均方 | *F*值 | 均方 | *F*值 | 均方 | *F*值 |
| 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 | |
| 残差标准差 | | 0.75 | | 1.71 | | 1.08 | | 0.67 | |
| [a]显著性水平=0.05时显著。<br>[b]显著性水平=0.05时不显著。 | | | | | | | | | |

**表 A.4　评价员个体的偏差和残差标准差**

| 评价员 | 偏差 | 残差标准差 |
|---|---|---|
| 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**：偏差是评价员个体均值和小组整体均值的差值,两者均在表A.1中给出。 | | |

**表 A.5　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 |

表 A.5　样品偏差项（续）

| 样品 | 评价员 | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 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 |
| 偏差标准差 | 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　小组整体表现——对统计结果的解释

由表 A.2 可见，“样品间”效应显著（$\alpha=0.05$），表明评价小组在不同产品的差异区分上表现出较好的一致性。同时，交互效应在 $\alpha=0.05$ 水平时并不显著，表明小组成员之间在产品差异上没有表现出显著的不一致性。

“评价员间”效应显著，表明评价员给产品（所有产品中）的平均分值存在差异。评价员均值的变化程度可用评价员的标准差表示。

本示例中，如果评价员效应是随机的、非固定的，也会得出相同结论。但是，如果产品和评价员的交互效应显著时，可能会有不同的解释。

## A.4　评价员个体表现——对统计结果的解释

### A.4.1 概述

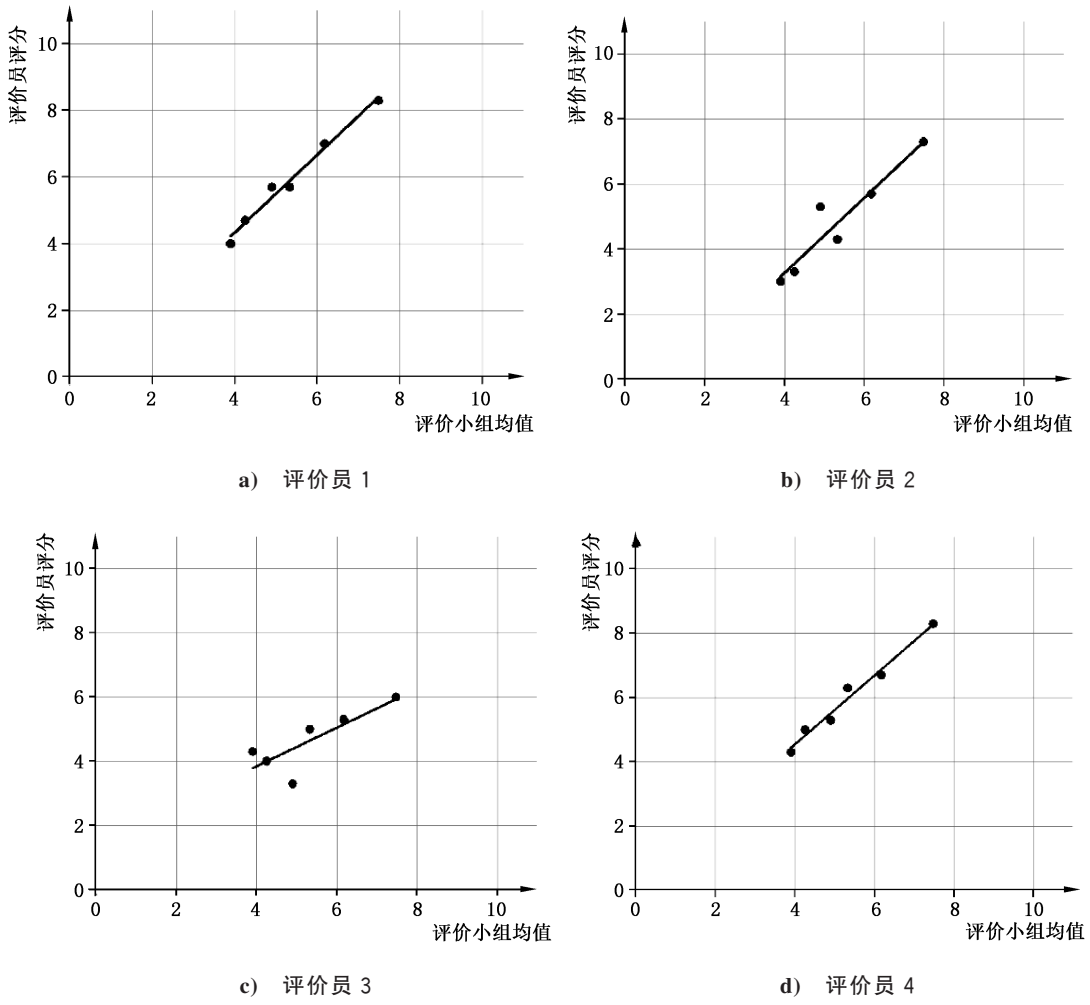
评价员 2 和评价员 3 的残差标准差最高（见表 A.4），表明在同一样品的重复评价中，他们的重复性低于评价员 1 和评价员 4。

评价员 3 具有较高的负偏差，表明倾向于给出比小组其他成员更低的分值。该评价员的一致性也不如其他的小组成员，偏差项分布范围从−1.58～0.42（见表 A.5），同时具有最大的偏差标准差（0.78）。

评价员 4 具有较高的正偏差（0.65），但其偏差标准差仅为 0.24。由于评价员 1 和评价员 4 的评分一致性好且偏差值低，因此他们的评分是可信的。同时，评价员 2 和评价员 3 的结果降低了小组均值，因此评价员 4 的“偏差”无需关注。

### A.4.2 回归性和相关性的统计分析

图 A.1 显示了针对 6 个产品每个评价员评分与小组均值的散点图。




a) 评价员 1　　b) 评价员 2

c) 评价员 3　　d) 评价员 4

**图 A.1　评价员 1/2/3/4 与小组平均值关系图**

在本示例中,评分没有“真”值。小组均值作为每个评价员的参照分值。理想的图谱是一个显示评价员分值和小组均值完全一致的图,数据点接近在一条斜率为 1、截距为 0 的直线上,相关系数宜接近 1。

4 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,其相关系数接近于 1,斜率接近于 1,截距最小。

评价员 3 的斜率较小,表明相比其他评价员,其使用的标度范围更窄。

评价员 2 的截距为负,表明存在负偏差。

## A.5 表现评估的其他问题

### 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 评价员个体

图 A.2～图 A.4 展示了小组中评价员个人表现比较的 3 个示例。

图 A.2 表明除了 1 名评价员，其他评价员对样品的区分具有良好的一致性。评价员 10 在样本之间几乎没有辨别能力，其余评价员在除样品 A 之外的所有样品中均表现出良好的一致性。

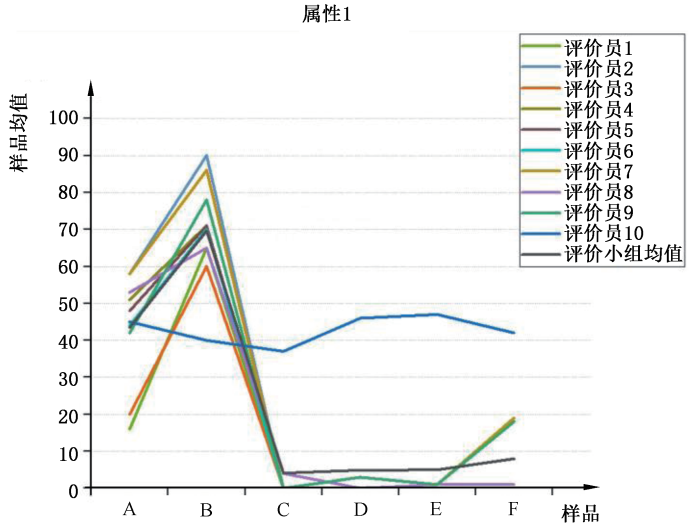



**图 A.2 10位评价员组成的感官小组在对6个样品的某一属性(属性1)的评分**

图 A.3 结果表明大多数评价员对 6 个样品的顺序结果均达成一致，但评价员 10 存在辨别力较差、标度使用范围小的情况。

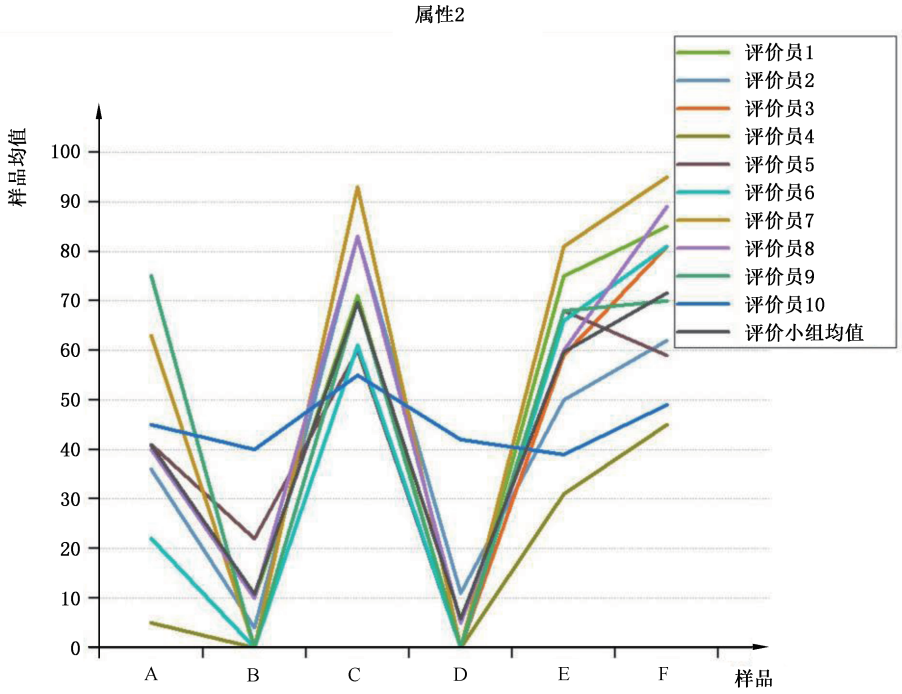



**图 A.3　10位评价员组成的感官小组在对6个样品的某一属性(属性2)的评分**

图 A.4 表明所有评价员在样品区分和标度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,其中有 2 名评价员对所有样品的评分都很低。

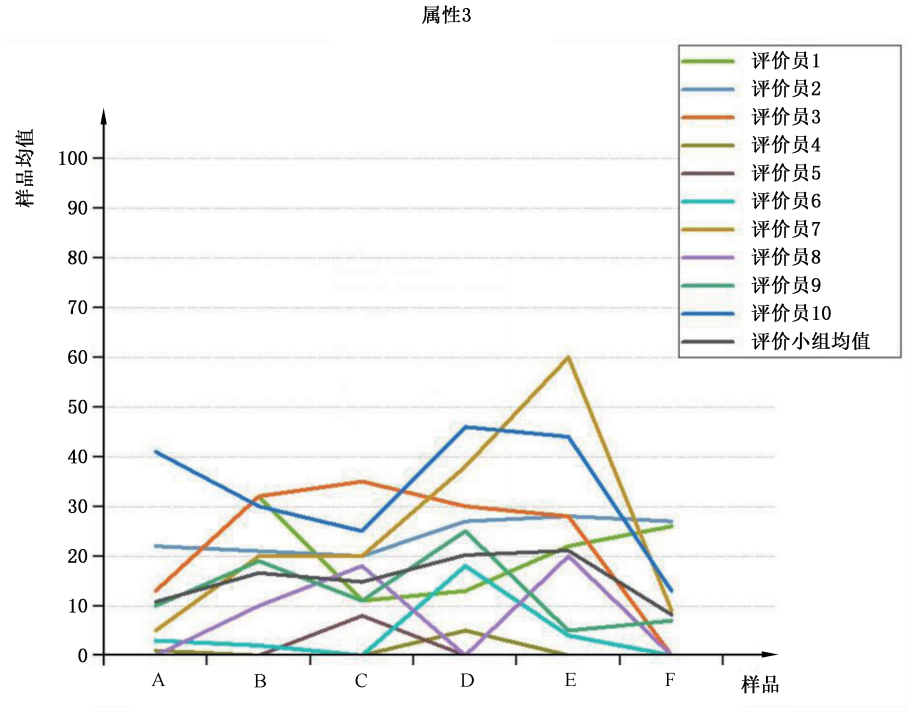



**图 A.4　10位评价员组成的感官小组在对6个样品的某一属性(属性3)的评分**

# 参 考 文 献

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

[2] ISO 8586 Sensory analysis—Selection and training of 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] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

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

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