# Investigating Sensory Equivalence of Reformulated Products

Benoît Rousseau[1] & John Cowden[2]
[1]The Institute for Perception, Richmond, VA
[2]General Mills, Minneapolis, MN

IFT | Wellness 13

# Five Factors for Sensory Matching

Benoît Rousseau & John M. Ennis
The Institute for Perception, Richmond, VA

IFT | Wellness 13

# Why Discrimination Testing?

- Measure the size of the difference between products
- Two main objectives
  - Prove products are different
    - "New and improved", "Fresher, crisper taste"

  - Prove products are equivalent
    - Ingredient change, new supplier, government regulation (e.g., salt or sugar reduction)

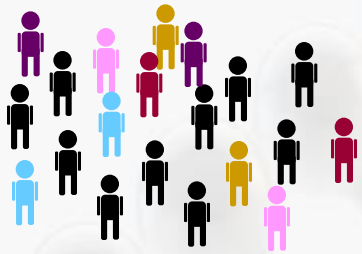
- Use discrimination testing to measure small sensory differences

# How Discrimination Testing?

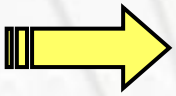
- Get samples to be compared

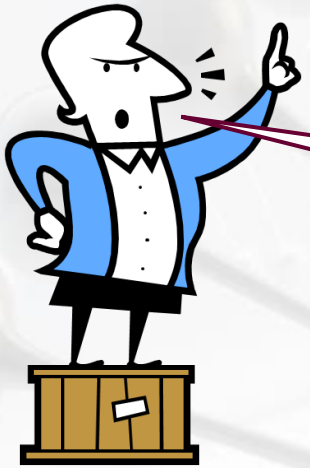

***"Which one is more salty?"***

- Get panelists



- Get results (e.g., $^{15}/_{20}$ correct)



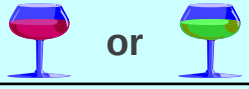



***"There are three kinds of lies: lies, damned lies, and statistics"***

# Common Sensory Discrimination Methods

- Many different protocols are available:

| *Information regarding the sample to be selected not required* | | |
|---|---|---|
| ❖ Triangle |  | Which one is different? |
| ❖ Duo-trio |  | Which one is the same as the reference ? |
| ❖ Tetrad |  | Group the samples into 2 groups of 2 identical samples |
| ❖ Same-different |  | Are they the same or different ? |
| ❖ … | | |

| *Information regarding the sample to be selected required* | | |
|---|---|---|
| ❖ 2-AFC |  | Which one is *more* … ? |
| ❖ Identification |  | Is it A or B ? |
| ❖ … | | |

# Sensory Discrimination Program

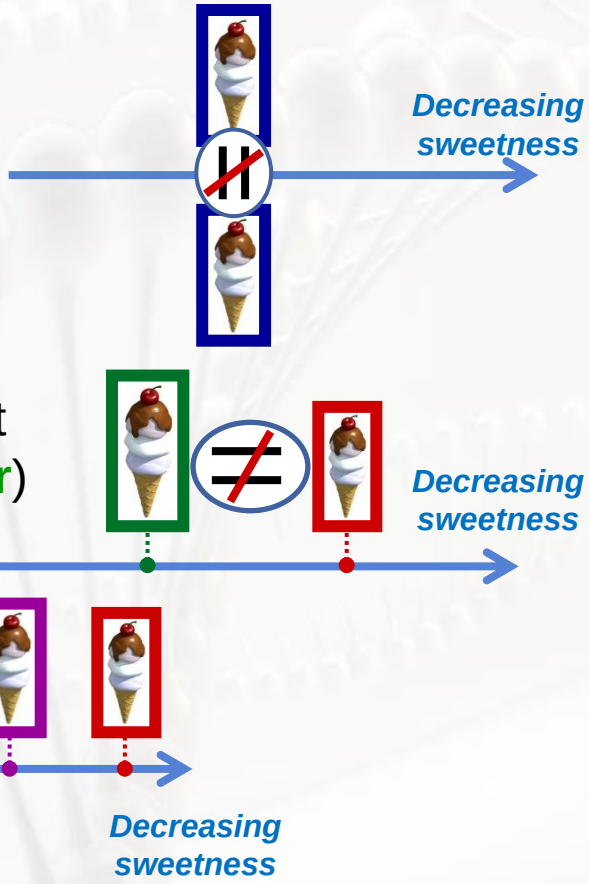
- 5 linked components:

α: Probability of a Type I error (wrongly concluding that a difference exists between the products)

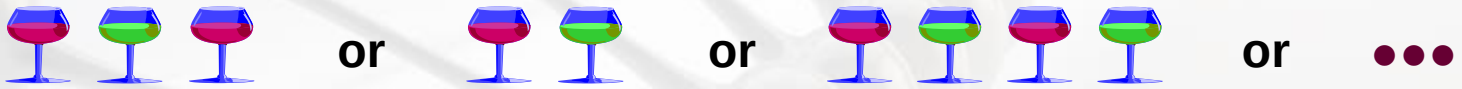
β: Probability of a Type II error (wrongly concluding that no difference exists between the products = **1-power**)

δ: Size of the difference of interest

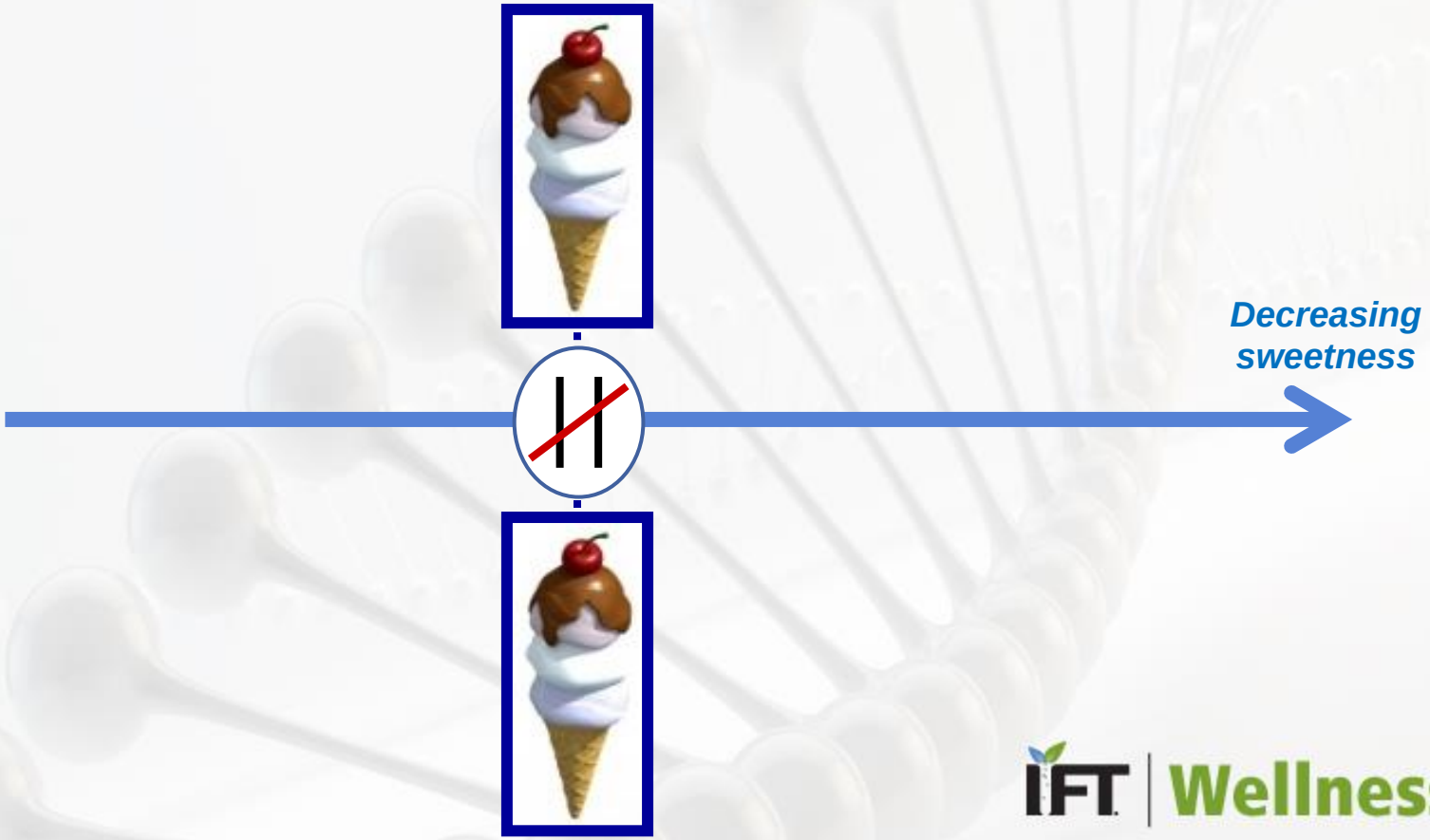
N: Sample size




**Testing Protocol**



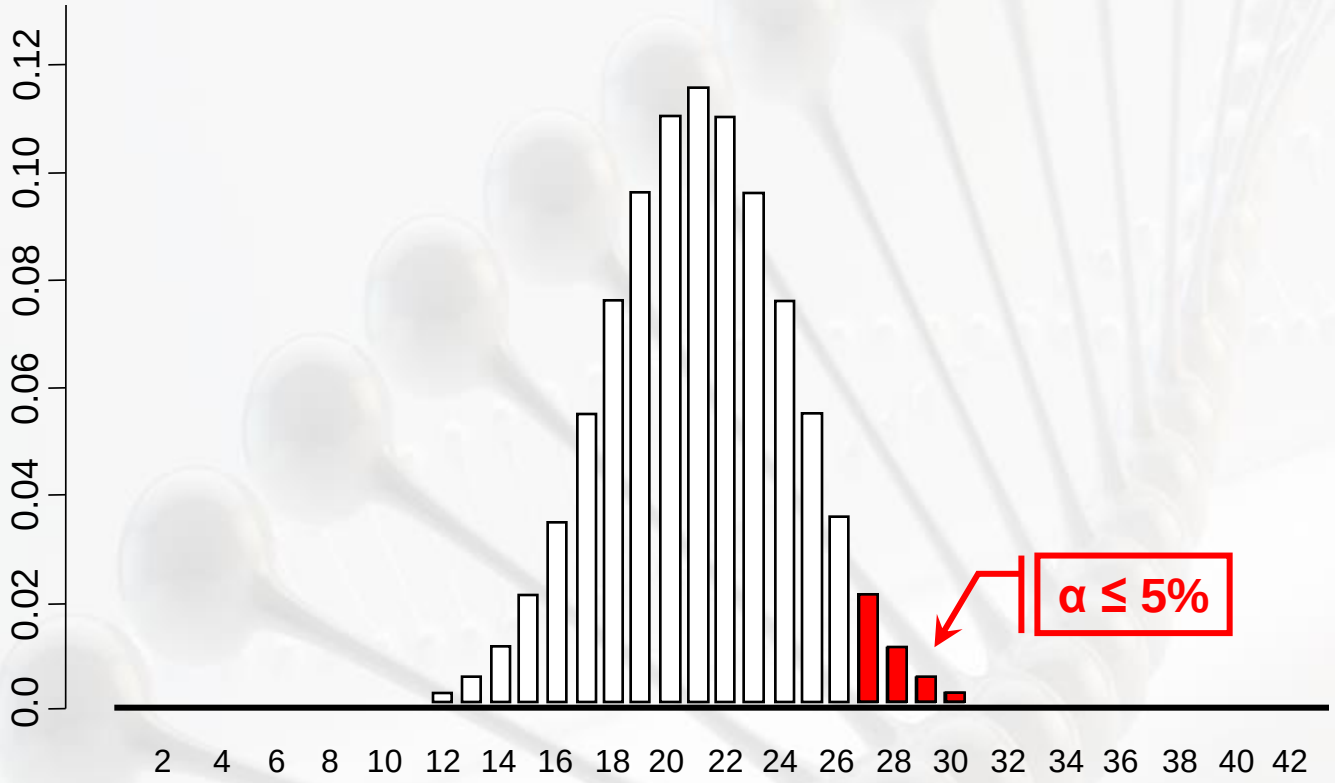
# α: Type I Error



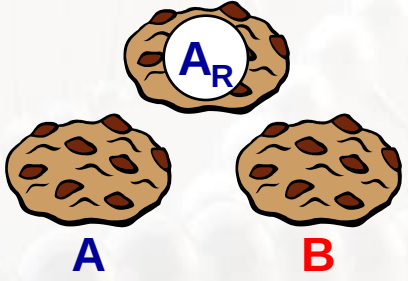
# Binomial Theory

**n = 42**

**Chance probability: $p = {}^1/_2$**

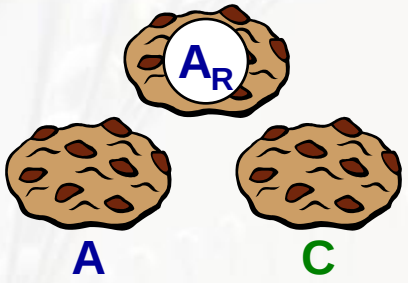


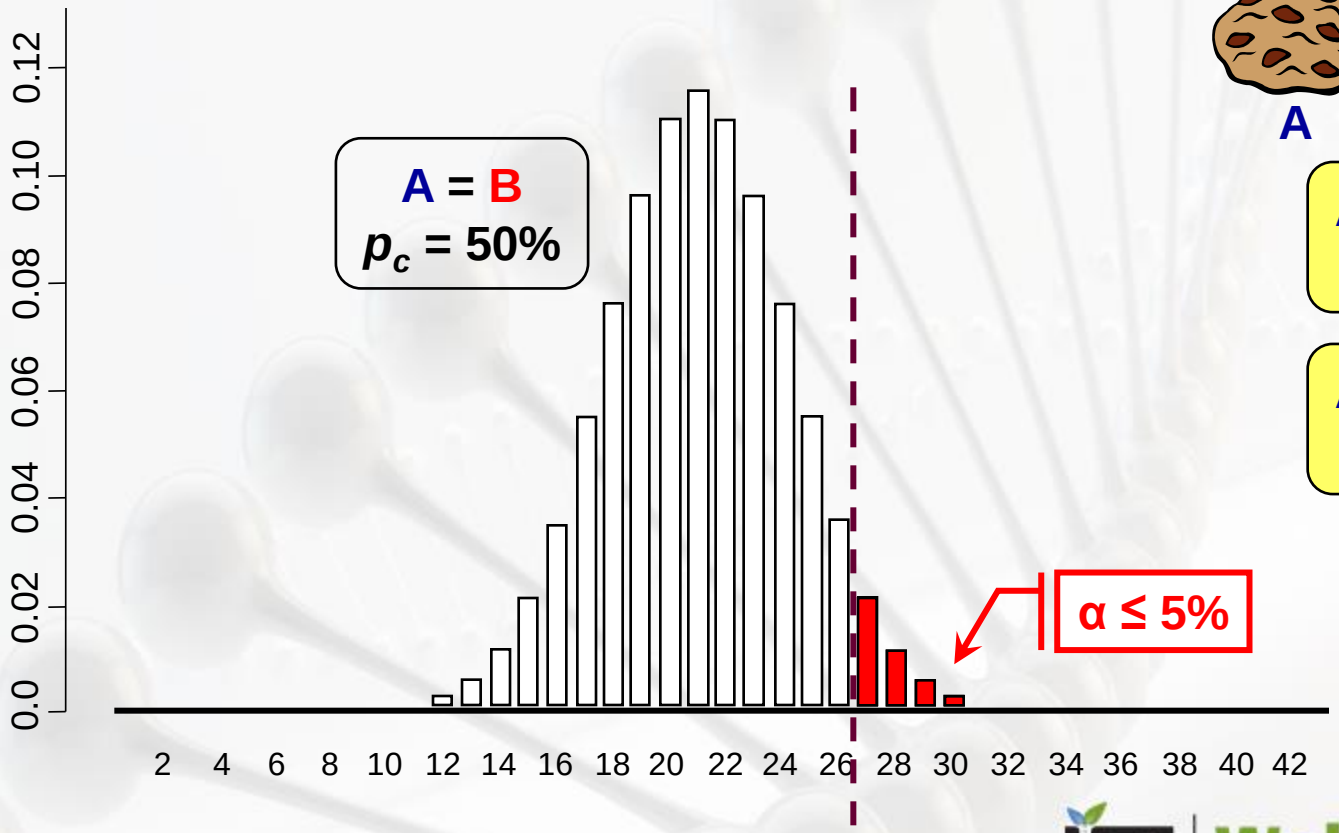
# What is α?




**n = 42**

**Chance probability: $p = {}^{1}/_{2}$**



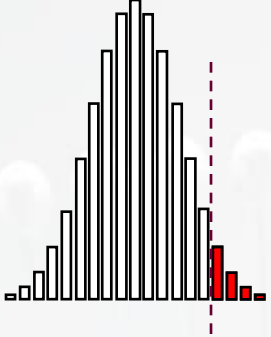

**A vs. C** $^{28}/_{42}$

**A vs. B** $^{27}/_{42}$

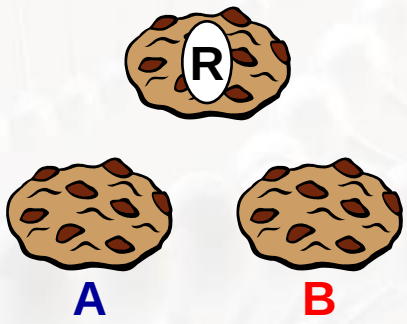
***"Different"***

**Type I error**

***If 27 correct or more***



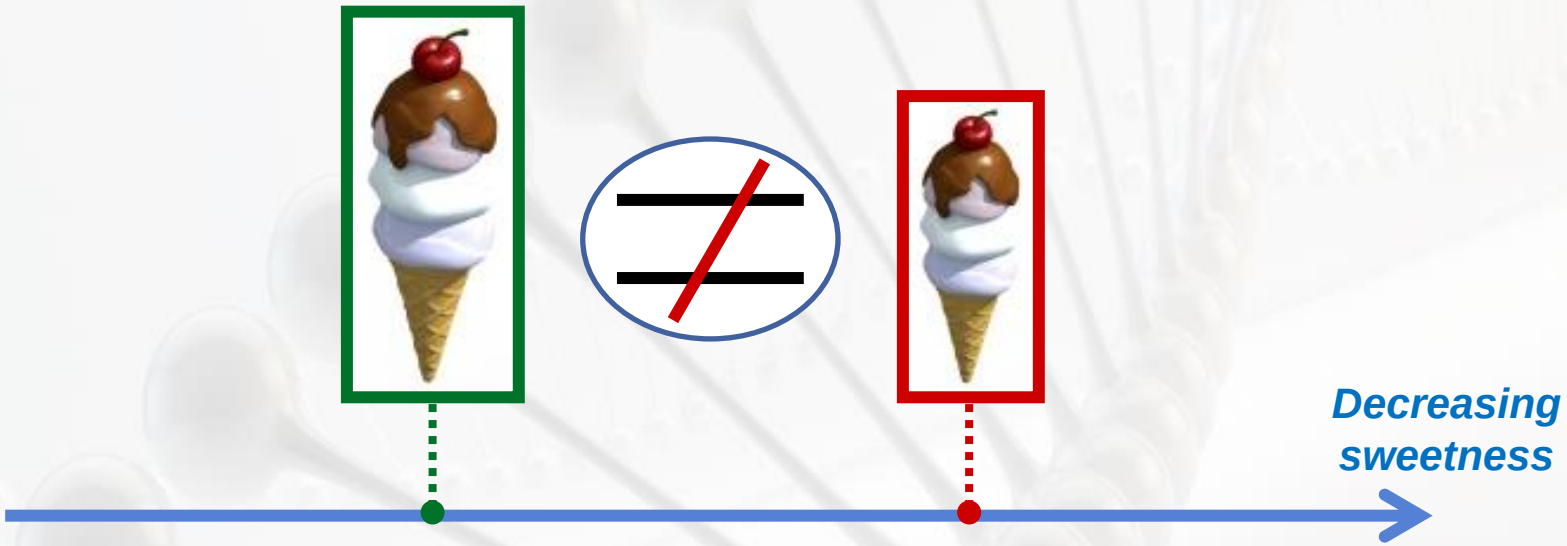
# Type I Error



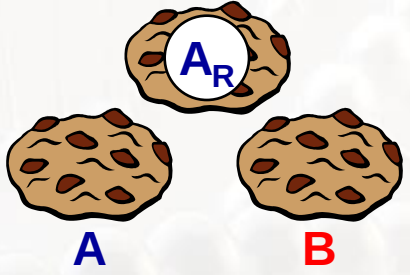

- Falsely concluding that the products are different when they are not
- Consequence:
  - Prove products are equivalent
    - Ingredient change, new supplier, government regulation (e.g., salt or sugar reduction)
    - *Conclude a difference*

      → *Missed an opportunity for change*

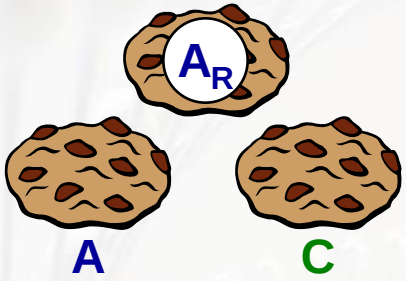
# β: Type II Error
## *(1 – power)*



# What is β?




**n = 42**

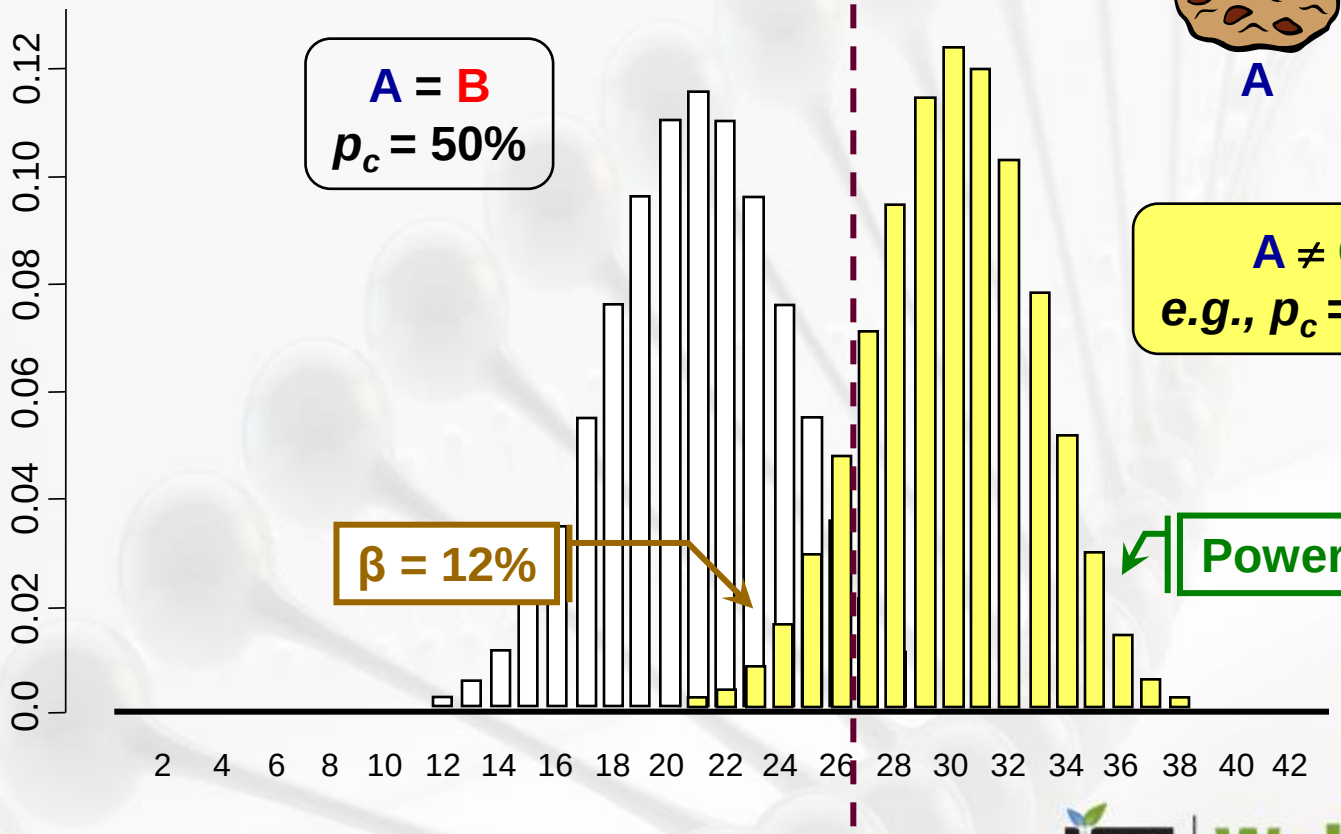



**A = B**
**$p_c$ = 50%**

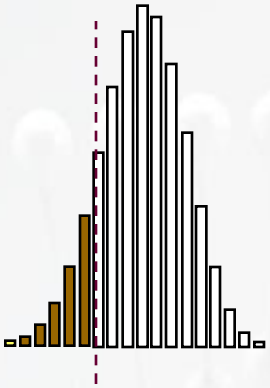
**A ≠ C**
***e.g.*, $p_c$ = 71%**

**β = 12%**

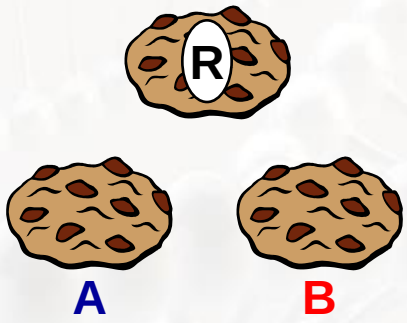
**Power = 88%**




IFT | Wellness 13

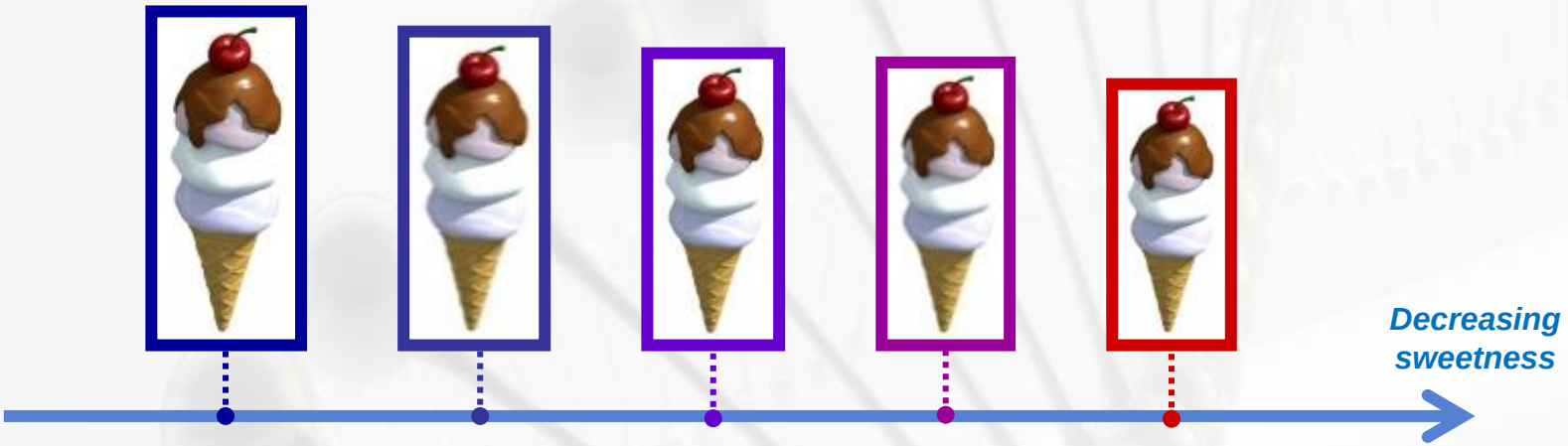


# Type II Error




- Falsely concluding that the products are not different when they are; failing to find a difference
- Consequence:
  - **Prove products are equivalent**
    - Ingredient change, new supplier, government regulation (e.g., salt or sugar reduction)
    - ***Conclude similarity → Release on the market of a sensorially different product***

# Size of the Relevant Difference



# Comments on Sensory Differences

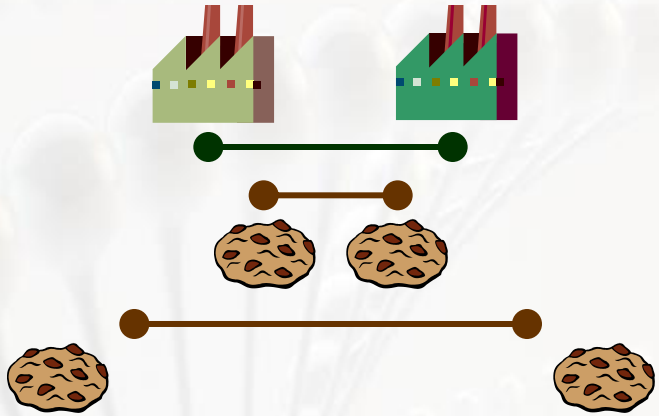
**Two different products will never be identical**

**A lack of significance difference does not mean that two products are equivalent**

**A significant difference can always be found, provided that the sample size is large enough**

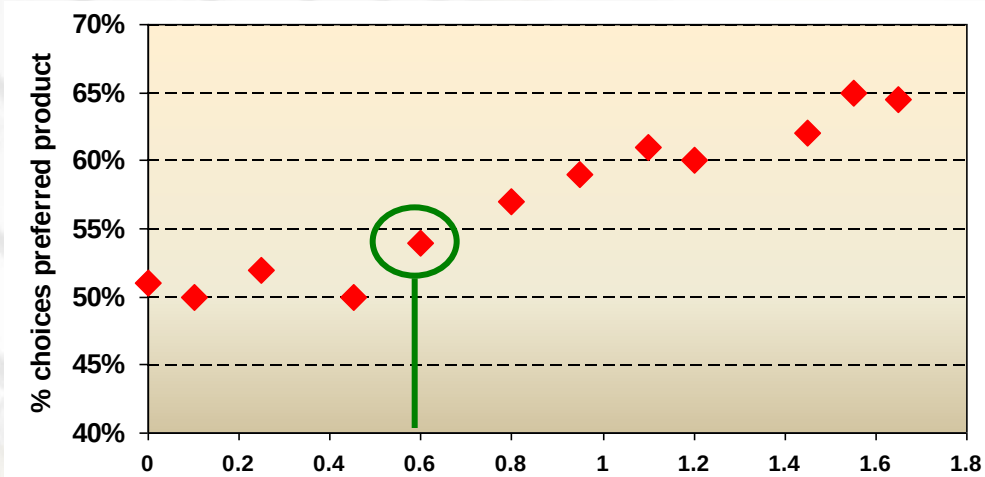
# Estimating the Size of a Relevant Difference

- No universal answer
- Research is necessary
- Various options
  - ***Measuring inter factory differences***



  - ***Use the same-different test***
    - *Estimate the size of the sensory difference above which consumers call the products 'Different"*
  - ***Build a relationship between perceived differences and consumer preferences***

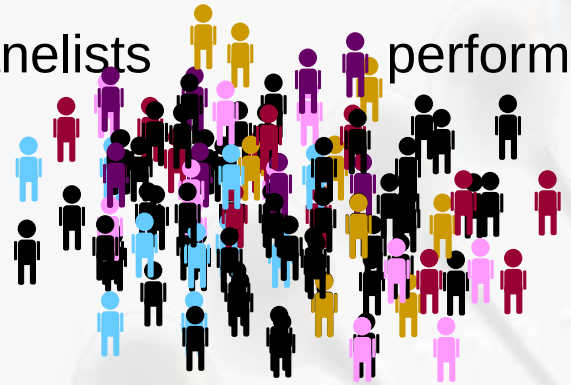
*"Are they the same or different?"*



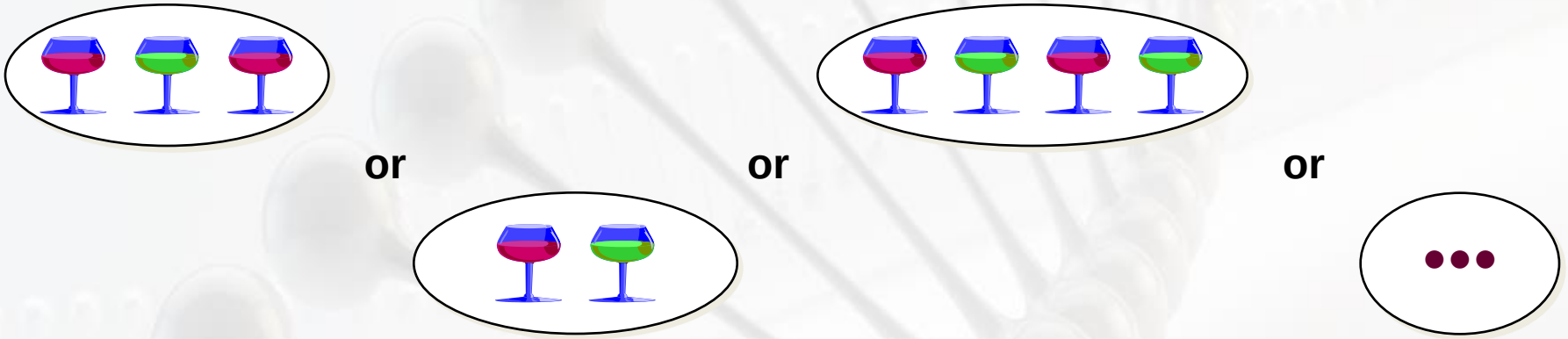
# Sample Size

# Effect of Sample size on Discrimination

- Greater sample sizes provide greater statistical power
- Example: One study, no significance difference found
  - 5 panelists  performing one triangle 

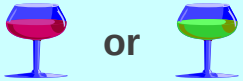
  vs.
  - 500 panelists  performing one triangle 
- Larger sample sizes are less likely to miss sensory differences
- The sample size is a function of **α**, **β**, the **size of the difference** of interest and the **test methodology**

# Test Protocol



# Common Sensory Discrimination Methods

- Many different protocols are available:

| *Information regarding the sample to be selected not required* | | |
|---|---|---|
| ❖ Triangle |  | Which one is different? |
| ❖ Duo-trio |  | Which one is the same as the reference ? |
| ❖ Tetrad |  | Group the samples into 2 groups of 2 identical samples |
| ❖ Same-different |  | Are they the same or different ? |
| ❖ … | | |
| *Information regarding the sample to be selected required* | | |
| ❖ 2-AFC |  | Which one is *more* … ? |
| ❖ Identification |  | Is it A or B ? |
| ❖ … | | |

# Tetrad vs. Triangle
## *An Industry Perspective*

John Cowden, Suzanne Pecore, Nort Holschuh, Amalie Kurzer,
General Mills
February 27, 2013

IFT | Wellness 13℠

# Talk Overview

22

- Background/test primer
- Sensitivity comparison
- Peek into reproducibility
- Tetrad to manage risk

# Background

23

- General Mills has a long standing history of using discrimination testing to guide product and processing reformulation changes.
- To date, triangle testing has been used to manage risk for blind product changes.
- Though the triangle method is inexpensive, obtaining adequate product and maintaining a large pool of motivated panelists is challenging.
- Tetrad shows promise to replace triangle methods and overcome current challenges of triangle.

# Test Primer, Triangle vs. Tetrad

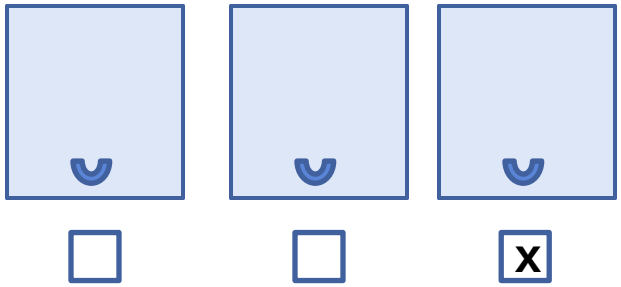


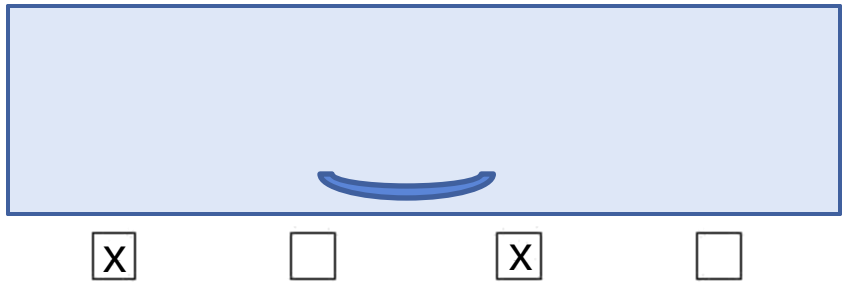
**Triangle**

- Panelists select the different sample
- **Sequential monadic**
- Outcome is who got pairing correct, p=1/3 by luck

**Tetrad**

- Panelists group 4 samples into 2 pairs
- **Side-by-side comparison**
- Outcome is who got pairing correct, p=1/3 by luck

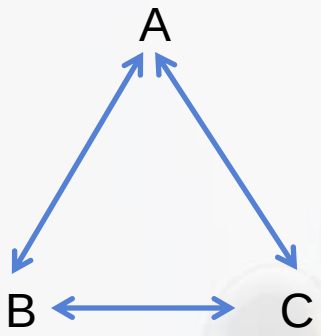


# Comparing The Psychological Task

- Triangle

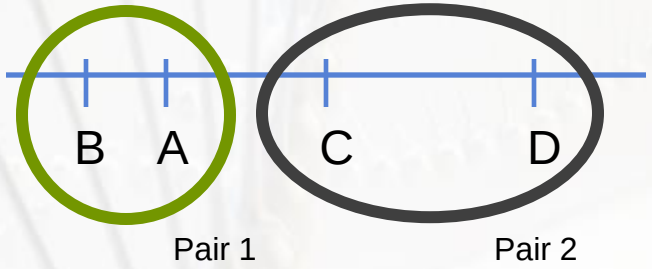
  **Relies on comparisons**




VS

- Tetrad

  **Uses an easier linear approach**

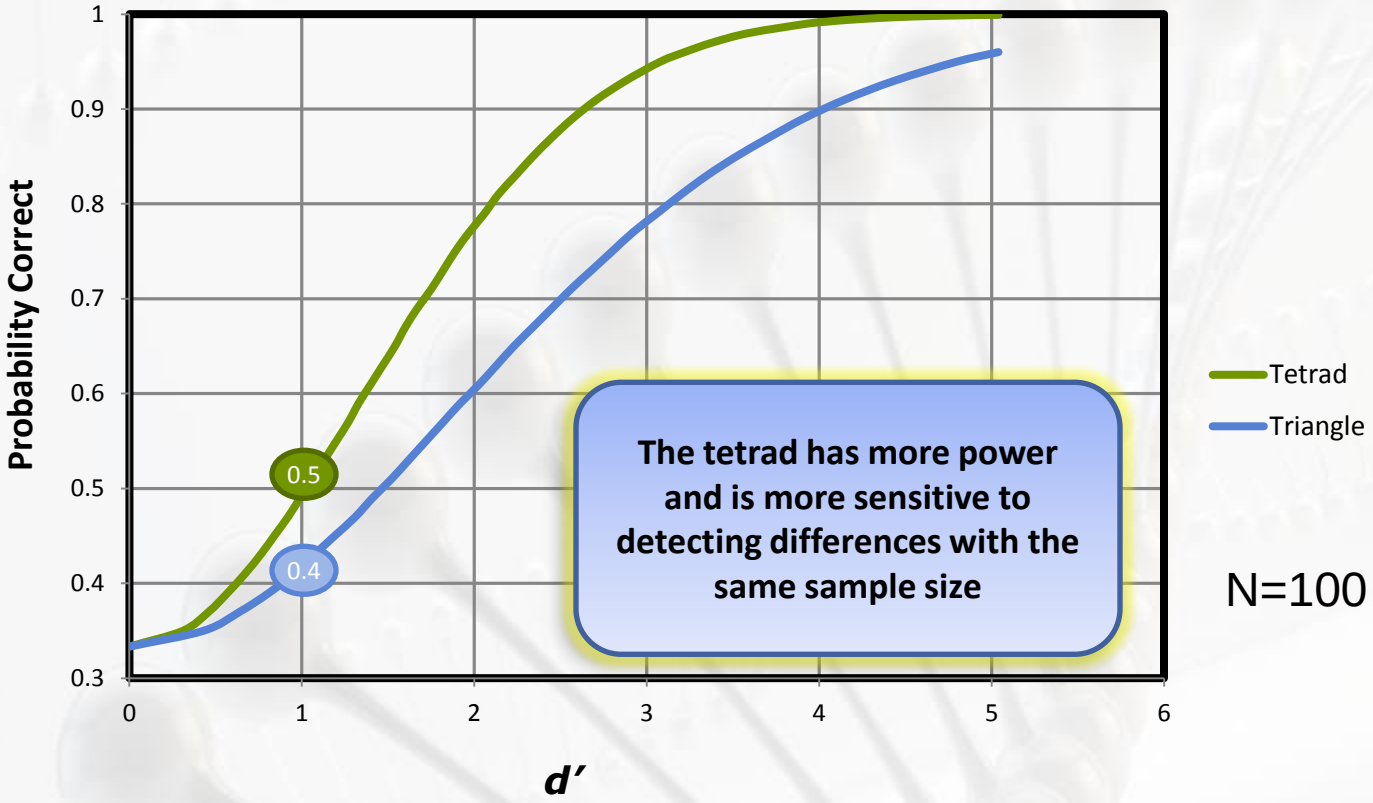



Which pair is most alike?

What is the order?

# Respondents More Likely To Find A Difference In Tetrad






*d'* is the way to compare multiple discrimination tests to one another and can be thought of as the amount of difference between test and control products

- David HA, Trivedi MC. Blacksburg, Va.: Virginia Polytechnic Insti; 1962. Pair, triangle and duo–trio tests. Technical report nr 55, Dept. of Statistics
- Ennis, J. M., Ennis, D. M., Yip, D. and O'Mahony, M. (1998). Thurstonian models for variants of the method of tetrads. *British Journal of Mathematical and Statistical Psychology*, **51**(2), 205-215.

# Comparing Sensitivity in Practice

## Products Run as Both A Triangle (Δ) and a Tetrad (□)

27

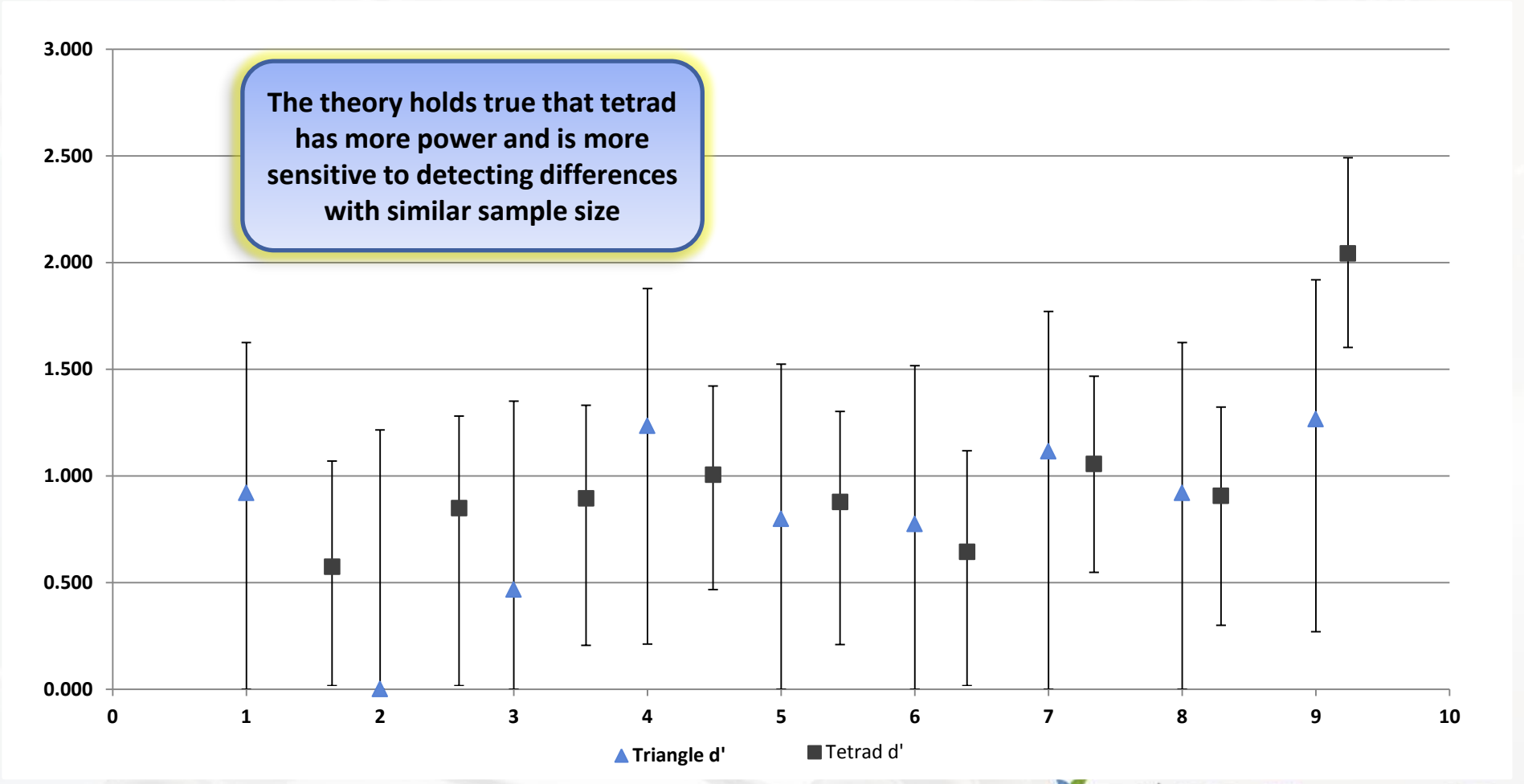
| Product | Difference |
|---|---|
| Cereal 1 | Flavor |
| Cereal 2 | Texture/Flavor |
| Cereal 3 | Texture |
| Baked Good 1 | Texture |
| Baked Good 2 | Flavor |
| Dairy 1 | Flavor |
| Dairy 2 | Flavor |
| Spicy Meal 1 | Flavor |
| Spicy Meal 2 | Flavor |

# Tetrad Is Consistently More Sensitive Than Triangle

| Product | Method | True Discriminators | Sample Size |
|---|---|---|---|
| Cereal 1 | Triangle Δ | 11% | 69 |
| | Tetrad □ | 8% | 72 |
| Cereal 2 | Triangle Δ | 0% | 67 |
| | Tetrad □ | 18% | 72 |
| Cereal 3 | Triangle Δ | 3% | 68 |
| | Tetrad □ | 19% | 67 |
| Baked Good 1 | Triangle Δ | 19% | 70 |
| | Tetrad □ | 24% | 69 |
| Baked Good 2 | Triangle Δ | 8% | 72 |
| | Tetrad □ | 19% | 72 |
| Dairy 1 | Triangle Δ | 8% | 70 |
| | Tetrad □ | 10% | 72 |
| Dairy 2 | Triangle Δ | 8% | 70 |
| | Tetrad □ | 26% | 72 |
| Spicy Meal 1 | Triangle Δ | 11% | 69 |
| | Tetrad □ | 20% | 63 |
| Spicy Meal 2 | Triangle Δ | 19% | 67 |
| | Tetrad □ | 68% | 66 |

# Tetrad No More Fatiguing Than Triangle



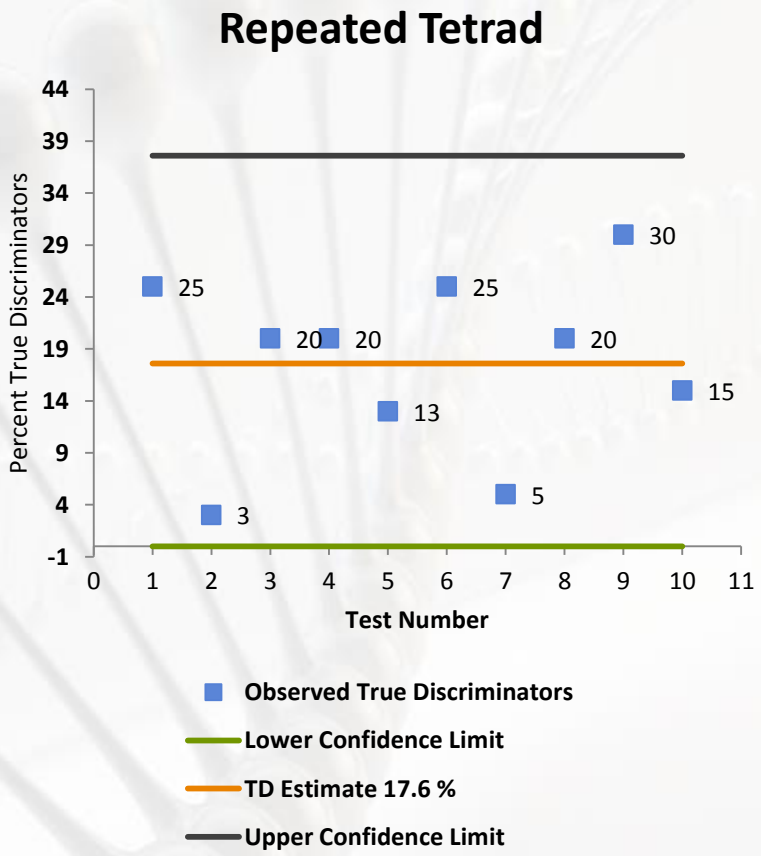


# Tetrad is Repeatable and Conservative



| Repeated Tetrad | | | |
|---|---|---|---|
| Rep | N | Correct | True Discriminators |
| 1 | 30 | 15 | 25 |
| 2 | 31 | 11 | 3 |
| 3 | 30 | 14 | 20 |
| 4 | 30 | 14 | 20 |
| 5 | 31 | 13 | 13 |
| 6 | 30 | 15 | 25 |
| 7 | 30 | 11 | 5 |
| 8 | 30 | 14 | 20 |
| 9 | 30 | 16 | 30 |
| 10 | 30 | 13 | 15 |
| Average | | | 17.6 |

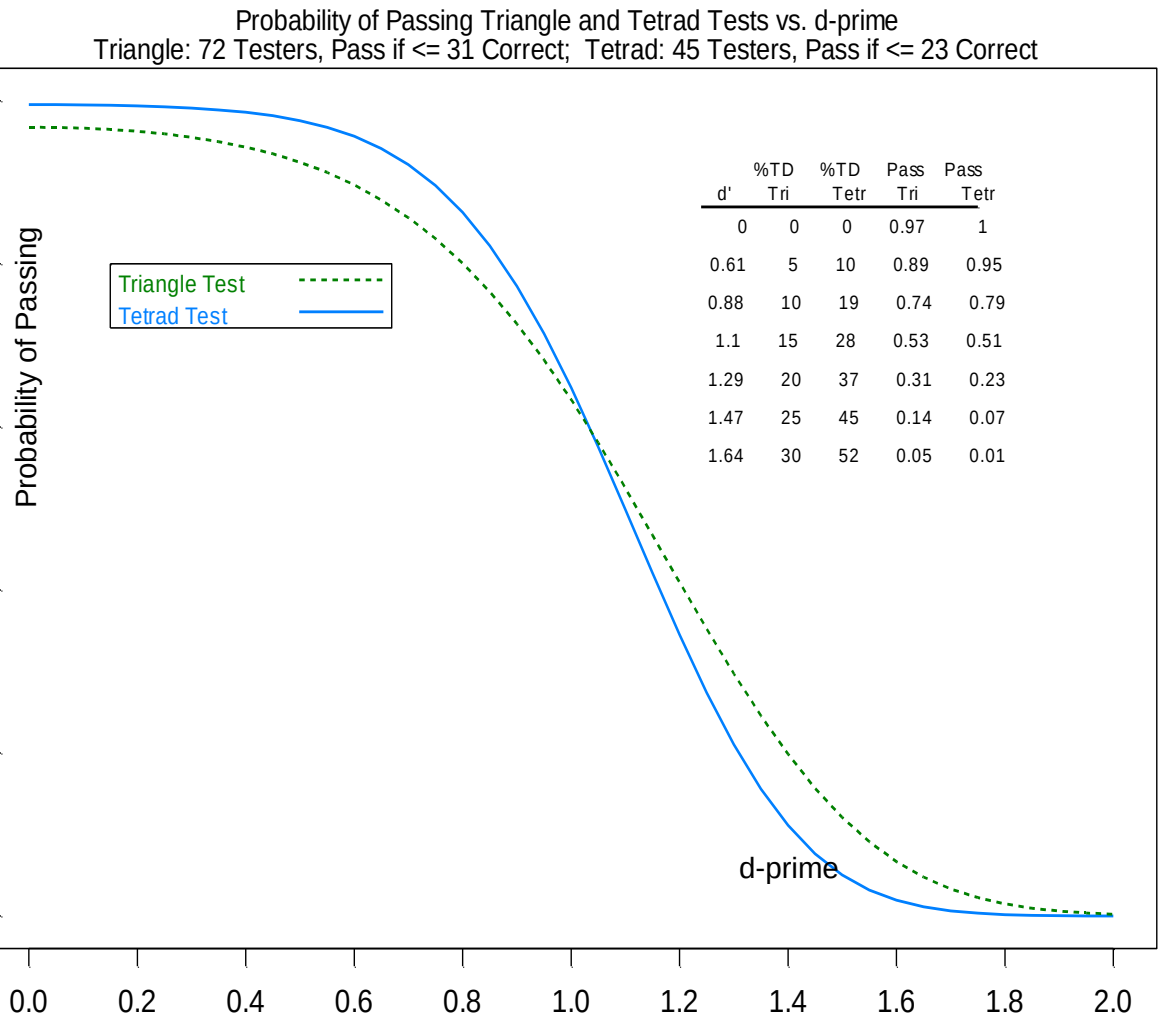
Chi-square test for differences in true discriminator across tests is not significant (p =0.9314)



# Tetrad Manages Risk Better Than Triangle With Fewer Respondents

Tetrad less likely to send R&D back to reformulate when they don't need to (type I error)

Tetrad less likely to pass a change when it should have failed (type II error)

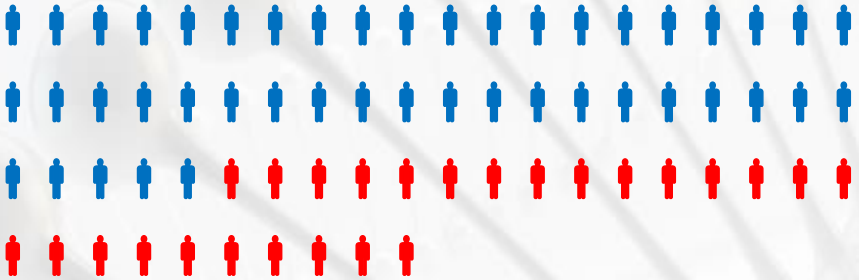



| d' | %TD Tri | %TD Tetr | Pass Tri | Pass Tetr |
|---|---|---|---|---|
| 0 | 0 | 0 | 0.97 | 1 |
| 0.61 | 5 | 10 | 0.89 | 0.95 |
| 0.88 | 10 | 19 | 0.74 | 0.79 |
| 1.1 | 15 | 28 | 0.53 | 0.51 |
| 1.29 | 20 | 37 | 0.31 | 0.23 |
| 1.47 | 25 | 45 | 0.14 | 0.07 |
| 1.64 | 30 | 52 | 0.05 | 0.01 |

# Implications for Replacing Triangle

- **The tetrad requires <u>fewer panelists</u> for the same risk profile as current triangle.**
  - Less product required for test = easier for R&D to make samples
  - Fewer respondents = less employee panelist and testing time
  - Less complex for lab to execute = increased testing capacity

<u>Benefit of Tetrad Fewer Respondents</u>



**Tetrad n=45** **Triangle n=72**

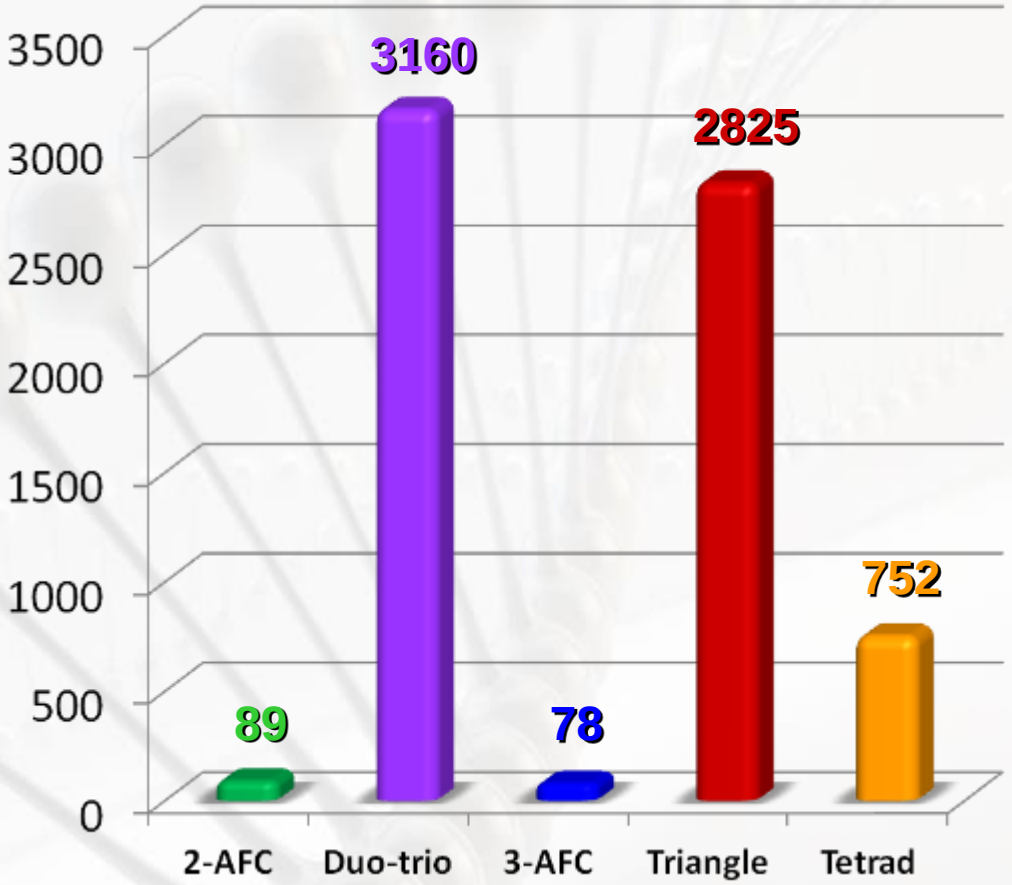
# Conclusion

## *Establishing Sensory Equivalence*

# The Cost of Decision Rules (1)

**Scenario 1**

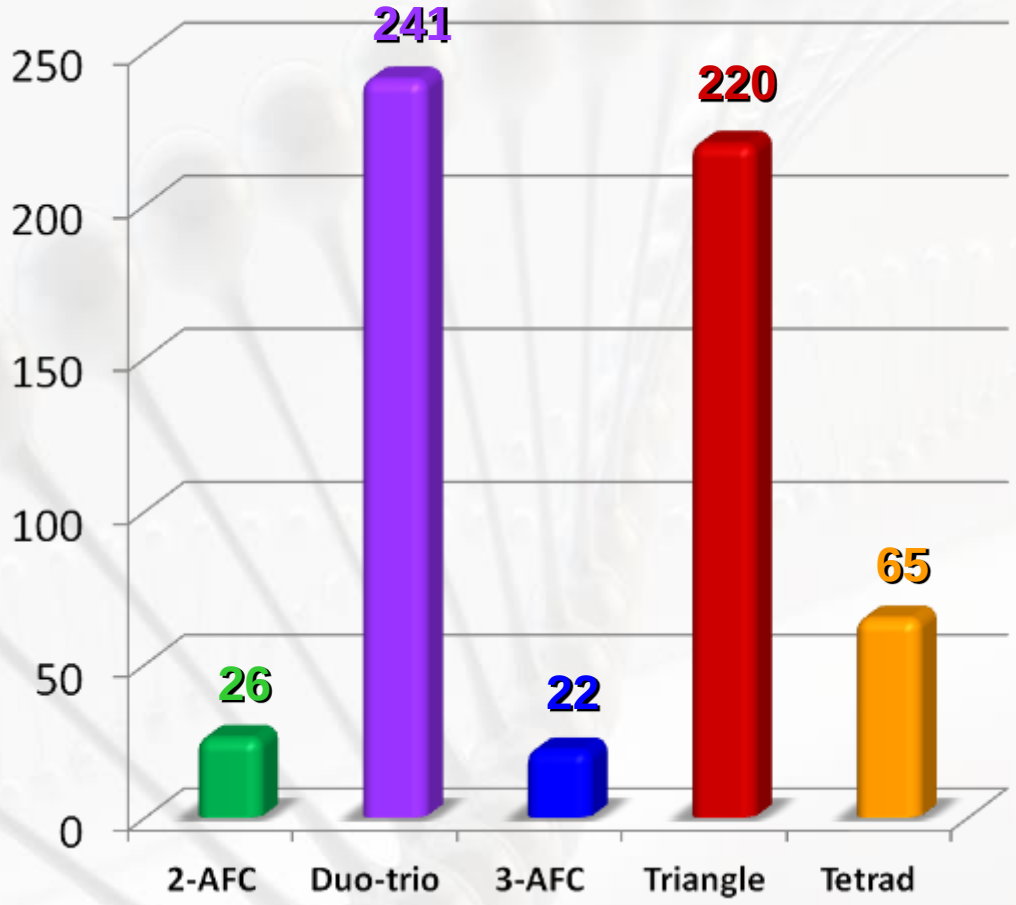
- **Size of the difference**: 64:36 in a 2-AFC ***(δ of 0.5)***
- **Power:** 80% chance of detecting the difference (β=20%)
- **α level**: 5%
- ***Sample size needed***



# The Cost of Decision Rules (2)

**Scenario 2**

- **Size of the difference**: 76:24 in a 2-AFC ***(δ of 1)***
- **Power:** 80% chance of detecting the difference (β=20%)
- **α level**: 5%
- ***Sample size needed***



# Sensory Discrimination Program

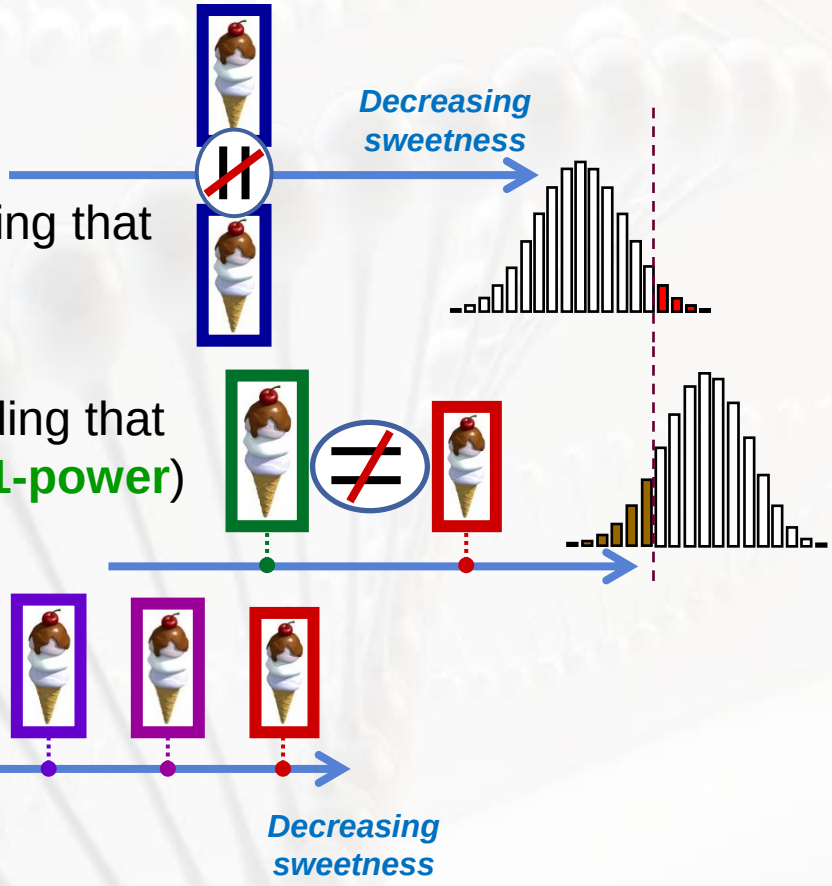
- 5 linked components:

α: Probability of a Type I error (wrongly concluding that a difference exists between the products)

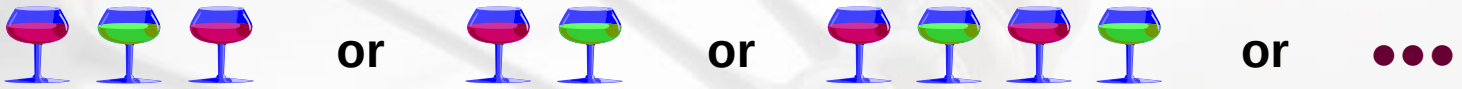
β: Probability of a Type II error (wrongly concluding that no difference exists between the products = **1-power**)

**δ**: Size of the difference of interest

**N**: Sample size




**Testing Protocol**






IFT | Wellness 13

# Investigating Sensory Equivalence of Reformulated Products

Benoît Rousseau[1] & John Cowden[2]
[1]The Institute for Perception, Richmond, VA
[2]General Mills, Minneapolis, MN

IFT | Wellness 13

## Thank You Very Much!
## Any Questions?