What's the Difference?

Understanding and selecting sensory difference test methods

Lindsay Guerdrum, New Belgium Brewing Co.
Meghan Peltz, Oregon State University

The Science of Beer

Workshop Outline

Goal: Participants will learn how to report results of a difference test, which one to choose for your situation, and how to utilize the panel while minimizing biases.

- Introduction to Sensory to aid Quality Assurance
- Types of Sensory Difference Tests
- Understanding the Discrimination Report

Scenarios:
Break out session
30mins

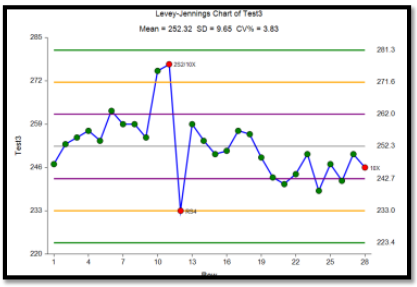
Practical Scenarios
Reporting Practice

- Conclusions

The Science of Beer

Thoughtfully charting and analyzing brewing results can ensure that a process and product are within control.

Time
BUs
ABV
Temperature
Color
Calories
Vol. CO_2
Gravity
VDKs
pH
AE

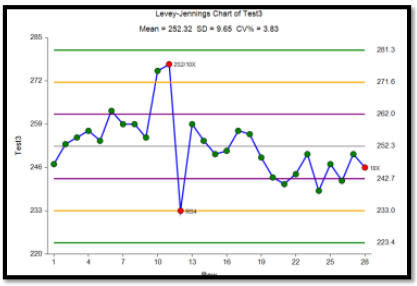


Sensory
Difference from Control
Competitive Analysis
Triangle
Release
True to Type



The Science of Beer

Maintenance and Methodology

- Panelist Training
- Validating your panel
- Maintaining a sensory panel

Webinar Resources
asbcnet.org

The Science of Beer

How Sensory is Used

- **Discrimination tests** – **"Is it different?"**
  - Quality assurance monitoring
  - Qualify process changes
  - Ingredient substitution testing
- **Descriptive tests** – **"How is it different?"**
  - Produces a *relative* product fingerprint
  - Uncover reason for difference for R&D or QA
- **Affective tests** – **"Does the difference matter?"**
  - Evaluate consumer response to a recipe or process change
  - Can be used for marketing purposes

Overview of Discrimination Tests

- **When to perform discrimination tests**
  - First step in your experimental design should be to answer the question: "Is there a statistically significant difference?"
    - » If there is no difference, descriptive and hedonic tests are irrelevant.
  - Specified v. unspecified difference tests
- **Assumptions**
  - All assessors have the same probability of discriminating
  - Observations are independent
- **Best Practices**
  - **Minimize bias**
    - » Randomized sample block
    - » Consider visual bias

Overview of Discrimination Tests

- **More Best Practices**
  - **Communication**
    - » Clear instructions without being too through
    - » "Right and wrong" vs. "correct and incorrect"
  - **Validation**
    - » Panelists must understand the question and have the skills to give an accurate answer.
  - **Choice of test**
    - » Number of panelists available
    - » Question to be answered
    - » Type of stimulus
  - **Keep subjectivity out of the sensory lab**

The Science of Beer

Duo-Trio Test

- **ASBC Sensory Analysis – 8**
- Presented a Reference and two coded samples.
  - A = Reference & 831
  - B = 355
- Option to have constant or balanced reference.
- 20+ assessors is recommended.
- Indicate the "sample that is different from the reference", even if it is only a guess.
- Guessing probability = **1/2**

"Identify the Reference."

The Science of Beer

Triangle Test

- **ASBC Sensory Analysis – 7**
- Three coded products, two are identical.
  - A = 230, 831
  - B = 648, 493
- Each samples takes a turn being the different sample.
  - AAB vs. ABB
- Typically, 18-36 assessors
- Indicate the "sample that is different", even if it is only a guess.
- Guessing probability = **1/3**

"Which is different?"

The Science of Beer

Alternative Force Choice (AFC) Test

- **ASBC Sensory Analysis – 6**
- 2-AFC or 3-AFC testing
- Two or three coded products, one is stronger.
  - A = 974
  - B = 223, stronger
- >15 assessors is recommended.
- Ask assessor to identify sample that has the most of a specified attribute. (ie. "saltiest", "sweetest")
- The guessing probability = **1/2** or = **1/3**

"Which is more ___?"

The Science of Beer

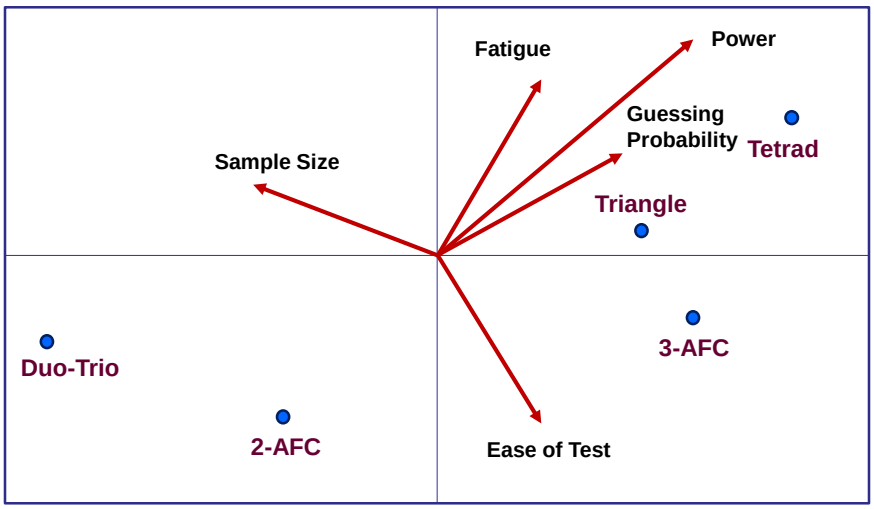
Tetrad Test

- **ASBC Sensory Analysis – NEW!**
- Four coded products, identically paired.
  - A = 940, 521
  - B = 782, 269
- Incorporates natural product variation into grouping exercise.
- 15-33 assessors is recommended.
- Ask assessor to "group samples into two groups of two based on similarity", even if it is only a guess.
- Guessing probability = **1/3**

"Group into two groups of two."

The Science of Beer

Choosing Between Discrimination Tests

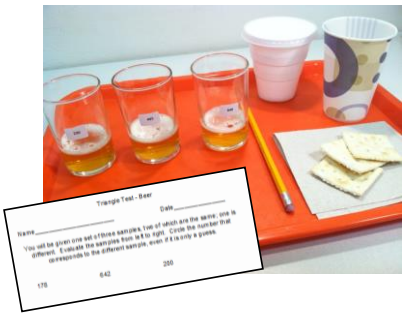


**Not measured statistically, for comparison purposes only*

The Science of Beer

Understanding Reported Values

- Hypothesis testing
- Type I and II error
- Power / confidence intervals
- Calculating P-values



Webinar: Intro to Sensory Statistics
asbcnet.org

Hypothesis Testing

- **Null Hypothesis (H_0)**
  - Typically, the simplest answer
  - $H_0 = M_A - M_B = 0$
  - No difference between beer **A** and **B**
- **Research/Alternative Hypothesis (H_a)**
  - A rejection of the null hypothesis

A ≠ = **B**

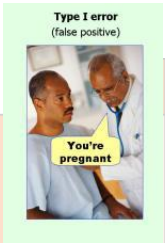
  - Be careful these two results are not equivalent!

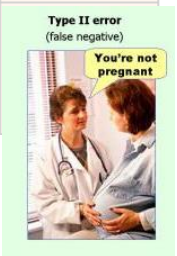
Type I and II Error

	H_0 True	H_0 False
Reject H_0	Type I Error, α ☹	Correct Rejection ☺
Fail to Reject H_0	Correct Decision ☺	Type II Error, β ☹

- **Type I error** – saying the samples are different when they are not. **Beer A and B truthfully were not different but said different.**
- **Type II error –** saying the samples are similar when they are not. **Beer A and B truthfully were different but said not different.**

Type I and II Error





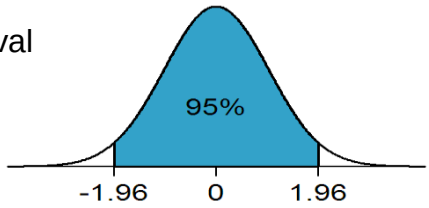
Source: flowingdata.com

- **Type I error** – saying the samples are different when they are not. **= Precision**
- **Type II error –** saying the samples are similar when they are not. **= Accuracy**

Power / Confidence Intervals

P-value is the probability of the Null Hypothesis occurring by chance

- 95% Confidence interval
- 5% Probability
- Alpha = 0.05



The Science of Beer

Statistical Tables

Number of Replies	Minimum Number of Correct Replies for a Significance Level of α ≤		
	0.05	0.01	0.001
18	10	12	13
19	11	12	14
20	11	13	14
21	12	13	15
22	12	14	15
23	12	14	16
24	13	15	16
25	13	15	17
26	14	15	17
27	14	16	18
28	15	16	18
29	15	17	19
30	15	17	19
31	16	18	20
32	16	18	20
33	17	18	21
34	17	19	21
35	17	19	22
36	18	20	22
48	22	25	27
54	25	27	30
60	27	30	33
66	29	32	35
72	32	34	38
78	34	37	40

Guessing Probability = 1/3

- Count correct responses compared to total.
- Sensory scientist determines α, β, P_d based on objective and business situation.
  - α, concluding a difference exists when there is none (typically, 5%)
  - β, risk of not finding an existing difference (typically, 5%)
  - P_d, percentage of population that can detect a difference (typically 25% or 35%)
- Can be calculated using a t-statistic, provides p-value.
- A chart can be used to determine significance, does not provide p-value.

The Science of Beer

Statistical Tables

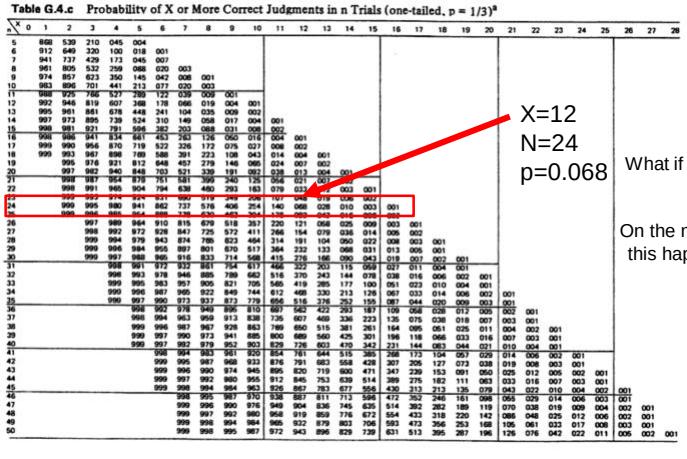
Number of Replies	Minimum Number of Correct Replies for a Significance Level of $\alpha \leq$		
	0.05	0.01	0.001
18	10	12	13
19	11	12	14
20	11	13	14
21	12	13	15
22	12	14	15
23	12	14	16
24	13	15	16
25	13	15	17
26	14	15	17
27	14	16	18
28	15	16	18
29	15	17	19
30	15	17	19
31	16	18	20
32	16	18	20
33	17	18	21
34	17	19	21
35	17	19	22
36	18	20	22
48	22	25	27
54	25	27	30
60	27	30	33
66	29	32	35
72	32	34	38
78	34	37	40

- **More than** 13/24 assessors correctly identify the "different" sample.
  - Products A and B are perceivably **different** at a 95% confidence level.
- **Less than** 13/24 assessors correctly identify the "different" sample.
  - Products A and B are **not** perceivably **different** at a 95% confidence level.
  - This does not mean they are the same!!!
- Need 48-78 assessors for similarity.
- **Less than** 21/48 assessors correctly identify the "different" sample.
  - Products A and B are perceivably **similar** at a 95% confidence level.

The Science of Beer

Calculating P-values

Table G.4.c Probability of X or More Correct Judgments in n Trials (one-tailed, p = 1/3)[a]



X=12
N=24
p=0.068

What if 12/24 identifications of the odd sample occur?

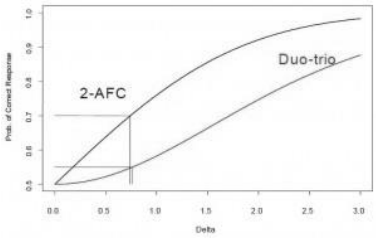
On the null hypothesis the probability of this happening on chance is **6.8%**…is this significant?

The Science of Beer

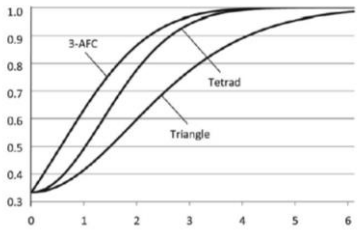
Advanced Statistical Methods

P-values cannot be used to compare across sensory discrimination test methods.

- Proportion of Distinguishers
  - Takes into account guessing probability
  - Cannot account for systematic differences between testing methods
- Thurstonian Scaling
  - Measures underlying sensory difference
  - Signal detection theory through the use of more tables



Ennis, Institute of Perception, Online, 2014

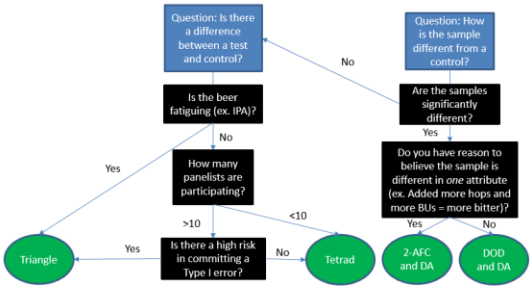


Ennis and Jesionka, Journal of Sensory Studies, 2011

The Science of Beer

Types of Discrimination Tests

The number of panelists needed is proportional to the power of the sensory method, assuming the size of the difference (effect size) and tolerable amount of risk in making a type 1 error (alpha) are the same.



	Power = 0.8			
Delta	2AFC	DT	TET	TRI
0.30	368	37,631	8,548	33,489
0.35	274	20,509	4,688	18,236
0.40	216	12,140	2,801	10,811
0.45	170	7,655	1,786	6,834
0.50	139	5,102	1,202	4,547
0.55	114	3,532	837	3,156
0.60	101	2,537	608	2,255
0.65	85	1,864	454	1,665
0.70	76	1,411	346	1,265
0.75	64	1,099	272	979
0.80	57	862	215	769
0.85	52	693	175	614
0.90	47	563	143	499
0.95	42	464	122	415
1.00	40	383	101	343
1.05	38	328	86	291
1.10	33	274	71	245

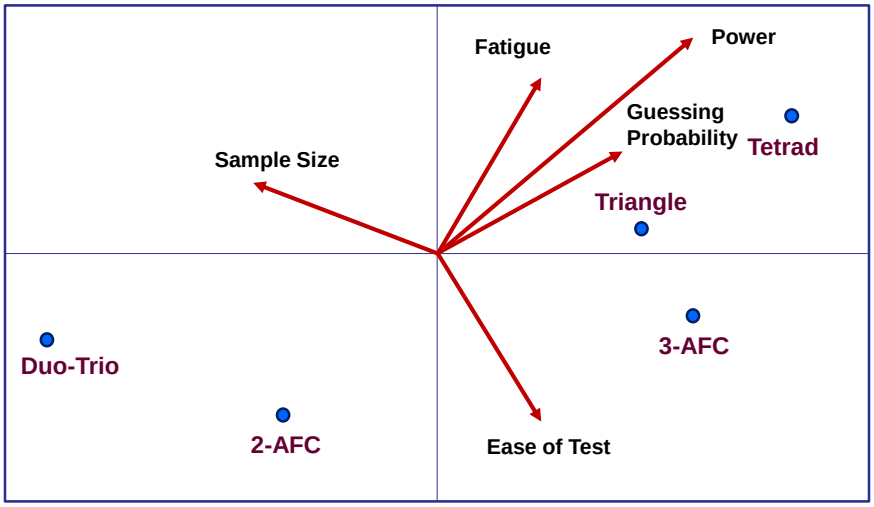
THE POWER OF SENSORY DISCRIMINATION METHODS REVISITED

JOHN M. ENNIS[1,3] and VIRGINIE JESIONKA[2]

[1]The Institute for Perception, 7629 Hull Street Road, Richmond, VA 23112
[2]O. P. & P. Product Research B.V., Utrecht, The Netherlands

The Science of Beer

Choosing Between Discrimination Tests



*Not measured statistically, for comparison purposes only

The Science of Beer

Discrimination Testing Examples

Packaging Change

Supplier/Raw Material Change

Processing conditions change

Ingredient Substitution

Ensure difference before further sensory tests

To select, train and monitor panelists

Shelf Life

Scenarios:
Break out session
30mins

The Science of Beer

Duo-Trio Scenario

- **Objective:** To qualify an alternative chocolate malt supplier
- **Results:**
  - **Set 1: 5/7** → **Not different** (need **7/7** for significant result!)
  - **Set 2: 6/7** → **Not different**
  - **Combined: 11/14** → **Statistically different**

- **Discussion:**
  - **Testing Choice**

    Need **20+** assessors for the duo-trio test
  - **Pooling Multiple Evaluations**

    "Avoid replicate evaluations by the same assessor whenever possible. However, if replications are needed to produce a sufficient number of total evaluations, every effort should be made to have each assessor perform the same number of replicate evaluations." - ASTM

Triangle Test Scenario

- **Objective:** Investigate effects of reduced boiling times
- **Results:**
  - **11/24** → **Not different**

    (need **13/24** for significance)

- **Discussion:**
  - **Testing Choice**
  - **Balanced Testing**
    - **7/12** identified the reduced boil time; **4/12** identified the 90 min boil when presented as the odd sample.
    - "Skewed" results are an indication of strength difference.
  - **Practical Significance**
    - Assessor's comments
    - **80 min boil** - "DMS" "Sulfur" "Canned corn" "Stronger" "More pronounced"
    - **90 min boil** - "More bitter" "Lingering"

Tetrad Test Scenario

- **Objective:** Observe the impact the crown caps (or don't) have on the rate of oxidation in your flagship amber ale.
- **Results:**
  - **19/32** Sorted the samples correctly → **Significantly different**
- **Discussion:** **Difference with the amber but not the IPA**
  - **Panelist fatigue**
    - Addition of the 4th stimulus should be considered when choosing a test especially when the samples are fatiguing.

"While it exhibits a greater power, the tetrad can potentially suffer from a decrease in performance linked to the addition of a fourth stimulus, compared to the three stimuli comprised in the triangle test." – Rie Ishii

  - **Rejecting the null hypothesis**
    - Failure to reject Ho should not be considered sufficient evidence to accept it by saying there is "no difference"

The Science of Beer

2-AFC Test Scenario

- **Objective:** Understand if adding double the amount of dry hops in the IPA has a significant positive (additive) impact on overall hop aroma.
- **Results:**
  - **10/20** Sorted the samples correctly → **Not Significantly different**
- **Discussion:** **No difference using the 2-AFC but difference using the triangle method**
  - **Specified v. non-specified**
    - The attribute to which the panel was being pointed may not be the source of the difference.
  - **Clear instructions**
    - Be as specific as possible with the attribute direction. For example, "Hoppy" could mean bitterness and/or aroma.
  - **Panel validation**
    - The panel may not understand the attribute to which they are being pointed.

The Science of Beer

Conclusions

Goal: Participants will learn how to report results of a difference test, which one to choose for your situation, and how to utilize the panel while minimizing biases.

- **Discrimination testing…**
  - Will tell you whether or not two samples are perceptibly different
  - Will *not* tell you the driver of the difference (unless specified)
  - Should be used *before* descriptive or hedonic tests are performed.
- **Considerations when choosing a test**
  - Power and risk of rejecting the null hypothesis
  - Panel size and replications
  - Type of product being tested and fatigue

The Science of Beer

Conclusions

- **Best Practices**
  - Sample order
  - Minimizing panelist bias
  - Clear instructions
  - Communication of results
- **Results**
  - Non-significant results can still be meaningful
  - Communication

Sensory

Results

Quality Assurance

Relevance

Next Steps

Product Release

The Science of Beer

Thank you!