The accuracy and repeatability of untrained laboratory consumer panelists in detecting differences in beef longissimus tenderness[1,2,3]

T. L. Wheeler[4], S. D. Shackelford, and M. Koohmaraie

Roman L. Hruska U.S. Meat Animal Research Center, ARS, USDA, Clay Center, NE 68933-0166

ABSTRACT: The objective of this study was to determine the accuracy and repeatability of untrained laboratory consumer panelists in detecting differences in beef longissimus tenderness. At 14 d postmortem, slice shear force was measured on one steak from 192 strip loins and used to select 54 strip loins and assign 18 of the strip loins to each of three tenderness classes (tender = <15 kg, intermediate = 15 to 27 kg, and tough = >27 kg). Sixty-eight untrained, laboratory consumer panelists evaluated paired steaks from each tenderness class in each of two sessions (12 total observations per panelist). Mean slice shear forces for "tender," "intermediate," and "tough" were 11.1, 21.0, and 32.2 kg, respectively. Mean tenderness ratings of the untrained laboratory consumer panel were different ($P < 0.05$) among tenderness classes (mean of 16 panelists = 6.2, 4.9, and 3.3 for tender, intermediate, and tough, respectively), and these differences were similar regardless of how many untrained panelists were averaged to determine the panel mean (4, 8, 12, or 16). The correlations ($P < 0.01$) between slice shear force and the mean untrained consumer panel tenderness rating (mean of 4, r = –0.82; mean of 8, r = –0.89; mean of 12, r = –0.91; and mean of 16, r = –0.92;) were similar. Overall repeatability of the untrained consumer panel was 0.80. Repeatability of individual untrained consumer panelists for tenderness rating was highly variable: 31% were >0.80, 36% were 0.60 to 0.79, and 33% were <0.60. Thirty-two percent of the consumers were both accurate (correlation to slice shear force = –0.75 to –1.00, $P < 0.01$) and repeatable (repeatability >0.75). There is wide variability in the ability of untrained laboratory consumer panelists to detect differences in beef tenderness. Nonetheless, untrained consumer panels can accurately and repeatedly detect differences in beef tenderness under controlled laboratory conditions. An untrained laboratory consumer panel may be able to provide as effective an evaluation of beef longissimus tenderness as a trained descriptive attribute panel.

Key Words: Beef, Consumer, Grade, Quality, Sensory, Tenderness

 J. Anim. Sci. 2004. 82:557–562

Introduction

The beef industry has made it a priority to address inconsistency in beef tenderness. Until it is possible to ensure that all beef is acceptably tender, one way to deal with the variation in tenderness is to identify the tenderness of meat from each carcass and to market it accordingly. One of many critical issues in determining whether it would be profitable for the industry to market beef based on tenderness is the ability of consumers to consistently recognize tenderness differences.

Several studies have concluded that consumers can detect differences in beef tenderness using in-home (Miller et al., 1995; Boleman et al., 1997; Shackelford et al., 2001), supermarket intercept (Miller et al., 2001), simulated restaurant (Miller et al., 1995; Huffman et al., 1996), and laboratory (Wheeler et al., 2002; Wylie et al., 2003) approaches. We have attempted a national consumer evaluation of tenderness-classified beef using in-home data from 320 consumers each in Chicago and Philadelphia (our unpublished data). However, these data indicate that consumers could not detect differences in longissimus tenderness. That result is not consistent with previous consumer data (Boleman et al., 1997; Miller et al., 2001; Shackelford et al., 2001), and it seems from thorough examination of our data from Chicago and Philadelphia that the consumers may have made numerous data recording errors; thus, we are skeptical of those results. Therefore, we conducted the

[1]Names are necessary to report factually on available data; however, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by USDA implies no approval of the product to the exclusion of others that may also be suitable.

[2]The authors express their gratitude to K. Mihm and P. Tammen for technical assistance and to M. Bierman for secretarial assistance.

[3]Partial funding provided by the USDA Fund for Rural America grant number 97-36200-5197.

[4]Correspondence—phone: 402-762-4229; fax: 402-762-4149; e-mail: wheeler@email.marc.usda.gov.

Received February 6, 2003.

Accepted September 24, 2003.

present experiment because repeatability of consumer evaluations of beef tenderness has not been determined and because untrained laboratory consumer studies, where most conditions can be controlled, may be preferable to in-home studies for testing the inherent ability of consumers to detect differences in beef tenderness. Thus, the objective of this experiment was to determine the accuracy and repeatability of untrained consumer panelists, individually and as a panel, in detecting differences in beef longissimus tenderness under controlled laboratory conditions.

Materials and Methods

Experimental Samples

The Roman L. Hruska U.S. Meat Animal Research Center (MARC) Animal Care and Use Committee approved the use of animals in this study. Fifty-four North American Meat Processors (NAMP, 1997) #180 strip loins (longissimus lumborum) were used. Forty-eight of these carcasses were from a fifth year of the study described by Wheeler et al. (2001) that included 131 purebred Angus, purebred Hereford, or Piedmontese crossbred steers. Six of the 54 carcasses were from Phase II of the study described by Wheeler et al. (2002), which included 400 carcasses. These 54 included the strip loins from both sides of eight carcasses and from one side of 38 carcasses. At 2 d postmortem, a 2.54-cm-thick steak was removed from the anterior end of each strip loin and used for another experiment. The remainder of the strip loin was vacuum-packaged, stored at 2°C, and then frozen (–30°C) at 14 d postmortem. Five 2.54-cm-thick steaks were cut from the anterior end of the frozen strip loins with a band saw.

Cooking

Steaks were thawed and cooked as described by Wheeler et al. (1998) with the following exceptions. The preheat platen on the belt grill was set at 149°C, rather than disconnected. That change required that the cook time be reduced from 5.7 min to 5.5 min.

Slice Shear Force

Steak 3 was used to measure slice shear force as described by Shackelford et al. (1999). Slice shear force values were used to categorize the strip loins into "tender," "intermediate," and "tough" classes (n = 18/class). The tender, intermediate, and tough classes ranged from 7 to 14.9 kg, 15 to 26.9 kg, and 27 to 42 kg of slice shear force, respectively.

Untrained Consumer Panel

Steaks 1, 2, 4, and 5 were used for replicate tenderness rating measurements of the same strip loin by untrained panelists. The same four untrained panelists evaluated steaks 1 and 2 and another four untrained panelists evaluated steaks 4 and 5 on the same day. Each panelist evaluated six samples (duplicate samples of one strip loin from each tenderness class) on each of 2 d (n = 12 total/panelist). Each panelist evaluated each strip loin twice and each tenderness class four times. Each strip loin was evaluated a total of 16 times by untrained consumer panelists (four steaks × four panelists/steak). Within tenderness class, strip loins were ranked by slice shear force and assigned to panelist in order so that the difference between classes was approximately the same for all panelists (i.e., not assigned randomly within class to prevent a panelist from getting assigned, for example, a "tender" steak with 14 kg of slice shear force and an "intermediate" strip with 15 kg of slice shear force).

Untrained consumer panelists were recruited to participate from among MARC, University of Nebraska (stationed at MARC), Great Plains Veterinary Education Center (located at MARC headquarters), and University of Nebraska South Central Research and Education Center (located at MARC headquarters) employees. The only additional criterion for participation was availability (researchers working in the area of meat palatability and sensory laboratory technicians were excluded). Sixty-eight panelists volunteered for the study. Demographic characteristics for the consumer panelists were as follows: 71% male, 29% female; 18% 21 to 35 yr old, 57% 36 to 50 yr old, 24% 51 to 65 yr old, and 1% >65 yr old. These untrained laboratory consumer panelists were not intended to represent U.S. consumers. They were only intended to be a sample of untrained consumers.

Tenderness Evaluation

Panelists evaluated samples in one session on each of 2 d, 1 wk apart. Panelists selected one of five sessions for each day. All evaluations occurred on Thursday and Friday of two consecutive weeks. Sample presentation order was randomized for each session. In each session, panelists were provided two warm, 1.3 cm × 1.3 cm × steak thickness cubes of each of six samples (included duplicate steaks from one strip loin for each of the three tenderness classes) in a labeled paper cup. Panelists were asked to evaluate both cubes and then record a final score for the sample's tenderness based on an eight-point scale (8 = extremely tender to 1 = extremely tough). Panelists were provided an unsalted cracker and room-temperature distilled water for cleansing the palate between samples. Panelists were allowed to either chew and swallow the sample or expectorate after their evaluation of a sample was completed.

Statistical Analysis. Data were analyzed by ANOVA for a completely randomized design using the GLM procedures of SAS (SAS Inst., Inc., Cary, NC) for the main effect of tenderness class (tender, intermediate, tough) for slice shear force and overall untrained consumer panel. The main effect of tenderness class was tested for untrained consumer panel data after averaging ei-

Table 1. Means and SD for slice shear force, and consumer panel tenderness ratings across tenderness classes

	Tender[a]		Intermediate[a]		Tough[a]	
Trait	Mean	SD	Mean	SD	Mean	SD
Slice shear force, kg	11.1	2.3	21.0	3.9	32.2	4.9
Consumer sensory panel[bc]						
4[d]	6.5[e]	0.88	4.8[f]	0.95	3.2[g]	0.99
8[d]	6.5[e]	0.68	4.8[f]	0.75	3.2[g]	0.83
12[d]	6.5[e]	0.62	4.8[f]	0.72	3.2[g]	0.76
16[d]	6.2[e]	0.58	4.9[f]	0.75	3.3[g]	0.71

[a]Tender = 14-d slice shear force of 7 to 14.9 kg, intermediate = 14-d slice shear force of 15 to 26.9 kg, tough = 14-d slice shear force of 27 to 42 kg.
[b]Mean of 18 strip loins/tenderness class. 1 = extremely tough, 8 = extremely tender.
[c]Means for 4, 8, 12, or 16 consumer panelists per strip loin.
[d]Means within tenderness class did not differ ($P > 0.05$) due to the number of consumer ratings averaged to obtain the mean.
[e,f,g]Means in a row that do not have a common superscript differ ($P < 0.05$).

ther 4, 8, 12, or 16 panelists to obtain the panel mean. The PROC FREQ procedure and Mantel-Haenszel chi-squared analysis were used on the frequencies of panelists for detecting differences among tenderness classes, and of the frequencies of panelists' tenderness ratings within tenderness classes (SAS Inst., Inc.). The PROC CORR procedure of SAS was used to determine the degree of association between strip loin means for untrained consumer panel tenderness ratings and slice shear force. Repeatability of mean untrained consumer panel and individual untrained panelist tenderness ratings were calculated using PROC VARCOMP (SAS Inst., Inc.) for the random effect of sample to get the estimated variance components (σ^2_{sample} and σ^2_{error}):

$$\text{Repeatability} = \frac{\sigma^2_{sample}}{\sigma^2_{sample} + \sigma^2_{error}}$$

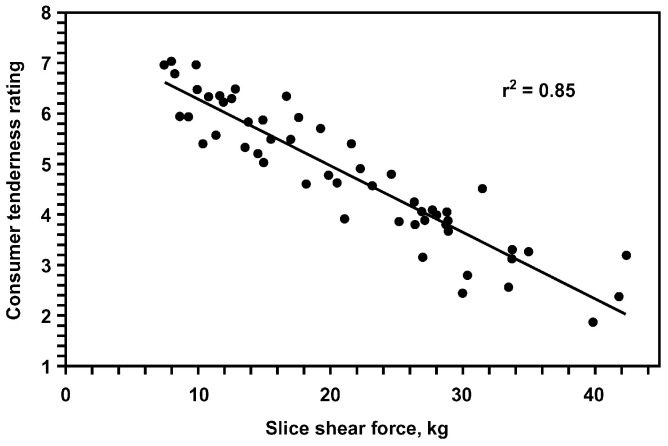

Figure 1. Regression of untrained laboratory consumer panel tenderness (1 = extremely tough to 8 = extremely tender) rating (mean of 16 observations/strip loin; n = 54 strip loins) on slice shear force.

Results

Slice shear force was used to create three classes of strip loins that were different in tenderness (Table 1). The laboratory consumer panel detected all three tenderness classes as different ($P < 0.05$) from one another, regardless of whether 4, 8, 12, or 16 consumer ratings were averaged to obtain the tenderness rating for each strip loin (Table 1). In addition, within each tenderness class, there were no differences ($P > 0.05$) in mean consumer tenderness ratings due to the number of consumer ratings used to obtain the means. The regression of slice shear force on untrained laboratory consumer panel tenderness rating indicated that the untrained consumer panel tenderness rating was strongly associated with the instrumental measure of meat tenderness (Figure 1).

The correlation of slice shear force with individual consumer tenderness ratings was lower ($P < 0.05$) than the correlations with mean consumer panel ratings, regardless of the number of consumers averaged to obtain the consumer panel mean (Table 2). Correlations between slice shear force and consumer panel tenderness rating increased ($P < 0.05$) as the number of consumers in the average increased up to eight.

The repeatability of the consumer panel tenderness ratings on duplicate steaks was 0.80 (Figure 2). The

Table 2. Correlation of slice shear force with individual consumer ratings and mean ratings of 4, 8, 12, and 16 consumers per strip loin[a]

	r
Individuals	–0.68
4	–0.82
8	–0.89
12	–0.91
16	–0.92

[a]All correlations were significant at $P < 0.01$.

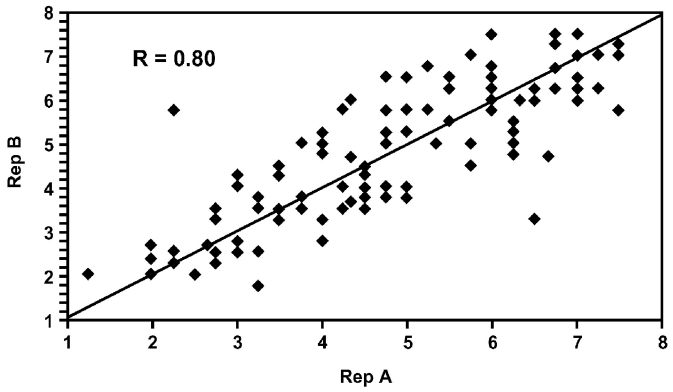

Figure 2. Consumer panel repeatability (R) for duplicate tenderness (1 = extremely tough to 8 = extremely tender) ratings (n = 108). The same consumers evaluated steaks 1 and 2 (Rep A and B) and another group of consumers evaluated steaks 4 and 5 (Rep A and B) from 54 strip loins.

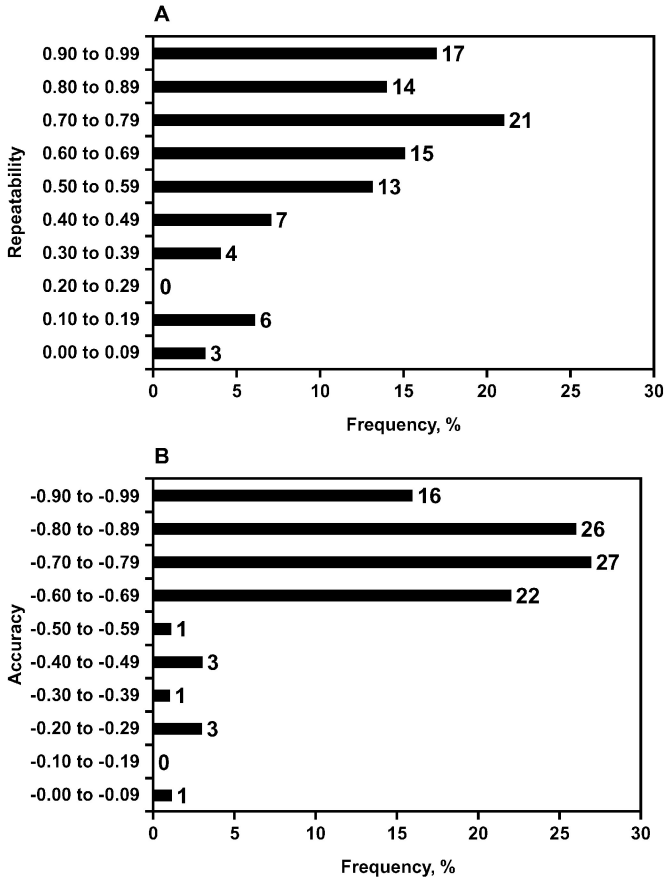

Figure 3. A) Distribution of individual consumer repeatabilities for tenderness rating and B) distribution of correlations between consumer tenderness ratings and slice shear force (accuracy).

repeatabilities of individual consumer tenderness ratings were highly variable and ranged from 0 to 0.99 (Figure 3A). Thirty-one percent of individual consumer repeatabilities for tenderness ratings were ≥0.80, 36% were 0.60 to 0.79, and 33% were <0.60. The accuracy of individual consumer tenderness ratings (defined as the correlation to slice shear force) was less variable than repeatability (Figure 3B). Forty-two percent of accuracy correlations were –0.80 to –1.00, 49% were –0.60 to –0.79, and 9% were 0.00 to –0.59. Thirty-two percent of individual consumers were both accurate (r = –0.75 to –1.00) and repeatable (R ≥ 0.75). The distribution of individual consumer tenderness ratings within each tenderness class spanned most of the tenderness scale, although ratings were concentrated at the high, middle, and low ends of the scale for tender, intermediate, and tough, respectively (Figure 4).

Discussion

To meet consumer expectations, the beef industry has become increasingly interested in implementing strategies for improving and reducing variation in beef quality. It has been suggested by industry leaders that sorting and marketing beef based on tenderness would result in increased consumer satisfaction with beef by enabling the industry to manage and reduce the variation in tenderness. The success of this approach partially depends on the existence of a segment of consumers that is capable of recognizing tender beef as superior in palatability and that is willing to pay a premium for guaranteed tender beef.

The present study has established that, under controlled laboratory conditions, there is wide variability in the ability of individual untrained consumers to accurately and repeatedly detect differences in beefsteak tenderness. However, a panel of untrained laboratory consumers (consisting of 4 to 16 panelists/strip loin) detected differences between "tender," "intermediate," and "tough" steak categories. In fact, even a panel of the worst eight consumer panelists (based on accuracy and repeatability of their tenderness evaluations) detected a 1.7-unit difference between "tender" and "tough" on an eight-point scale (data not shown). For a panel of the best eight consumers, the difference between "tender" and "tough" was 3.8 units, and for a panel of all consumers, the difference was 2.9 units. Thus, the ability of the individual consumers on the panel to evaluate beef tenderness affected the magnitude of the differences detected between tenderness classes by the panel, but even a panel of the worst consumer beef tenderness evaluators detected significant differences among all three tenderness classes. These results should not be extrapolated to include consumer evaluation of other sensory traits. It has been shown that texture traits of meat are the sensory traits that untrained or less trained panelists evaluate as well as trained panelists, but this may not be true of juiciness and flavor traits (Chambers et al., 1981).

Historically, instrumental methods and trained sensory panels have been used by meat science researchers

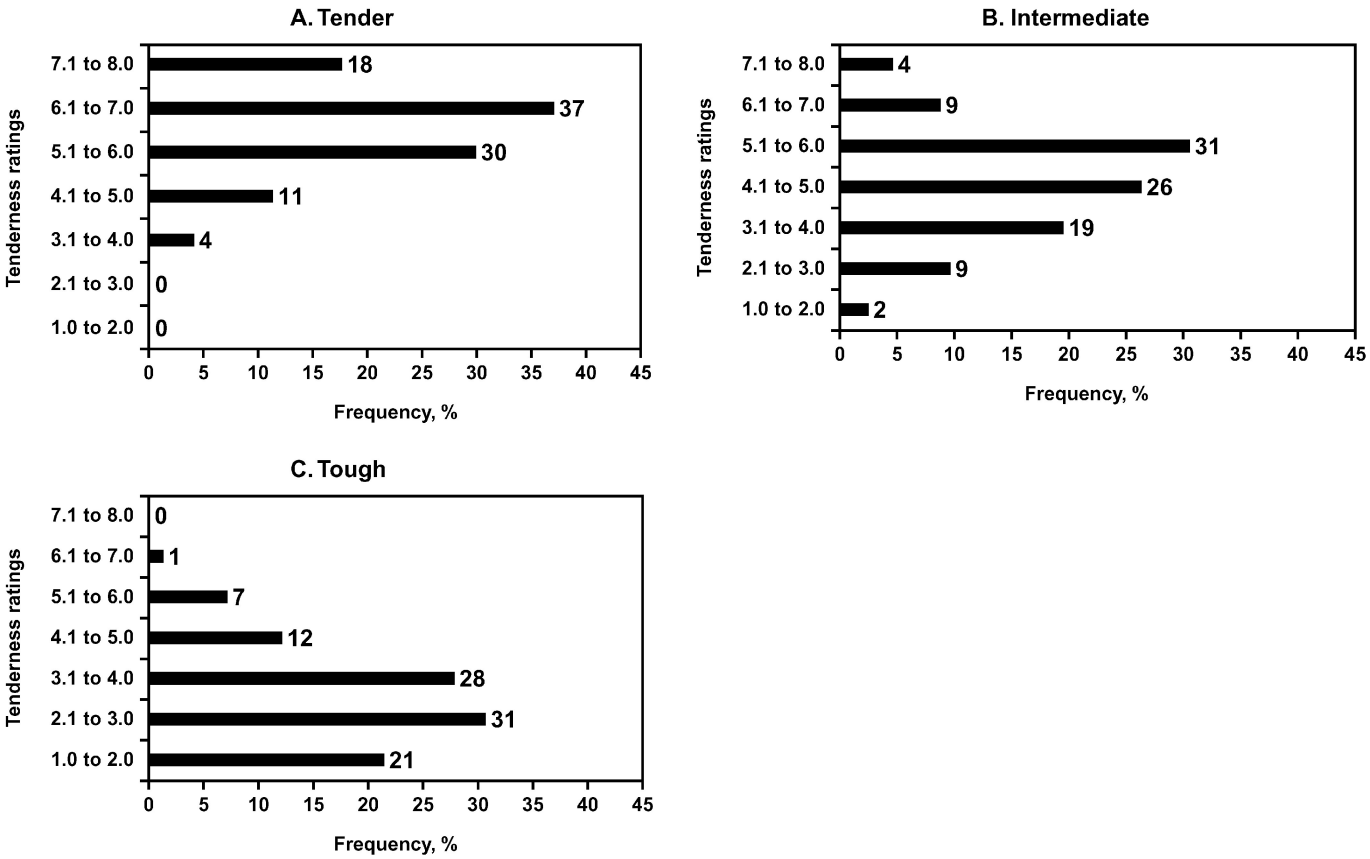

Figure 4. Distribution of tenderness ratings (1 = extremely tough to 8 = extremely tender) by individual untrained consumers for longissimus classified as A) tender, B) intermediate, and C) tough, based on slice shear force (tender = <15 kg, intermediate = 15 to 27 kg, and tough = >27 kg).

to determine differences among samples for tenderness. Consumer evaluation is usually employed to determine relative satisfaction, acceptability, or desirability among meat samples (Munoz, 1998), and to confirm that differences detected by objective methods could be detected by consumers. A number of studies of meat tenderness have utilized untrained consumers in a variety of approaches to evaluate beef tenderness.

Brooks et al. (2000) reported that the lower Warner-Bratzler shear force value of USDA Prime ribeye and Top Choice top sirloin foodservice steaks were not detected by a laboratory consumer panel. However, for retail ribeye steaks, neither Warner-Bratzler shear nor the laboratory consumer panel detected any differences in tenderness among quality grade groups. Wheeler et al. (2002) reported that the magnitude of the difference in tenderness ratings between "certified tender" and "not certified tender" longissimus was similar for a laboratory consumer panel and a trained panel. The correlation of laboratory consumer panel tenderness rating to trained panel tenderness rating was –0.56 (Wheeler et al., 2002). Branson et al. (1986) reported that the magnitude of the differences among quality grade groups that was detected by the laboratory consumer panel was closer to that detected by the trained sensory panel than was the magnitude of differences detected by the in-home consumer panel. This result was likely due to the greater control over cooking method and degree of doneness, and a higher probability of properly completed data sheets for the laboratory consumer evaluations compared with in-home consumer evaluations where cooking was left up to consumer preference and recording errors were more likely to occur.

In-home consumer evaluations also have been shown to detect differences in tenderness classes. Boleman et al. (1997) were the first to demonstrate that consumers could detect differences in beef tenderness classes that had been selected based on shear force. In a study of Denver metropolitan area consumers, Shackelford et al. (2001) reported that guaranteed tender (low slice shear force) USDA Select loin steaks were rated more favorably for all consumer traits than were high slice shear force Select loin steaks. In addition, using the supermarket intercept approach, Lusk et al. (2001) found that 69% of consumers preferred a low slice shear force steak to a high slice shear force steak based solely on their eating experience from the two steaks. That percentage increased to 84% when the consumers were informed they were evaluating a "guaranteed tender" and a "probably tough" steak.

Thus, available data indicate that some proportion of consumers is capable of detecting differences in steak

tenderness. Depending on the specific objectives of an experiment, a laboratory consumer panel may be preferable to an in-home consumer panel. Furthermore, it is very expensive to train and maintain a trained descriptive attribute sensory panel and increasingly difficult to find people capable of, and willing to, serve on a panel, even when compensated. The present experiment and previously reported results indicate it may be possible to use an untrained laboratory consumer panel to obtain an evaluation of meat tenderness similar to that obtained with a trained descriptive attribute panel.

Implications

A large proportion of untrained consumers can accurately and repeatedly detect differences in beef tenderness. Despite wide variability in the ability of untrained consumers to detect differences in beef tenderness, a consumer panel can accurately and repeatedly detect differences in beef tenderness under controlled conditions. An untrained laboratory consumer panel may be able to provide as effective an evaluation of beef longissimus tenderness as a trained descriptive attribute panel.

Literature Cited

Boleman, S. J., S. L. Boleman, R. K. Miller, H. R. Cross, T. L. Wheeler, M. Koohmaraie, S. D. Shackelford, M. F. Miller, R. L. West, D. D. Johnson, and J. W. Savell. 1997. Consumer evaluation of beef of known tenderness levels. J. Anim. Sci. 75:1521–1524.

Branson, R. E., H. R. Cross, J. W. Savell, G. C. Smith, and R. A. Edwards. 1986. Marketing implications from the National Consumer Beef Study. West. J. Agric. Econ. 11:82–91.

Brooks, J. C., J. B. Belew, D. B. Griffin, B. L. Gwartney, D. S. Hale, W. R. Henning, D. D. Johnson, J. B. Morgan, F. C. Parrish, Jr., J. O. Reagan, and J. W. Savell. 2000. National Beef Tenderness Survey—1998. J. Anim. Sci. 78:1852–1860.

Chambers IV, E., J. A. Bowers, and A. D. Dayton. 1981. Statistical designs and panel training/experience for sensory analysis. J. Food Sci. 46:1902–1906.

Huffman, K. L., M. F. Miller, L. C. Hoover, C. K. Wu, H. C. Brittin, and C. B. Ramsey. 1996. Effect of beef tenderness on consumer satisfaction with steaks consumed in the home and restaurant. J. Anim. Sci. 74:91–97.

Lusk, J. L., J. A. Fox, T. C. Schroeder, J. Mintert, and M. Koohmaraie. 2001. In-store valuation of steak tenderness. Am. J. Agri. Econ. 83:539–550.

Miller, M. F., M. A. Carr, C. B. Ramsey, K. L. Crockett, and L. C. Hoover. 2001. Consumer thresholds for establishing the value of beef tenderness. J. Anim. Sci. 79:3062–3068.

Miller, M. F., L. C. Hoover, K. D. Cook, A. L. Guerra, K. L. Huffman, K. S. Tinney, C. B. Ramsey, H. C. Brittin, and L. M. Huffman. 1995. Consumer acceptability of beef steak tenderness in the home and restaurant. J. Food Sci. 60:963–965.

Munoz, A. M. 1998. Consumer perceptions of meat. Understanding these results through descriptive analysis. Meat Sci. 49:S287–S295.

NAMP. 1997. The Meat Buyers Guide. NAMP, Reston, VA.

Shackelford, S. D., T. L. Wheeler, and M. Koohmaraie. 1999. Evaluation of slice shear force as an objective method of assessing beef longissimus tenderness. J. Anim. Sci. 77:2693–2699.

Shackelford, S. D., T. L. Wheeler, M. K. Meade, J. O. Reagan, B. L. Byrnes, and M. Koohmaraie. 2001. Consumer impressions of Tender Select beef. J. Anim. Sci. 79:2605–2614.

Wheeler, T. L., S. D. Shackelford, E. Casas, L. V. Cundiff, and M. Koohmaraie. 2001. The effects of Piedmontese inheritance and myostatin genotype on the palatability of longissimus thoracis, gluteus medius, semimembranosus, and biceps femoris. J. Anim. Sci. 79:3069–3074.

Wheeler, T. L., S. D. Shackelford, and M. Koohmaraie. 1998. Cooking and palatability traits of beef longissimus steaks cooked with a belt grill or an open hearth electric broiler. J. Anim. Sci. 76:2805–2810.

Wheeler, T. L., D. Vote, J. M. Leheska, S. D. Shackelford, K. E. Belk, D. M. Wulf, B. L. Gwartney, and M. Koohmaraie. 2002. The efficacy of three objective systems for identifying beef cuts that can be guaranteed tender. J. Anim. Sci. 80:3315–3327.

Wyle, A. M., D. J. Vote, D. L. Roeber, R. C. Cannell, K. E. Belk, J. A. Scanga, M. Goldberg, J. D. Tatum, and G. C. Smith. 2003. Effectiveness of the SmartMV prototype BeefCam system to sort beef carcasses into expected palatability groups. J. Anim. Sci. 81:441–448.