Reprinted from IFPress (2006) 9(3)2,3

Indices of Sensory Difference: *R*-Index and *d'*

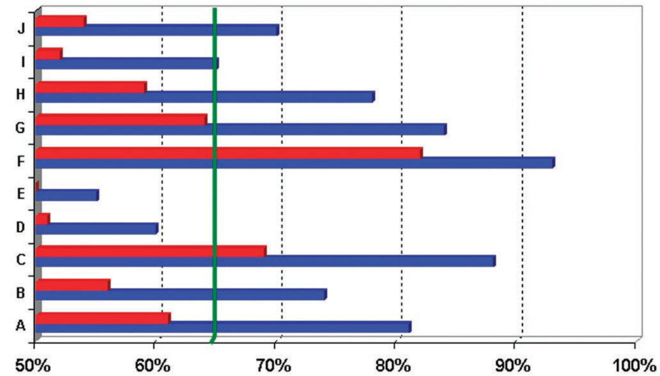
Benoît Rousseau

Background: Discrimination methodologies are used to investigate similarities and differences among variants of manufactured products. Applications typically involve cost savings, storage and product improvements. A researcher often requires more information than a simple test of significance which can be readily changed by simply altering the sample size used. In many cases, estimation of the degree of difference between pairs of products is more important. Two indices obtained from discrimination methods have been largely used in industry: *d'*[1] and the *R*-Index[2]. *R*-Indices are usually obtained from the Same-Different and A/Not A (Yes/No task) with sureness judgments or using a ranking procedure. *d'* values can be estimated from the same procedures as well as from a variety of discrimination and scaling methods (triangle, duo-trio, 2-alternative forced choice or 2-AFC, category scaling, etc.) Both indices measure degrees of difference for which a greater value indicates a larger sensory difference. *R*-Indices take values between 50% (no difference) and 100% (maximum difference), while *d'* values vary between 0 and infinity. Usually *d'* values between 0 and 2 are encountered when using discrimination and scaling methodologies. Both measures have the same goal. However, are they interchangeable?

Scenario: Your company's two main peanut butter manufacturing plants have been using sensory evaluation in their quality control program. *R*-Indices allow the specification of a common accept/reject decision rule. Upon preliminary research using the same-different method, an *R*-Index of 65% was set as the standard so that *R*-indices greater than 65% indicate that a product is out of specifications and should be rejected. In this scenario, assume that most or all of the variability observed is due to run-run differences and that the sample size within each trial is sufficiently large to make within trial variability small. Shortly after implementation of this QC program, one of the plants (Plant A) is found to reject a greater proportion of samples than the other (Plant B) consistently. Concerned about this trend, your management asks you to investigate the source of this greater proportion of rejection.

Differences in Methodology: One possible reason for inter-plant differences may be due to greater process variability in Plant A. Before investing valuable time identifying variation in the Plant A process, you first look at the methodology used in each location. Both locations estimate *R*-Indices between a 'Gold Standard' and production samples. However, methodologies are different: panelists in Plant A have been trained to recognize the Gold Standard as a reference and use the 'A-Not A' methodology in which they evaluate each sample monadically. Panelists rate presented samples as "Reference sure", "Reference not sure", "Not Reference sure", "Not Reference not sure." In the B location, a same-different method is used in which panelists evaluate pairs of samples (the production sample and Gold Standard) and state whether the two samples are "Different sure", "Different not sure", "Same sure" or "Same not sure".

In order to study the differences measured in each location, you decide to have the same 10 peanut butter samples evaluated by each panel against the Gold Standard. The results are shown in Figure 1.



	A	B	C	D	E	F	G	H	I	J
Plant B, SD *R*-Index	61%	56%	69%	51%	50%	82%	64%	59%	52%	54%
Plant A, A/NA *R*-Index	81%	74%	88%	60%	55%	93%	84%	78%	65%	70%

Figure 1. *R*-Indices between the Gold Standard and each peanut butter sample for each plant.

The results you obtain confirm the discrepancies that had been observed earlier as the difference between the plants is highly significant. Using Plant A, seven samples would be rejected, while Plant B would reject only two samples. Since the samples evaluated were the same, the variability must come from differences between the panels or the manner in which testing is conducted. It is possible that the Plant A panelists are more sensitive than those of Plant B. Another possibility is that the methodologies might actually not yield the same estimate of the underlying degree of difference as measured by the *R*-Index.

***R*-Index and *d'*:** *R*-Indices were introduced by Brown (1974)[2] to investigate various memory mechanisms. An *R*-Index obtained with the A-Not A (yes-no) method corresponds to the area under a Receiver Operating Characteristic curve (ROC curve), which in turn equals the predicted proportion of tests correct in a 2-AFC[3]. *R*-Indices can also be obtained with a same-different procedure. However, in this case, they no longer correspond to a proportion of tests correct in the 2-AFC. They correspond to a predicted proportion of correct decisions obtained using an alternative decision rule in the same-different method[4]. Therefore, for the same underlying sensory difference, the A-Not A and same-different with sureness judgments will not yield the same *R*-Index. This difference has also been observed with rating and ranking *R*-Indices[5].

R-Indices are method-dependent, as are proportions of correct response for the triangular and 3-AFC or the 2-AFC and duo-trio protocols[6]. In order to establish the statistical significance of *R*-Index values, a table has been published[7]. However, this table does not take into account the specific decision rule associated with the various protocols. Any *R*-Index and its associated significance level will depend on how the data was generated. The binomial approach to analyzing data from a variety of discrimination methodologies suffers from the same limitation.

As mentioned earlier, *d′* values can also be calculated using a variety of procedures. Using the appropriate Thurstonian model (ratings with 4 categories for the A-Not A with sureness judgments and degree of difference with 4 categories for the same-different with sureness judgments), the relationship between *d′* and *R*-Indices can be generated. For each set of scale boundaries, a different relationship can be drawn. An example is shown in Figure 2.

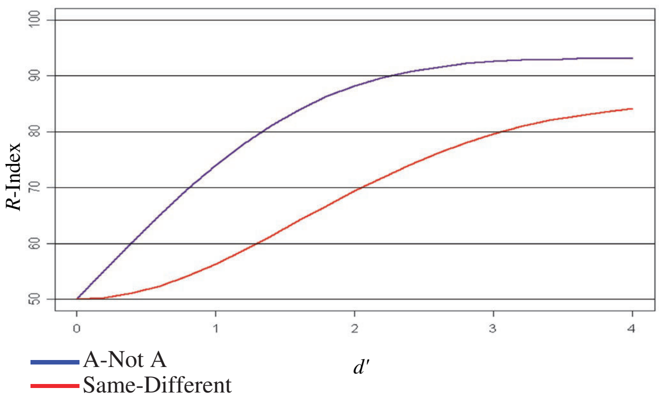
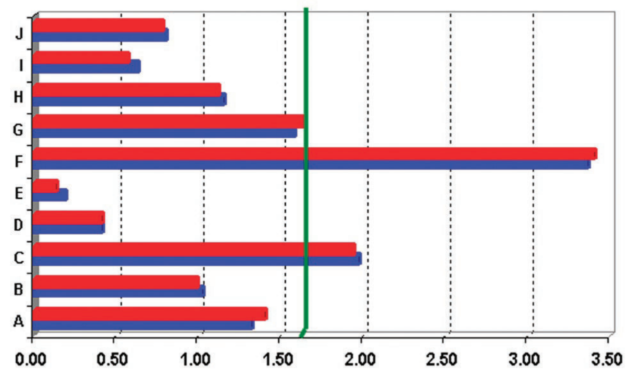


Figure 2. Relationship between the underlying sensory difference measured in terms of *d′*- and *R*-Indices calculated from the A-Not A and same-different with sureness judgments.

This relationship shows that the A-Not A method consistently yields a greater *R*-Index value than the same-different method, indicating that it is statistically more powerful. However, the main finding is that *R*-Indices do not provide method-free measures of underlying sensory differences. Similar effects exist among other sensory methodologies.

Another particularity of *R*-Indices is that they can be calculated using different numbers of categories. They can be estimated using 4, 6, 8, etc. categories. For the same underlying difference (*d′*), the calculated *R*-Indices will actually vary, not because the sensory difference is larger or smaller, but because of the task itself. In addition, simply changing the way subjects space the response categories mentally will lead to different *R*-Indices. For instance, for the same *d′* of 1.2, simulations show that the estimated *R*-Index can vary from 70% to 78%.

Estimating degrees of difference in terms of *d′* values: You revisit the data and estimate the *d′* values from the A-Not A and Same-Different. The results are shown in Figure 3. You realize that both panels were actually measuring the same degree of difference (for each sample, *d′* values are not significantly different.) The difference was simply due to the fact that *R*-Index calculations do not take into account each specific decision rule. As mentioned earlier, the size of an acceptable difference estimated from research conducted with the same-different, was an *R*-Index of 65% or less. You translate this *R*-Index into *d′* and find the reference *d′* to be 1.65. This value can then be used both with the A/Not A and same-different, but could also be used if other procedures were employed (for example the triangular method, duo-trio method, 2-AFC, ratings, and ranking)[6,8].



	A	B	C	D	E	F	G	H	I	J
Plant B, SD *d′*	1.40	0.99	1.94	0.41	0.13	3.40	1.64	1.12	0.57	0.78
Plant A, A/NA *d′*	1.32	1.02	1.97	0.41	0.19	3.36	1.58	1.15	0.63	0.80

Figure 3. *d′* values between the Gold Standard and each peanut butter sample for each plant.

Conclusion: *R*-Index and *d′* are two ways of estimating sensory differences. However, *R*-Indices are actually method-dependent, and results within the same-method will not be constant if the mental boundaries used by the subjects, while constant within an experiment, shift from one experiment to another. It is possible to use an *R*-Index to measure the size of differences between products, but one needs to be aware of the context in which they are being used. Estimating *d′* values can be a good alternative as Thurstonian models have been developed and are available for many protocols.

References

1. Ennis, D.M. (1993). The power of discrimination methods. *Journal of Sensory Studies*, **8**, 353-370.
2. Brown, J. (1974). Recognition assessed by rating and ranking. *British Journal of Psychology*, **65**, 13-22.
3. Green, D.M. (1964). General prediction relating yes-no and forced-choice results. *Journal of the Acoustical Society of America*, **36**, 1042
4. Irwin, R., Hautus, M.J, and Butcher, J.C. (1999). An area theorem for the same-different experiment. *Perception & Psychophysics*, **61**, 766-769.
5. Ishii, R., Vié, A., and O'Mahony, M. (1992). Sensory difference testing: Ranking *R*-Indices are greater than rating *R*-Indices. *Journal of Sensory Studies*, **7**, 57-61.
6. Ennis, D.M. (1998). Thurstonian scaling for difference tests. *IFPress*, **1**(3), 2-3.
7. Bi, J. and O'Mahony, M. (1995). Table for testing the significance or the *R*-Index. *Journal of Sensory Studies*, **10**, 341-347.
8. Ennis, D.M. (1999). Thurstonian models for intensity ratings. *IFPress*, **2**(3), 2-3.