The Institute for Perception

For technical reports from The Institute for Perception or for information about short courses, please visit www.ifpress.com or email us at mail@ifpress.com

Reprinted from IFPress (2020) 23(3) 3,4

Can Larger Sample Sizes Result in Missed Opportunities?

Benoît Rousseau and Daniel M. Ennis

Background: The design of sensory and consumer tests relies on a combination of factors including statistical risks, methodological choices, and the size and nature of the respondent population. The research sample size is often the most discussed aspect among project team members because it requires finding a balance among representation, statistical power, timing, and costs.

Discrimination testing to investigate equivalence (or similarity) is used extensively by CPG companies, typically using the traditional concept of statistical power[1,2] and a null hypothesis of no difference. However, if we rely on a nonsignificant difference to establish equivalence, then increasing the sample size may work against our objective[3]. In this report, we illustrate how this can lead to missed opportunities and we describe a more direct and stable test of equivalence based on a pre-established consumer relevant action standard.

Scenario: You manage the sensory testing program at a national condiment manufacturer with dominant offerings in the mayonnaise, mustard, and salad dressing product categories. Aware of the importance of incorporating consumer sensitivity in reformulation decisions, your management has supported a focus on conducting consumer-driven research to establish δ_R, a standardized sensory difference action standard. Above this standard, consumers will reject a product reformulation[4,5]. Your consumer-based research found a value of 0.96 across your product categories. By comparing consumer sensitivity to the sensitivity of your internal panel[6] of 26 panelists using the tetrad method, you establish that your internal panel is 30% more sensitive than consumers for the same underlying product differences. Consequently, you conclude that your program's δ_R value should be $1.30 \times 0.96 = 1.25$. Your program's risk profile has a Type I error of 5% (α), a power of 95% ($1 - \beta$), and a corresponding sample size of 52 using a double replication of the tetrad test. For major brands, you use a quadruple replication ($N = 104$), a company standard to further increase power. Your routine investigations follow this testing pattern: Reject formulation/ingredient changes that result in a statistically significant difference and accept changes that do not.

	Study 1	Study 2
Number correct	44	47
Total *N*	104	104
Proportion correct	0.423	0.452
p-value (H_0: P_0 = 1/3)	0.04	0.01
Decision	***Reject***	***Reject***

Table 1. Summary of two full-fat ranch dressing tetrad tests with the null hypothesis of no difference.

Research on the company's best-selling full-fat ranch salad dressing involved two quadruple replications and both studies resulted in a statistically significant difference as shown in Table 1. This led to the recommendation to reject the corresponding salad dressing changes. Since these two reformulations would translate into significant cost-savings and manufacturing simplification, you decide to look back at the research more critically.

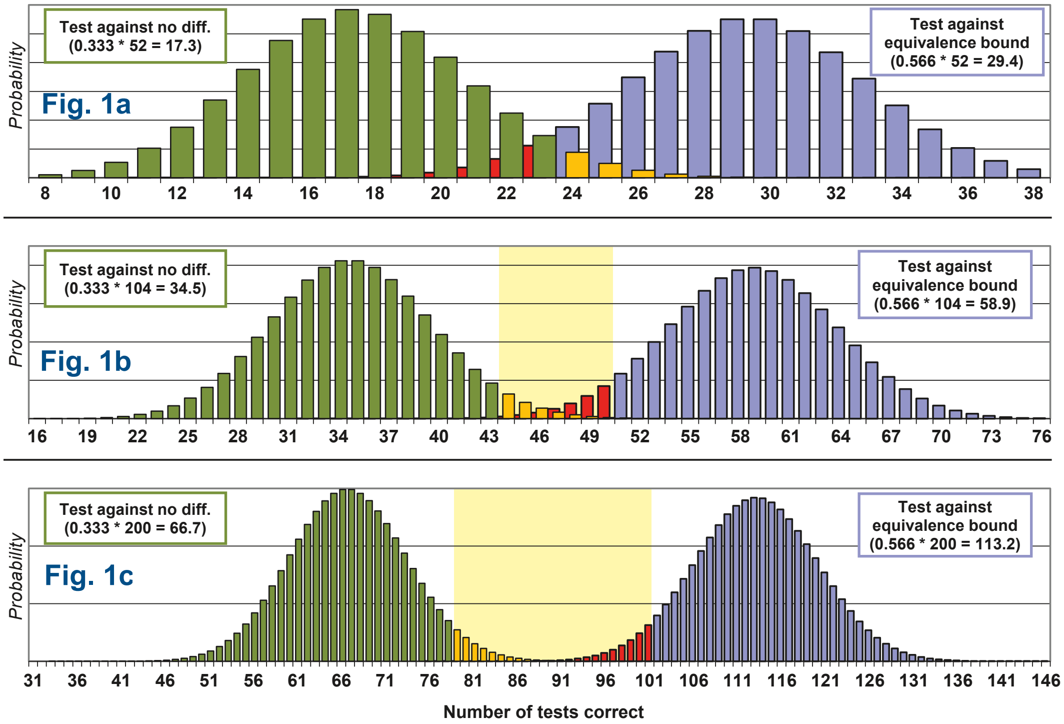
Difference vs. Equivalence Testing: In a previous technical report[3], we showed that changes rejected using a null hypothesis of no difference could be acceptably less than the action standard and that improving the power of the test exacerbated this finding. This result relates to an earlier report[7] that showed how superiority and equivalence, which occurs in advertising claims, could be justified using the same data. Since a difference will always exist, basing equivalence decisions solely on *p*-values, as shown in Table 1, is misdirected. In equivalence testing, the concept of a difference action standard comes into play and, when properly used, is a more direct and satisfying approach because it is focused on whether a perceived difference is significantly lower than an established action standard, δ_R. Equivalence testing has been described using action standards based on the concept of proportion discriminators which has been shown to be flawed[8] because it does not account for different methodologies. With an action standard set as a Thurstonian δ value, we translate this value into the corresponding proportion of correct responses for a given discrimination protocol[9]. Using binomial theory, an experimental result can then be compared to this action standard, which we call P_{cR}, and two products will be deemed equivalent if the observed proportion of correct responses is significantly lower than P_{cR}.[10]

Applying Equivalence Principles: Looking back at your data, you translate your action standard, δ_R, into the probability of a correct response in the method of tetrads to 0.566. This is your P_{cR} value. You then compare the two experimental proportions, 0.423 and 0.452, respectively, to your action standard using a one-tailed binomial test for a sample size of 104. A binomial test using the null hypothesis of 0.566 results in significant evidence for equivalence for both studies (Table 2).

	Study 1	Study 2
Number correct	44	47
Total N	104	104
Proportion correct	0.423	0.452
p-value (H_0: P_{cR} = 0.566)	0.002	0.01
Decision	***Accept***	***Accept***

Table 2. Summary of two full-fat ranch dressing tetrad tests with the null hypothesis, $P_{cR} = 0.566$.

Reprinted from IFPress (2020) 23(3) 3,4



Figures 1a-c.

Left Curve: Binomial distribution at $p = 1/3$ (chance level for the tetrad test). In orange, number of tests correct needed for a significant difference.

Right Curve: Binomial distribution at $p = 0.566$ (δ_R of 1.25 for the tetrad test). In red, number of tests correct needed for a significant equivalence test.

Yellow Zone: Region where both tests are significant.

Accordingly, even though one analysis of these two studies found a significant difference using the null hypothesis of no difference (Table 1), an equivalence analysis also found that these differences were significantly lower than the action standard (Table 2.) This led to the conclusion to reject the change in the first case and accept it in the second. The reason behind this difference is illustrated in Figures 1a-c. Figures 1a-c show that a difference between the approaches will occur when the test result is significantly greater than the null difference of 1/3 and also significantly less than a null action standard of 0.566. While the conclusions will be identical if α and β are the same and the corresponding sample size is used (Fig. 1a), if the sample size is increased, the likelihood of rejecting a planned change when it was suitable will increase using the null difference of 1/3. Figures 1a-c show binomial distribution outcomes for sample sizes of 52 (No conflicting cases), 104 (7 cases) and 200 (23 cases).

Conclusion: Since the sensory differences are smaller than the established action standard, you report to your management that the reformulations should be considered for the next stage of development, avoiding missed cost-saving opportunities for your company. This investigation makes it clear that the power approach you have been using in your quality testing program, with a null hypothesis of no difference, is not optimal. Switching to an equivalence testing approach would provide more stability and consistency in the decisions you reach using your discrimination testing research. You recommend that future risk profiles for ingredient and other quality assurance changes be established using δ_R for the null hypothesis instead of a hypothesis of no difference. Then, testing for differences less than δ_R can be implemented using standard binomial statistical testing.

References and Notes

1. Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies,* **8**(4), 353-370.

2. Ennis, J. M. and Jesionka, V. (2011). The power of sensory discrimination methods revisited. *Journal of Sensory Studies*, **26**(5), 371-382.

3. Ennis, J. M. and Ennis, D. M. (2018). Consumer relevant confidence in sensory difference tests. *IFPress,* **21**(2) 3-4.

4. Rousseau, B. (2015). Sensory discrimination testing and consumer relevance. *Food Quality and Preference,* **43**, 122-125.

5. Rousseau, B. and Ennis, D. M. (2013). When are two products close enough to be equivalent? *IFPress,* **16**(1) 3-4.

6. Ishii, R., Kawaguchi, H., O'Mahony, M., Rousseau, B. (2007). Relating consumer and trained panels' discriminative sensitivities using vanilla flavored ice cream as a medium. *Food Quality and Preference,* **18**(1), 89-96.

7. Ennis, D. M. (2017). Claiming superiority and equivalence simultaneously. *IFPress,* **20**(2) 3-4.

8. Rousseau, B. and Ennis, D. M. (2007). Why proportion of discriminators is method specific. *IFPress,* **10**(3), 2-3.

9. All the IFPress references are part of the most recent edition of *Tools and Applications of Sensory and Consumer Science* (2020) edited by Daniel M. Ennis and Benoît Rousseau. They can also be found on the website, www.ifpress.com.

10. Bi, J. (2011). Similarity tests using forced-choice methods in terms of Thurstonian discriminal distance, *d'*. *Journal of Sensory Studies,* **26**(2), 151-157.