Postharvest assessment of fruit quality parameters in apple using both instruments and an expert panel

Nnadozie Oraguzie a,*, Peter Alspach b, Richard Volz a, Claire Whitworth a, Chandra Ranatunga b, Rosemary Wesket a, Roger Harker c

a The Horticulture & Food Institute of New Zealand (HortResearch), Private Bag 1401, Havelock North 4157, New Zealand
b HortResearch, Old Mill Road, RD 3, Motueka 7198, New Zealand
c HortResearch, Private Bag 92169, Auckland Mail Centre, Auckland 1142, New Zealand

1. Introduction

Apple fruit texture and flavor (including sugars, acids and aroma) are important traits that guide consumer preference (Dailant-Spinnler et al., 1996; Jaeger et al., 1998). In fact, apples that have crisp, juicy texture and that maintain these characteristics during postharvest life are highly favored by consumers. Sweet and acid tastes are also included in assessments of postharvest quality (Smith, 1985; Mitcham, 1997). According to Mann et al. (2005), the sensory attributes that define apple fruit texture include firmness (force required to bite into the fruit), crispness (amount and pitch of sound generated when the fruit is first bitten with the front teeth), mealiness (degree to which the flesh breaks down to a fine lumpy mass) and juiciness (amount of juice released from the fruit in the first three chews, when chewing with the back teeth). This complex set of sensory attributes is derived from a varied set of physical and biochemical properties of the fruit cell wall and is quantitatively inherited (Durel et al., 1998; King et al., 2000; Oraguzie et al., 2001; Alspach and Oraguzie, 2002). Due to the influence of both genetic and environmental factors, estimates of heritability of texture traits in apple are often low to moderate (Oraguzie and Currie, 2004), although machine measurements generally give higher heritabilities than sensory scores. This is due to the higher magnitude of the residual variance component in comparison to the additive genetic component, with assessor differences and/or errors partly contributing to the environmental component (Alspach and Oraguzie, 2002). Low heritabilities impede genetic improvement due to the difficulty in identifying progenies from bi-parental crosses that have desired attributes.
Instrumental measurements of fruit texture have become one of the cornerstones of fruit quality assessment. The industry often sets quality standards that are based on instrumental texture measurements made using a penetrometer (Washington State Dept of Agriculture, 1999). Also, Brix and titratable acidity are often measured as proxy for sweetness and acidity (Harker et al., 2002a). The relevance of these machine measurements however will depend on how well they are able to predict sensory attributes.

This study was initiated in particular to gain an understanding of the components of the phenotypic variation for apple fruit texture due to evaluation methodology to facilitate designing phenotyping methods that will enhance the accuracy of trait values to maximise the genetic potential of a genotype. The objectives were to: (1) determine how many fruit samples that can be reliably assessed in a day by an expert panel; (2) determine consistency within and between assessments; (3) examine the relationships of machine measurements and sensory evaluations, which used a 10-point scale. The study differs from typical consumer and sensory studies that are commercially oriented and use trained/consumer panels and a few cultivars or advanced selections. We have used an expert panel who themselves are breeders with 40 years combined experience in sensory evaluations and a typical population used for postharvest studies in an apple breeding program.

2. Materials and methods

2.1. Experimental material

A seedling apple population from a half diallel with six parental cultivars (viz. ‘Cox’, ‘Pacific Rose’, ‘Prima’, ‘Liberty’, ‘Red Delicious’ and ‘Dulmener Rosen’) was used in this study. The trial consisted of a total of 324 trees representing four replicates of each of 15 families and the parents, planted in a randomised incomplete block design in a research orchard at the HortResearch centre in Havelock North, New Zealand. Only trees that were fruiting and had more than eight fruit were selected for the study. The aim was to randomly select 3 trees per plot (in instances where there were more than 3 trees from which to choose) giving 12 trees per family. In the event, this was not possible and fruit were phenotyped from 141 trees in total: 128 progeny (median 9 per family) and 1–4 from each of the parents (except for ‘Cox’ which was not represented).

2.2. Panel of experts

The panel of four experts (A–D) was set up using the following criteria: (1) membership of the pipfruit breeding team; (2) at least 5 years of experience in pipfruit sensory evaluation and; (3) participation in a sensory training exercise. The HortResearch sensory and consumer science group provides sensory training each year before apple harvest. The aim is for assessors to familiarise themselves with known reference standards for different apple texture attributes. The panel also re-oriented the day before the present study was undertaken for 2 h on specific texture attributes using reference standards and a selection of test samples. Throughout the training panelists were provided with feedback on scoring of attributes to ensure panel consistency. Artificial solutions containing mixtures of sucrose (7–14% Brix) and malic acid (0.08–0.2%, w/v) were also provided for evaluation of sweetness and acidity, respectively. For a detailed description of the reference standards and how they are applied for apple fruit sensory analysis, see Alspach and Oraguzie (2002).

2.3. Fruit picking, storage and evaluations

In a study designed to assess postharvest fruit quality differences among seedlings from a bi-parental apple population (Brookfield, pers. comm.), half of the fruit were harvested before ‘commercial maturity’ and the other half at ‘eating ripeness’. The results showed more significant differences among seedlings for fruit stored at 0.5 °C for 10 weeks than for fruit assessed immediately after harvest irrespective of time of harvest. Based on these data, we chose one picking date at ‘commercial maturity’ and cold storage at 0–4 °C for 8 weeks for this study to save time and costs, and to also mirror the research closely to the approach used to evaluate commercial apple breeding lines to identify individuals with desired postharvest attributes for more advanced testing and release. Trees were harvested once a week from 4 February to 25 March 2003 using background colour, taste and seed colour to ascertain maturity. Eight fruit were harvested from each tree, weighed and transferred to the cold room on the same day. A day before sensory assessment the fruit samples were removed from cold storage and held at 20 °C. Fruit evaluations took place between 1 April and 20 May 2003.

Each fruit was weighed before puncture measurements using a hand-held Effegi penetrometer (Model FT327, McCormick Fruit Tech, Washington) fitted with an 11.1-mm probe were made on opposite sides of the fruit in the blush and shaded portions. Maximum force was recorded. A slice of skin was removed from the skin prior to the puncture test. Juice was squeezed from the holes made with the penetrometer and collected onto a digital refractometer (Model 2110-W07, Atago, Tokyo, Japan) for measurement of Brix. Fruit were cut after completion of instrumental measurements, peeled, cored, coded and presented to the experts in a random order which was either recorded or able to be inferred from the data. Two experts independently scored segments from each fruit, and each expert assessed two fruit per tree. Assessments were for firmness, juiciness, crispness, acidity and sweetness on a 0–9 scale (0, meaning nil and 9 being the extreme value). Tap water was provided as a palate cleanser and reference standards were available for re-calibration in between samples. Experts worked in sessions, with a maximum duration of 2 h, during which they assessed 10–34 fruit. All eight evaluation dates had morning sessions, two had one afternoon session and one had two afternoon sessions. The juice was titrated to an end-point of pH 8.1 with 0.1N NaOH, using an automatic titrator (Metrohm Ion analysis, CH-9101 Herisau/Switzerland).

3. Statistical data analysis

3.1. Consistency among and within experts and instrumental readings

The design of the experiment allowed for comparison between different experts’ assessment of the same fruit, and an expert’s assessment of different fruit from the same tree (and harvest and evaluation dates). For each pair of experts, the correlations between their scores were computed, and the numbers of fruit for which their scores differed by every possibility (i.e., −9, −8, . . . , 9) were tabulated. Similarly for each expert, the correlations between the scores for the pairs of fruit from the same tree were computed, and the numbers of instances for which the scores differed by every possibility tabulated. Since the four fruit from each tree were also subject to instrumental measurements, it was possible to calculate similar tables for the instrumental measurements. For flesh firmness, which was measured twice per fruit, it was also possible to derive a within-fruit comparison.

3.2. Relationship between expert assessment and instrumental readings

The relationships between expert assessments and instrumental readings were examined using scatterplots, augmented with
non-parametric smoothed fits (loess curves), and modeled using ordinary least squares regression with experts, the instrumental reading and the interaction included in the model. For sweetness, more detailed modeling was undertaken using both soluble solids and titratable acidity as well as experts and all interactions.

### 3.3. Fatigue

After assessing a large number of fruit, experts could conceivably become fatigued. This might manifest in two ways.

#### 3.3.1. Reduced accuracy

The assessments might become less accurate. Hence, the average magnitude of the difference between two experts scoring the same fruit would increase as the session progressed. This was explored graphically, and by calculating the correlation coefficient between the inter-expert difference and the assessment order for each trait during each session.

#### 3.3.2. Reduced discrimination

Experts might find it more difficult to discriminate between different levels of the sensory traits; that is, all fruit would tend to seem similar. Under this circumstance, the variation between experts assessing the same fruit would not necessarily change (in fact, it might even decrease). However, the scatter about the relationship between the expert’s assessment and the instrumental reading would increase, and there would be less variation in the scores of a particular expert. This possibility was also examined graphically using scatterplots of the standard deviation of the residuals of the loess smoothing of expert assessment on instrumental readings against the order of fruit assessment.

The method used for reduced accuracy differed from that used to examine reduced discrimination in two ways: (a) the method for reduced accuracy looks at a trend in absolute difference whereas the method for reduced discrimination looks at a trend in standard deviation; (b) the method for reduced accuracy breaks up the data according to session whereas the method for reduced discrimination breaks it up according to expert.

### 3.4. Bias from previous fruit

The characteristics of the previous fruit, or the expert’s assessment of them, could lead to bias in assessing the current fruit in a couple of ways.

#### 3.4.1. Contrast error

If the previous fruit were near the extreme, then an expert’s assessment of the current fruit might be biased. For example, a fruit tasted after a particularly sour one might be perceived as sweeter than if the same fruit were tasted after a sweet one. This was investigated by cross-correlation plots of difference between the current and previous fruit on the error in the assessment of the current fruit. The difference was taken from difference in instrumental reading, and the error as the residual from the loess smooth trend with the instrumental reading.

#### 3.4.2. Habituation error

Two similar fruit in succession are likely to receive the same score. Thus, if the first was scored incorrectly, then it is plausible that the second will also be similarly rated incorrectly. Such an effect could manifest itself in successive errors being correlated. This was examined by autocorrelation plots, where the error was taken as the residual from the loess smooth trend with the instrumental reading.

### 3.5. Halo error

A halo error occurs when an assessor scores a fruit positively in one trait because it is positive in another unrelated trait. For example, a firm fruit could be perceived as sweeter than a soft fruit of the same sweetness. Such an error would result in a tendency to over-estimate the score for one trait when the influencing trait is high. This was examined using scatterplots, where the residuals from the loess smooth trend of the trait under consideration with its corresponding instrumental reading provided the measure of over/under-estimation, and the value for the influencing trait was provided by its instrumental reading. Halo error was examined for distortions in perceived sweetness attributable to flesh firmness, and distortions in perceived firmness attributable to soluble solids.

All analyses and graphics were undertaken in R 2.5.0 (R Development Core Team, 2007).

### 4. Results

#### 4.1. Consistency among and within experts and instrumental readings

The correlations between paired experts assessing the same fruit were highest for firmness ($r > 0.7$) and lowest for sweetness ($r < 0.5$) (Table 1). For firmness, almost all the paired assessments were within two points of one another, and even for the sweetness about 90% of the paired assessments were similarly close. It was also noticeable that some experts tended to score higher (or lower) than others. For example, the difference between the highest and lowest expert mean firmness score was 1.32. Of the instrumental readings, only flesh firmness was undertaken more than once on each fruit. The correlation between these paired readings was 0.87 (Table 2), which was considerably higher than that between paired

<table>
<thead>
<tr>
<th>Expert pair</th>
<th>$r$</th>
<th>Mean S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:B</td>
<td>0.69 86 98</td>
<td>4.53 0.124</td>
</tr>
<tr>
<td>A:D</td>
<td>0.78 84 99</td>
<td>4.60 0.126</td>
</tr>
<tr>
<td>B:C</td>
<td>0.45 67 89</td>
<td>4.04 0.117</td>
</tr>
<tr>
<td>C:D</td>
<td>0.61 68 94</td>
<td>4.24 0.120</td>
</tr>
</tbody>
</table>

### Table 1

Left portion of table, comparison between two experts’ assessment of the same fruit, for each assessed trait: the correlation coefficient ($r$), and the percentage of fruit ($n=147$) for which the two experts scored within one ($r < 0.7$) and lowest for sweetness ($r < 0.5$) (Table 1). For firmness, almost all the paired assessments were within two points of one another, and even for the sweetness about 90% of the paired assessments were similarly close. It was also noticeable that some experts tended to score higher (or lower) than others. For example, the difference between the highest and lowest expert mean firmness score was 1.32. Of the instrumental readings, only flesh firmness was undertaken more than once on each fruit. The correlation between these paired readings was 0.87 (Table 2), which was considerably higher than that between paired

<table>
<thead>
<tr>
<th>Expert pair</th>
<th>$r$</th>
<th>Mean S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:B</td>
<td>0.69 86 98</td>
<td>4.53 0.124</td>
</tr>
<tr>
<td>A:D</td>
<td>0.78 84 99</td>
<td>4.60 0.126</td>
</tr>
<tr>
<td>B:C</td>
<td>0.45 67 89</td>
<td>4.04 0.117</td>
</tr>
<tr>
<td>C:D</td>
<td>0.61 68 94</td>
<td>4.24 0.120</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Expert pair</th>
<th>$r$</th>
<th>Mean S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:B</td>
<td>0.69 86 98</td>
<td>4.53 0.124</td>
</tr>
<tr>
<td>A:D</td>
<td>0.78 84 99</td>
<td>4.60 0.126</td>
</tr>
<tr>
<td>B:C</td>
<td>0.45 67 89</td>
<td>4.04 0.117</td>
</tr>
<tr>
<td>C:D</td>
<td>0.61 68 94</td>
<td>4.24 0.120</td>
</tr>
</tbody>
</table>

* Two pairings, A:C and B:D, were not included in the design.
expert assessments of firmness (Table 1). Only about 7% of the fruit had paired flesh firmness readings that differed by more than 20% of the overall mean, which was less than the percentage of paired expert assessments that differed by more than one point (Table 1).

The correlations between the same expert’s assessments of different fruit from the same tree were generally higher than those for firmness (Table 1). For firmness, the correlations were only between different experts’ assessment of the same fruit, except for firmness that differed by more than one point (Table 1). There was no evidence of inter-expert agreement decreasing during a session for any of the five traits (data not presented). This is confirmed by the correlation coefficients between inter-expert difference and the average assessment order during each session (Table 4).

4.3. Fatigue

4.3.1. Reduced accuracy

There was no evidence of inter-expert agreement decreasing during a session for any of the five traits (data not presented). This is confirmed by the correlation coefficients between inter-expert difference and the average assessment order during each session (Table 4).

4.3.2. Reduced discrimination

There was no convincing evidence of fatigue resulting in the experts’ discriminatory ability reducing during a session (Fig. 2). Although the standard deviations for two of the experts trend upwards at the end for sweetness perception, those for the other two trend downwards. Both these ‘trends’ are likely to be noise. It is interesting to note that there was no evidence of higher variance for the first one or two readings of a session. The superior ability of expert A in predicting soluble solids content is noticeable in the lower standard deviations.

Table 4  Correlation coefficients between inter-expert difference and average assessment order for each trait during each session.

<table>
<thead>
<tr>
<th>Day</th>
<th>Time</th>
<th>Trait</th>
<th>Sweetness</th>
<th>Acidity</th>
<th>Firmness</th>
<th>Crispness</th>
<th>Juiciness</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 April</td>
<td>am</td>
<td></td>
<td>0.02</td>
<td>0.17</td>
<td>0.05</td>
<td>0.30</td>
<td>0.15</td>
</tr>
<tr>
<td>08 April</td>
<td>pm</td>
<td></td>
<td>0.12</td>
<td>0.00</td>
<td>−0.30</td>
<td>−0.15</td>
<td>−0.10</td>
</tr>
<tr>
<td>15 April</td>
<td>pm</td>
<td></td>
<td>−0.29</td>
<td>0.19</td>
<td>0.06</td>
<td>0.05</td>
<td>−0.14</td>
</tr>
<tr>
<td>23 April</td>
<td>am</td>
<td></td>
<td>−0.05</td>
<td>−0.16</td>
<td>0.15</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>23 April</td>
<td>pm</td>
<td></td>
<td>−0.24</td>
<td>0.07</td>
<td>0.02</td>
<td>0.36</td>
<td>0.14</td>
</tr>
<tr>
<td>29 Apr</td>
<td>am</td>
<td></td>
<td>0.33</td>
<td>0.01</td>
<td>0.22</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>29 Apr</td>
<td>pm</td>
<td></td>
<td>−0.07</td>
<td>−0.14</td>
<td>−0.03</td>
<td>0.04</td>
<td>−0.10</td>
</tr>
<tr>
<td>06 May</td>
<td>am</td>
<td></td>
<td>0.01</td>
<td>−0.13</td>
<td>0.23</td>
<td>0.06</td>
<td>−0.10</td>
</tr>
<tr>
<td>06 May</td>
<td>pm1</td>
<td></td>
<td>0.00</td>
<td>−0.04</td>
<td>0.35</td>
<td>−0.09</td>
<td>−0.15</td>
</tr>
<tr>
<td>06 May</td>
<td>pm2</td>
<td></td>
<td>−0.16</td>
<td>−0.13</td>
<td>−0.08</td>
<td>0.03</td>
<td>−0.07</td>
</tr>
<tr>
<td>14 May</td>
<td>am</td>
<td></td>
<td>−0.05</td>
<td>0.36</td>
<td>0.58</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>20 May</td>
<td>am</td>
<td></td>
<td>−0.04</td>
<td>−0.30</td>
<td>0.22</td>
<td>0.23</td>
<td>−0.46</td>
</tr>
</tbody>
</table>
4.4. Bias from previous fruit

4.4.1. Contrast error
The correlations between the difference between the current and previous fruit as gauged by the instrumental reading, and the experts’ score of the current fruit were occasionally significantly different from zero. For firmness these were no more than would be expected by chance given the number of correlations examined, but for sweetness on soluble solids and acidity on titratable acidity there were six and seven (out of 48), respectively, such correlations for zero lag (data not shown). This is slightly more than would be expected, but the correlations were both positive and negative. For all traits by instrumental reading pairs (i.e., sweetness on soluble solids, acidity on titratable acidity, sweetness on titratable acidity and firmness on mean flesh firmness), the residuals tended to be highly correlated with the magnitude of the expert’s score—particularly for sweetness on soluble solids and firmness on mean flesh firmness.

4.4.2. Habituation error
For each scored trait by instrumental reading pair (i.e., sweetness on soluble solids, acidity on titratable acidity, sweetness on titratable acidity and firmness on mean flesh firmness), there were four experts by 12 sessions and up to five lags were examined. That is, 240 (auto) correlations. Thus, one would expect some to differ from zero at the 5% level by chance alone. There were no more ‘significant’ correlations than would be expected overall, or for a particular expert or lag number (data not shown). Neither was there any indication that particular lags tended to have either negative or positive correlations. Thus, by this criterion, there was no evidence of error perpetuation for any of the experts’ scores of any of the traits.

4.5. Halo error
No evidence was found for fruit soluble solids content influencing the assessment of firmness by any of the assessors (data not shown). However, there was a slight indication that one of the experts (C) may have assessed firm fruit as sweeter than less firm fruit (Fig. 3). The effect was small (perceived sweetness score increase of 0.10, standard error 0.049, per unit increase in flesh firmness; \( P = 0.059 \)). This could simply be a false positive result.

5. Discussion
The postharvest evaluation of apples has always relied upon some element of tasting by humans, whether they have been the
experimenters themselves or other participants. For example, the early and seminal research on mealiness, a dry and floury texture that develops during storage of some cultivars of apples, was based on the tasting by the sole researcher (Fisher, 1943). In more recent years the approaches to tasting of fruit have been greatly influenced by the development of the formalized sensory evaluation techniques (Lyon et al., 1992; Lawless and Heymann, 1999; Meilgaard et al., 2007). In particular, the concept of Quantitative Descriptive Analysis® (QDA), in which a group of 10–12 participants are trained to evaluate selected foods against a set of tightly defined flavor and texture attributes using quantitative scales that are often fixed using a series of reference foods, has had a great influence on the way fruit have been evaluated. For example, generic forms of QDA have been used to establish sensory–instrumental relationships for texture and taste of apples (Harker et al., 2002a,b) and to help identify flavor and texture attributes that may be targets for apple breeding programs (MacFie and Beyts, 1995; Daillant-Spinnler et al., 1996; Alston et al., 1996). Generally, these types of apple studies are smallish and discrete with no more than 12 treatments, which can be evaluated over a number of days or weeks. The number of fruit treatments that need to be evaluated by apple breeding programs far exceeds this number of treatments. Despite this, a few apple breeding initiatives have attempted to use trained panels to evaluate fruit from selected progeny. King et al. (2000) evaluated two batches of 16 genotypes using a 11-member trained sensory panel in a procedure aimed at identifying sensory-texture QTLs, while Hampson et al. (2000) used 12 in-house panelists from a pool of 60 to evaluate a mixture of descriptive and hedonic attributes as part of the evaluation of apples from a Canadian apple breeding program. Even so, the first screening was by a team of three (Hampson et al., 2000). The use of small teams of tasters from breeding programs has continued to be the mainstay of germplasm evaluation. Yet consideration of this form of flavor evaluation in relation to the QDA literature focuses attention on three main questions: (1) is it reasonable to use small numbers of tasters to evaluate germplasm in this way? (2) How do these expert tasters compare with trained panelists used in QDA? (3) How does the large number of samples being tasted affect the accuracy of scoring?

### 5.1. Number of tasters

While the four experts used in the current study were less than the usual 10–12 panelists used in QDA, they are not so different from the number of panelists required for the Flavor Profile® approach (FP; Meilgaard et al., 2007). Note that the choice of four experts in this study was not intended in anyway as an indication of what breeding programs should aim for. Rather it was arrived at based on the number of experts available, the amount of fruit and time available and the need to achieve inter-expert comparisons. However, as few as four panelists are required for FP assessments.
The FP methodology has well recognised shortcomings in terms of consistency and reproducibility; however, these problems can be overcome by training and through the process in which panelists reach a consensus on the scores for flavor (Meilgaard et al., 2007). Consensus in the sensory literature relates to how a trained panel as a result of discussions will agree on a single score for a particular flavor attribute, but is often considered inferior as it artificially removes variability from the data.

The small number of assessors in the current study, in itself, is not an issue so long as assessors are operating in a reproducible fashion (see later) and according to robust protocol which allows performance to be evaluated against other assessors or tested within the protocol using blinded reference samples. Indeed, it is not unusual to find individuals with sole responsibility for tasting of foods and beverages in commerce. For example, an individual tea taster would evaluate 500–1000 teas in a day when deciding on raw product to purchase for subsequent blending (Greenwood, pers. comm.; Wright, 2005). A professional tea taster would ensure s/he is regularly audited for sensory capability so that they remain confident about the flavors they can perceive, they have an excellent appreciation of the diversity of flavors within the product space, and they often need to have an obsession with the consistency of an individual brand (Greenwood, pers. comm.). This ‘obsession’ with the product and product space is perhaps not that different to what might be expected from someone devoting their life to fruit breeding.

5.2. Comparisons of expert and trained panelists

The relative merits of using expert or trained panelists have been considered in the wine literature (e.g., Gawel, 1997; Zamora and Guirao, 2004; Perrin et al., 2007). While trained panelists are subjected to detailed training programs (Meilgaard et al., 2007), experts often accumulate similar knowledge through experience. Gawel (1997) described this accumulation of experience as being related to: "(1) a familiarity with a product class resulting from long term exposure to a wide variety of members representing that class and (2) where that exposure occurs in conjunction with considered thought as to the products sensory characteristics." An overview of this wine literature suggests that there is little difference in ability of either group to discriminate and/or describe differences among wines (Gawel, 1997; Zamora and Guirao, 2004; Perrin et al., 2007). While acknowledging the many nuances in human perception of odor, taste and texture, it would seem difficult to argue, a priori, that

---

**Fig. 2.** Scatterplots of the standard deviation of the residuals of the loess smoothing of expert assessment on instrumental reading against the order of assessment. Points are for all experts at each assessment order, and the thick line shows loess smooth trend through these. Thinner lines show the loess smooth trend for each expert (A, – · –; B, – – – –; C, . . . ; D, – – – –) separately based on standard deviations derived from a moving window of five adjacent assessment orders.
either expert or trained panelist would be better able to perceive and discriminate flavor among products they are familiar with. In the current study, the assessors were not only experts in terms of Gawel’s description above, but also went through basic sensory training (see Section 2) so that they might be considered to be well prepared for the project. While there is no direct evidence for how they performed relative to any trained panel, such inferences can be made by comparing their results to those published in the literature. Harker et al. (2002a,b) used a 15–20-member trained panel to explore the relationships between instrumental and chemical measurement and human perception of texture and taste. Correlations between sweet taste and SSC, acid taste and TA and crispness and puncture force (i.e., flesh firmness) ranged from –0.2 to 0.8, 0.3 to 0.9 and 0.7 to 1.0, with median correlations of 0.41, 0.86 and 0.87, respectively. The correlations obtained by the experts for similar instrumental–sensory relationships in the current study (Table 1 and Fig. 1a and b) were close to the median values obtained previously for trained panelists, even though the task was more challenging since it was over a much longer time period and occurred in less than an optimal testing environment.

Another aspect of assessor performance that must be considered is the extent that they are influenced by psychological factors such as habituation, contrast and halo effects. The error of habituation relates to a panelist’s reluctance to give the current sample a flavor score that is different to the previous sample even though the stimulus has changed; contrast errors relate to the way a panelist will exaggerate the weakness or strength of flavor intensity of the current sample when it provides a weaker or stronger stimulus compared with the previous sample; and a halo error occurs when a panelist transfers positive experience across a number of different flavor attributes, for example a firm apple is also rated as being very sweet even though its taste has not changed (Lawless and Heymann, 1999; Meilgaard et al., 2007). In the current study there was no indication that these psychological errors (except for the weak contrast effect for acidity and sweetness) were affecting assessor’s evaluation of the fruit samples. Therefore we can be confident that not only were the assessors able to perceive differences in flavor and texture attributes being considered, but they were also able to undertake these evaluations in an objective fashion. Although these results are encouraging, the differences between the experts do point to potential problems if more than one expert is used to assess a particular germplasm population. The assessor should always be recorded and, importantly, the information should be used in making selection decisions after it has been adjusted for assessor differences. Also, the differences between the experts, and the superior ability of expert A to distinguish different levels of sweetness in the presence of varying acidity, suggests that further training and/or standardization of experts could be undertaken to improve the precision and accuracy of the scores, increase the concordance between expert’s scores and enhance selection efficiency.
5.3. Number of samples tasted

The number of fruit that needs to be assessed by sensory evaluation in these types of fruit-heritability studies has the potential to overwhelm tasters. The subject of sensory fatigue is well recognised in the literature on QDA and related sensory methodologies. The greater the number of samples that are assessed the greater the risk of fatigue of the taste and odor sensers (Lyon et al., 1992). The extent that fatigue will affect assessors will depend on (1) the nature of the product being tasted, (2) the number of flavor/texture attributes that need to be evaluated for each sample, (3) the amount of sample that needs to be consumed during assessment, and (4) the experience and motivation of individual panelists (Lyon et al., 1992). In the current study, there was no evidence that the protocol used for sensory evaluation was fatiguing to the sensory acuity of participants. This is perhaps not surprising given the high motivation of the participants who were aware of the objective of the study rather than simply being given a role in the evaluation of flavor. Furthermore, apples are not a food that would expect to be overly fatiguing to taste—they do not have strong flavors that provide lingering after-tastes and contribute to carry-over effects.

6. Conclusion

The experimental design and the data analysis herein represent a fairly extensive search for different sources of error that could potentially inflate the variability in experts’ postharvest assessment of fruit quality. We found no evidence of fatigue while the carry-over effect was insignificant despite the experts assessing up to 34 fruit in a session. The smaller intra- than inter-expert variation points to the utility of the data for making selection decisions after some adjustment for inter-expert differences. The data also confirm previous findings that instrumental readings are more accurate than sensory scores. However, the differences between experts and the superior ability of expert A to distinguish varying levels of sweetness in the presence of varying levels of acidity suggest that further training and/or standardization of experts may be necessary.

Acknowledgments

This research was partially funded by the New Zealand Foundation for Research, Science and Technology (FRST). We thank Dr. Anne Gunson for initial assistance with data analysis and Dr. Vincent Bus for setting up the partial diallel trial. Our thanks also go to Matt Greenwood, the master tea taster for Bell Tea, Huttinington Park, Manukau, New Zealand, for providing information on expert assessment of tea quality, and Paul Brookfield, HortResearch, New Zealand, for sharing his postharvest fruit quality research data.

References