**The impact of multiple immersion levels on data quality and panelist engagement for the
evaluation of cookies under a preparation-based scenario**

Running Title: Minimal Immersive Environments

Drew Hathaway¹ and Christopher T. Simons^{1,*}

¹ The Department of Food Science & Technology, The Ohio State University, Columbus, OH
43210-1007

*To whom correspondence should be addressed:

Christopher T. Simons Ph.D.
The Department of Food Science & Technology
The Ohio State University
2015 Fyffe Rd., Columbus, OH 43210-1007
Tel: 614-688-1489
Fax: 614-292-0218
email: simons.103@osu.edu

Keywords: immersive technology, consumer testing, preparation scenario, cookies

Abstract

Prior research indicates using immersive technologies to restore relevant contextual cues during product acceptance testing leads to more discriminating and reliable data likely due to improved ecological validity and heightened panelist engagement. However, as the use of immersive technologies in food and consumer product testing is emerging, the opportunities and limitations associated with such paradigms remain largely unexplored. To gain further insight, panelists assessed cookie liking in a traditional testing environment devoid of contextual information and two immersive environments in which audiovisual and olfactory cues depicting the baking of cookies in a home kitchen were presented. In the mixed immersion condition contextual information was presented to panelists via computer screen, head phones and localized aroma dispersion while evaluating cookies in a sensory booth. In the full immersion condition, the same information was presented via video wall, surround sound ceiling speakers and hidden scent dispersal. Following a three-week hiatus, panelists re-assessed the same cookies in the same three conditions. Despite panelists indicating similar levels of engagement in the two immersive conditions, hedonic data proved to be more discriminating and reliable in the fully immersive environment. These data indicate that scenarios depicting food preparation provide relevant contextual information that can influence product liking. Moreover, as the level of immersion becomes more complete, the discriminability and reliability of consumer acceptance data improves. Whether similar findings are observed with other product-scenario combinations remains to be explored.

Introduction

New product development is critical to the success and growth of most companies manufacturing and selling consumer packaged goods. In the food and beverage industry, prior studies have indicated that 60-80% of new product launches fail (Costa & Jongen 2006; Redmond 1995), costing companies billions of dollars in lost resources and potential revenue. Although multiple factors may contribute to these failed product launches, we recently posited that the poor reliability of traditional consumer testing methodologies may bear some responsibility (Bangcuyo et al., 2015).

In traditional consumer testing scenarios, panelists are asked to evaluate products in isolated booths where relevant non-product contextual information is purposefully minimized. Whereas these environments enable strict control over product testing, they lack ecological validity and dismiss the role of context in shaping product perceptions and acceptance (Bell, Meiselman, Pierson & Reeve 1994; Delarue & Boutrolle 2010; Ferber & Cabanac 1987; Kasof 2002; King et al., 2004; Petit & Siefferman 2007; Stroebele & De Castro 2004; Sommer & Steele 1997; Westerterp-Platenga 1999;). Recently, we utilized immersive technologies to depict a virtual coffeehouse in which important visual, audio and olfactory information was restored during consumer testing (Bangcuyo et al., 2015). We found significant differences in preference order and liking for coffees evaluated by the same people when evaluations occurred in the virtual coffeehouse compared to traditional testing booths. The hedonic data collected in the virtual coffeehouse was also more discriminating and a more reliable predictor of future coffee liking unlike data collected in traditional sensory booths. Furthermore, we found consumers to be more engaged in the testing when evaluating coffees in the virtual coffeehouse, an outcome that likely also contributed to improved data quality.

These results suggest that methodological changes to current testing strategies have the potential to improve the power and reliability of consumer data, providing food and consumer product companies significant savings on product development costs and failed launches. However, as the use of immersive technologies in consumer testing paradigms is new, questions still remain regarding the utility and best practices incorporating these techniques. In particular, it is unclear whether limitations exist as to the product types or scenarios where

immersive technologies prove useful. In our prior study, we depicted a virtual coffeehouse—a scenario that represents a generally positive consumption occasion. However, food liking can also be influenced by contextual information unrelated to consumption, for instance preparation scenarios (Dohle, Rall & Siegrist 2016). The sights, sounds and smells associated with the preparation of foods have been previously shown to influence food wanting and liking (Yeomans 2006; Kantono et al., 2016). In addition, it is presently unclear how immersive an environment must be to realize the heightened test power and reliability. It is possible that data quality can be substantially improved simply by providing relevant contextual information in a traditional sensory environment. Indeed, some studies have indicated that simply providing verbal context in the form of a story can have an impact on liking results (Hein, Hamid, Jaeger & Delahunty 2010; 2012; Jaeger & Meiselman 2004; Kim, Lee & Kim 2016) although the effect is inconsistent (Lusk, Hamid, Delahunty & Jaeger 2015; Jaeger et al., 2017).

To address these questions, we asked consumer panelists to evaluate cookie liking in a traditional testing booth devoid of environmental contextual cues and compared results from those same panelists when evaluating the cookies in an immersive environment depicting a domestic kitchen scenario in which homemade cookies are being prepared by the homeowner. Two levels of immersion were assessed—a fully immersive environment as used previously (Bangcuvo et al., 2015) and a mixed immersion environment in which relevant contextual information was conveyed to panelists via computer screen, headphones, and localized scent dispersion while sitting in a sensory booth. Cookies were evaluated in all three environments during a single session. The ability of consumer data to reliably predict future cookie liking was also measured by having panelists evaluate the same cookies in the same three conditions during a second session of testing that occurred following a three week hiatus. Moreover, in both testing sessions, we assessed the level of panelist engagement in each environment. We hypothesized that improving the ecological validity of consumer hedonic testing by incorporating relevant visual, audio and olfactory cues through the use of immersive technologies would result in a more engaging testing experience and hedonic assessments that are more discriminating and reliable compared to those derived from traditional testing paradigms. We further hypothesized that the different levels of immersion utilized in the full

immersion and mixed immersion environments during testing would differentially impact how much improvement was observed in the sensitivity, power and reliability of hedonic data and the subjective assessment of engagement when compared to results from the traditional sensory testing environment.

Materials & Methods

Subjects

Fifty nine subjects (21 male and 38 female) ranging in age from 18 to 69 years old were recruited for this study using The Ohio State University Sensory Evaluation Center's recruitment database. All 59 subjects reported consuming cookies (of any type) at least once a month and approximately half of them baked cookies (of any type) at least once a month. All participants were enrolled under IRB-approved informed consent (2013B0585). Subjects were asked to refrain from eating, drinking or smoking for at least 2 hours prior to the start of the experiment. Each subject participated in two experimental sessions (referred to as replication 1 and replication 2) approximately 3-weeks apart. For each panelist, the start time of both experimental sessions was consistent and each session lasted approximately 30-min. At the conclusion of their second session, participants received a \$20 cash incentive for their participation.

Stimuli

Four brands of store bought soft chocolate chip cookies were used in this study: Chips Ahoy Chewy Chocolate Chip Cookies (Mondelēz International Group, East Hanover, NJ), Kroger Bakery Chocolate Chip Cookies (Cincinnati, OH), Pepperidge Farm Soft Baked Montauk Chocolate Chip Cookies (Norwalk, CT) and Keebler Soft Batch Chocolate Chip Cookies (Kellogg's, Battle Creek, MI). Cookie samples were purchased regularly at a local grocery store throughout the experiment and were stored in sealed containers at room temperature when not in use. These four stimuli were selected based on the preliminary assessment that they differed enough in quality to obtain discriminating hedonic responses, were similar enough in appearance to prevent panelists from being able to memorize samples, and were all store bought, chocolate chip, soft-style cookies. To further standardize their appearances, a circular

cookie cutter with a 1.5 inch diameter was used on all of the samples prior to their evaluation to give them a uniform size and shape. All sample preparation took place within one day of the testing sessions to prevent samples from becoming stale. Each sample was packaged in a translucent plastic 2 oz. soufflé cup (Georgia-Pacific, Atlanta, GA) and covered with a translucent 2 oz. portion cup lid (Georgia-Pacific, Atlanta, GA) to prevent panelists from being able to see the appearance of the cookie until it was time for them to evaluate it. Each stimulus was assigned a 3-digit blinding code displayed by a sticker on the sample cup's lid and new blinding codes were reassigned to each cookie for each environment in each session. All four cookie samples were presented simultaneously at room temperature. The order the samples were presented in was randomized among the panelists but remained the same for each individual across all three testing environments for both sessions to avoid changes in hedonic scores due to order effects (Mead & Gay 1995) and to help ensure that the main variable influencing the data were the environmental differences.

Procedure

At the beginning of each testing session an instructions sheet was read to the panelists that contained general information regarding proper sample evaluation methods and how to input answers using the tablet device provided. This general instructions sheet was made available in all three of the testing environments so that participants could refer back to it at any point of the testing session. After the general instructions sheet was read, panelists confirmed their understanding of the testing procedures and were directed to their first testing environment. Once each panelist arrived in their first testing environment for their first testing session they began by signing the informed consent form and filling out a brief demographic questionnaire. This was the only time this information was gathered so upon entrance to any subsequent testing environment panelists immediately began product evaluation. Panelists were then instructed to rate the acceptability of each of the four cookie samples using the 9-point hedonic scale. They were instructed to evaluate the samples in a serial monadic fashion, without re-tasting previous samples, taking a drink of water between each sample to cleanse their palate, and making sure to rate the correct cookie by ensuring that the number on the lid matched the number on the screen. While rating the samples, panelists had access to their

previous ratings for that testing environment, but they did not have access to ratings given in previous testing environments or testing sessions.

Following sample evaluation, panelists completed a 19-item Engagement Questionnaire (Supplementary Table 1). The questionnaire was largely based on one used in our previous study (Bangcuvo et al., 2015) which was originally derived based on prior testing instruments developed by O'Brien and Toms (2010) to measure general engagement and by Witmer and Singer (1998) to specifically measure engagement within a virtual environment. The Engagement Questionnaire was composed of 19 questions designed to measure the level of agreement of consumers to specific statements comprising eight dimensions of engagement that included Usability, Environmental Aesthetics, Novelty, Involvement, Sensory Awareness, Immersion, Distraction and Realism. Level of agreement to statements related to Usability, Environmental Aesthetics, Novelty, Involvement, and Immersion were collected using a 5-point Likert scale (coded -2 to +2) whereas responses related to the Sensory Awareness, Realism, and Distraction dimensions were collected using a 7-point categorical scale ranging from 0 (none/not) to 6 (very). For each panelist, all of their dimensional responses were combined linearly to derive a singular measure of engagement referred to as the total engagement score.

In each of the two experimental sessions panelists were asked to complete the same sample evaluation and engagement questionnaire in all three of the environmental settings. All responses were recorded on a tablet device (8" Samsung Galaxy Tab 4, Suwon, South Korea) using Compusense At-Hand software (Ontario, Canada). Three panelists were recruited for each testing session although not every testing session contained three panelists due to scheduling issues. Each panelist started in one of the three testing environments and was then rotated to the other two testing environments as instructed by the two test supervisors upon completion of their sample evaluation/engagement questionnaire. Panelists were typically able to be moved to their next testing condition within 1-3 mins of completing the test. Which environment the panelist started in and their order of rotation to the other two testing environments was randomized and counterbalanced across panelists and across the two different sessions. Before panelists entered their first testing environment, their cupped cookie samples, the computer tablet, a filled water cup and the general instruction sheet were placed

on the counter top. When rotating between environments panelists brought their water cup and assigned tablet device with them to the next environment. The test supervisors ensured that the panelist's next samples were ready upon entrance to their next testing environment so that they could immediately begin their evaluation. The only procedural difference between the three testing environments was that panelists were instructed to start the video component in the mixed immersion and full immersion environments before beginning their sample evaluation/engagement questionnaire. For all three testing environments, samples were evaluated under white light at an ambient temperature. Subjects returned after approximately three weeks and completed their second session of testing in an identical manner.

Testing environment - Traditional

Testing in the traditional environment took place in one of two private sensory booths located in a laboratory room adjacent to the Immersive Technologies Laboratory at The Ohio State University. The sensory booth consisted of a 5'x7' room with a counter top containing a spit sink and computer monitor with the only door leading to a small waiting area between the two sensory booths. A picture of the traditional testing environment can be seen in Figure 1A. Upon entrance panelists were seated at the counter, reminded of the instructions, and then began their sample evaluation and engagement questionnaire. The door to the booth was closed during their evaluation and only opened upon their completion.

Testing environment - Mixed immersion

Testing in the mixed immersion environment took place in the private sensory booth adjacent to the traditional sensory booth. The traditional and mixed immersion testing environments were separated by a 14'x8' waiting area through which both booth entrances were located. The mixed immersion sensory booth was physically identical to that of the traditional testing booth. The only difference between the two testing environments was the addition of visual, audio, and olfactory context cues in the mixed immersion environment. A picture of the mixed immersion testing environment can be seen in Figure 1B. The visual and audio context cues were provided by video footage of a man in the act of making cookies in a domestic kitchen setting. The video begins with the man combining common cookie ingredients in a tabletop mixer. As the video continues an oven timer goes off indicating a

previous batch of cookies were finished so he pauses mixing to remove a tray of cookies from the nearby oven and places it on a cooling rack. The video ends with the man placing the newly created tray of cookies in the oven. The video was displayed to the panelist by a LG Flatron E2211 22" computer monitor (Seoul, South Korea) situated on the counter directly in front of the panelist. The audio from the video was relayed to the panelist using headphones (Sennheiser HD 280 Professional headphones, Wedemark, Germany). The video was previously recorded using a Nikon D5100 video camera (Tokyo, Japan) in 1080p HD at 30 frames per second with simultaneous recording of audio using a Tascam TM-2X stereo directional microphone (TEAC, Tokyo, Japan). The final video was edited to a length of 14.5 minutes to ensure that it would not end before panelists finished their evaluation. In addition to the audiovisual stimulus a cookie aroma (flavor code XL-879-653-9, Givaudan Flavors, Cincinnati, OH) was dispersed into the testing environment by bubbling air through the neat flavor at a constant rate (0.25 LPM) to fill the room with a subtle but noticeable room aroma before panelists' entry. Although unlikely, this aroma could potentially serve as an interfering variable. However, we believe it is an important and salient cue when baking homemade cookies and contributes to the improved ecological validity of this scenario. Since the traditional and the mixed immersion testing environments shared a small waiting area outside of their entrances, an air purifier (Austin Air HealthMate HM400, Buffalo, NY) was placed directly outside of the two testing environments and was constantly running during testing to ensure that the aroma being dispersed in the mixed immersion booth did not enter the waiting area or the traditional testing environment. Upon entrance panelists were seated at the counter, instructed to put on the headphones and start the video, and then begin their sample evaluation and engagement questionnaire. The door to the booth was closed during their evaluation and only opened upon completion to minimize possible odor contamination of the other two rooms. A testing supervisor was always present in the waiting space outside of the traditional and mixed immersion testing environment to answer any questions that panelists had, swap out trays of cookie samples between panelists, and to coordinate the proper rotation of panelists.

Testing environment - Full immersion

Testing in the full immersion environment took place in the Immersive Technologies Laboratory at The Ohio State University which is located directly adjacent to the room containing the traditional and mixed immersion environments. The Immersive Technologies Laboratory consists of a testing area containing a video wall composed of nine 48-inch high-definition LCD screens (model 47VS20-BAA, LG Electronics, Seoul, South Korea), surround sound tile speakers (Quam-Nichols, Chicago, IL) mounted in the ceiling and a tall chair and table placed directly in front of the video wall (see Figure 1C). The same video used in the mixed immersion testing environment was displayed on the video wall with the corresponding audio being played through the speaker system while the panelists completed their evaluation as in the mixed immersion environment. The same cookie aroma used in the mixed immersion testing environment was also dispersed into the full immersion testing environment in an identical fashion. Upon entrance panelists were seated at the table in front of the video wall and the testing supervisor started the video prior to panelists' beginning their sample evaluation and engagement questionnaire.

Data Analysis

All data analyses were performed using SPSS 22 statistical software (IBM, Armonk, NY). Data transformations and calculation of the dimensional engagement scores, total engagement score, and change in mean liking scores between replicates for each cookie in each testing environment were accomplished using Excel (Microsoft Corp., Redmond, WA). All data are reported as means with standard errors.

Liking

In each replicate, differences in cookie liking under each condition (testing environment) were assessed using a 3-way, repeated measures Analysis of Variance (ANOVA) with subject, cookie, and testing condition as main effects. With a significant condition*cookie interaction, we subsequently analyzed data from each condition separately using 2-way repeated measures ANOVA with post-hoc Tukey's tests; subject and cookie were evaluated as main effects under both analyses. For completeness, data from each replicate were combined and analyzed as above using 3-way, repeated measures ANOVA with subject, cookie, and testing condition as main effects.

Preference

To assess preference, each subject's liking scores were converted to rankings. In the case of ties, cookies were given the mean ranking. Converted data for each condition were analyzed using Friedman's test with Wilcoxon signed-ranks test used post hoc to identify significant differences between cookie rankings.

Reliability

The ability of cookie liking scores obtained in each testing condition to reliably predict cookie liking following a ca. 3-week hiatus was assessed based on the change in liking scores between replicates for each cookie in each condition. To do so we calculated the difference in each individual's liking score between the two testing sessions for each cookie under each condition and then took the average across all participants. In addition, we combined the average change in mean liking scores across all cookies to further illustrate the differences in reliability between conditions. Further analyses on these data were conducted using a 3-way, repeated measures ANOVA (subject, cookie, and testing condition as main effects), with Tukey's post-hoc analysis to identify the specific differences in reliability found between conditions. Finally, data unreliability may be considered most drastic when, for a given testing condition, a cookie liked (hedonic rating 6-9) in one session is disliked (hedonic rating 1-4) in another. Thus, the proportions of "flipped" responses obtained in each condition were calculated and compared using a related samples proportions test.

Engagement

Likert data from the engagement questionnaire were coded -2 to 2 whereas categorical responses were coded 0 to 6. Coded data from relevant questions were averaged across panelists to generate dimensional scores for each testing condition. Dimensional scores were then linearly summed to generate the Total Engagement Score (TES) that represents a composite level of engagement. To determine whether dimensionality scores or the TES differed according to the testing environment, data from each dimension were subjected to 2-way ANOVA with subject and testing condition as main effects. To determine whether the engagement disparity between the three testing environments was stable over time, a 3-way repeated measures ANOVA (subject, condition and replicate as main effects) was used to assess

the average of the TES across sessions for each condition in addition to the condition*replicate interaction term. When main effects were found to be significant, Tukey's post-hoc analysis was utilized to identify the specific differences between conditions.

Results

Panelists' cookie consumption/preparation habits.

At the beginning of each panelist's first testing session they reported their average cookie consumption and preparation habits as part of the demographic questionnaire. All subjects reported to consume cookies at least once a month with 44% eating cookies an average of one to two times a month, 37% eating cookies once to twice a week, 17% eating cookies three to four times a week and one panelist reporting to eat them more than five times a week. All panelists reported baking cookies at least once in the previous year with 53% reporting to bake cookies less than once per month, 44% baking cookies one to two times a month and 3% reporting to bake cookies one to two times a week on average.

Liking and Preference

The use of immersive technologies to provide relevant contextual information during product evaluation affected the resulting consumer hedonic responses to the cookies. Results from the first session of testing for all panelists revealed a significant immersion level main effect ($F_{2,312}=6.535$; $p=0.002$) indicating that the average combined liking score for all cookies differed according to the environment in which they were evaluated. Post-hoc Tukey's test showed that cookies received significantly higher hedonic ratings when evaluated in the full immersion condition compared to the traditional testing environment with average liking scores obtained from the mixed immersion condition falling between the other two (Figure 2A; left hand panel). A non-significant cookie*immersion level interaction ($F_{6,312}=0.326$; $p=0.923$) indicated this pattern of rating was consistent across all cookies (Table 1 and Figure 2B; left hand panel). This finding may reflect a specialized case of evaluative conditioning whereby the relatively pleasant experience of testing in contextually relevant immersive conditions has the general effect of increasing liking for all the cookies (De Hower et al., 2001). When panelists' liking data were converted to ranks and analyzed separately for each condition, differences in preference order were observed (Figure 2C, left hand panel). When evaluated in the traditional

environment, the rank order of cookie preference was Keebler>Pepperidge Farm>Kroger>Chips Ahoy although no significant preference differences were observed ($S=4.67$, $p=0.198$). However, in the mixed and full immersion conditions, the rank order of cookie preference changed with Pepperidge Farm being the most preferred followed by Keebler, Kroger and lastly Chips Ahoy. In both cases, the Chips Ahoy cookie was preferred significantly less than the other products (Mixed Immersion: $S=9.84$, $p=0.020$; Full Immersion: $S=10.65$, $p=0.014$). Note in some cases the preference order obtained from the ratings data did not match what was expected from mean liking scores. For instance, according to Table 1, in the mixed immersion condition, the Keebler cookies would be expected to be most preferred due to the slightly higher mean hedonic rating. However, as the ranking data do not account for the magnitude of liking, the Pepperidge Farm cookies are slightly more preferred due to the greater number of times this product received a higher ranking.

A somewhat similar overall pattern for liking was seen during the second testing session when the same panelists evaluated the same four cookie samples following a ca. 3-week hiatus. Again a significant immersion level main effect was found ($F_{2,312}=5.837$; $p=0.003$) indicating that the average combined liking score for all cookies differed according to the environment in which they were evaluated. However, for this session, the post-hoc Tukey's test showed that both the full immersion and mixed immersion environments generated average liking scores that were significantly greater than those obtained in the traditional environment ($p=0.05$ and $p=0.003$, respectively; Figure 2A, middle panel). The cookie by immersion level interaction was not significant ($F_{6,312}=0.813$, $p=0.561$) again indicating a consistent pattern of liking across all immersion levels (Table 1 and Figure 2B, middle panel). When liking data obtained in the second testing session were converted to ranks, differences in preference order were once again observed between testing environments (Figure 2C, middle panel). In the traditional environment, cookie preference order was Pepperidge Farm>Keebler>Kroger>Chips Ahoy with Pepperidge Farms cookies being significantly preferred over Kroger brand and Chips Ahoy while Keebler cookies were significantly preferred over Chips Ahoy ($S=11.23$, $p=0.011$). In the full immersion condition, the same preference order was observed as in the traditional environment with the Chips Ahoy cookie being less preferred than all other brands ($S=10.74$;

p=0.013). However, in the mixed immersion environment, a different preference order was observed as the Keebler cookies were most preferred followed by Pepperidge Farm, Kroger and Chips Ahoy cookies. The Chips Ahoy cookies were significantly less preferred than the Keebler and Pepperidge farm cookies but not the Kroger cookies ($S=10.93$, $p=0.012$).

When data were combined across replicates, unsurprisingly there was a main effect of immersion ($F_{2,948}=6.06$, $p=0.002$) as liking scores were significantly higher in the mixed and full immersion conditions compared to the traditional condition (Figure 2A, right hand panel). This pattern of liking was consistent for all cookies (cookie X immersion interaction term $F_{6,948}=0.308$, $p=0.933$; Table 1 and Figure 2B, right hand panel). As expected preference orders also differed between testing environments (Figure 2C, right hand panel).

Power

The ability to resolve differences in cookie liking or preference was modestly improved by the inclusion of relevant contextual information in a fully immersive environment. In the first replicate, significant differences in liking were observed between the cookies in all three environments, however, the test power was slightly greater in the full immersion condition ($F_{3,156}=4.261$, $p=0.006$) where 3 significant pairwise product differences were identified using Tukey's post-hoc test (Figure 2B; left hand panel). In contrast, the power to resolve differences in cookie liking was less successful in the minimal immersion ($F_{3,156}=3.630$, $p=0.014$) and traditional environments ($F_{3,156}=3.941$, $p=0.010$), where only 1 or 2 pairwise product differences were identified in each case, respectively (Figure 2B; left hand panel). Product ranks derived from the hedonic scores reflect this trend as well. Whereas no significant cookie preferences were observed when evaluations were completed in traditional testing booths ($S=4.67$, $p=0.198$; Figure 2C, left hand panel), 3 significant pairwise comparisons were obtained when cookies were evaluated in the mixed or full immersion environments ($S=9.84$, $p=0.020$ and $S=10.65$, $p=0.014$, respectively; Figure 2C, left hand panel). These results suggest that inclusion of relevant contextual information enhances the power of hedonic testing albeit in the present study, the effect was relatively modest.

During the second testing session, significant differences in cookie liking were not seen in any of the testing conditions (traditional $F_{3,156}=0.891$, $p=0.447$; minimal immersion

$F_{3,156}=2.33$; $p=0.076$; full immersion $F_{3,156}=1.71$, $p=0.167$). Significant differences in cookie preferences however, were observed in each of the three testing conditions. Specifically, 3 significant pairwise comparisons were identified when cookies were evaluated in the traditional environment ($S=11.23$, $p=0.011$; Figure 2C, right hand panel). In the mixed immersion condition, only 2 significant product differences were identified ($S=10.93$, $p=0.012$; Figure 2C, right hand panel) whereas in the full immersion condition, 3 significant product differences were found ($S=10.741$; $p=0.013$; Figure 2C, right hand panel). Thus, during the second replicate of testing, increasing levels of immersion did not result in increased test power.

Reliability

To determine the reliability of hedonic data, the consumers re-evaluated the same cookies in all three of the testing environments following a ca. 3-week hiatus. In general, reliability between the two sessions was relatively high for all three of the testing environments. Although during the second replicate of testing, no significant liking differences were observed between cookies, the hedonic scores attributed to the products remained reasonably stable for all conditions. This high degree of consistency across replications may have reflected obvious quality differences between the cookies used. However, the full immersion condition appeared to generate more reliable data as the average change in cookie liking between replicate 1 and 2 was closer to zero (indicating no change in liking) in this condition compared to the mixed immersion or traditional environments (Figure 3). When change in mean liking scores was averaged across all cookies, we found the change to be marginally significantly smaller in the full immersion environment compared to the mixed immersion condition ($F_{2,312}=2.460$, $p=0.087$; Figure 4). Similarly, although there were more instances of flipped hedonic responses (liking to disliking or vice versa) between sessions in the mixed (20.2%) immersion environment compared to the traditional (18.3%) or full immersion (14.6%) conditions, these differences were only marginally significant ($p=0.083$).

The stability of cookie preference order was also assessed across the two replicates. Whereas the preference order was consistent when the cookies were evaluated in the fully immersive environment, this was not the case when the products were evaluated in the mixed immersion or traditional environments (Figure 2C). Indeed, although the two least preferred

products were stable in all conditions, the most preferred product differed from replication 1 to replication 2 in the mixed immersion and traditional testing conditions.

Engagement

The use of immersive technologies to provide relevant contextual information during product evaluation had a significant effect on panelist engagement. A significant difference was found between testing environments for every dimension during at least one of the two testing sessions (see Table 1 and Figure 5A and B). In particular, the presence of relevant contextual information during product evaluation in the two immersive environments generally had a significantly positive impact on engagement data when compared to the traditional testing environment. During both testing sessions, panelists found the two immersive environments to be significantly more appealing (Environmental Aesthetics) by engaging multiple senses (Sensory Awareness) to increase enthusiasm for the test (Involvement) and provide a testing environment that panelists felt was more ecologically valid than the traditional environment (Immersion).

Three of the dimensions (Usability, Novelty, and Realism) did not produce any significant differences between environments during the first testing session but did so for the second session (see Table 2). Specifically, during the second session, panelists felt both of the immersive environments to be more consistent with their real-world experiences associated with cookie preparation and consumption compared to the traditional environment (Realism) and that the full immersion environment significantly assisted in their evaluations of the cookies (Usability) and enhanced curiosity associated with testing (Novelty) compared to the traditional environment. Interestingly, during the first session panelists found the immersive environments to be significantly more distracting compared to the traditional environment (Distraction; $F_{2,104}=5.23$, $p=0.007$) but only found them to be marginally more distracting during the second testing session (Distraction; $F_{2,104}=2.79$, $p=0.066$). This may reflect the fact that panelists were used to the potential distractions during their second testing session. The only dimension that was significantly different between all three of the testing environments was Immersion (see Table 2) suggesting that the level of immersion used might have produced varying degrees of ecological validity.

For each panelist, dimensional responses were combined to derive the TES. When averaged across both testing sessions, significant differences ($F_{2,260}=94.75$, $p<0.001$) in the TES were observed between each testing environment. The largest average score was observed in the full immersion environment (6.44 ± 0.61), followed closely by the mixed immersion environment (5.09 ± 0.59), and a negative average score for the traditional testing environment (-1.07 ± 0.68). This result indicates that both of the immersive environments were significantly more engaging on average than the traditional environment with the full immersion environment slightly more engaging than the mixed immersion environment. Similar findings were observed when TES data from each session were analyzed separately (Session 1: $F_{2,159}=38.00$, $p<0.001$; Session 2: $F_{2,104}=62.91$, $p<0.001$), however the difference between the mixed immersion and full immersion conditions did not reach significance. The TES data produced a marginally significant testing environment*testing session interaction ($F_{2,260}=2.79$, $p=0.064$) which would normally suggest that the level of engagement was not consistent between testing sessions for the three environments, but further investigation showed this effect was a result of a significant decrease in the TES for the traditional testing environment from the first testing session to the second while the TES for the mixed and full immersion environments showed a slightly positive but not significant increase from the first session to the second.

Discussion

The present study investigated the effect of utilizing immersive technologies to provide relevant contextual information within a sensory testing environment on the resulting discriminability and reliability of hedonic data and panelist engagement. This was accomplished by comparing results in which subjects evaluated the same cookie samples in a traditional sensory testing booth, an identical sensory booth complete with visual, auditory, and olfactory cues associated with cookie preparation, and a full immersion testing room complete with the same three contextual cues on two separate occasions. We found the presence of contextual information and the degree of immersion employed to affect the panelists' evaluations of the cookie samples and their subjective assessment of engagement. Cookies received significantly higher hedonic ratings in the immersive conditions compared to the traditional environment

(Figure 2A), and hedonic or preference data collected in the fully immersive environment were found to be slightly more discriminating than data collected in the traditional and/or mixed immersion testing environments, at least for one of the two testing sessions (Figure 2B and 2C). Similarly, data collected in all three testing conditions were relatively reliable in that hedonic scores obtained in session 1 and 2 were similar. However, scores obtained in the fully immersive condition tended to be more consistent compared to those obtained in the mixed immersion condition (Figure 4). In the case of product preference, rankings were consistent only when obtained in the full immersion environment (Figure 2C). Prior studies have indicated that with repeated testing, product liking can change over time as liked products become disliked and vice versa (Koster et al., 2003; Dalenberg et al., 2014). Our results are inconsistent with these findings as only a small percentage of responses flipped in any of the conditions.

The two immersive environments also proved to be more engaging to consumer panelists (Figure 5) which may have contributed to the higher average hedonic ratings given in these environments versus the traditional testing condition. In addition, the two immersive environments remained equally engaging during both sessions whereas the traditional testing environment was found to be less engaging during the second session. This finding suggests that the impact of immersive environments, as designed presently, was enduring, at least over the approximately 3-week test period evaluated here. Interestingly, although more engaging, the mixed and full immersion conditions were also found to be more distracting than the traditional environment, at least in the first testing session. In our prior study, the virtual coffee house was found to be less distracting than the traditional testing environment (Bangcuyo et al., 2015). Recognizing and minimizing potential distractions associated with immersive testing is needed to fully optimize the utility of these methods.

Immersion

In the Food and Beverage Industry approximately 60-80% of new food product launches fail to survive in the marketplace (Costa & Jongen 2006; Redmond 1995). While multiple causes contribute to this high failure rate, the inability of current consumer testing strategies to reliably measure product acceptance and predict resulting consumer behavior is believed to be a major contributing factor (Bangcuyo et al., 2015; Deliza & MacFie 1996; Koster & Mojet 2007;

Rosas-Nexticapa et al., 2005; Weiss et al., 2010). Despite numerous studies having found that extrinsic factors can change a consumers' resulting hedonic assessment of a product (Bell, Meiselman, Pierson & Reeve 1994; Delarue & Boutrolle 2010; Ferber & Cabanac 1987; Zellner, Stewart, Rozin & Brown 1988), commonly used consumer sensory testing practices place panelists in environments intentionally designed to lack contextual information. A consequence of these sterile testing conditions and monotonous testing procedures is a general lack of consumer engagement during testing (Bangcuyo et al., 2015). This lack of engagement can result in uninformative hedonic data where panelists provide the same or similar ratings to all of the samples (Koster 2009). Therefore, we have proposed the lack of ecological validity and consumer engagement associated with typical highly controlled sensory testing methodologies underpins the inaccuracy and unreliability of data collected using this approach (Bangcuyo et al., 2015).

Since virtual environments, by their very nature, are designed to be engaging and capable of providing extrinsic contextual information, they have the potential to be an ideal alternative to traditional sensory testing environments. In our prior study, the inclusion of relevant contextual cues depicting a coffeehouse within a sensory testing environment had a positive effect on the resulting hedonic data quality and consumers' engagement level when compared to results obtained in a traditional sensory testing environment devoid of context (Bangcuyo et al., 2015). These results, along with the present data, suggest that incorporating immersive technologies in the field of consumer sensory testing improves the ecological validity of testing conditions by restoring relevant contextual information resulting in data that can be more discriminating and reliable when compared to results obtained in a traditional sensory testing environment.

One of the main objectives of the present study was to determine if implementation of the same immersive elements could be utilized in a traditional sensory testing environment (i.e. the mixed immersion environment) to positively impact the quality of hedonic data and panelist engagement. In so doing, we wanted to determine the minimum immersive requirements needed for delivering impactful contextual cues. Such information is insightful because the effort and cost required for adoption by industrial sensory professionals would be minimized if

fully immersive installations were not required. In the present experiment, the two immersive environments were found to be significantly more engaging than the traditional environment with the mixed immersion environment only slightly less engaging than the full immersion environment (Figure 5B). We believe this suggests that the presence of context cues may be more important than the scale of them as it relates to improving panelist engagement during consumer testing. However, despite this heightened level of panelist engagement, the mixed immersion condition failed to show the same increase in test power and reliability compared to the fully immersive condition. These data suggest that for a preparation scenario depicting the baking of cookies, minimally immersive environments do not restore ecological validity to a level observed with a more complete immersive environment. Whether this finding is common to all testing scenarios or specific to the virtual baking of cookies as tested presently is unknown.

The choice of environment depicted in the immersive conditions is likely to have an influence on the results. Home kitchens reflect personal and/or private space and, in the present study, the kitchen presented in the video was not recorded from any participating subject's home. As such, the degree of contextual relevancy associated with the testing environment may not have been as fitting as if we had recorded from each subject's personal environment. Although improvements to test power, reliability and panelist engagement were identified presently, larger effects might be possible if the virtual environment is more personally meaningful to each panelist. Although logistic limitations prevent this as a viable practice, interesting questions remain regarding the saliency of the virtual environments. For instance, how close to a subject's actual experience must the contextual environment be to elicit favorable outcomes? Similarly, the attractiveness of the environment and/or persons within the video may also differentially impact the results. Prior studies have shown that environmental cleanliness (Rozin, Fallon & Mandell 1984), ambiance or atmosphere (Delarue & Boutrolle, 2010; Stroebele & De Castro, 2004), and server attractiveness (Lynn & McCall 2016) can impact food liking and meal enjoyment. These questions remain the focus of future studies.

Preparation Scenario

A retail based scenario, such as a coffee house or restaurant, represents only one of the many ways that food products are regularly purchased and consumed. Indeed, many food products purchased at grocery stores or other outlets typically require some sort of preparation prior to consumption. Therefore the act of preparation likely plays a role in the expectation and resulting opinions about a food product before it is able to be consumed (Yeomans 2006; Rozin & Tuorila 1993; Dohle, Rall & Siegrist 2016). Presently, we sought to determine whether the inclusion of context cues relevant to the preparation of cookies in a home kitchen setting would positively impact the quality of hedonic data and level of panelist engagement as was seen using consumption-based cues (Bangcuyo et al., 2015). Our results showed this to be true, although the effects are rather modest, and provides further support to the theory that sensory cues experienced during the preparation of a food can modulate product liking (Yeomans 2006; Rozin & Tuorila 1993; Dohle, Rall & Siegrist 2016). Therefore, implementation of this consumer testing methodology, or a similarly modeled one, might also be beneficial to the large number of companies who produce food products meant to be prepared and consumed at home. However, since this category represents a very diverse and wide variety of food products, further research would be necessary to confirm the effectiveness of this approach in other foods or beverages and ensure the costs of implementation are offset by more actionable product guidance. It is possible that preparation scenarios are simply less relevant to consumer testing compared to consumption scenarios as tested previously with coffees and alcohol (Sester et al., 2013; Velasco, Jones, King & Spence 2013; Bangcuyo et al., 2015). As such, it may be speculated that restoring contextual information in these types of situations is less impactful thus leading to relatively smaller effect sizes than observed in our coffeehouse study (Bangcuyo et al., 2015). Further studies are needed to identify the limits associated with using minimally immersive environments in consumer testing.

Perspectives

The design used in this study represents one of many ways that immersive technology could be incorporated into the field of consumer sensory testing to provide relevant context cues to subjects during product evaluation. Due to the very limited amount of previous research exploring this concept (Sester et al., 2013; Ischer et al., 2014; Bangcuyo et al., 2015;

Kim, Lee & Kim 2016) the present study was intended to further understand the potential benefits and shortcomings associated with immersive testing methodologies.

One potential limitation in the present study was the choice to have panelists evaluate store bought products. The video shown in the two immersive environments displayed cookies being made from scratch and freshly taken out of the oven which might create an expectation of freshly made warm cookies. The four cookies that panelists actually evaluated were all store bought products served at room temperature. Whereas use of these cookies reduced the within sample variability associated with baking cookies ourselves, the products used presently might not have met the expectation created by the video shown. Serving temperature has previously been shown to influence expectations and liking associated with beverages (Zellner et al., 1988). Additionally, quality differences between the cookie samples may have been obvious, making it relatively easy for consumers to be consistent in their preferences. The combination of these two factors might have confounded the quality of the resulting hedonic data potentially making it less effective at illustrating potential differences between the three testing environments. Future studies on the topic should attempt to select products for evaluation that are as close as possible to the expectation being created by the context cues employed while still maintaining logistical testing feasibility.

Another potential limitation of this study's design was the means by which the video was utilized in the two immersive environments. One of the objectives of this study was to evaluate if the inclusion of context cues during consumer acceptance testing would still be effective for food products that are prepared and consumed at home as opposed to those purchased and consumed at a retail location. Therefore the scenario revolved around a preparation based food product, in this case cookies, which was represented in the video showing someone making cookies from scratch. However, we had panelists watch the video while they consumed and evaluated the cookie samples which meant that the task they were performing was consumption based while the video and audio context cues shown to them were preparation based. While this method still had some congruency since both factors revolved around cookies it might have been more effective to have had the panelists watch the entire preparation based video before beginning their evaluation of the cookies. Such a design

could have ended with the man pulling out a finished batch of fresh cookies right as the panelists received their cookie samples in order to maximize congruency between the two events. Ideally this approach could increase the ecological validity of the testing methodology and resulting data. Future studies are needed to fully evaluate the effect of such testing designs.

Ultimately the approach of adding relevant context cues to previously existing sensory testing booths could make adoption by the food industry much easier, so future research should further investigate this approach. Furthermore, the use of immersive technologies in consumer testing of food products is novel and to date, there has been no previously published research delineating the approaches or obstacles associated with the implementation of such testing within the food industry. Therefore significant research is still needed to identify the parameters necessary for implementing immersive technologies in consumer sensory testing and to develop new methodologies that produce more reliable and discriminating hedonic data compared to traditional sensory testing paradigms.

Acknowledgments

The authors thank Michael Arato, Ron Bangcuyo, Blansh del Portal, Gretchen Guttman, Rob Kaufman, Brianne Linne, Rebecca Liu and Alex Pierce for technical assistance and Jim and Patti Hathaway for allowing us to film in their kitchen. This study served in partial fulfillment of the thesis requirement for DH. Research support provided by the USDA National Institute of Food and Agriculture, Hatch project [233590], state and federal funds appropriated to The Ohio State University, Ohio Agricultural Research and Development Center, and by a SEEDS grant from the Center for Advanced Processing and Packaging Studies.

References

- Bangcuyo, R.G., Smith, K. J., Zumach, J. L., Pierce, A. M., Guttman, G. A., & Simons, C. T. (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.
- Bell, R., Meiselman, H. L., Pierson, B. J., & Reeve, W. G. (1994). Effects of adding an Italian theme to a restaurant on the perceived ethnicity, acceptability, and selection of foods. *Appetite*, 22, 11–24.

Costa, A. I., & Jongen, W. M. (2006). New insights into consumer-led food product development. *Trends in Food Science & Technology*, 17, 457–465.

Dalenberg, J. R., Nanetti, L., Renken, R. J., de Wijk, R. A., & ter Horst, G. J. (2014) Dealing with Consumer Differences in Liking during Repeated Exposure to Food; Typical Dynamics in Rating Behavior. *PLoS ONE* 9(3): e93350. doi:10.1371/journal.pone.0093350

De Houwer, J., Thomas, S., & Baeyens, F. (2001) Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, 127(6), 853-869.

Delarue, J., & Boutrolle, I. (2010). The effects of context on liking: Implications for hedonic measurements in new product development. In S. R. Jaeger & H. MacFie (Eds.), *Consumer-driven innovation in food and personal care products* (pp. 175–218). Cambridge: Woodhead Publishing Ltd.

Deliza, R., & MacFie, H. J. H. (1996). The generation of sensory expectation by external cues and its effect on sensory perception and hedonic ratings: A review. *Journal of Sensory Studies*, 11, 103–128.

Dohle, S., Rall, S., & Siegrist, M. (2016). Does self-prepared food taste better? Effects of food preparation on liking. *Health Psychology*, 35(5), 500-508.

Ferber, C., & Cabanac, M. (1987). Influence of noise on gustatory affective ratings and reference for sweet or salt. *Appetite*, 8, 229–235.

Hein, K. A., Hamid, N., Jaeger, S. R., & Delahunty, C. M. (2010). Application of a written scenario to evoke a consumption context in a laboratory setting: Effects on hedonic ratings. *Food Quality and Preference*, 21, 410–416.

Hein, K. A., Hamid, N., Jaeger, S. R., & Delahunty, C. M. (2012). Effects of evoked consumption contexts on hedonic ratings: A case study with two fruit beverages. *Food Quality and Preference*, 26, 35–44.

Ischer, M., Baron, N., Mermoud, C., Cayeux, I., Porcherot, C., Sander, D., and Delplanque, S. (2014). How incorporation of scents could enhance immersive virtual experiences. *Frontiers in Psychology* 5, e736.

Jaeger, S. R., Fiszman, S., Reis, F., Chheang, S. L., Kam, K., Pineau, B., Deliza, R., and Ares, G. (2017). Influence of evoked contexts on hedonic product discrimination and sensory characterizations using CATA questions. *Food Quality and Preference*, 56, 138-148.

Kantono, K., Hamid, N., Shepherd, D., Lin, Y. H. T., Yakuncheva, S., Yoo, M. J., Grazioli, G. and Carr, B. T. (2016). The influence of auditory and visual stimuli on the pleasantness of chocolate gelati. *Food Quality and Preference*, 53:9-18.

689 Kasof, J. (2002). Indoor lighting preferences and bulimic behavior. An individual differences
690 approach. *Personality and Individual Differences*, 32, 383–400.

691 Kim S.-E., Lee S.M., & Kim K.-O. (2016). Consumer acceptability of coffee as affected by
692 situational conditions and involvement. *Food Quality and Preference*, 52(1), 124-132.

693 King, S. C., Weber, A. J., Meiselman, H. L., & Lv, N. (2004). The effect of meal situation, social
694 interaction, physical environment and choice on food acceptability. *Food Quality and*
695 *Preference*, 15, 645–653.

696 Koster, E. P. (2009). Diversity in the determinants of food choice: A psychological perspective.
697 *Food Quality and Preference*, 20, 70–82.

698 Koster, E. P., Couronne, T., Léon, F., Lévy, C., & Marcelino, A. S. (2003). Repeatability in hedonic
699 sensory measurement: a conceptual exploration. *Food Quality and Preference*, 14(2), 165-176.

700 Koster, E. P., & Mojet, J. (2007). Boredom and the reasons why some new food products fail. In
701 H. MacFie (Ed.), *Consumer-led food product development* (pp. 262–268). Cambridge:
702 Woodhead Publishing Ltd.

703 Lusk, K. A., Hamid, N., Delahunty C. M. and Jaeger, S. R. (2015). Effects of an evoked refreshing
704 consumption context on hedonic responses to apple juice measured using best–worst scaling
705 and the 9-pt hedonic category scale. *Food Quality and Preference*, 43:21-25.

706 Lynn, M., & McCall, M. (2016). Beyond gratitude and gratuity: A meta-analytic review of the
707 predictors of restaurant tipping [Electronic version]. Retrieved August 4, 2016, from Cornell
708 University, SHA School site: <http://scholarship.sha.cornell.edu/workingpapers/21>

709 Mead, R., & Gay, C. (1995). Sequential design of sensory trials. *Food Quality and Preference*,
710 6(271–280), 14.

711 O’Brien, H. L., & Toms, E. (2010). The development and evaluation of a survey to measure user
712 engagement. *Journal of the American Society of Information Science and Technology*, 61, 50–
713 69.

714 Petit, C., & Siefferman, J. M. (2007) Testing consumer preferences for iced-coffee: Does the
715 drinking environment have any influence? *Food Quality & Preference*, 28, 23–31.

716 Redmond, W. (1995). An ecological perspective on new product failure: The effects of
717 competitive overcrowding. *Journal of Product Innovation Management*, 12, 200–213.

718 Rosas-Nexticapa, M., Angulo, O., & O’Mahony, M. (2005). How well does the 9-point hedonic
719 scale predict purchase frequency? *Journal of Sensory Studies*, 20, 313–331.

720 Rozin, P., Fallon, A., & Mandell, R. (1984). Family resemblance in attitudes to foods.
721 *Developmental Psychology*, 20(2), 309-314.

722 Rozin, P. & Tuorila, H. (1993) Simultaneous and temporal contextual influences on food
723 acceptance. *Food Quality and Preference*, 4, 11–20.

724 Sester, C., Deroy, O., Sutan, A., Galia, F., Desmarchelier, J., Valentin, D., & Dacremont, C. (2013).
725 “Having a drink in a bar”: An immersive approach to explore the effects of context on drink
726 choice. *Food Quality & Preference*, 28, 23–31.

727 Sommer, R., & Steele, J. (1997). Social effects on duration in restaurants. *Appetite*, 29, 25–30.

728 Spence, C. (2012). Auditory contributions to flavour perception and feeding behavior.
729 *Physiology and Behavior*, 107, 505–515.

730 Stroebele, N., & De Castro, J. M. (2004). Effect of ambience on food intake and food choice.
731 *Nutrition*, 20, 821–838.

732 Weiss, B. H., O’Mahony, M., & Wichchukit, S. (2010). Various paired preference tests:
733 Experimenter effect on “take home” choice. *Journal of Sensory Studies*, 25, 778–790.

734 Westerterp-Platenga, M. S. (1999). Effects of extreme environments on food intake in human
735 subjects. *Proceedings of the Nutrition Society*, 58, 791–798.

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

738 Yeomans, M. R. (2006). Olfactory influences on appetite and satiety in humans. *Physiology and*
739 *Behavior*, 87, 800–804.

740 Zellner, D. A., Stewart, W. F., Rozin, P., & Brown, J. M. (1988). Effects of temperature and
741 expectations on liking for beverages. *Physiology and Behavior*, 44, 61–68.

742

Figure 1. Testing environments. A. Traditional sensory testing booth (no immersion) with counter top, spit sink and cookie samples presented to a panelist. B. Mixed immersion environment in which contextual information depicting the baking of homemade cookies was reintroduced into the sensory booth testing environment using computer screen, head phones and localized cookie aroma diffusion. C. Full immersion condition in which the same audiovisual information was displayed on a video wall comprised of high-definition monitors and a subtle cookie aroma was dispersed into the space. Cookie samples were served to panelists sitting on a barstool at a high-top kitchen table.

Figure 2. Cookie liking and preference results obtained from panelists evaluating samples in each of the three immersive conditions. Tasting replicates were obtained by having the same subjects evaluate the cookies in each of the three testing environments a second time following a ca. 3-week hiatus. A. Combined average hedonic ratings across cookies separated by condition and replicate. Right hand panel shows combined data across both replicates. Letters above bars denote significant differences ($p < 0.05$) found between conditions using Tukey's post-hoc analysis. B. Average hedonic ratings for each cookie separated by condition and replicate. Right hand panel shows combined data across both replicates. Letters above bars denote significant differences ($p < 0.05$) between cookies using Tukey's post-hoc analysis. C. Cookie preference order separated by condition and replicate. Figure shows rank order of preference from most preferred to least preferred. Right hand panel shows combined data across both replicates. Letters above logos indicate homogenous subsets determined by post hoc Wilcoxon signed rank test.

Figure 3. Change in mean liking scores for each cookie in each testing environment. For all cookies, the smallest change in mean liking score from replicate 1 to replicate 2 occurred in the full immersion condition.

Figure 4. Average change in mean liking scores combined for all cookies in each testing environment. Letters above bars denote marginally significant differences ($p = 0.087$) found between conditions using Tukey's post-hoc analysis.

Figure 5. Subjective assessment of panelist engagement in each of the three testing environments collected using the provided questionnaire. A. Average ratings for engagement dimensions for each testing environment separated by replicate. Letters above bars denote significant differences ($p < 0.05$) found between conditions using Tukey's post-hoc analysis. B. Average total engagement score for each testing environment separated by replicate. Letters

779 above bars denote significant differences ($p < 0.05$) found between conditions using Tukey's
780 post-hoc analysis.

Table 1

Table 1. Average hedonic ratings (mean ± SE) for each cookie separated by testing environment and replicate or combined across replicates. Superscript letters denote significant differences in cookie liking within a testing environment using Tukey’s post-hoc analysis.

Environment	Cookie	Replicate 1		Replicate 2		Combined	
		Mean±SE	Statistics	Mean±SE	Statistics	Mean±SE	Statistics
Traditional	Keebler	5.94±0.23 ^A	F _{3,156} =3.94 p=0.010	5.91±0.25	F _{3,156} =0.891 p=0.447	5.93±0.17 ^{AB}	F _{3,212} =8.51 p<0.001
	Pepperidge Farms	5.93±0.25 ^A		6.03±0.28		5.99±0.19 ^A	
	Kroger	5.40±0.23 ^{AB}		5.72±0.24		5.56±0.16 ^{BC}	
	Chips Ahoy	5.06±0.24 ^B		5.57±0.27		5.31±0.18 ^C	
Mixed Immersion	Keebler	6.15±0.23 ^A	F _{3,156} =3.63 p=0.014	6.21±0.23	F _{3,156} =2.33 p=0.076	6.18±0.16 ^A	F _{3,212} =10.8 p<0.001
	Pepperidge Farms	5.94±0.27 ^{AB}		6.32±0.25		6.13±0.19 ^A	
	Kroger	5.56±0.27 ^{AB}		6.13±0.24		5.45±0.18 ^A	
	Chips Ahoy	5.17±0.24 ^B		5.59±0.27		5.38±0.17 ^B	
Full Immersion	Keebler	6.22±0.20 ^A	F _{3,156} =4.26 p=0.006	6.23±0.22	F _{3,156} =1.71 p=0.167	6.23±0.17 ^A	F _{3,212} =12.2 p<0.001
	Pepperidge Farms	6.18±0.24 ^A		6.23±0.27		6.21±0.19 ^A	
	Kroger	5.81±0.27 ^A		5.89±0.27		5.85±0.19 ^A	
	Chips Ahoy	5.26±0.25 ^B		5.60±0.25		5.43±0.17 ^B	

Table 2

Table 2. Average engagement scores for each dimension separated by testing environment and replicate. Superscript letters denote significant differences found between dimensions within a testing environment using Tukey's post-hoc analysis.

Dimension	Replicate 1			Replicate 2		
	Environment	Mean±SE	Statistics	Environment	Mean±SE	Statistics
Usability	Traditional	-0.09±0.16 ^A	$F_{2,104}=0.90$ $p=0.409$	Traditional	-0.21±0.15 ^A	$F_{2,104}=3.51$ $p=0.034$
	Mixed Immersion	0.08±0.13 ^A		Mixed Immersion	0.17±0.12 ^B	
	Full Immersion	0.11±-0.14 ^A		Full Immersion	0.23±0.14 ^B	
Environment Aesthetics	Traditional	-0.38±0.15 ^A	$F_{2,104}=50.02$ $p<0.001$	Traditional	-0.81±0.12 ^A	$F_{2,104}=85.50$ $p<0.001$
	Mixed Immersion	0.94±0.09 ^B		Mixed Immersion	0.96±0.10 ^B	
	Full Immersion	1.00±±0.10 ^B		Full Immersion	0.98±0.12 ^B	
Novelty	Traditional	0.31±-0.09 ^A	$F_{2,104}=0.79$ $p=0.456$	Traditional	0.18±0.07 ^A	$F_{2,104}=7.78$ $p<0.001$
	Mixed Immersion	0.36±0.11 ^A		Mixed Immersion	0.44±0.09 ^B	
	Full Immersion	0.46±0.09 ^A		Full Immersion	0.64±-0.10 ^B	
Involvement	Traditional	0.35±0.10 ^A	$F_{2,104}=22.12$ $p<0.001$	Traditional	0.03±0.09 ^A	$F_{2,104}=39.68$ $p<0.001$
	Mixed Immersion	0.85±0.07 ^B		Mixed Immersion	0.74±0.08 ^B	
	Full Immersion	0.97±0.07 ^B		Full Immersion	0.88±0.08 ^B	
Immersion	Traditional	-1.43±0.12 ^A	$F_{2,104}=64.11$ $p<0.001$	Traditional	-1.83±0.05 ^A	$F_{2,104}=82.03$ $p<0.001$
	Mixed Immersion	-0.23±0.15 ^B		Mixed Immersion	-0.58±0.13 ^B	
	Full Immersion	0.30±0.16 ^C		Full Immersion	0.17±0.17 ^C	
Sensory Awareness	Traditional	1.58±0.15 ^A	$F_{2,104}=95.53$ $p<0.001$	Traditional	1.08±0.12 ^A	$F_{2,104}=126.6$ $p<0.001$
	Mixed Immersion	3.71±0.15 ^B		Mixed Immersion	3.77±0.16 ^B	
	Full Immersion	3.99±-0.16 ^B		Full Immersion	3.96±0.17 ^B	
Distraction	Traditional	1.62±0.12 ^A	$F_{2,104}=5.23$ $p=0.007$	Traditional	1.62±0.12 ^A	$F_{2,104}=2.79$ $p=0.066$
	Mixed Immersion	2.06±0.14 ^B		Mixed Immersion	1.94±0.12 ^A	
	Full Immersion	2.08±0.13 ^B		Full Immersion	1.93±0.11 ^A	
Realism	Traditional	1.28±0.13 ^A	$F_{2,104}=2.65$ $p=0.075$	Traditional	1.06±0.15 ^A	$F_{2,104}=11.16$ $p<0.001$
	Mixed Immersion	1.48±0.09 ^A		Mixed Immersion	1.50±0.09 ^B	
	Full Immersion	1.54±0.08 ^A		Full Immersion	1.65±0.09 ^B	
Total	Traditional	-0.01±0.70 ^A	$F_{2,104}=38.01$	Traditional	-2.13±0.60 ^A	$F_{2,104}=62.91$

Engagement Score	Mixed Immersion	5.13±0.57 ^B	p<0.001	Mixed Immersion	5.06±0.60 ^B	p<0.001
	Full Immersion	6.30±0.58 ^B		Full Immersion	6.57±0.62 ^B	

Figure 1

Figure 2 revised

A.

B.

C.

Figure 3

Figure 4

Figure 5

A.

B.