AN OBJECTIVE NUMERICAL METHOD OF ASSESSING THE RELIABILITY OF TIME-INTENSITY PANELISTS

K. BLOOM

Department of Psychology
University of Waterloo
Waterloo, ON. Canada N2L 3G1

and

L.M. DUIZER and C.J. FINDLAY

Compusense Inc.
150 Research Lane
Guelph, ON, Canada N1G 4T2

Received for Publication October 16, 1994

ABSTRACT

A measure of the reliability (T-IR) of time-intensity measurements was developed based on the concept of standard deviation as a measure of panelist variability. The T-IR measure was applied to time-intensity data collected from 10 panelists evaluating the sweetness of 4 model sweetener solutions on horizontal and vertical time-intensity line orientations. T-IR scores showed that the panelists were similarly reliable across the sweeteners and orientations. As well, independent of scale orientation, responses to sweeteners were similarly reliable. The T-IR measure can be used to maintain a high level of performance by monitoring time-intensity panelists. T-IR also provides an objective method of selecting panelists for time-intensity panels.

INTRODUCTION

Many sensory evaluation tests measure perception of food flavor and texture as static events. However, intensity perception does not occur at a single point

[1]Send correspondence to L. Duizer at Compusense at the address listed above.

Journal of Sensory Studies 10 (1995) 285–294.

in time. Both texture and flavor intensity changes as the food moves through the mouth and is prepared for swallowing. The time-intensity test, which measures changes in perception of product attributes over time, is gaining wider application. Time-intensity is defined as a measurement of the rate, duration and intensity of stimulation of a single attribute through the collection of responses over an established period of time (Amerine *et al.* 1965). A typical time-intensity curve is illustrated in Fig. 1. The parameters and their definitions extracted from each time-intensity curve are found in Table 1. Each attribute evaluated by time-intensity testing has a temporal profile which represents changes in perception. The time-intensity curve graphically displays attribute intensity from peak intensity (*IMAX*) to its disappearance. The onset of perception of an attribute (*RX*), the length of time required to reach maximum intensity (*TMAX*), as well as the duration (*DUR*) of perception, can also be obtained from the time-intensity curve. The increase and decrease angles (*INC ANGLE* and *DEC ANGLE*) extracted from the time-intensity curve give an indication of the rate of onset and decline of the attribute being studied.

Time-intensity testing can play a valuable role in product development. With the use of this method, the time course of texture and flavor attributes of any food product may be optimized. As well, time-intensity curves allow researchers

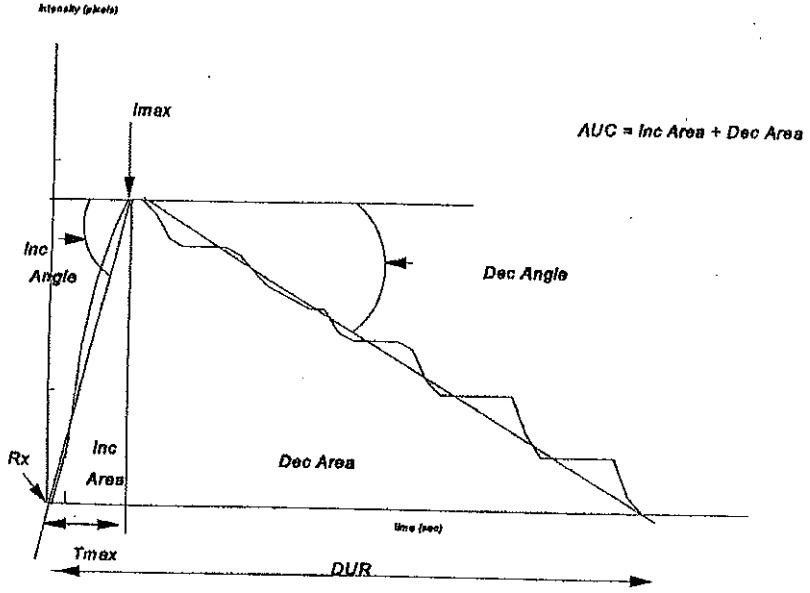

FIG. 1. TIME-INTENSITY CURVE AND PARAMETERS

TABLE 1.
TIME-INTENSITY PARAMETERS AND THEIR DEFINITIONS

PARAMETER	ABBREVIATION	DEFINITION
Maximum intensity	*IMAX*	The maximum sweetness intensity (up to 60 pixels) of each sample.
Time to maximum intensity	*TMAX*	The time (in seconds) at maximum intensity.
Duration	*DUR*	The time (in seconds) for sweetness perception (from first perception to the end of the perception).
Increase angle	*INC ANGLE*	The angle of increase to maximum intensity. This can be interpreted to be the rate of onset of sweetness of the sample.
Increase area	*INC AREA*	The area under the increasing portion of the curve.
Decrease angle	*DEC ANGLE*	The angle of decrease from maximum intensity. This can be interpreted to be the rate of decrease of sweetness perception.
Decrease area	*DEC AREA*	The area under the decreasing portion of the curve.
Area Under the Curve	*AUC*	The total area under the time-intensity curve.

to explore changes in sensory processes of a variety of foods, including changes in texture during mastication and flavor release over time. Time-intensity testing is also applicable to nonfood products. The fragrance changes in perfume, color change in lipsticks, and the effectiveness of hair spray over time can all be easily measured by time-intensity testing.

The time-intensity data reported in this paper was collected using the CSA$_{TPA}$™ program developed by Compusense. The CSA$_{TPA}$™ program allows the panelist to use a mouse to move a cursor along labelled horizontal or vertical lines, 60 pixels in length. Upon ingestion of the sample, the panelist initiates the time-intensity test and continuously moves the cursor along the line as the attribute intensity changes. The panelist ends the test by moving the cursor to the zero point of the line when the perception of the attribute ceases. With the CSA$_{TPA}$™ program, the sampling rate can be set to obtain the maximum amount of information about the attribute being evaluated. For this research, the computer was programmed to collect the data at half second time intervals.

During testing, it is necessary to include replicates of individual judgments of each sample to provide an estimate of the experimental error. The experimental error is defined as the unexplainable, natural variability of the population being studied (Meilgaard *et al.* 1991). The measurement of perceptions will in all cases have "error." Some of the error will be due to physiological and psychological variations in the panelist from trial to trial, and some due to variations in the motor responses that express the perceptions from trial to trial. The number of

trial replications required to calculate the experimental error varies depending on the product being tested. For this study, 4 replications of testing on each sample were completed.

Time-intensity evaluations of taste are therefore derived from a composite of repeated trials for each panelist. The average time-intensity curves of panelists are further combined to yield a group curve for each sample. This resultant curve is used to draw conclusions about food product attributes. Although they will rarely be identical, if the individual curves are not similar, the averaged curve may not accurately reflect the time-course of taste perception. Given the scientific and practical implications of the averaged time-intensity curve, it is imperative that the curve is composed of reliable and reproducible perceptions. Investigators must have confidence that individual panelists have similar perceptions of identical samples, and that panelists are similar as a group in their perceptions. How can the reliability of time-intensity curves be evaluated within or amongst panelists? Attempts have been made to develop a reliability measurement for single point evaluations (Mangan 1992). However, there has been no published work regarding reliability measurements of time-intensity data.

ANALYSIS OF INDIVIDUAL COMPONENTS OF THE TIME-INTENSITY CURVE

Reliability of repeated time-intensity curves can be measured by evaluating the variation in scores on each parameter of the curve for a group of panelists. In this way, a correlation coefficient would represent the relationship between the panelists' rank order on one trial versus a second trial. To evaluate the reliability of the time-intensity response, correlation coefficients may be calculated for all parameters of the curve (e.g., *IMAX, DUR, INC ANGLE*, etc.). However, there are problems associated with this method of evaluating reliability of taste perceptions. Firstly, correlation coefficients are themselves reliable only when they are derived from a sufficient number of participants. Given the training and time required, time-intensity panels are usually small. Secondly, correlation coefficients for the parameters of the curve are not independent of each other. For example, area under the curve is calculated using the combined measures of duration, maximum intensity, angle of increase, etc. For this reason, the magnitudes of correlations are not readily interpretable. Thirdly, typical time-intensity studies present more than two trials to each panelist. Although the Cronbach alpha coefficient can be used in these cases, problems such as adequate sample size are magnified with multivariate correlational procedures. Finally, correlational techniques are useful for evaluating the reliability of individual parameters across small numbers of repeated trials. We conclude, therefore, that the measurement

of reliability of time-intensity curves demands a method for evaluating the curve as a whole, and of evaluating both individual panelists and groups of panelists.

QUANTIFICATION OF RELIABILITY OF TIME-INTENSITY CURVES

We have developed a measure of reliability (T-IR) of a set of time-intensity curves, based upon the concept of standard deviation as a measure of variability. The calculation of reliability is simple and straightforward, and the resulting value, T-IR, is a standardized unit of measurement which can be used comparatively. T-IR is an inverse measure; the lower the score, the more reliable the time-intensity curves. T-IR is the absolute mean of a set of standard deviations. The standard deviations represent the variability at each sampling point in a set of repeated time-intensity trials. The number of points depends on the sampling rate used in the study and on the duration of the time-intensity response of the panelist(s).

Figure 2 graphically illustrates the calculation of T-IR for three hypothetical time-intensity curves. For illustration, each of the 10 points on the abscissa

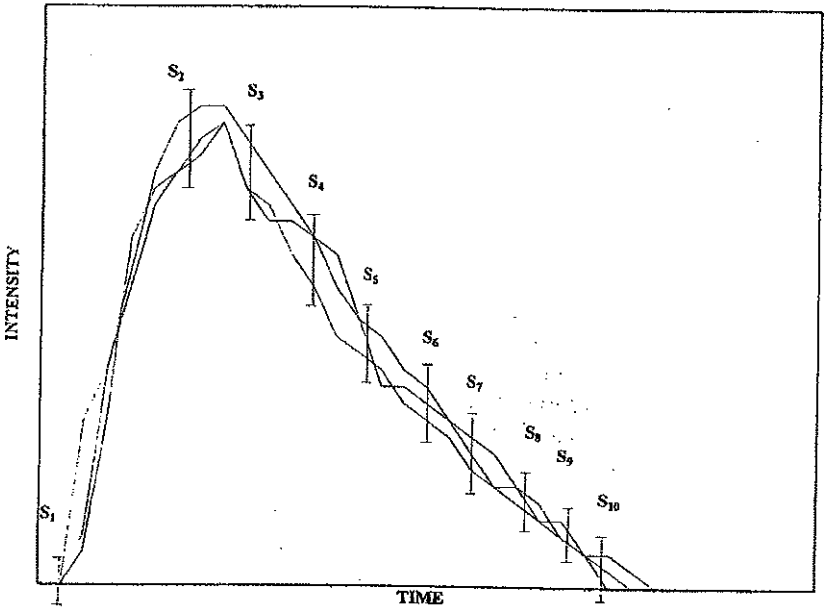

FIG. 2. GRAPHIC ILLUSTRATION OF T-IR CALCULATION

represents an instance of response sampling. Of course, computer sampling is considerably faster, and $CSA_{TPA}{}^{TM}$, is set to sample every half-second. This yields up to 120 responses for one minute of taste perception.

Reliability is calculated using the following formula.

$$s_i = \sqrt{\frac{\Sigma (x - \bar{X})^2}{n-1}} \tag{1}$$

$$z_{s_i} = \frac{s_i - \bar{S}}{s_{s_i}} \tag{2}$$

where $\bar{S}$ is the mean of all standard deviations (s_i) collected along the time-intensity curve, and s_{s_i} is the standard deviation of the collection of s_i values.

$$T\text{-}IR = \frac{\Sigma \mid z_{s_i} \mid}{N} \tag{3}$$

where N is the number of normalized standard deviations, determined by sampling rate of the time-intensity curve.

The standard deviation (s_i) amongst the three intensity values is determined at each sampling time (e.g., s_1–s_{10}) (formula 1). The resultant standard deviations are then normalized (z_{s_i}) so that reliabilities can be compared amongst time-intensity responses of various magnitudes (e.g., substances of differing concentrations, individuals of varying sensitivities, etc.). Normalization is accomplished by calculating the mean and standard deviation of the standard deviations, then converting each standard deviation to a z-score (formula 2). Reliability, T-IR, is obtained by computing the absolute mean of the normalized standard deviations (formula 3). The absolute value of the mean is computed so that negative standard deviations add to the positive standard deviations to produce a measure of magnitude of the degree of variability. T-IR, therefore, is a normalized value representing the degree of variability across the entire curve based upon deviations in responding at each instance of sampling. T-IR cannot be smaller than zero and, like all standard scores, will rarely exceed 3.0.

T-IR is a u nit of measure that can be compared from individual to individual or from panel to panel. Of course, comparisons will be most meaningful when they are made amongst reliability scores derived from comparable numbers of trials. That is, if we compare the reliability of two panelists, we should calculate T-IR from a similar number of trials for each panelist, because standard devia-

tions are by nature smaller as the number of scores on which they are based increase. It would be unfair to compare the reliability of 30 trials for one panelist with the reliability of 3 trials for another panelist. One final consideration in the calculation of T-IR is the starting point. Panelists could differ individually and from trial to trial in reaction time. Since T-IR is based upon the spread of scores at each unit of sampling, some investigators may want to synchronize the curves by using the panelist's first response to the stimulus as "time 1." Other investigators may want to incorporate differences in reaction times as part of the reliability measure. In this way, if a panelist is slower to respond to one trial, there may be greater deviations amongst curves, and T-IR will be increased. In our own work, we have found panelists to be remarkably consistent in reaction time, the onset of response after ingestion.

AN EXAMPLE WITH DATA FROM TESTS OF SWEETENER AND ORIENTATION

Ten panelists were presented with eight identical samples of four sweeteners (sucralose, aspartame, acesulfame-k, and sucrose). Panelists indicated their sweetness perceptions on a horizontal scale for four of the samples, and they indicated their sweetness perceptions on a vertical scale for the other four samples. Therefore, each panelist produced four time-intensity curves for each of 8 stimulus conditions (four sweeteners × two orientations of response scale). Using the formula listed above, T-IR was calculated for the four replicated trials of each condition for each panelist.

Table 2 lists the T-IR scores for each panelist under each condition. It can be noted that the panelists differed in variability with T-IR ranging from a low (high reliability) of .550 for panelist 5 (perception of sucralose recorded on a vertical time-intensity s cale) to a high of .970 for panelist 10 (perception of sucralose recorded on a horizontal time-intensity scale). The mean of the 80 T-IR scores was .83 with a standard deviation of .06. The scores of reliability in Table 2 also suggest a degree of consistency amongst panelists, i.e., some panelists were consistently less variable than others. To determine the magnitude of this consistency, a Cronbach's alpha correlation of reliability was calculated for the eight T-IR scores of the ten panelists. The resultant Cronbach's alpha of .77 confirmed that panelists who are low in variability for one condition, tend to be similarly low in variability when presented with other conditions, relative to their counterparts.

How did reliability change as a result of sweeteners and orientation of the time-intensity scale? To answer this question a 2-way repeated-measures analysis of variance was conducted with Sweeteners (4) as one variable and Orientation (2)

TABLE 2.
RELIABILITY SCORES FOR 10 PANELISTS WITH FOUR SWEETENERS AND TWO ORIENTATIONS (HORIZONTAL AND VERTICAL) OF TIME-INTENSITY SCALES

	Horizontal				Vertical			
Panelist	Sucralose	Aspartame	Acesulfame k	Sucrose	Sucralose	Aspartame	Acesulfame k	Sucrose
1	.960	.890	.840	.840	.900	.760	.870	.740
2	.770	.850	.900	.820	.920	.780	.790	.790
3	.970	.840	.860	.850	.920	.920	.890	.930
4	.820	.880	.850	.760	.620	.810	.810	.850
5	.820	.770	.880	.960	.930	.950	.910	.840
6	.770	.780	.810	.900	.930	.930	.830	.810
7	.830	.720	.790	.710	.550	.620	.790	.810
8	.630	.930	.940	.810	.610	.810	.730	.900
9	.970	.840	.820	.900	.830	.900	.900	.890
10	.710	.700	.800	.710	.770	.880	.660	.770

as the other variable, and T-IR as the dependent variable. There were no group differences in T-IR amongst the four sweeteners (F (3,27) = .33, p = .807), no differences between the two orientations (F (1,9) = .28, p = .611), and no Sweetener X Orientation interaction (F (3,27) = .39, p = .764). This means that panelists were similarly reliable across the four sweeteners and two orientations of time-intensity scales, and that their responses to the sweeteners were similarly reliable independent of the orientation of the scale on which they recorded their perceptions (see Figure 3).

APPLICATION FOR TIME-INTENSITY RESEARCH AND PANEL TRAINING

The analyses of T-IR scores add confidence to our published research reports on differences amongst sweeteners and comparisons of time-intensity scale orientation (Duizer *et al.* 1995). Although the panelists may have a greater preception of maximum intensity, for example, when responding to one sweetener over another, they are no less or more reliable in their response to that sweetener.

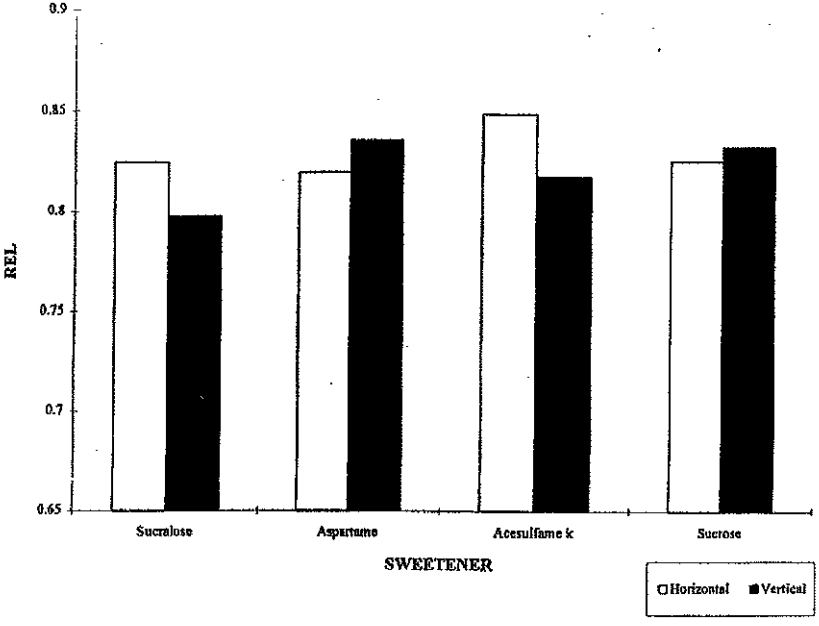

FIG. 3. MEAN T-IR SCORES FOR FOUR SWEETENERS AFFECTED BY TWO ORIENTATIONS

Therefore, the calculation of T-IR and its analysis for research data is important for establishing and confirming the reproducibility of research findings. The analysis of reliability also has an important role in the development, refinement, and maintenance of time-intensity panels. During training, panelists can be evaluated based upon their T-IR scores, remembering that the lower the T-IR scores, the more reliable (less variable) the panelist. Panels can be refined by eliminating panelists who have consistently high T-IR scores, and by setting the decrement in T-IR scores as a goal in panel training. Finally, the reliability of panelists can be charted over months and years to confirm the maintenance of standards (reliability levels) for time-intensity evaluations in research and industry.

REFERENCES

AMERINE, M.A., PANGBORN, R.M. and ROESSLER, E.B. 1965. In *Principles of Sensory Evaluation of Food*. p. 563, Academic Press, New York.

DUIZER, L.M., BLOOM, K. and FINDLAY, C.J. 1995. The effect of line orientation on the recording of time-intensity perception of sweetener solutions. J. Food Quality and Preference. In Press.

MANGAN, P.A.P. 1992. Performance assessment of sensory panelists. J. Sens. Stud. 7, 229–252.

MEILGAARD, M., CIVILLE, G.V. and CARR, B.T. 1991. In *Sensory Evaluation Techniques*. p. 257. CRC Press, Boca Raton.