Chapter 6

A-Not-A Test

Antoine G. de Bouillé
Philip Morris Products S.A., Neuchâtel, Switzerland

In analytical sensory testing, two main types of test can be defined: descriptive testing and discriminative testing.

In the latter type, the aim of conducting such tests is usually for the sensory scientist or product developer to find out whether there is a perceptible difference between two (or sometimes more) stimuli (Stone and Sidel, 1993). Assessors performing the test would usually receive one or several samples (depending on the test used) and asked questions such as: which sample is the different one, which one is the most bitter, is it the reference sample, and how different is it from the reference sample.

Scientists nowadays have many tools at their disposal to achieve their goals, but it can be confusing when it comes to choosing the right one. In this chapter, attention will be given to the A-not-A test, focusing on its principle, the type of assessors who can participate in the test, and the analysis of the data generated. Examples are given in two case studies at the end of the chapter.

1. WHAT IS THE A-NOT-A TEST?

The A-not-A test is a discriminative sensory test that requires assessors to identify whether a sample is "A" or "Not-A"; "A" is considered as a reference sample (or signal) and defined before the test. The A-not-A test has been defined in the literature as a rating method with two categories (Bi and Ennis, 1999).

Usually, it is recommended for assessors to be familiar with the reference sample A. This makes the A-not-A a relevant test when assessors have a high degree of exposure to the same sample over time, which is often the case in quality control (QC) and quality assurance (QA) environments (Van Hout et al., 2011). For example, in a production environment, the A-not-A test will give a quick answer as to whether a produced batch has the sensory properties it is supposed to have. However, the A-not-A test would not give insights into the nature of the difference if, say, a batch would happen to be perceived sensorially different and therefore rejected. Methods such as the difference

Discrimination Testing in Sensory Science. http://dx.doi.org/10.1016/B978-0-08-101009-9.00006-X

from control (DFC) would, in that case, give direction on which sensory aspect the tested sample has problems with.

Although not widely used in consumer studies, the A-not-A test can present benefits when the question to be answered is whether heavy users of a product can identify if a new version of it is different or similar to the old one. When the A-not-A test is performed with several test samples, it would be then possible to identify which candidate/prototype sample is sensorially closest to the reference sample.

In the literature, the A-not-A test can be described as a single sample presentation where assessors receive one sample in a session and are asked to identify if it is "A" or "not-A." In that case, a reference (or in this case reminder) is not provided. In this scenario, assessors have only their own internal reference (or from previous training) to establish whether the tested sample is indeed A or not-A. When assessors are presented with only one sample, two types of designs can be considered: a monadic design (the number of assessors getting "A" and getting "not-A" is decided in advance) or a mixed design (assessors are randomly allocated either sample "A" or "not-A" to evaluate).

Recently, Stocks et al. (2013) discussed the concept of a reminder in discrimination testing. This is meaningful when a company does not have an adequate training/familiarization procedure (Bi et al., 2013a,b) and does not have the time or resource to develop one. A reminder sample can also be useful when the assessors selected to take part in the test have limited knowledge about the product and therefore need to be (re)familiarized with the reference sample.

In the context of the A-not-A test, including one or several reminder samples will help to ensure that assessors are evaluating the samples on fairly similar grounds regarding the A reference. The chosen reminder can be either A or not-A, and assessors are presented with it before each test sample. When the total number of samples evaluated in a session is counted up, this approach reduces the time necessary for familiarization/training sessions (Stocks et al., 2013). When reminder samples are used, the A-not-A test is often referred as "A-not-AR". Table 6.1 displays the different possible variants of the A-not-A method including advantages and disadvantages.

2. PROCEDURE

2.1 Familiarization

When a reminder sample is used, assessors are first given the reference "A" sample and asked to get familiarized with it. Assessors can do this step either individually (e.g., in a sensory booth) or as a discussion with a panel leader where the sensory properties of the sample are discussed.

During the familiarization step, assessors are given the reference "A," but it is also a good practice to give them the sample(s) "not-A." Depending on the

TABLE 6.1 Summary of the A-not-A Protocols With Advantages and Disadvantges

	Reminder Before Session				
	"A" as Reminder	"A" and "Not-A" as Reminder	Reminder Throughout Session	Advantages	Disadvantages
A-not-A (version 1)	No	No	No	• Fast • Relies on the assessors' true internal reference • Beneficial in consumer studies with heavy users of the product	• Needs assessors to be very familiar with the product space • Assessors' memory of the reference is challenged
A-not-AR (version 2)	Yes	No	No	• Light training involved • Assessors are aligned on the A sample • Assessors are more able to detect differences • Good compromise between versions 1 and 3	• Assessors do not fully know what sort of difference to expect between A and not-A (e.g., sample to sample variation of real difference detected)
A-not-AR (version 3)	Yes	Yes	No	• Assessors evaluate the samples on similar grounds • Assessors are aware of differences to expect • Does not stretch assessors' memory like version 1	• Longer training • Assessors can be biased • Assessors answer less impulsive
A-not-AR (version 4)	Yes	Yes	Yes	• Efficient if assessors do not have knowledge of the product	• Longer training and session • Not recommended for samples with high carryover effects

training level of the panel, this can help them get familiar with the sensory space of the product category. It also helps them to be aware of the type of differences that are to be expected between the "A" and "not-A" samples.

Van Hout et al. (2011) showed that assessors needed to get familiarized with the "A" sample in a training session in addition to familiarization with the method itself. It was also found that the learning curve for the A-not-A method was shallow, as the performance of the panel was still improving after six testing sessions of the A-not-A when compared with the nonattribute-specified 2-alternative forced choice (2-AFC) and 2-alternative forced choice with reminder (2-AFCR).

Familiarization with the method itself can be done using a set of samples displaying large/obvious differences to start with. This would be an easy step for assessors so that they can be used to and familiar with filling in the questionnaire. From the second familiarization session, those differences can already be reduced as the assessors get more and more used to the method.

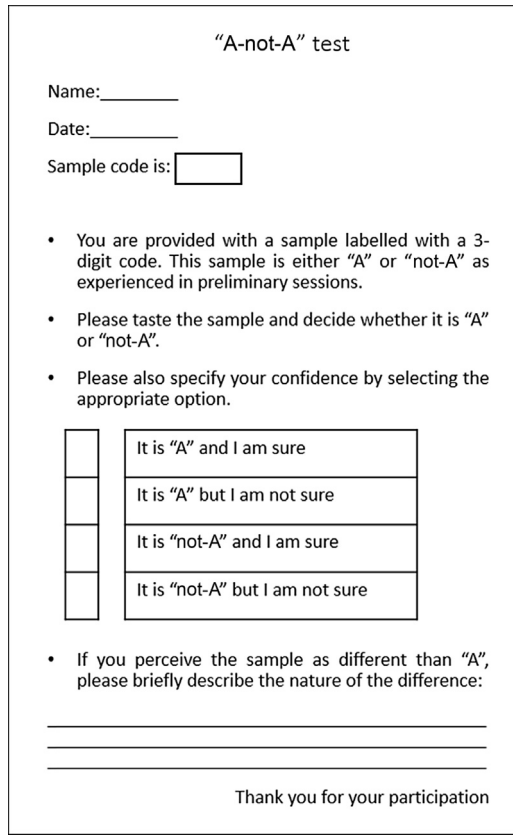
In the context of QC/QA, when assessors have received extensive training on the sensory properties of the sample A, a refamiliarization step might not be necessary each time they perform an A-not-A assessment. However, as part of the QA/QC program implemented, it is useful to have planned several refamiliarization sessions throughout the year to make sure assessors are confident on the sensory properties of the reference A.

2.2 Testing

When carrying out the test, if a reminder "A" was given, it is removed and assessors are given a three digit-coded sample and asked to evaluate it. This sample can either be "A" or "not-A." Assessors must determine if the tested sample is the reference "A" or not the reference: "not-A." It is generally recommended to ask assessors about their confidence level, which typically includes the following options: absolutely sure, fairly sure, not very sure, and just guessed. Asking about confidence can be helpful in the context of training to monitor assessors' sureness in addition to their answer. Testing with sureness also allows for R-index computation (see Section 4).

Depending on the nature of the sample (taking into account strength, carryover effect), subsequent samples can be evaluated by the assessors. In some cases, and especially if the type of samples tested allow for it, it is possible to include one or several reminder samples in between tested samples to keep assessors aware of the sensory characteristics of sample A. This, however, will have to be taken into account in the analysis of the data.

Fig. 6.1 gives an example of an A-not-A testing sheet. The experimenter can also include a comment box so that assessors can briefly indicate why a sample is different than the reference "A." This can be a convenient way to identify why "A" can be perceived differently than "not-A" in the case of statistically significant difference. In the case where assessors carry out the


"A-not-A" test

Name:_____

Date:_____

Sample code is: ☐

- You are provided with a sample labelled with a 3-digit code. This sample is either "A" or "not-A" as experienced in preliminary sessions.
- Please taste the sample and decide whether it is "A" or "not-A".
- Please also specify your confidence by selecting the appropriate option.

☐	It is "A" and I am sure
☐	It is "A" but I am not sure
☐	It is "not-A" and I am sure
☐	It is "not-A" but I am not sure

- If you perceive the sample as different than "A", please briefly describe the nature of the difference:

_____

_____

_____

Thank you for your participation

FIGURE 6.1 Example of a tasting sheet for the A-not-A method.

A-not-A test on a regular basis and when some specific differences are expected, the response sheet can also include a grid displaying possible sensory attributes that could differentiate both tested samples. However, this table should not influence assessors when deciding whether they are tasting "A" or "not-A." Collecting sensory information about possible differences between the two samples should also be taken from assessors correctly identifying the "not-A" sample as being indeed "not-A". It is important to stress that collecting explanations or reasons from assessors when performing the test is for guidance only, and it should not replace a sensory descriptive test for which the primary aim is to describe products whereas the aim of the A-not-A test is to detect if products are different.

2.3 Type of Assessors

In any sensory experiment, in addition to how many assessors to select (Meilgaard et al. recommend between 10 and 50), comes the question of what

type of assessor to recruit/use to perform the test. Usually, the recommendation is to select either naïve or trained assessors. There are fundamental differences in the way trained assessors and consumers perform sensory experiments. Trained assessors will adopt an analytical/objective approach while consumers will adopt an affective/subjective approach. When consumers are heavy users of a product and also have an emotional link to it, they can be more sensitive to small changes and can be more discriminative than a sensory trained panel (Lee, 2010). It is important to stress that a mix of both trained assessors and consumers should not be used when selecting the assessors for the test (BS ISO 8588:1987).

Table A2.8, in Appendix 2 (Bi, 2006) helps determine how many assessors to select for the A-not-A test based on estimated probability P_A (probability of response "A" when sample "A" given) and P_N (probability of response "A" when sample "not-A" given) for a power of 0.8 and a significance level $\alpha \leq 0.1$ and 0.05. For example, assuming P_N 0.4 and a sensory difference $\delta = 1$, in a monadic design, the sample size required would be 21 for a significance level $\alpha \leq 0.1$ and 26 for a significance level $\alpha \leq 0.05$.

Defining which parameters to select before the test can be tricky for the experimenter. The level of risk α (also called type I error) is defined as the probability of saying that samples are different when in fact they are the same. In difference testing (when we want to check whether samples are different), the α risk should be minimized. The following levels for the α risk can be interpreted as:

- 10%−5%: slight evidence that a difference was apparent
- 5%−1%: moderate evidence that a difference was apparent
- 1%−0.1%: strong evidence that a difference was apparent
- Less than 0.1%: very strong evidence that a difference was apparent

The power of the test can be defined as the probability of detecting a difference when it really exists. The closer the value to 1, the more we will be able to detect a difference when it exists. The power of the test has a direct impact on the number of assessors to select. Usually, for difference testing, a power of 0.8 is acceptable.

P_A and P_N would be set up based on previous similar experiments. δ represents an index of sensory difference or similarity (Bi, 2006). Its level expresses the size of the expected difference between A and not-A. Practically, a $\delta = 1$ is equivalent to 76% of discriminators in a 2-AFC test or 42% in a triangle test.

3. WHEN TO USE THE A-NOT-A TEST

There are important aspects related to tasting that must be taken into account when designing a sensory study. Because of the nature of samples used, several problems can arise such as carryover effects for strongly flavored

samples or limits due to the effects of consuming a specific sample (e.g., tobacco or alcohol products). To counter those issues, assessors are usually given palate cleansers and time breaks for them to rest between samples. Those solutions are usually effective, but sometimes the number of given samples merely needs to be reduced. In that case, methods such as the A-not-A help, as the number of presented samples can be as low as only 1 (assuming assessors are familiar with the reference A) or more if reminders are used or if the test is replicated. The A-not-A as a single presentation test can be useful in giving directions about the difference between two samples while keeping the number of samples presented in a single tasting session very low. This makes the A-not-A method usable not only with high carryover effect samples but also with less intense samples. This method is generally regarded as suitable for most types of products (Lee et al., 2007).

The A-not-A test can also be used when assessors are often exposed to the same sample. This is the case in a production environment where few variants of the sample are produced making assessors very familiar with the same sample, in this case, the reference sample: "A." Even if assessors know its sensory characteristics, it is a good practice to regularly retrain them by presenting them with it before the test, especially if assessors did not receive extensive training.

The A-not-A test is often used when there is a slight visual difference between two samples (color, size, shape) and an objective comparison is needed (Rogers, 2010; BS ISO 855:1987, Lawless and Heymann, 1999). It is, however, important to keep in mind that if the visual difference is too important, assessors are likely to remember it and will be biased during the evaluation and make their judgment on unwanted stimuli (Lawless and Heymann, 1999).

4. ANALYSIS OF A-NOT-A RESULTS

4.1 Chi-Squared Model

Data generated after an A-not-A experiment can be summarized as presented in Table 6.2.

TABLE 6.2 Output Example of A-Not-A Method With 150 Assessors

		Sample Presented Is		
		"A"	"not-A"	Total
Number of responses identifying tested sample as	"A"	50	30	80
	"not-A"	25	45	70
Total		75	75	150

In this example, 150 assessors are given one sample: either "A" or "not-A." Therefore, 75 "A" and "not-A" samples are tested by assessors. Among those, sample "A" was described as "A" 50 times and as "not-A" 25 times, while sample "not-A" was described rightly as "not-A" 45 times and as "A" 30 times. The aim now is to know whether we can conclude that "A" and "not-A" samples are different or not.

Such a design is called monadic (Bi and Ennis, 1999) as assessors are only given one sample to evaluate, and in addition, the number of assessors getting A and the number of assessors getting not-A is known in advance. To analyze these types of data, Pearson Chi-squared (χ^2) test for homogeneity is usually used.

This test is described in ISO BS 5929-5:1988 and its statistic is:

$$\chi^2 = \sum_{j=1}^{2}\sum_{i=1}^{2}\frac{\left(n_{i,j} - E_t\right)^2}{E_t}$$

where $n_{i,j}$ is the observed value in cell $(i;j)$ of the contingency table; E_t is, for each cell, the product of the sum of the row, times the sum of the column given, divided by the total number of answers. For example, for a cell expressing the number of correct answers given when "A" was presented (i.e., 50), it is equal to the multiplication of the total number of "A" answers by the total number of "A" presented divided by the total number of answers (i.e., $75 \times 80/150$).

Therefore, we have:

$$\chi^2 = \frac{\left(50 - \frac{(75\times 80)}{150}\right)^2}{\frac{(75\times 80)}{150}} + \frac{\left(25 - \frac{(70\times 75)}{150}\right)^2}{\frac{(70\times 75)}{150}} + \frac{\left(30 - \frac{(80\times 75)}{150}\right)^2}{\frac{(80\times 75)}{150}} + \frac{\left(45 - \frac{(75\times 70)}{150}\right)^2}{\frac{(75\times 70)}{150}} = 10.714$$

Applying the formula, the calculated χ^2 is 10.714. This calculated value (or observed value) need to be compared to critical value that can be found in the χ^2 critical value table (Table A2.9, in Appendix 2). For 1 degree of freedom (defined by number of tested samples minus 1) and a significance level $\alpha \leq 0.05$, the critical value is 3.84. As our observed value is above the critical value, we conclude that there is a statistically significant difference between the two samples "A" and "not-A."

Most statistical packages include a chi-squared test for homogeneity in their available analysis. For example, this computation can easily be done with

the R software using the `chisq.test()` formula. Below is the syntax the user could use:

```
> x <- matrix(c(50, 25, 30, 45), ncol = 2)
> chisq.test(x, correct=F)

        Pearson's Chi-squared test

data: x
X-squared = 10.714, df = 1, p-value = 0.001063
```

Continuity correction (necessary for low values in the contingency table) can be applied by replacing FALSE by TRUE in the function arguments. The continuity correction should be applied when at least one cell on the contingency table (for example Table 6.2 above) is less than 5.

This function returns both the observed chi-squared as well as the *p*-value, which would be for a 95% confidence level interpreted as:

- Less than 0.05: a statistically significant difference exists
- Above 0.05: no statistically significant difference

The analysis of the A-not-A test will depend on the type of design used. Bi and Ennis (1999) have detailed different statistical models for the analysis of the data generated depending if the test design used during the sensory testing is monadic (Pearson χ^2 test for homogeneity—example above), mixed (Pearson χ^2 test for independence), or paired (McNemar χ^2 test for correlated proportion test).

In a mixed design, the number of assessors getting "A" and the number of assessors getting "not-A" are not known in advance but distributed randomly. To do so, a randomized design has to be made either in advance or before the test, where assessors randomly pick a sample to evaluate. While for the monadic design, the aim of the statistical test is to compare the proportion of "A" responses from assessors initially getting "A" versus assessors initially getting "not-A," in the mixed design, the aim of the test is to estimate whether the presentation of "A" or "not-A" to the assessors have an effect on the number of "A" answers (Bi, 2006).

In both monadic and mixed design, the compared proportions are independent. However, in a paired design, those two proportions are not independent anymore as assessors are given both "A" and "not-A" to evaluate. In such a design, it is better not to tell assessors in advance that they will be evaluating both "A" and "not-A" samples.

4.1.1 Note on Replicated Testing

In the case of replicated testing, assessors receive several samples during one session. The number of samples to evaluate should be decided prior the beginning of the sessions. Depending on the level of training of the assessors,

it is a good practice to (re)present assessors with a reminder "A" in between tested samples to avoid confusion. In replicated testing, instructions given to assessors are unchanged, but analysis of the data will differ slightly. In the case of monadic and mixed design, adjustments to the Pearson χ^2 test must be made [Beta Binomial model (Bi, 2006) and Dirichlet Multinomial model (Ennis and Bi, 1999), respectively].

It is common that companies do not have access to the recommended number of assessors for the desired risk that they are willing to take. To "increase" the sample size, replications are usually made, especially if no additional resources are available. While it is acceptable to do replications, it is recommended to do it on a different tasting session. This would avoid additional sensory fatigue and unwanted familiarization of the assessors with the samples.

4.2 Thurstonian Distance

Another approach to interpret data from the A-not-A test is to apply Thurstonian modeling. In Thurstonian models developed by Louis Leon Thurstone, the perception of a stimulus varies in intensity in a probabilistic way. In other words, it describes and takes into account that when most assessors are perceiving a stimulus at an average score, some also perceive it weaker while others perceive it to be stronger. This is also the case on an individual level as the perception of the stimuli may change over repeated consumption (ASTM E2262). The variability is observed because of many factors such as not only psychological and physiological reasons but also product variation (illustration in Fig. 6.2).

In the case of discrimination tests, we often are interested in knowing whether there is a perceptible difference between two samples. The chi-squared test described above, as well as widely used binomial statistics, tells us whether assessors have done better than guessing.

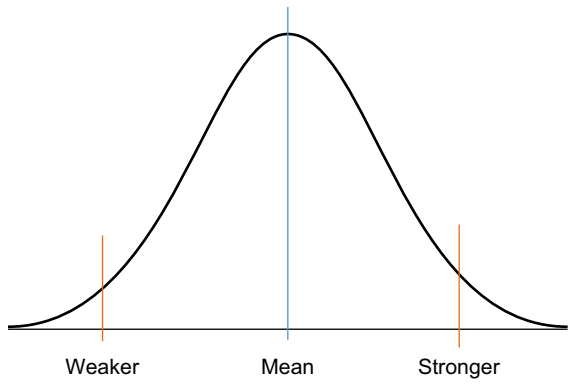


FIGURE 6.2 Probabilistic representation of the perception of a sensory stimulus.

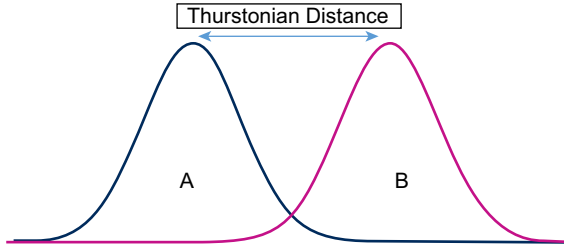


FIGURE 6.3 Representation of the Thurstonian distance (δ) between two sensory stimuli.

Thurstonian modeling gives insights about the magnitude of the sensory difference between the two samples rather than only a different/not different answer.

When two samples are compared in a discrimination test, the magnitude of the difference between them can be expressed as a Thurstonian distance. As shown in Fig. 6.3, sample B is on average perceived with a higher intensity than sample A, although this is not the case in a very few situations, as seen by the small portion of the left tail of the B curve overlapping with the right tail of the A curve.

In Fig. 6.3, A and B represent the intensity of two sensory stimuli. It shows that sample B tends to be perceived more intense than sample A. The size of the difference between A and B can be quantified in Thurstonian modeling and expressed as Thurstonian distance or δ.

The statistic associated with the Thurstonian distance δ is called d' (d prime). Theoretically, the d' measure is independent from the sensory test used (ASTM E-2262) unlike binomial statistics where the number of correct answers is biased by the discrimination test used (Brockhoff and Christensen, 2009). Therefore, d' is a useful tool to compare results from different tests even if those were performed under different conditions. For example, it is possible to compare data coming from two different panels using two different methods or even with different levels of training such as an expert panel and a consumer panel. In addition, as it is possible to compare if two d' values are significantly different from each other, Thurstonian modeling can also be applied to monitor the performance of a panel over time if, for example, they are given the same set of samples to compare at different time points.

4.2.1 Decision Rule for the "A-Not-A"

The decision rule or decision strategy aims at understanding the cognitive process of assessors when performing a sensory discrimination test. In m-AFC test ($m > 2$) the decision rule is called the skimming strategy where the assessor compares the perceived intensity of all samples and selects (skims off) the appropriate one (for example, the most bitter). In tests such as triangle, or duo-trio, the cognitive strategy is a comparison of sensory distances where the most distant sample will be selected as the odd one. In the A-not-A test, the cognitive strategy is neither of them: it is the β criterion. This criterion can be

seen as the assessors' internal reference to which they will base their answer during the test. It is assumed that this criterion is fixed for one assessor over several repetitions but will change from assessor to assessor. In some cases and in practice, relying on assessors' internal references can be a problem because of poor memory, resulting in lower discrimination between the samples (Van Hout et al., 2011). Giving a reminder sample of the reference to assessors not only aligns assessors' judgments, but also helps them in recognizing if a presented stimulus is close or not to the reference A. This can have a significant impact in increasing the test performance when assessors become more familiar and confident about the reference A over multiple sessions.

ASTM E2262 describes how to compute the d' and its variance for A-not-A tests with a monadic design.

The first step is to compute the following two proportions:

- P_A: number of "A" responses when given the "A" sample
- P_{nA}: number of "A" responses when given the "not-A" sample.

Using Table A2.10, in Appendix 2, the d' value can be found at the intersection of both P_A and P_{nA}. The d' alone can be misleading as here the number of assessors taking part into the test has not been taken into account. It is important to have an idea of the variance of the d' and therefore have an idea of its actual range. To calculate the variance of the d', Table A2.11, in Appendix 2 must be used. The use of Table A2.11 is quite similar to the use of Table A2.10, as the value to find in the table (B value) is at the intersection of P_A and P_{nA}. Once the B value found, the standard deviation S^2 can be expressed as:

$$S^2 = \frac{B}{n}$$

with n being the number of assessors receiving either "A" or "not-A."

Taking the data from the example in Table 6.2, P_A and P_{nA} can be calculated as follows:

$$P_A = \frac{50}{75} = 0.67; \quad P_{nA} = \frac{30}{75} = 0.4$$

Using Table A2.10 and A2.11, in Appendix 2, the associated d' can be found such as $d' = 0.693$ as well as the associated B value (3.294). Hence, the variance of the d' is equal to:

$$S^2 = \frac{3.294}{75} = 0.044$$

Calculating the variance of the d' is necessary to compute its confidence interval. The upper and lower confidence intervals can be calculated as below at 95% level:

$$Lower\ CI = d' - Z_{\frac{\alpha}{2}} \times \sqrt{S^2} = 0.693 - 1.96 \times \sqrt{0.044} = 0.28$$

$$Upper\ CI = d' + Z_{\frac{\alpha}{2}} \times \sqrt{S^2} = 0.693 + 1.96 \times \sqrt{0.044} = 1.10$$

The computed d' value can be as low as 0.28 and as high as 1.10 at 95% confidence level.

The sensR package (Christensen and Brockhoff, 2016) (available from the R free software) includes a function called `AnotA()`, which computes the d' and its variance for the A-not-A with monadic design. Among arguments that need to be inputted are the number of "A" responses when "A" was presented and the number of "A" responses when the "not-A" sample was presented such as:

```
> AnotA (50,75,30,75)

Call:   AnotA(x1 = 50, n1 = 75, x2 = 30, n2 = 75)

Results for the A-Not A test:

           Estimate Std. Error       Lower      Upper        P-value
d-prime  0.6840744  0.2094062   0.2736459   1.094503   0.0008885696
```

The function `AnotA()` also returns the p-value associated with the one tailed Fisher exact test also mentioned in Bi (2006). By typing "`?AnotA`" in the R console, the user get access to the help page of the function, providing details and examples.

4.3 R-Index

Another way of looking at the data generated from an "A-not-A" experiment is to compute the R-index. The R-index was developed by J. Brown (1974) and can be interpreted as the predicted proportion of correct responses in a 2-AFC test. One interesting aspect of the R-index is that in addition to considering assessors' sureness when performing the test, it is also assumption free of the underlying sensory difference between the two samples (Ennis et al., 2014).

Typically, to compute the R-index, respondents will be presented with a sample ("A" or "not-A") and their possible answer will be one of the following four:

- A sure (A!)
- A not sure (A?)
- Not-A sure (not-A!)
- Not-A not sure (not-A?)

By considering which assessors received samples "A" and "not-A", generated data can be put as shown in Table 6.3:

TABLE 6.3 A-Not-A Data Matrix for R-Index Computation

	A!	A?	Not-A!	Not-A?
Sample A presented	a	b	c	d
Sample not-A presented	e	f	g	h

The R-index is then calculated as:

$$R\text{-}index = \frac{a(f+g+h)+b(g+h)+ch+\frac{1}{2}(ae+bf+cg+dh)}{(a+b+c+d)(e+f+g+h)}$$

The computed R-index value varies from 50% (no discrimination) to 100% (full discrimination).

The R-index has the advantage of being easy and intuitive to interpret as it directly expresses an estimation of the percentage of people being able to discriminate between two samples in addition to be a powerful nonparametric test (Bickel and Doksum, 1977). However, the R-index is method dependent (Ennis et al., 2014), making it difficult to compare outcomes from two different methods.

5. CONCLUSION

The A-not-A method is a relatively simple method in appearance, but it has many subtleties from the design, to the analysis and interpretation of the data generated. However, it is a simple task for assessors to carry out and can be applied in both analytical sensory (e.g., trained/expert panels) and consumer studies involving naïve subjects. Even if underused with consumers, the A-not-A can provide useful insight with heavy users of a type of products or of a particular brand as, in essence, it relies on assessors' internal reference. The method is less recommended when assessors are untrained and/or with no experience on the products, and tests such as the 2-AFC or triangle test may be more suitable.

6. CASE STUDY

Case Study 1: Use of the R-Index

As part of the expansion of its main manufacturing site, a company is trying to assess whether the newly added production line has an effect on the sensory characteristics of its products. The sensory scientist is asked to check if there is a perceptible difference between the products manufactured on the new line and the products manufactured on the current line.

An A-not-A test is set up with 50 assessors; all company employees familiar with the sample produced on the current line. To ensure that assessors assess samples on similar grounds, a preliminary tasting session was organized so that assessors could get refamiliarized with the samples produced on the current line.

A monadic design was used, so 25 assessors got the "A" sample while 25 assessors got the "not-A" sample. Assessors were also asked about their sureness when deciding if the tasted sample was "A" or "not-A".

The table below summarizes the results obtained after tasting:

	A!	A?	Not A?	Not A!	Total
Sample "A" presented	$a = 10$	$b = 5$	$c = 8$	$d = 2$	25
Sample "not-A" presented	$e = 6$	$f = 6$	$g = 3$	$h = 10$	25

From the 25 assessors who received sample "A", 15 of them identified it as "A" with 10 being sure and 5 not sure and 10 identified it as "not-A" with 2 being sure and 8 not sure.

From the 25 assessors who received sample "not-A", 12 of them identified it as A with 6 being sure and 6 not sure and 13 identified it as "not-A" with 10 being sure and 3 not sure.

To communicate the results, the sensory scientist chooses to compute the R-index value as it is relatively straightforward to interpret.

The R-index value is defined as:

$$R\text{-}index = \frac{a(f+g+h) + b(g+h) + ch + \frac{1}{2}(ae + bf + cg + dh)}{(a+b+c+d)(e+f+g+h)}$$

Therefore,

$$R\text{-}index = \frac{10 \times (6+3+10) + 5 \times (3+10) + 8 \times 10 + \frac{1}{2}(10 \times 6 + 5 \times 6 + 8 \times 3 + 2 \times 10)}{(10+5+8+2) \times (6+6+3+10)} = 0.64$$

As the computed R-index is 0.64, the sensory scientist concludes that if given side by side, 64% of assessors could distinguish between "A" and "not-A". However, if there were truly no differences between those two samples, the proportion of correct answers would be close to 50% by chance.

Based on the table for testing the significance of the R-index, Table A2.3 in Appendix 2 (Bi and O'Mahony, 2007), for $n = 50$ and $\alpha = 0.05$ (one-tail test), the computed R-index should be higher than 59.33% (50 plus the table value of 9.33) to claim that there is a perceptible difference between the two samples.

The sensory scientist can claim that there is a perceptible difference between samples produced on the current line and samples produced on the new line at 95% confidence level.

Case Study 2: Similarity Testing Based on Bi (2006)

Note on Similarity Testing

Stating that there are no statistically significant differences between two samples is not equivalent to saying that the two samples are similar. This would be the case if an ingredient replacement is taking place (e.g., change of supplier) and samples should be interchangeable without consumers noticing a difference. Bi (2006) suggested a χ^2 for similarity based on Dunnett and Gent (1977).

In this similarity approach, expected proportions must be computed taking into account the limit for which "A" and "not-A" can be claimed to be similar. Those expected proportions consider a value called Δ_0, which expresses the

maximum allowable difference that can be observed between the two proportions P_A and P_{nA} to claim similarity. P_A and P_{nA} are, respectively, the proportion of "A" answers when sample A was presented and the proportion of "A" answers when "not A" was presented.

A company is changing supplier for a key ingredient in their recipe. They want to know whether this supplier change will affect their product's sensory properties and if consumers would notice a difference compared to the existing product.

A monadic A-not-A test for similarity was set up with 200 consumers (100 receiving "A" and 100 receiving "not-A"—A being the original product and not-A being the reformulated sample with the ingredient from the new supplier). The maximum allowable difference that can be observed between the two proportions P_A and P_{nA} to claim similarity was set up to 0.2.

At the end of the test, from the 100 assessors who received A, 50 of them identified it as A and from the 100 assessors who received not-A, 42 of them identified it as A. As per Bi (2006), the expected proportion of A samples is calculated as:

$$\widehat{\pi}_A = \frac{x + y + n_N\Delta_0}{n_A + n_N} = \frac{50 + 42 + 100 \times 0.2}{100 + 100} = 0.56$$

with n_A and n_N being, respectively, the number of assessors receiving A and receiving "not-A" and x and y being the observed number of responses "A" when presented with "A" and "not-A," respectively.

The χ^2 for similarity is then calculated as:

$$\begin{aligned}\chi^2 &= (x - x')^2\left[\frac{1}{x} + \frac{1}{m - x'} + \frac{1}{n_A - x'} + \frac{1}{n_N - m + x'}\right] \\ &= (50 - 56)^2\left[\frac{1}{50} + \frac{1}{92 - 56} + \frac{1}{100 - 56} + \frac{1}{100 - 92 + 56}\right] \\ &= 3.10\end{aligned}$$

with $m = x + y$ and the expected number of assessors finding A (noted x') calculated as $100 \times 0.56 = 56$.

For one degree of freedom, the p-value (one-sided) associated for the χ^2 test is equal to **0.039**. This p-value can easily be computed using the R software using the formula:

```
(1-pchisq(3.10,1))/2
```

As the computed p-value is lower than 0.05, the sensory scientist can claim that the product reformulated with the ingredient from the new supplier is perceived similarly to consumers compared with the original product.

REFERENCES

ASTM International, E2262 − 03, Standard Practice for Estimating Thurstonian Discriminal Distances.

Bi, J., Ennis, D., 1999. The power of the "A"-"NOT-A" method. Journal of Sensory Studies 16 (1), 343−359.

Bi, J., 2006. Sensory Discrimination Tests and Measurements, Statistical Principles, Procedures and Tables. Blackwell Publishing Ltd., Oxford, UK, pp. 01−02.

Bi, J., Lee, S.H., O'Mahony, M., 2013a. Statistical analysis of receiver operating characteristic (ROC) curves for the ratings of the A-NOT-A and the same-different methods. Journal of Sensory Studies 28 (1), 34−46.

Bi, J., O'Mahony, M., 2007. Updated and extended table for testing the significance of the R-Index. Journal of Sensory Studies 22 (6), 713−720.

Bi, J., O'Mahony, M., Lee, H.S., 2013b. Nonparametric estimation of d′ and its variance for the A-NOT-A with reminder. Journal of Sensory Studies 28 (1), 381−386.

Bickel, P.J., Doksum, K.A., 1977. Mathematical Statistics: Basic Ideas and Selected Topics. Holden-Day, Inc., San Francisco, CA, pp. 350−353.

Brockhoff, P.B., Christensen, R.H.B., 2009. Thurstonian models for sensory discrimination tests as generalized linear models. Food Quality and Preference 21 (1), 330−338.

Brown, J., 1974. Recognition assessed by rating and ranking. British Journal of Psychology 65 (1), 13−22.

BSI, BS 5929-5:1988, ISO 8588:1987, Sensory Analysis − Methodology − "A"-"not-A" Test.

Christensen, R.H.B., Brockhoff, P.B., 2016. sensR − an R-Package for Sensory Discrimination. R package version 1.4-7. http://www.cran.r-project.org/package=sensR/.

Dunnett, C.W., Gent, M., 1977. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. Biometrics 33, 593−602.

Ennis, D.M., Bi, J., 1999. The Dirichlet-multinomial model: accounting for inter-trial variation in replicated ratings. Journal of Sensory Studies 14 (3), 321−345.

Ennis, J.M., Rousseau, B., Ennis, M., 2014. Sensory difference tests as measurement instruments: a review of recent advances. Journal of Sensory Studies 29 (1), 89−102.

Lawless, H., Heymann, H., 1999. Sensory Evaluation of Food. Springer, pp. 79−100.

Lee, H.S., Van Hout, D., O'Mahony, M., 2007. Sensory difference tests for margarine: a comparison of R-indices derived from ranking and A-not-A methods considering response bias and cognitive strategies. Food Quality and Preference 18 (1), 675−680.

Lee, H.S., 2010. Measuring Food or Consumers? Latest Ideas and Methodological Issues in Difference Tests, 10th Sensometrics, Rotterdam.

Rogers, L.L., 2010. Sensory methods for quality control. In: Kilcast, D. (Ed.), Sensory Analysis for Food and Beverage Quality Control a Practical Guide. Woodhead Publishing Limited, Cambridge, pp. 49−74.

Stocks, M.A., Van Hout, D., Hautus, M.J., 2013. Cognitive decision strategies adopted in reminder tasks by trained judges when discriminating aqueous solutions differing in the concentration of citric acid. Journal of Sensory Studies 28 (3), 217−229.

Stone, H., Sidel, J., 1993. Sensory Evaluation Practices. Elsevier.

Van Hout, D., Hautus, M.J., Lee, H.S., 2011. Investigation of test performance over repeated sessions using signal detection theory: comparison of three nonattribute-specified difference tests 2-AFCR, A-NOT-A and 2-AFC. Journal of Sensory Studies 26 (1), 311−321.

FURTHER READING

Bi, J., Ennis, D., 2001. Statistical models for the a-not-a method. Journal of Sensory Studies 16 (1), 215−237.

Hautus, M.J., Shepherd, D., Peng, M., 2011. Decision strategies for the A Not-A, 2AFC and 2AFC-reminder tasks: Empirical tests. Food Quality and Preference 22 (1), 433−442.

Morten, C., Meilgaard, M.C., Carr, T., Civille, G.V., 2006. Sensory Evaluation Techniques, fourth ed. CRC Press.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Stone, H., Sidel, J., 2004. Sensory Evaluation Practices. Elsevier.