

## Review 1,

Dear Reviewer, thanks for comments. We have taken them all into account and revised the manuscript. See below text in red for amendments.

The authors present a paper that is succinct and has relevance for the journal – the authors assess consumer liking and sensory appraisal of apples presented with combinations of optimal and sub-optimal colour, shape and damage characteristics (presented as images), in the context of understanding food waste from aesthetic differences in a fresh-produce. Although the content is relevant, the paper is not suitable for publication in its current form. More information on the methods, study design and statistical analyses is required in order to properly assess the quality of the research. Please see my specific comments below. I would be happy to re-consider the manuscript once these clarifications have been made.

**Optimal/Suboptimal characteristics** – The damage and shape scales seem appropriate but it is unclear why green was considered a suboptimal colour? Apples are typically red or green, so it is possible that the colour was not viewed as sub-optimal. I would consider brown or dull colouration a suboptimal colour, not green. Can the authors explain their choice here?

Thank you for notifying us! This is absolutely right. The subject is discussed in the discussion paragraph. There is, however, a need to clarify the red-green color already in the introduction. In line 58-59 the following sentence is thus added: It has further been shown that red apples are more liked than green ones, and that red colored apples may be seen as more optimal than for example green.

**Figure 1** – This figure is quite nice. I am not sure if this is just the PDF I received, but the images of the apple should be in colour, not greyscale.

The figure is in color.

**Methods** – More detail is required to understand the study design.

Firstly, there appears to be 8 picture combinations paired with 8 different apple-samples... does this mean that there were (8 x 8) 64 combinations tasted and rated by all participants? Please clarify.

A clarification is certainly needed to avoid any misunderstanding. A table showing the eight samples in the cubic design is added (Table 1). The text is clarified and now reads (line 89-94): The study was based on a cubic full factorial design of visual appearance (color, shape and damage) presented in a total of eight samples, each consisting of a combination of pictures describing the optimality-suboptimality. The samples were ranging from optimal to sub-optimal in all three dimensions (based on combinations of color (red-to-green), shape (normal-to-odd) and damage (none-to-damage). The design is presented in Table 1 and visualized in Figure 1. The picture combinations for each sample were combined with the tasting of eight randomly selected apples.

Secondly, how did these 8 samples of apples differ? Were they different types of apples? Different colours? How were they selected? What quantity (g) was served as a sample and how were they prepared? skin-on?

The eight categories/samples were 1. the pictures (categories) in the design and 2 taste samples. The pictures differed according to color, shape and damage. The tasting samples were random apples of different cultivars, all grown in Sweden. The text in the material and method section has been rewritten to better explain the test procedure and the included samples.

Thirdly, did all participants evaluate all sample combinations? Between-group comparisons are mentioned on line 102-103, but it is not clear if a between-groups design was employed (and how), or

if this was a completed repeated-measures/within-groups study. The study design needs to be fully described.

All consumers evaluated the same samples. This is clarified in line 105. Between group comparison is due to liking and part of the statistical evaluation. The text in the statistical evaluation is changed to clarify this. Please see below (Statistical Analyses).

Finally, the participants were not trained panellists, but it is not clear whether any descriptors were provided for each of the 13 attributes assessed. Please add this.

The questionnaire was self-instructive with no descriptors of the attributes. This is added in line 104.

**Statistical Analyses** - The analysis plan requires clarification in order to understand the study design and whether the data have been treated correctly. For example, the authors state that Tukey-Kramer tests were used alongside 1-way ANOVAs, and that 14 tests were completed in total. However, it is unclear to me which variables (both independent and dependent variables) are assessed in these analyses? There are 8 pictures, 8 apple samples and 14 liking/sensory outcomes. Does this mean that separate within-groups ANOVAs with 'picture' as a factor (8 levels) were conducted for each of the 14 outcome attributes, and the data were collapsed across the 8 apple samples? Please clarify in the text. Rather than assessing the different apple-attributes (colour/damage/shape) within one variable, these could have been included as three different Factors (2 levels each – optimal vs suboptimal) within the same statistical models. Why was this not done?

We performed a one-way analysis using 8 groups, and we performed this analysis first for liking and then for all 13 perceived sensory attributes. It is corrected that we could have used a three-way analysis with two layers per grouping variable; however, given the small number of groups (8) we felt that it is actually more clear to look at the "whole" grouping directly without adding the complexity of n-way ANOVA. We did confirm that the differences and significance of differences between groups remains exactly the same, and since this is what we aim to study (rather than the importance of pure vs interaction terms), we decided to settle for the simplest ANOVA analysis. We added a comment on this in the Statistical evaluation section.

Another query is that it is not clear whether any correction for multiple comparisons was applied. If it was not, it is possible that the significant differences reported may not remain significant. Please add in this detail and justify why corrections were not applied, if they were not.

The reviewer makes a good point here, and between different ANOVA tests for the perceived sensory attributes we do not perform an additional multiple comparison correction. We have however changed that and introduced an extra Bonferroni correction for the sensory attributes, effectively dividing our alpha by 13 in that study.

We comment on this in the Statistical evaluation section, and furthermore, this correction removes the only significant finding we had on sensory attributes, a difference in earthiness. We have changed the text in the Results and Discussion sections accordingly.

Finally, the authors state that between-groups differences were assessed for liking and the 13 different attributes, but it is not clear how/why this can be as I was under the impression that this was a repeated-measures designed. What variables were assessed between-groups? Please clarify.

With "between-group" we refer to groups of apples, not groups of respondents. We changed "between group" to "between apple categories" to clarify this further.

**Results** – The results of the statistical analyses (e.g. actual f-values and p-values) should be reported or summarised where possible, even for non-significant findings. Currently the authors primarily describe the results in text and it is unclear which analyses are being reported where and how

'singificant' and 'non-significant' these differences were without any indication of the relevant f-values and p-values.

We choose here to describe to visualize the results in Figures 2 and 3. Furthermore, we added p-values for the significant differences in terms of liking, and the lowest p-values for each sensory attribute to indicate "how insignificant" the differences were in those cases.

We may add a collection of tables indicating all results for the statistical analyses, but we feel that this would consume too much space in the article. If the editor feels that this would be appropriate we will add these tables.

Table 1 legend –  $P = 0.05$  should be  $P > 0.05$ ?

Thank you, this is corrected!