Implementing immersive technologies in consumer testing : Liking and Just-About-Right ratings in a laboratory, immersive simulated café and real café

Food Quality and Preference

Zandstra, E.H.; Kaneko, D.; Dijksterhuis, G.B.; Vennik, E.; Wijk, R.A.

https://doi.org/10.1016/j.foodqual.2020.103934

Food Quality and Preference 84 (2020) 103934

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Short Communication

Implementing immersive technologies in consumer testing: Liking and Just-About-Right ratings in a laboratory, immersive simulated café and real café

E.H. Zandstra[a,b,*], D. Kaneko[c], G.B. Dijksterhuis[d], E. Vennik[b], R.A. De Wijk[d]

[a] *Unilever Innovation Centre Wageningen, The Netherlands*
[b] *Human Nutrition & Health, Wageningen University and Research, The Netherlands*
[c] *Kikkoman Europe R&D Laboratory B.V. Wageningen, The Netherlands*
[d] *Wageningen Food and Biobased Research, Wageningen University and Research, The Netherlands*

ARTICLE INFO

Keywords:
Context
Immersive technology
Engagement
Liking
Just-About-Right ratings

ABSTRACT

Initial research indicates that the use of immersive technologies may improve the predictive validity and reliability of liking scores in consumer testing. However, how immersive technologies impact Just-About-Right ratings is hardly known. Forty-five participants took part in three tasting sessions, each in a different context: 1) laboratory, 2) immersive context simulating a café using audiovisual cues, and 3) real café. Each session, participants tasted four tomato soups varying in salt content preceded by a warm-up sample. Liking, optimal levels of sensory attributes (JAR) and engagement were measured. Results showed that there were no differences in liking or JAR ratings on sensory attributes of the soups across the three contexts. Nevertheless, participants felt more engaged in the real café and simulated café than in the laboratory. These results contribute to a better understanding of how sensory differences as assessed in a laboratory or immersive context relate to sensory differences that consumers would notice when they use the products in real-life.

1. Introduction

In traditional central location tests, participants evaluate products in isolated sensory booths where everything is as standardized as possible (*i.e.,* temperature, light conditions, sound *etc.*) and non-product contextual information is intentionally minimized. This is completely different from real-life situations in which you drink or eat a product in different contexts (*e.g.*, at home or work, in a café or restaurant) together with other people (*e.g.*, with family, friends and colleagues). Although sensory booths enable a strict control over product testing, they may not be representative of what happens in the real world as they do not take into consideration the role of context in shaping product perceptions and acceptance (Galiñanes Plaza, Delarue, & Saulais, 2019).

A solution would be to simulate the real-life context in the laboratory via the use of immersive technologies. *i.e.*, re-creating the physical context of a consumption situation in a laboratory with visual, auditory and olfactory cues. Initial research indicates that incorporating immersive technologies into sensory and consumer testing may improve the predictive validity and reliability of liking scores (Bangcuyo, Smith, Zumach, Pierce, & Guttman, 2015; Delarue, Brasset, Jarrot, & Abiven, 2019; Hathaway & Simons, 2017; Sinesio et al., 2018). However, little is known how immersive technologies impact optimal levels of sensory attributes (JAR ratings) of products. The aim of the present study was therefore to measure JAR ratings next to liking in a laboratory context, an immersive simulated context and a real-life context.

2. Material and methods

2.1. Participants

In total, 54 Dutch participants were recruited to participate in three taste sessions from the Wageningen Food & Biobased Research database according to the following selection criteria: 1) 18–60 years of age, 2) consuming tomato soup at least once a month, and 3) healthy (self-reported). Of the 54 recruited participants, nine participants were excluded from the study as they did not show up (eight participants) or dropped out after the first taste session (one participant). This resulted in a data set of 45 participants (eight males; mean age 37.9 ± 14.9 years). Participants signed an informed consent and received a monetary compensation of €15 for their participation. Participants were not informed about the actual purpose of the study but were told beforehand that the taste sessions consisted of tasting and eating tomato soup. The Social Sciences Ethics Committee of

* Corresponding author.
E-mail address: liesbeth.zandstra@unilever.com (E.H. Zandstra).

https://doi.org/10.1016/j.foodqual.2020.103934
Received 4 September 2019; Received in revised form 17 March 2020; Accepted 18 March 2020
Available online 24 March 2020

Wageningen University and Research approved the study.

2.2. Product

Three tomato soups (Dutch brand: Unox Cup-a-Soup) were included in the study. The variant Tomato Cream was chosen as it is a common variant that was expected to be accepted by most people. Also, this variant does not contain any croutons or other big pieces that could lead to variation in taste perception per bite. Salt was added to the soup to make three distinguishable samples: 1) benchmark tomato soup as available on the Dutch market at the time of the study (340 mg Na/100 ml prepared soup), 2) tomato soup with 25% extra salt (425 mg Na/100 ml prepared soup), and 3) tomato soup with 50% extra salt (510 mg Na/100 ml prepared soup). These levels were based on the preliminary assessment that they differed enough to obtain discriminating scores while still being acceptable. The recipe for all soups was the same (per 100 ml prepared soup: energy 149 kJ, protein 1.0 g, total fat < 0.5 g, saturated fat 0.2 g, carbohydrates 6.8 g, sugar 4.0 g); only the salt content was different. A duplicate of the benchmark tomato soup was used as a fourth sample to check the reliability of the subjects. As warm-up sample, the regular tomato soup variant (Dutch brand: Unox Cup-a-Soup) was used. The regular tomato soup was a different product from the Tomato Cream soup, but not too different as it was still within the same product category. Participants followed the same procedure for the warm-up sample as for the other soups. The results of the warm-up samples are not reported here.

2.3. Contexts

Participants took part in three tasting sessions, each in a different context: 1) laboratory, 2) immersive context simulating a café using audiovisual cues, and 3) real café. Fig. 1 shows pictures of the testing environments.

Laboratory - Testing in the laboratory setting took place in the so-called Experience Room in the Helix building of Wageningen University & Research, however, none of the immersion tools were activated. The lighting of the room was tailored to daylight illumination. Two standard office tables and 6 chairs were used. There were no dividers: subjects were separated simply by distance. Therefore, it was just an ordinary room with tables and chairs. Participants were asked to take a seat, three persons per table, and were asked to be quiet during the whole session (Fig. 1, panel A).

Immersive simulated café - Testing in the immersive simulated context took place in the same Experience Room. Photographic footage of the Grand Café of another building of Wageningen University & Research (Forum building) was projected on all four walls using eight projectors. A recording of the sounds in the Grand Café was played through the speaker system. The room was lit with spots that were set to give somewhat yellow light, similar to the lighting in the Grand Café. Six white chairs were used consistent with the theme of the Grand Café. The same two tables as those used in the laboratory setting were used. Participants were asked to take a seat, three persons per table. They were allowed to talk during the whole session, as long as they would not discuss the samples (Fig. 1, panel B).

Real café - Testing in the real café setting took place in the Grand Café at the Forum building of Wageningen University & Research. The barista and other customers were present during the sessions, causing a buzz of people talking. Also, music was playing in the background. The space was lit with white lighting; however, the lights were in orange lampshades, making the light somewhat yellow. Participants were asked to take a seat, three persons per table. They were allowed to talk during the whole session, as long as they would not discuss the samples (Fig. 1, panel C).

2.4. Procedure

Table 1 shows the study planning with the distribution of contexts over the testing days and weeks. The study took place on three weekdays for three weeks. Each day, three taste sessions of 20 min took place starting at 5 pm, 6 pm and 7 pm. Participants were assigned to a day and timeslot, which were consistent each week per individual to avoid biases due to different satiation states between sessions. Six participants could participate simultaneously in each session, however, due to dropouts and illnesses, most sessions consisted of four to five participants. At the beginning of each taste session, participants were given verbal and written instructions. The soups were served (100 ml) at 70–75 °C in an insulating white foam cup to keep the soup warm together with a white plastic spoon to eat it. Soups were presented one by one together with an evaluation ballot to make sure participants would not re-taste the soups or look at previous scores. Participants were instructed to eat as much of the soup as necessary to evaluate it properly and to take a sip of water between each sample to cleanse their palate. The order in which the four soups were presented was randomized among the participants, however it remained the same for each individual across the three sessions. This was done to avoid changes in hedonic scores and attribute evaluation due to order effects (Mead & Gay, 1995) and to help ensure that the main variable influencing the data was the testing environment. After finishing the last soup, participants completed an engagement questionnaire (Hathaway & Simons, 2017).

2.5. Measurements

Participants first scored the tomato soups on liking using a 9-point hedonic scale ranging from 1 “Dislike extremely” to 9 “Like extremely”. Participants then evaluated the soups on six sensory attributes (Fat, Salt, Sour, Sweet, Thick and Tomato flavor) using a 9-point Just-About-Right (JAR) scale ranging from −4 “Not nearly [attribute] enough” to +4 “Much too [attribute]” and 0 “Just about right” in the middle. After finishing the last soup, panelist engagement was measured with the 19-item engagement questionnaire as used by Hathaway and Simons (2017) which was translated to Dutch by the researchers. This engagement questionnaire is largely based on the one used by Bangcuyo et al. (2015), which was originally derived from the Presence

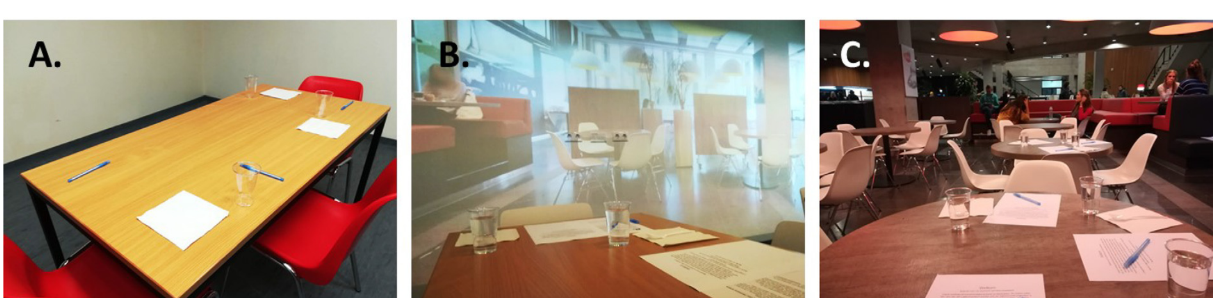

Fig.1. Testing environments from left to right: A) Laboratory setting (no immersion), B) Immersive simulated café, C) Real café.

Table 1
Study planning. Distribution of contexts over the testing days and weeks.

	Week 1	Week 2	Week 3
Day 1	Laboratory	Real café	Immersive simulated café
Day 2	Immersive simulated café	Laboratory	Real café
Day 3	Real café	Immersive simulated café	Laboratory

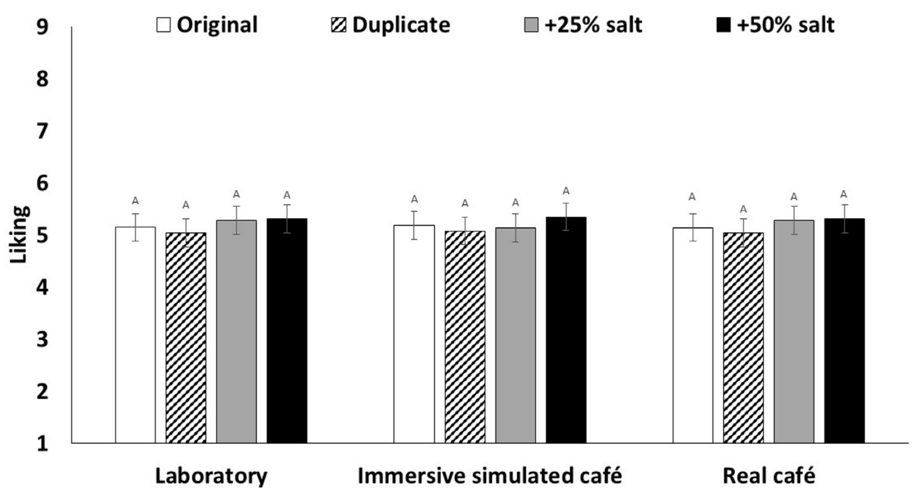

Fig. 2. Mean liking scores ± SE (from 1 'dislike extremely' to 9 'like extremely') for the four soup samples in the three testing environments, *i.e.*, laboratory, immersive simulated café and real café. Mean values for soups with unlike upper-case letters were significantly different from each other ($p < 0.05$).

Questionnaire (Witmer & Singer, 1998) and the User Engagement Scale (O'Brien & Toms, 2010). The 19 items measured the level of agreement of participants to specific statements comprising eight dimensions of engagement. For the statements on the dimensions *Usability*, *Environmental aesthetics*, *Novelty*, *Involvement* and *Immersion*, level of agreement was measured using a 5-point Likert scale ranging from "Strongly disagree" to "Strongly agree". Statements on the dimensions *Sensory awareness*, *Distraction* and *Realism* were measured using a 7-point categorical scale ranging from "None" to "Very".

2.6. Statistical analysis

Statistical analyses were performed using the statistical software package R version 3.5.2 (R Core Team, 2018). A $p < 0.05$ was used as the criterion for statistical significance. Data are reported as means ± standard error. Normality and homogeneity of variance were visually assessed with Q-Q plots and residual plots. Non-independencies were resolved by using a linear mixed model analysis with subject as a random factor.

Liking and JAR ratings – A linear mixed model analysis was performed using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) to assess differences in liking and JAR ratings between the four samples and three contexts. Soup Sample, Context and the interaction Sample × Context were entered as fixed factors in the model and Subject was entered as a random factor. Also, Gender, Day of week and Timeslot were added to the model as fixed factors, to take any possible differences into account. *p*-values were obtained by performing an analysis of variance (ANOVA) on the full model. The model with the effects in question (Sample or Context) was compared to the model without the effect in question. When a significant effect was found, a post hoc Tukey test was performed using the emmeans package (Lenth, 2018) to see which samples and/or contexts differed from each other.

Engagement – Scores per dimension of the engagement questionnaire were calculated following Bangcuyo et al. (2015). The dimensions *Usability*, *Environmental Aesthetics*, *Novelty*, *Involvement* and *Immersion* were coded from −2 to +2 and the dimensions *Sensory Awareness*, *Distraction* and *Realism* from 0 to 6. For each individual, the dimensional scores were summed leading to the Total Engagement Score (TES), with a theoretical range from −15.5 (distracting/not engaging at all) to + 18.5 (extremely engaging). To test whether the dimension scores and the TES differed between the three contexts, a linear mixed model analysis was performed. As for the liking and JAR analysis, Gender, Day of week, Timeslot and Context were entered as fixed factors in the model and Subject was entered as a random factor. *p*-values were obtained by performing an ANOVA on the full model against the model without the effect in question, in this case, the context. When a significant effect was found, a post hoc Tukey test was performed to see which contexts differed from each other.

3. Results

Visual assessment of the Q-Q plots and residual plots indicated that normality and homogeneity of variance assumptions were met. As we observed no statistically significant differences for Gender, Day of week and Timeslot ($p > 0.05$), we report only the means of the experimental groups for all analyses.

3.1. Liking

The four soups were neither liked nor disliked, with means ranging from 5.16 ± 0.25 for the benchmark soup, to 5.05 ± 0.25 for the duplicate, 5.29 ± 0.25 for the soup with 25% extra salt and 5.32 ± 0.25 for the soup with 50% extra salt (see Fig. 2). The main effects of Sample ($p = 0.36$) and Context ($p = 0.95$) and the interaction effect between Sample and Context ($p = 0.93$) were not significant. In other words, liking did not differ among the four soups within each context, nor across the three contexts.

3.2. JAR ratings

The saltiness JAR ratings significantly differed across the soups ($p < 0.001$; Fig. 3). The 50% extra salt soup scored significantly higher

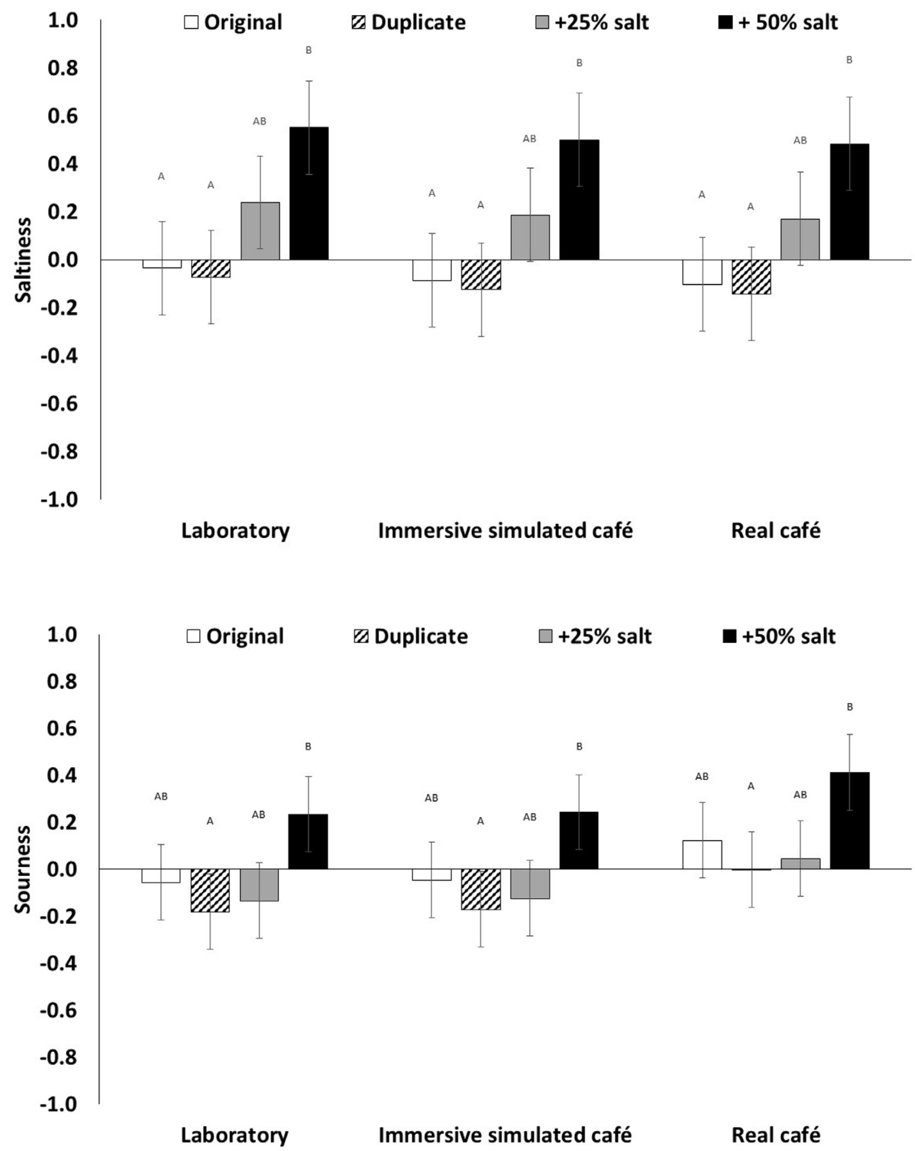

Fig. 3. Mean JAR scores of salt ± SE (upper panel) and sour ± SE (lower panel) for the four soup samples in the three testing environments. JAR scores ranged from −4 ('not nearly [attribute] enough') to + 4 ('much too [attribute]') and 0 ('just about right'). Mean values for soups with unlike upper-case letters were significantly different from each other ($p < 0.05$).

on saltiness than the benchmark and duplicate soup (JAR ratings 0.51 ± 0.19 vs. JAR ratings −0.07 ± 0.19 and −0.11 ± 0.19, respectively), with the 25% extra salt soup in between (JAR rating 0.20 ± 0.19). A non-significant Sample × Context effect ($p = 0.78$) indicated that the saltiness JAR ratings of the soups were consistent across the thee contexts. The sourness JAR ratings follow similar patterns as the saltiness JAR ratings, with significant differences among ideal sour concentrations of the soups ($p < 0.001$; Fig. 3). The 50% extra salt soup scored significantly higher in sourness (JAR rating 0.30 ± 0.16) than the benchmark soup, the duplicate and the 25% extra salt soup (JAR ratings 0.01 ± 0.16, −0.12 ± 0.16 and −0.07 ± 0.16, respectively). A non-significant Sample × Context interaction ($p = 0.75$) indicated that the sourness JAR ratings of the soups did not differ according to the context in which they were evaluated. Results from the linear mixed model analysis showed no significant differences between the four soups for the JAR ratings on the attributes Fat ($p = 0.47$), Sweet ($p = 0.99$), Thick ($p = 0.97$) and Tomato flavor ($p = 0.05$). Also, no significant differences were found among the test settings (laboratory, immersive simulated café, real café) for the JAR ratings on all descriptive sensory attributes (Fat: $p = 0.34$, Salt: $p = 0.79$, Sour: $p = 0.12$, Sweet: $p = 0.41$, Thick: $p = 0.12$, Tomato flavor: $p = 0.80$).

3.3. Engagement

The engagement dimensions significantly differed among the three contexts (see Table 2). The laboratory setting proved better in assisting the subjects with their evaluations (*Usability*) compared to the immersive simulated café and real café setting. Participants found the immersive simulated café and the real café setting to be more appealing (*Environmental Aesthetics*), more involving (*Involvement*) by engaging multiple senses (*Sensory Awareness*) compared to the laboratory setting, although not necessarily better in enhancing their curiosity associated with testing (*Novelty*). The real café setting, unsurprisingly, felt most like a real café (*Immersion*); the laboratory setting scored lowest on this dimension with the immersive simulated café in between. The real café setting was most consistent with participants' real-world experiences of a café (*Realism*). The relatively low mean values of the dimension *Distraction* indicate that the contexts were distracting from the task to a limited extent, albeit with the laboratory setting being slightly more

Table 2
Mean engagement dimension scores (± SE) and Total Engagement Score (± SE) for each testing environment, *i.e.*, laboratory, immersive simulated café and real café. Dimensions *Usability*, *Environmental Aesthetics*, *Novelty*, *Involvement* and *Immersion* have a theoretical range from −2 to +2, whereas *Sensory Awareness*, *Realism* and *Distraction* have a theoretical range from 0 to 6. The *Total Engagement Score* is the sum of the dimensionality scores and can range from −15.5 to 18.5. Dissimilar upper-case letters above the bars indicate significant differences ($p < 0.05$) between the testing environments.

Dimension	Laboratory	Immersive simulated café	Real café
Usability[1]	0.63^{A} ± 0.16	-0.26^{B} ± 0.15	-0.38^{B} ± 0.13
Environmental aesthetics[1]	-0.63^{A} ± 0.17	0.49^{B} ± 0.17	0.57^{B} ± 0.16
Novelty[1]	0.36^{A} ± 0.18	0.01^{B} ± 0.16	0.21^{A} ± 0.13
Involvement[1]	0.29^{A} ± 0.09	0.79^{B} ± 0.09	0.73^{B} ± 0.12
Immersion[1]	-1.77^{A} ± 0.09	-0.28^{B} ± 0.19	0.69^{C} ± 0.16
Sensory awareness[2]	2.21^{A} ± 0.14	3.74^{B} ± 0.17	3.99^{B} ± 0.17
Realism[2]	2.64^{A} ± 0.16	2.92^{A} ± 0.16	3.47^{B} ± 0.13
Distraction[2]	0.42^{A} ± 0.16	-0.47^{B} ± 0.17	-0.87^{B} ± 0.19
Total Engagement Score	-0.74^{A} ± 0.60	3.18^{B} ± 0.66	4.61^{B} ± 0.55

[1] 5-point Likert scale ranging from 1 “Strongly disagree” to 5 “Strongly agree”.
[2] 7-point category scale ranging from 1 “‘None” to 7 “Very”.

distracting (*Distraction*) than the immersive simulated café and real café. Finally, participants felt most engaged in both the real café and immersive simulated café, and least in the laboratory (*Total Engagement Score*).

4. Discussion

This study investigated taste perception and liking of tomato soups in a laboratory context, an immersive simulated café and a real café. To our surprise, we observed no significant differences in liking and JAR ratings for the soups between the three contexts. Nevertheless, participants felt most engaged in the real café and immersive simulated café, and least in the laboratory.

Overall, we can conclude that the more realistic scenarios (*i.e.,* real café and immersive simulated café) increased participants’ engagement compared to the laboratory setting, with the immersive simulated café being a good trade-off between standardization and still enabling a relatively high participants’ engagement –hence encouraging panelist involvement– (Hehn, Lutsch, & Pessel, 2019). This is in line with previous research that also found higher consumer engagement in immersive simulated contexts (Bangcuyo et al., 2015; Hathaway & Simons, 2017; Holthuysen, Vrijhof, de Wijk, & Kremer, 2017; Sinesio et al., 2018). However, counterintuitively, the laboratory setting appeared to be more distracting than the other environments. The reason for this finding is not clear, although it could be that people felt more vulnerable – as in exposed- in the laboratory setting as no context was provided. In spite of the latter, standardized testing situations with induced immersive contexts may bring together the best of both worlds – laboratory (high degree of standardization) and real life (higher consumer engagement).

Despite a higher consumer engagement, the more realistic context scenarios (*i.e.*, real café and immersive simulated café) did not allow for greater hedonic product differentiation as we observed no differences in liking among the three contexts. Previous research showed either a higher hedonic discrimination when incorporating immersive settings in testing (Bangcuyo et al., 2015; Hathaway & Simons, 2017), or no such effect (Delarue et al., 2019; Hannum, Forzley, Popper, & Simons, 2019; Sinesio et al., 2018) as found in the present study. As the use of immersive technologies in product testing is still in its infancy, *i.e.*, only a few products have been tested in limited immersive simulated contexts, we recommend replicating the current findings using a broader range of products and (immersive) contexts to validate this way of contextualization in product testing. It may also be worthwhile for sensory/consumer research communities to agree on ways to compare different types of immersive environments. The types of simulated environment can differ considerably over different research groups. Some use video-audio projections on wall screens (a 180° screen), others use videoclips on computer screens, in combination with other stimuli or not, *e.g.*, odors, sounds, congruent furniture, while we used surround video-audio projection (on all four walls of a room) in the present study. Still the quality (brightness, resolution, etc.) and type of projections in such immersive rooms will differ, which may affect results. This will hold for the visual projection, but it may also apply for the auditory component, and in particular for ambient odors.

Incongruence of the product tasted with the café environment could be another explanation for the homogeneous liking scores obtained in this study. Tomato soup may be a less common product to consume in a café, as opposed to for example a coffee and a muffin, even though tomato soup was catered in the café at the time of the study as well. Cross-modal perception research showed that products are especially liked if they are congruent with the context (Raudenbush, Meyer, Eppich, Corley, & Petterson, 2002; Zandstra & Lion, 2019), *e.g.*, champagne is better liked in a virtual winery setting than in a bar, while beer is better liked in a bar than in a winery. Recent research using immersive technologies to examine the effects of (in-)congruent contextual cues (*i.e.*, visual, auditory and olfactory) on liking, showed a significant decrease in liking for incongruent product-context combinations (Liu, Hannum, & Simons, 2019). However, based on Liu’s research, we would then expect a decreased liking in our study for the café environments, instead of no change in liking. We therefore assume that congruency did not have a large impact on the results of this study.

With respect to the JAR ratings, the addition of contextual effects did not change the ratings in this study. We anticipated a higher product differentiation ability in the laboratory setting than in the immersive simulated café and real-life café, as participants would be more analytical and more focused on the taste of the product in the laboratory setting (*i.e.*, less distracted). However, we found that subjects discriminated only between the sensory attributes *Salt* and *Sour* of the soup samples, and that these perceived differences were similar in each context. Research showed that salt and sour can enhance each other at moderate concentrations (Breslin, 1996), which could explain the higher sourness JAR ratings found for the soup with 50% extra salt. A possible explanation for the lack of finding differences in JAR ratings for saltiness across contexts could be that participants took only a few small sips from the soups in each setting and might not have experienced the soups’ saltiness properly as they would have with consuming a full portion (Zandstra, De Graaf, Van Trijp, & Van Staveren, 1999). These results may have important implications for product development. In case of products with small sensory differences (*e.g.*, when reformulating products to reduced sugar, fat or salt levels), it is important to know whether consumers notice a sensory difference or not. Are laboratory tests predictive of what consumers will notice in real life or do we need to include more contextual cues to improve the validity in sensory and consumer testing in general?

Bisogni et al. (2007) characterized eating episodes using eight dimensions: location, food and drinks, time, activities, social setting, mental processes, physical condition and recurrence. For this study, the dimensions location and social setting were altered between contexts; the other dimensions were the same (*e.g.,* time of day, food, cup & spoon used, etc.). Participants were allowed to talk during soup evaluation in the immersive simulated café and the real café and were asked to be quiet during evaluation in the laboratory setting. Nevertheless, subjects actually were quiet in all three contexts. They did not know each other and were focused on the evaluation task, which could explain the silence. This could have influenced the results, as social interaction is suggested to be an important factor in natural consumption conditions. However, previous research (Bangcuyo et al., 2015;

Hathaway & Simons, 2017) asked subjects to evaluate the samples alone, without social interaction, and still found significant differences in liking between the laboratory and immersive settings. The lack of social interaction may therefore probably not a cause of the unexpected results in this study; however, we recommend future research to further investigate the effect of social interaction on taste perception and liking in immersive settings.

To recapitulate, the current study investigated the effect of contextual information provided by immersive technologies on liking and JAR ratings of soups differing in salt content. The immersive simulated setting was an attempt to get closer to a real-life situation, while allowing a complete control of the test conditions at the same time. The results indicate that we succeeded in creating an immersive and engaging simulated café experience that is similar to the real café experience, however, improvements could be made to may get closer to real life (*e.g.,* by using video footage instead of stills, adding an ambient odor). To our surprise, we observed no differences in overall liking or in JAR ratings of the soups between the three contexts. As the use of immersive technologies in sensory and consumer testing is still in its infancy, further research is needed to define approaches and barriers associated with such testing (*e.g.*, testing other products, different contexts, broader consumer segments). Ideally, this will lead to a standardization of immersive research methodologies for sensory and consumer testing with improved data quality and increased ecological validity.

Acknowledgments

This work was supported by a grant from the Dutch Top Consortium for Knowledge and Innovation (TKI) Agri&Food together with Unilever R&D Vlaardingen, Kikkoman Europe R&D Laboratory and Noldus Information Technology (TKI-AF-17005). We would like to thank Manouk van Zoggel for helping in the data collection, Saskia Meijboom for her help with the recruitment of subjects and João Paulo for her statistical advice.

Conflict of interest

E.H.Z. is an employee of Unilever Innovation Centre Wageningen, The Netherlands, which markets food, home and personal care products. D. Kaneko is an employee of Kikkoman Europe R&D Laboratory Wageningen, The Netherlands, which is a major supplier of naturally brewed soy sauce products.

Author statement

E.H.Z. was responsible for the design and manuscript. E.V. carried out the study and was involved in the data analyses and manuscript writing. R.A.W. contributed to the design, data analyses and manuscript writing. D.K and G.B.D. were involved in the design and earlier drafts of the manuscript.

References

Bangcuyo, R. G., Smith, K. J., Zumach, J. L., Pierce, A. M., & Guttman, G. A. (2015). The use of immersive technologies to improve consumer testing: The role of ecological validity, context and engagement in evaluating coffee. *Food Quality and Preference, 41*, 84–95.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bisogni, C. A., Falk, L. W., Madore, E., Blake, C. E., Jastran, M., Sobal, J., & Devine, C. M. (2007). Dimensions of everyday eating and drinking episodes. *Appetite, 48*(2), 218–231.

Breslin, P. A. S. (1996). Interactions among salty, sour and bitter compounds. *Trends in Food Science and Technology, 7*, 390–398.

Delarue, J., Brasset, A. C., Jarrot, F., & Abiven, F. (2019). Taking control of product testing context thanks to a multi-sensory immersive room. A case study on alcohol-free beer. *Food Quality and Preference, 75*, 78–86.

Galiñanes Plaza, A., Delarue, J., & Saulais, L. (2019). The persuit of ecological validity through contextual methodologies. *Food Quality & Preference, 73*, 226–247.

Hannum, M., Forzley, S., Popper, R., & Simons, C. T. (2019). Does environment matter? Assessments of wine in traditional booths compared to an immersive and actual wine bar. *Food Quality and Preference, 76*, 100–108.

Hathaway, D., & Simons, C. T. (2017). The impact of multiple immersion levels on data quality and panelist engagement for the evaluation of cookies under a preparation-based scenario. *Food Quality and Preference, 57*, 114–125.

Hehn, P., Lutsch, D., & Pessel, F. (2019). Inducing context with immersive technologies in sensory consumer testing. In H. L. Meiselman (Ed.). *Context: The Effects of Environment on Product Design and Evaluation* (pp. 475–500). Woodhead Publishing.

Holthuysen, Nancy T. E., Vrijhof, Milou N., de Wijk, René A., & Kremer, Stefanie (2017). "Welcome on board": Overall liking and just-about-right ratings of airplane meals in three different consumption contexts-laboratory, re-created airplane, and actual airplane. *Journal of Sensory Studies, 32*(2), e12254. https://doi.org/10.1111/joss.12254.

Lenth, R. (2018). Emmeans: Estimated Marginal Means, aka Least-Squares Means. In). https://cran.r-project.org/package=emmeans.

Liu, R., Hannum, M., & Simons, C. T. (2019). Using immersive technologies to explore the effects of congruent and incongruent contextual cues on context recall, product evaluation time, and preference and liking during consumer hedonic testing. *Food Research International, 117*, 19–29.

Mead, R., & Gay, C. (1995). Sequential design of sensory trials. *Food Quality and Preference, 6*(4), 271–280.

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology, 61*, 50–69.

R Core Team (2018). R: A language and environment for statistical computing. In http://www.r-project.org (Ed.), R Foundation for Statistical Computing). http://www.r-project.org: Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, B., Meyer, B., Eppich, W., Corley, N., & Petterson, S. (2002). Ratings of pleasantness and intensity for beverages served in containers congruent and incongruent with expectancy. *Perceptual and Motor Skills, 94*(2), 671–674.

Sinesio, F., Saba, A., Peparaio, M., Saggia Civitelli, E., Paoletti, F., & Moneta, E. (2018). Capturing consumer perception of vegetable freshness in a simulated reallife taste situation. *Food Research International, 105*, 764–771.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence, 7*, 225–240.

Zandstra, E. H., De Graaf, C., Van Trijp, H. C. M., & Van Staveren, W. A. (1999). Laboratory hedonic ratings as predictors of consumption. *Food Quality and Preference, 10*, 411–418.

Zandstra, E. H., & Lion, R. (2019). In-home testing – Bringing lab results to the real world. Chapter 4 In H. L. Meiselman (Ed.). *Context: The Effects of Environment on Production Design and Evaluation*(pp. 67–85). Woodhead Publishing.