J. Dairy Sci. 85:2075–2080

Consumer Acceptability Versus Trained Sensory Panel Scores of Powdered Milk Shelf-Life Defects

G. Hough,*,[1] R. H. Sánchez,*,[1] G. Garbarini de Pablo,* R. G. Sánchez*, S. Calderón Villaplana,† A. M. Giménez‡ and A. Gámbaro‡
*Instituto Superior Experimental de Tecnología Alimentaria,
(6500) Nueve de Julio, Buenos Aires, Argentina
†Centro Nacional de Ciencia y Tecnología de Alimentos, Ciudad Universitaria Rodrigo Facio,
(2060) San José, Costa Rica
‡Facultad de Química, Universidad de la República Oriental del Uruguay,
Av. General Flores 2124, Montevideo, Uruguay

ABSTRACT

The objective of the present work was to correlate consumer panel acceptability versus trained sensory panel scores for appearance and flavor defects likely to appear during storage of whole milk powder. Descriptors selected for the study were: acid, caramel, cooked, dark color, lipolysis, and oxidized. For each descriptor a set of nine samples with different intensities were measured for acceptability and likelihood to consume by a 50-member consumer panel and for sensory intensity by a trained panel. Linear correlations between sensory acceptability and trained sensory panel scores were used to determine the sensory failure cut-off point for each descriptor, except caramel and cooked, which were not critical from the consumer's point of view. Differences in acceptability were found between Argentine and Uruguayan consumers for oxidized samples, while for lipolysis flavor, Argentine and Costa Rican consumers behaved similarly. For the color descriptor, significant changes in acceptability measured on a hedonic scale did not mean that consumers refused to consume the product. In contrast, for flavor descriptors, as soon as a significant decrease in acceptability occurred approximately 30% of the consumers said they would not consume the product. The sensory failure cut-off points presented in this paper can be used as a guide in future studies on the shelf life of MP and can also be of value in establishing sensory specifications for quality control programs. The methodology of correlating consumer acceptability to sensory panel scores and, thus, defining sensory failure is an improvement over more arbitrary criteria presented in most shelf-life studies.

(**Key words:** milk powder, sensory, shelf life)

Abbreviation key: IDF = International Dairy Federation, **MP** = milk powder, **RMP** = reconstituted milk powder.

INTRODUCTION

Sensory evaluation of milk powder (**MP**) is the key factor for determining the shelf life of the product. It is not expected that a stored product should be exactly the same as a fresh standard, rather that the sensory differences be sufficiently small not to significantly alter the acceptability of the product. Such sensory differences would define the failure criterion. Labuza (1982) cites a number of authors who have used the failure criterion for pasteurized milk of a flavor score <35 to 37 in relation to a score of 40 for the fresh product. Vallejo-Córdoba and Nakai (1994) evaluated pasteurized milk with a panel of five assessors using a 10-point quality scale, and they defined failure as a score ≤5 from three of the assessors. Marsili (2000) used a similar criterion, using a panel of four assessors. Stapelfeldt et al. (1997) used a three-member expert panel who scored reconstituted MP on a 0- to 15-point scale, 0 being unacceptable and 15 being excellent; ≥10 indicated an acceptable sample. Nielsen et al. (1997) used a similar criterion, but defined 0 as extremely oxidized flavor and 15 as 'no oxidized flavor'. Studies as to how these criteria relate to consumer acceptability of the products have not been presented.

A consumer panel would be the most appropriate tool to determine when a food product reaches the end of its shelf life. However, to repeatedly assemble consumer panels for the multiple measurements needed during shelf-life studies would be impractical and expensive. A sensory panel is more appropriate for repeated assessments, but the panel measures analytical attri-

Received November 5, 2001.
Accepted February 7, 2002.
Corresponding author: Guillermo Hough; e-mail: guille@ghough.cyt.edu.ar.
[1]Authors Hough & Sánchez are research fellows of the Comisión de Investigaciones Cientıficas de la Provincia de Buenos Aires.

butes such as oxidized flavor or darkness, rather than directly assessing acceptability. How high does the oxidized flavor measured need to be for a trained panel's assessment of product acceptability to decrease? The answer to this question can be obtained by correlating data obtained from a consumer panel with data obtained from a trained panel. For products such as sunflower kernels (Fritsch et al., 1997) and sunflower oil (Ramírez et al., 2001), correlations have been published, but similar correlations have not been published for dairy products. Such correlations would be of tremendous benefit for sensory quality control programs.

The International Dairy Federation (**IDF**; 1997) lists the appearance and flavor defects to be expected in MP. The causes of these defects are predominantly due to raw milk quality, processing, and storage. To choose the defects most likely to appear due to prolonged storage, the work of a sensory panel trained in recognizing MP defects is desirable.

MP is often marketed internationally. The failure criterion determined with consumers from the producing country may not necessarily be the same as for the consumers of a client country. Comparison of failure criteria among consumers from different countries would help to determine the shelf life with a more appropriate perspective.

The objectives of the present work were to: 1) correlate consumer acceptability versus sensory panel scores for appearance and flavor defects likely to appear during the storage of milk powder, and 2) compare the failure criteria obtained from consumers in different countries.

MATERIALS AND METHODS

Milk Powder

The MP used in all the experiments was provided by SanCor Coop., Ltda. (Sunchales, Santa Fe, Argentina), in 1-kg packets packaged with nitrogen in the headspace, all from the same batch. Composition, as indicated by the manufacturer, was 26% protein, 26% fat, 3.5% moisture, 38% lactose, and 6.5% ash. Packets were stored at 18°C ± 2°C for <6 wk. Milk was reconstituted following the IDF Standard (1997). Both the quality control panel from SanCor and the panel used for the present study found the milk to be without defects.

Ethical Considerations

The Ethical Committee of our Institute concluded that all samples were acceptable for human testing in the concentrations and quantities to be served. A copy of the document approving the study was sent to the universities in Costa Rica and Uruguay as a reassurance for consumers recruited in these countries.

Trained Sensory Panel

A panel of seven assessors was selected and trained following the guidelines of the ISO (1993) standard, including the Ishihara color test. They all had a minimum of 18 mo of experience in discrimination and descriptive tests. For sensory testing, 30 ml of reconstituted MP (**RMP**) was served in 70-ml odorless plastic cups at room temperature. Water and unsalted crackers were used as palate cleansers between samples. The testing was performed in a sensory laboratory equipped with individual booths and artificial daylight (fluorescent lighting). For scoring, a 10-cm structured scale ranging from 0 and 100 was used.

For this project, assessors were initially trained to recognize the following flavors in RMP: cooked, feed, flat, burned, bitter, oxidized, metallic, lipolysis, salty, and acid. Above-threshold samples of these defects were prepared as indicated by Hough et al. (1992). These authors did not prepare the metallic flavor sample; 0.01 g of ferrous sulfate per liter of RMP was used. Labeled above-threshold samples were presented together with four or five unknown samples. Unknown samples had to be correctly identified in repeated sessions to ensure that each judge was properly trained.

To identify the flavors developed during storage, MP was stored at 37 and 45°C for a period of 3 mo. Every 3 wk, samples were removed from the ovens, reconstituted, and served to the trained panel paired with a fresh control sample. Assessors recorded the appearance and flavor descriptors that differentiated stored samples from the fresh control. The following were the descriptors used most frequently: acid, caramelized, cooked, dark color, and oxidized. Lipolysis flavor [called rancid in the IDF standard (1997)], although not developed during storage of the tested MP, was also included as it is a common flavor defect in dairy products.

Table 1 shows the preparation of stock solution for the selected descriptors, and Table 2 shows the concentrations used to correlate consumer acceptability versus sensory panel scores. The highest concentrations in Table 2 were designed to be above what would normally appear during prolonged storage of MP. To calibrate the trained panel for each descriptor, each panelist first received four samples corresponding to concentrations 1, 2, 6, and 9 (Table 2) labeled with K (control), A, B, and C, respectively, and the corresponding descriptor. Their task was to score these samples in the named defect, and following discussion, reach a consensus. The consensus score for each sample is in parenthesis in Table 2. In later sessions, the panel received the same

Table 1. Preparation of stock solutions used to prepare samples of reconstituted milk powder (RMP) with different defects.

Defect	Stock solution
Acid	1 ml lactic acid/1 L of RMP
Caramel	8 g flavoring[1]/1 L of RMP
Cooked	RMP heated 15 min in a boiling water bath
Dark color	100 ml of 2% coloring[2] solution completed to 1 l with RMP
Lipolysis	Fatty acid mixture[3] + 5 g of vaseline + 2 g Tween 80, heated to 50°C for dilution, added to 1 L RMP at 35°C.
Oxidized	1 ml of 1% copper sulfate solution + 6 × 6 cm copper foil strip added to 1 L RMP, stored 48 h at 4°C

[1]Givaudan Roure (Munro, Argentina) caramel essence code 73865-33.

[2]SICNA (Milan, Italy) caramel coloring.

[3]46 mg of butiric + 30 mg of caproic + 23 mg of caprilic + 28 mg of capric + 30 mg of lauric; all acids were analytical grade.

four samples with three-digit codes and scored them according to the consensus. A total of 12 45-min sessions were used to calibrate the panel, never presenting more than two descriptors per session.

Once the panel was calibrated, each descriptor was measured in triplicate, once per each session. Therefore, a total of 18 sessions were necessary for the six descriptors. In each session, the nine samples corresponding to the different concentration levels for the descriptor (see Table 2) were presented in random order to each assessor.

Consumer Panels

Consumers were recruited among students between the ages of 18 and 25 yr from the city of Nueve de Julio, Buenos Aires, Argentina. They completed a questionnaire asking about their frequency of consumption of different dairy products, and those who stated consuming milk at least once a week were chosen for the present study. Fifty consumers, approximately balanced between females and males, were used for each descriptor. Thus, a total of 300 consumers assessed the six descriptors. Each consumer tested only one descriptor. For lipolysis flavor, an additional 50 consumers were recruited among students from the University of Costa Rica in San José. Similarly, for oxidized flavor, 50 consumers were recruited among students from the Universidad de la República in Montevideo, Uruguay. In Costa Rica and Uruguay, the sample preparation differed from preparation in Argentina only in the water used to reconstitute the MP.

Each consumer received the nine samples corresponding to the nine concentration levels of one descriptor (see Table 2) presented monadically in random order. For each sample, the product was scored using a scale with nine boxes labeled 'dislike very much', 'indifferent,' and 'like very much'. They also answered the question, "Would you normally consume this product?" with a yes or a no response. Following the test, the consumers received a chocolate bar as a reward for their participation.

Statistical Analysis

Analysis of variance (ANOVA) was performed on the trained sensory panel data using sample, assessor, and their interaction as variation factors. On the consumer data using sample and consumer as variation factors,

Table 2. Concentrations (percentage of stock solutions; Table 1) in reconstituted milk powder) of different standards used to determine correlations between consumer acceptability and sensory panel scores. Numbers in parenthesis indicate the consensus score from the sensory panel on the 0 to 100 sensory scale.

Concentration	Acid	Caramel	Cooked	Dark color	Lipolysis	Oxidized
1	0.0 (10)	0.0 (0)	0.0 (10)	0.0 (10)	0.0 (0)	0.0 (0)
2	9.5 (20)	5.9 (40)	5.9 (20)	5.9 (20)	5.9 (30)	9.5 (10)
3	13.3	8.8	8.8	8.8	8.8	13.3
4	18.6	13.2	13.2	13.2	13.2	18.6
5	26.0	19.8	19.8	19.8	19.8	26.0
6	36.4 (50)	29.6 (70)	29.6 (60)	29.6 (50)	29.6 (70)	36.4 (50)
7	51.0	44.4	44.4	44.4	44.4	51.0
8	71.4	66.7	66.7	66.7	66.7	71.4
9	100 (100)	100 (100)	100 (90)	100 (90)	100 (100)	100 (90)

their interaction could not be calculated, as each consumer measured each sample only once.

To calculate the failure point from the consumer data, the following equation was used:

$$S = F - Z_{\alpha} \frac{\sqrt{2 \cdot MSE}}{N} \quad [1]$$

where:

S = minimum tolerable acceptability of stored sample,
F = acceptability of fresh sample,
Z_{α} = one-tailed coordinate of the normal curve for α significance level,
MSE = mean square of the error derived from the analysis of variance of the consumer data, and
N = number of consumers.

The coordinate of the normal curve is one-tailed because it was assumed that the stored product would have lower acceptability than the fresh product. Fritsch et al. (1997) and Ramírez et al. (2001) used expressions similar to equation (1) in their calculations but derived from two-tailed comparisons. If α (significance level) is small, we tend to increase the shelf life of the product, and if α is large, we tend to decrease the shelf life. The classical 5% significance level seems to be a sensible choice.

For each descriptor, a regression of consumer acceptability (averaged over consumers) versus sensory panel scores (averaged over assessors) was performed. The following equations were tested using linear and nonlinear regression facilities of Genstat (VSN International, Ltd., Oxford, UK):

$$\begin{aligned} linear&: A = a + b.T \\ exponential&: A = a + b.c^{T} \\ logistic&: A = \frac{c}{1 + \exp(-b(T - a))} \end{aligned} \quad [2]$$

where A = consumer acceptability, T = trained sensory panel score, and a-b-c = regression constants.

Once the regression equation was calculated, the cut-off point on the sensory scale was determined by entering the S value [equation (1)] in the acceptability scale. Subsequently, the corresponding sensory coordinate was calculated from the regression equation, as shown graphically in Figure 1. Finally, logistic regression (McConway et al., 1999) was used to correlate the proportion of consumers who rejected the sample ('no' answers) to the average sensory panel score for the same sample.

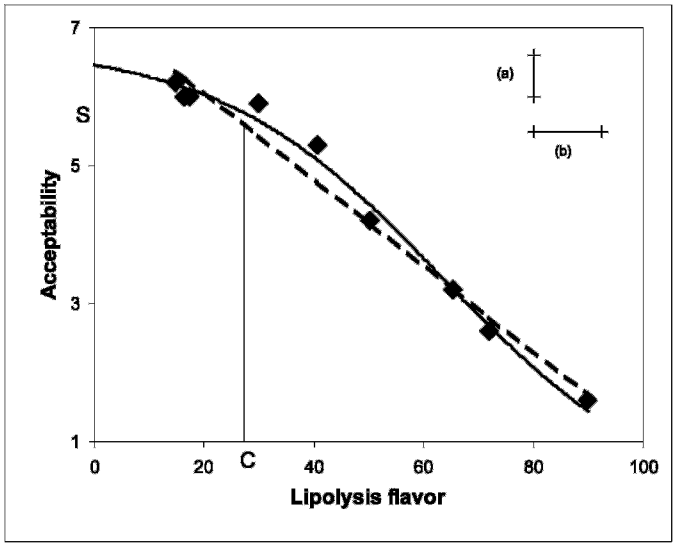

Figure 1. Sensory acceptability of Argentine consumers versus sensory panel scores for RMP with different levels of lipolysis flavor. Curve corresponds to logistic equation (equation 2) and straight line-to-linear equation. S = minimum tolerable acceptability of stored sample, C = sensory failure cut-off point. (a): least significant difference from acceptability ANOVA, (b): least significant difference from trained panel ANOVA.

RESULTS AND DISCUSSION

Figure 1 presents the results of Argentine consumers' acceptability versus sensory panel scores for lipolysis flavor. The logistic equation gave a slightly better fit than the linear equation [equations (2)], but it can be observed that the sensory failure cut-off point was similar using either equation. Figure 2 shows the Uruguayan consumers' acceptability versus sensory panel scores for oxidized flavor. In this case, the exponential equation gave a slightly better fit than the linear equation, but again, the difference between the determined cut-off points was small. In the rest of the regressions, (not shown), the differences in the cut-off point between using the logistic or exponential equations instead of the linear equation were also small; therefore, for simplicity, the linear equation was used in all cases.

For the cooked flavor, sensory acceptability did not decrease over samples; for caramel flavor, sensory acceptability decreased significantly for only the most concentrated sample. As mentioned above, the highest concentrations in Table 2 were designed to be above what would normally appear during prolonged storage of MP. Thus, our conclusion was that from a consumer's point of view, these descriptors would not normally be critical in defining shelf life of MP.

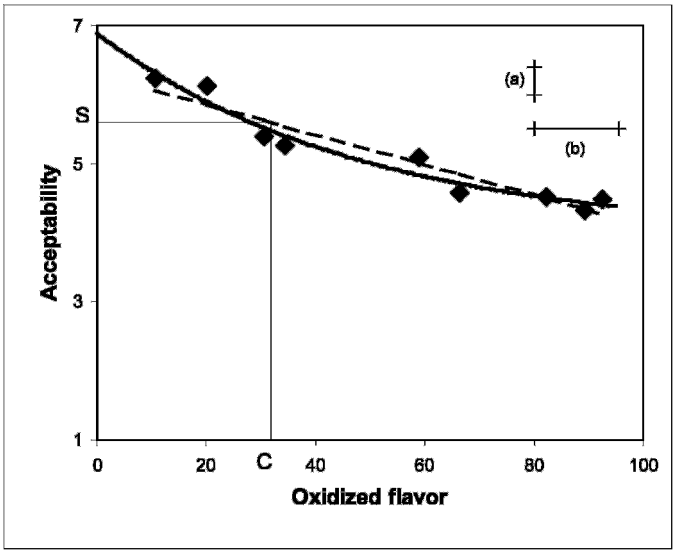

Figure 2. Sensory acceptability of Uruguayan consumers versus sensory panel scores for RMP with different levels of oxidized flavor. Curve corresponds to exponential equation (see equation 2) and straight line-to-linear equation. S = minimum tolerable acceptability of stored sample, C = sensory failure cut-off point. (a): least significant difference from acceptability ANOVA, (b): least significant difference from trained panel ANOVA.

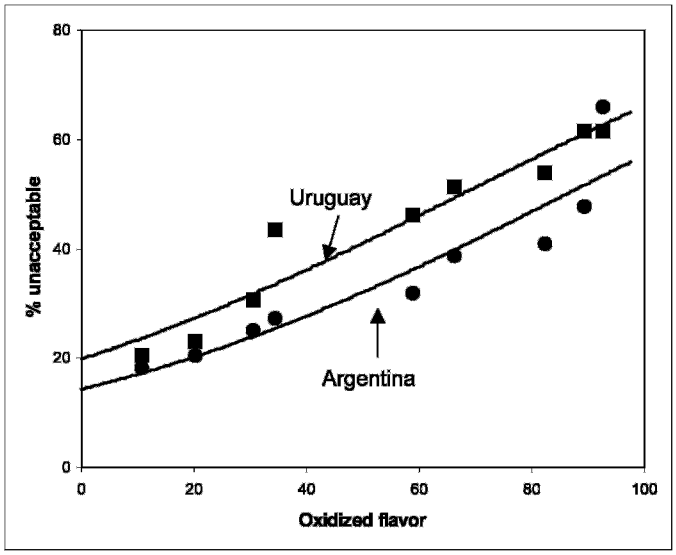

Figure 3. Percentage of Argentine and Uruguayan consumers who found samples unacceptable for consumption versus sensory panel scores of RMP with different levels of oxidized flavor. Curves were obtained by logistic regression.

Figure 3 shows the relationship for oxidized flavor between the percentage of consumers who found samples unacceptable versus sensory panel scores. For the same oxidized concentrations, a larger percentage of Uruguayan consumers rejected the sample than did Argentine consumers. However, the sample of consumers used in this study was small and as they were all students, they did not necessarily represent the general population. Therefore, this difference could be due to the particular samples and not strictly due to country differences. Nevertheless, it does show that results from one sample of consumers cannot be extrapolated to other consumers without additional research. For lipolysis flavor, (not shown), consumers from Argentina and Costa Rica behaved similarly.

Table 3 summarizes the results from the present study. The fresh sample was the same for all descriptors, yet ANOVA showed significant differences in acceptability for this sample. These results may be attributed to different descriptors being measured (color and flavor), different consumer samples, and a possible context effect.

For dark color, the percentage of consumers who would reject the sample at the sensory failure cut-off point was only 4%, compared with approximately 30% for the flavor descriptors (P column in Table 3). Acceptability of dark-colored samples close to the fresh sample was reduced significantly, yet few consumers refused to consume these samples. For flavor descriptors, once a significant decrease in preference occurred, consumers were more likely to reject the sample. Thus, rejection appears to be driven by flavor characteristics.

The sensory failure cut-off points (column C in Table 3) cannot be compared directly inasmuch as the sensory scales were different. For example, to state that consumers tolerated a higher intensity of acid flavor than lipolysis flavor, would imply that the acid and lipolysis samples from Table 2 were somehow equivalent from a sensory point of view. Complex multimodal sensory tests would have been necessary in order to equate the scales. For future studies on the shelf life of MP, the cut-off points from Table 3 can be used as a guide to define sensory failure, provided the panel is trained with the same samples and scale as in the present work.

CONCLUSIONS

Linear correlations between sensory acceptability and sensory panel scores were found adequate to determine the sensory cut-off point for acid, dark color, lipolysis, and oxidized descriptors. Caramel and cooked descriptors were not critical from the consumers' point of view. Differences in acceptability were found between Argentine and Uruguayan consumers for oxidized samples; for lipolysis flavor, Argentine and Costa Rican consumers behaved similarly. For dark color, significant changes in acceptability measured on a hedonic scale did not mean that consumers refused to consume the product, while for flavor descriptors, as soon as a

Table 3. Acceptabilities of fresh samples (F), minimum tolerable acceptabilities of stored samples (S), sensory failure cut-off points (C), percent variance of acceptability versus sensory panel linear regression, (R^2), and percent of consumers who would reject the samples at their sensory failure cut-off points (P).

Descriptor	F[1]	S[1]	C[2]	R^2	P(%)[4]
Acid-A[3]	6.0	5.3	42	0.88	34 ± 9
Dark color-A	7.4	7.0	9	0.99	4 ± 4
Lipolysis-A	6.2	5.6	27	0.97	32 ± 7
Lipolysis-CR	7.1	6.5	25	0.99	27 ± 7
Oxidized-A	6.8	6.0	45	0.80	30 ± 6
Oxidized-U	6.2	5.6	32	0.91	32 ± 8

[1]Measured on a 1 (dislike very much) to 9 (like very much) hedonic scale.

[2]Measured on a 0 to 100 sensory intensity scale.

[3]A: Argentine consumers, CR: Costa Rican consumers, and U: Uruguayan consumers.

[4]Percent ± 95% confidence interval.

significant decrease in acceptability took place, approximately 30% of the consumers said they would not consume the product. The sensory failure cut-off points presented in this paper can be used as a guide in future studies on the shelf life of MP, and they may also be of value in establishing sensory specifications for quality control programs. The methodology of correlating consumer acceptability to sensory panel scores to define sensory failure is an improvement over more arbitrary criteria presented in most shelf-life studies.

ACKNOWLEDGMENTS

We are grateful for financial help from the Agencia Nacional de Promoción Científica y Tecnológica (PICT 98), from the Ciencia y Tecnología para el Desarrollo (CYTED) program, and from SanCor Coop., Ltda.

REFERENCES

Fritsch, C. W., C. N. Hofland, and Z. M. Vickers. 1997. Shelf Life of Sunflower Kernels. J. Food Sci. 62, 425–428.

Hough, G., E. Martínez, and T. Barbieri. 1992. Sensory thresholds of flavor defects in reconstituted whole milk powder. J. Dairy Sci. 75:2370–2374.

International Dairy Federation. 1997. Sensory evaluation of dairy products by scoring. IDF Standard 99C. Int. Dairy Federation, Brussels, Belgium.

ISO. 1993. ISO Standard 8586—1. Sensory Analysis—General guidance for the selection, training, and monitoring of assessors. Part 1—Selected assessors. 1st ed. Int. Organization for Standardization, Geneva, Switzerland.

Labuza, T. P. 1982. Pages 201–206 in Shelf-Life Dating of Foods. Food & Nutrition Press, Westport, CT.

Marsili, R. T. 2000. Shelf-life prediction of processed milk by solid-phase microextraction, mass spectrometry, and multivariate analysis. J. Agric. Food Chem. 48:3470–3475.

McConway, K. J., M. C. Jones, and P. C. Taylor. 1999. Chapter 9 in Statistical Modeling Using Genstat. Arnold Publishers, London, UK.

Nielsen, B. R., H. Stapelfeldt, and L. H. Skibsted. 1997. Early prediction of the shelf-life of medium-heat whole milk powders using stepwise multiple regression and principal component analysis. Int. Dairy J. 7:341–348.

Ramírez, G., G. Hough, and A. Contarini. 2001. Influence of temperature and light exposure on sensory shelf life of a commercial sunflower oil. J. Food Quality 24:195–204.

Stapelfeldt, H., B. R. Nielsen, and L. H. Skibsted. 1997. Effect of heat treatment, water activity, and storage temperature on the oxidative stability of whole milk powder. Int. Dairy J. 7:331–339.

Vallejo-Cordoba, B., and S. Nakai. 1994. Keeping quality assessment of pasteurized milk by multivariate analysis of dynamic headspace gas chromatographic data. 1. Shelf life prediction by principal component regression. J. Agric. Food Chem. 42:989–993.