Food Quality and Preference 100 (2022) 104603

Contents lists available at ScienceDirect

Food Quality and Preference

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

The stability of self-reported emotional response and liking of beer in context

Marit Nijman [a], Qian Yang [a,*], Claire Hidrio [b], Rebecca Ford [a]

[a] Sensory Science Centre, Division of Food, Nutrition and Dietetics, University of Nottingham, Sutton Bonington Campus, United Kingdom
[b] Anheuser-Busch InBev, Belgium

ARTICLE INFO

Keywords:
Context
Emotional response
Test-retest Reliability
Liking
Beer

ABSTRACT

Collecting emotional response to products has been found to be more discriminating than liking alone, yet emotions are not static and are highly impacted by mood and our environment (or context). Therefore the question rises whether a consumer's emotional response to a product is consistent within different contexts, both throughout consumption of a product and with repeated exposure.

This study investigated 1) the impact of context (bar vs. central location test) and consumed amount (1 sip, half glass and full glass) on consumers emotional response and liking of beer, 2) test–retest reliability per consumed amount in both contexts, 3) repeated exposure across all sensory sessions including dummy session regardless of consumed amount and context. Beer consumers (N = 97) evaluated emotional response and liking to beer (3.5% ABV) after one sip, half a glass (≈142 ml) and a full glass (≈284 ml), using a short self-report questionnaire in a real life context (i.e. Bar) and central location test (CLT). A subset of 62 participants repeated the session in both contexts to determine test–retest reliability with intra-class correlation (ICC). Prior to the test sessions, emotional response and liking were measured for the same beer during a 'dummy' session. ANOVA revealed an overall effect of context where the emotions *Shocked*, *Content*, *Excited, Nostagic, Disgusted* and *Curious* were rated higher in the Bar than in the CLT ($p \leq 0.01$). Consumed amount had limited effect on emotional response and liking. Although the effects in this study were modest, they show no adverse effects of conducting consumer testing in realistic environments on the stability of consumer responses. The ICC scores indicated more stable emotion measurements after consuming a full glass of beer compared to half glass and one sip ($p < 0.05$) regardless of context, suggesting that stability of emotional response to beer may be slightly higher if representative amounts of product are consumed. *Content* and *Excited* emotions were rated significantly higher during the first 'dummy' session than the last session (5th) but the effects were modest indicating that a dummy session may not need to be considered in future studies.

1. Introduction

The majority of sensory and consumer studies collect hedonic ratings at one single time point, however, consumption of food and drink is a dynamic experience, involving multiple bites or sips. The validity of using single hedonic ratings to predict food preference is of general interest for sensory and consumer research. In addition, hedonic responses alone have not been found to be strong predictors of food choice leading researchers to explore alternative strategies including the measurement of emotional response (Kaneko, Toet, Brouwer, Kallen, & van Erp, 2018; Mora, Giussani, Pagliarini, & Chaya, 2019) and the impact of context (Jaeger & Porcherot, 2017; Kaneko et al., 2018; Worch et al., 2020). Emotional response has been shown to provide additional insights (Yang, Dorado, Chaya, & Hort, 2018) and the food choice prediction is achieved by both emotions and liking (Dalenberg et al., 2014). In addition, consumer responses (liking and emotion) collected in a realistic consumption context (bar) has been found to be more discriminating than when collected in a sensory booth (Nijman et al., 2019). However, a contradicting finding was reported by Worch et al. (2020), where a Central Location Test (CLT) was found to be more discriminating for emotions than any other context (in a pub, immersive room and two VR conditions).

Even though more realistic contexts are recommended for consumer research to optimise the ability of test results to predict consumer

* Corresponding author.
E-mail address: qian.yang@nottingham.ac.uk (Q. Yang).

https://doi.org/10.1016/j.foodqual.2022.104603
Received 2 December 2021; Received in revised form 3 March 2022; Accepted 6 April 2022
Available online 9 April 2022

behaviour in real life situations, the majority of sensory consumer research takes place in a controlled setting such as a sensory laboratory or CLT to minimise error and maximise reliability of results (Galiñanes Plaza, Delarue, & Saulais, 2019). The decision to test products in either a controlled setting or a natural consumption environment can result in a trade-off between ecological validity and reliability. Ecological validity is commonly used to indicate whether the test environment is representative of the context of interest (Delarue, Brasset, Jarrot, & Abiven, 2019; Galiñanes Plaza et al., 2019). Reliability refers to the consistency of the test (Tavakol & Dennick, 2011), commonly assessed with test–retest procedures using Pearson correlation (Cardello et al., 2012) or Intra-class correlation coefficient (Yen & Lo, 2002) to determine if repeating a test yields the same result. Cardello et al. (2012) studied the test–retest reliability of the consumers' emotional responses to food names and tasted foods in a controlled setting, finding high correlations for food names when measured at two instances, one week apart. Whereas, for tasted foods the reliability was less stable. This was thought to be due to food names reflecting consistent and broader cognitive images and associations, whereas tasted foods vary over time due to perceptual variability, changing expectations, and preceding appetitive contexts (Cardello et al., 2012). However, it is unclear what impact realistic contexts would have on the stability of emotional responses to tasted foods or drinks.

There is some evidence to suggest that self-reported emotional response to products may change over the course of multiple sessions. In a previous study by Nijman et al. (2019), participants completed three sessions, one or two weeks apart. It was found that emotional response to beer was generally more positive during the first session of a within-subject design than during subsequent sessions, suggesting there may be a 'first session' effect (Nijman et al., 2019). It is well known that first order effects can be reduced by inclusion of a 'dummy' sample, and this was also found to be the case for emotional response to beer (Dorado, Pérez-Hugalde, Picard, & Chaya, 2016). Therefore, it would be interesting to explore if the same can be applied to session effects by inclusion of a 'dummy' session.

In order to make the consumption context realistic, a true representative amount of product for testing is ideal, yet often impractical. Instead, consumer responses after a single bite or sip (Nijman et al., 2019) are often collected and the resulting data assumed to be representative, but may not accurately reflect true consumer experience (ecological validity) which has been found to be dynamic over time for beer (Silva et al., 2019).

This study investigated 1) the impact of context (bar vs. central location test) and consumed amount (1sip, half of glass and full glass) on consumers emotional response and liking of beer, 2) test–retest reliability per consumed amount in both contexts, 3) repeated exposure across all sensory sessions including dummy session regardless of consumed amount and context.

2. Materials and methods

2.1. Experimental design

Beer consumers (N = 97) evaluated their emotional response three times over the course of drinking a glass of beer (after first sip, after half a glass (142 ml) and after a full glass (284 ml)), using a self-report questionnaire in three separate sessions (a dummy session and two test sessions in two different context conditions; a real life context (Bar) and in a central location context (CLT). A subset of 62 participants voluntarily to repeat the study to allow an evaluation of test–retest reliability, taking part in a further two sessions (real life context (Bar repeat) and central location (CLT repeat).

All 97 participants completed the 'dummy' session. This was a session that took place prior to the test sessions during which the participants completed the test protocol and reported their responses in the same way as they would during the test sessions. The study was approved by the Faculty of Medicine & Health Sciences Research Ethics Committee of the University of Nottingham (Ethics reference number 123–1709).

2.2. Participants

Participants were recruited through the consumer database held at the Sensory Science Centre, University of Nottingham. In total, 97 participants (37% female) ranged between 18 and 62 years of age, with an average age of 23 (±6.3) years were recruited. Native English speakers made up 84.5% ($n = 82$) of the participants. Non-native English speakers were given extra time to familiarise themselves with the English questionnaire and to ask questions during the 'dummy' session. All participants drank beer at least once every two months and 73% stated they drank beer weekly. Individuals that had any reason to refrain from drinking beer (including self-declared health, pregnancy, religion or addiction) were excluded from participation. Participants received an inconvenience allowance for their participation.

2.3. Beer

A commercial lager beer (3.5% alcohol by volume) was used as the sensory stimulus. All beer was from the same production batch. Sample preparation commenced once participants were seated and logged into the questionnaire. For each participant approximately 284 ml (half pint) of refrigerated beer was poured from 500 ml cans into a standard unlabelled half pint glass (284 ml content) and served immediately.

2.4. Context conditions

The on-campus Student Union Bar at the University of Nottingham (Fig. 1a) was used as the Bar context. During experimental sessions, the bar was open for business as usual and thus there was no control over external variables such as music, room temperature and the presence of other customers that were not taking part in the experiment. Additionally, food was prepared and served in the bar, which added odours of food as an additional uncontrolled variable. At the start of the session, participants received verbal instructions about the study, after which they were given a token that they could use to pick up their beer sample at the bar when they were ready. Participants could choose to freely sit anywhere in the bar and were asked to complete a questionnaire on their smartphone. Tablet devices were available for participants that did not have a smartphone. Interaction with other people was allowed, but participants were instructed not to talk about the beer or the questionnaire.

The central location test (CLT) took place in a large computer room at the University of Nottingham (Fig. 1b). Participants were isolated from each other using portable booth dividers and instructed to complete the session in silence. Questionnaires were completed on individual desktop computers within each booth. Participants were not allowed to use their phone during the CLT sessions. Beer samples were distributed by the researcher and noise was minimised.

2.5. Sessions

All participants were invited to complete the first 3 sessions. Session 1 was the dummy session, whilst sessions 2 and 3 were the Bar and CLT contexts (randomised). During the dummy session, participants were familiarised with the emotion questionnaire and test protocol by consuming the beer and completing the questionnaire for the first time in the CLT location. In order to investigate the stability of emotional response and liking with subsequent exposures, the same beer was used for the dummy session as during the actual test sessions.

Participants' availability determined the order in which they were exposed to the two conditions after the dummy session. Forty-two participants (43%) completed the session in the Bar context first and 55

Fig. 1. On-campus student union bar used for Bar condition (a) and computer room used for Central Location Test (b) at the University of Nottingham.

participants (57%) started with the CTL context, with all 97 participants completing the first three sessions.

In order to determine test–retest reliability, a subset of 62 participants voluntarily came back to complete the Bar and CLT sessions again. Of this group, 35 participants (56%) started in the Bar context while the other 27 (44%) completed the CLT session first, with all 62 participants successfully completing all five sessions.

All sessions were scheduled at 30-minute time-slots between 5 pm and 9 pm on Mondays, Tuesdays, Wednesdays and Thursdays. Participants were encouraged to come in at the same time and on the same day of the week for all sessions, as far as their schedules allowed. Depending on participants' availability, there were one or two weeks between sessions, as well as between test and retest.

2.6. Emotion questionnaire

Data was collected using an online questionnaire via Compusense Cloud (Compusense, Canada). During each session, participants reported their emotional response and liking for the beer (I) after the first sip, (II) after consuming half of the beer in the glass (approximately 142 ml) and (III) after consuming all of the beer in the glass (approximately 284 ml). A reduced beer specific emotion lexicon (Eaton, Chaya, Smart, & Hort, 2019) was used to determine emotional response at each consumption stage. The emotion terms for each category were presented together and participants were asked to rate each category on ten separate continuous line-scales of which the ends were labelled from 'very low' to 'very high' at 5% and 95% of the scale. The emotion categories were randomised across participants. Overall liking was rated after emotional response on a continuous line scale anchored from 'dislike extremely' at 5% to 'like extremely' at 95%.

2.7. Data analysis

To determine the impact of context on emotional response and liking from the first sip through to finishing an entire glass of beer, a Mixed ANOVA (with interaction) with context and consumed amount as fixed factors and participants as random factor, and Tukey's HSD post hoc test was performed ($p < 0.05$).

Test-retest reliability (Intra-class Correlation Coefficients (ICC)) were calculated for consumer responses (emotion and liking data), at each consumed amount in each context from the 62 participants that completed both contexts (Bar and CLT) in duplicate. ICC estimates were calculated using SPSS statistical package version 24 (SPSS Inc, Chicago, IL) based on a mean-rating (k = 2), absolute-agreement, 2-way mixed-effects model. In order to examine the general effect of consumed amount, context and consumer responses on reliability, a three-factor ANOVA (with two-way interactions) was performed on the ICC values, with consumed amount, context and consumer response as fixed factors.

Repeated exposure was studied by using the data collected from the 62 participants who attended 5 sessions in total. For each participant the sessions were coded for chronologic session order, based on the date on which the participant took part in the session. The dummy session was coded as the first session for all participants, followed by the test sessions in the Bar and CLT context as either second or third, depending on individual session order. The re-test sessions were coded fourth and fifth. The data was averaged across the three consumed amounts (1 sip, half glass, full glass). The effect of session order on the responses were examined with a Mixed ANOVA using session as factor and participant as random effect. Statistical analyses were performed using XLSTAT (Addinsoft, NewYork, USA).

3. Results

3.1. Impact of context and consumed amount on consumer response

Table 1 displays the F and p-values from the Mixed ANOVA comparing the two contexts (Bar and CLT) and consumed amount of beer (one sip, half a glass and a full glass), as well as the interaction between context and consumed amount by 97 participants.

Context showed a significant overall effect on two negative emotion categories *Shocked* and *Disgusted*, and four positive emotion categories *Content, Excited, Nostalgic* and *Curious* ($p < 0.05$). As shown in Fig. 2, participants scored these emotion categories somewhat higher in the Bar than in the CLT, regardless of positive or negative emotions (Fig. 2). Limited effects of consumed amount on emotional response were found for the majority of emotions ($p > 0.05$), apart from *Content, Underwhelmed* and *Curious* ($p < 0.05$) (Table 2). In general, participants felt more *Curious* and *Underwhelmed* after taking one sip in comparison to drinking a glass of beer. Participants also felt more *Content* after drinking a glass of beer than after a single sip or half a glass (Fig. 3). However, no significant interaction effects were found between context and consumed amount ($p > 0.1$) for all variables (Table 1), indicating that the test environment did not impact changes in emotional response to beer over time and number of sips. Liking was not found to be significantly different between the Bar and CLT context, nor over the course of consuming a single sip up to a glass of beer ($p > 0.05$).

3.2. Test-retest reliability of response per consumed amount

Table 2 shows the Intra-class Correlation Coefficients (ICC) for

Table 1
F and *p*-values from a two-factor ANOVA on emotional response and liking scores from 97 participants with context (Bar or CLT) and amount (sip, half glass & entire glass) as fixed factors. Bold font highlights *p*-values < 0. 05.

		Context (DF = 1)		Amount (DF = 2)		Context*Amount (DF = 2)	
		F	*p*	*F*	*p*	*F*	*p*
Emotional response	Shocked	12.7	**< 0.01**	0.65	0.52	0.98	0.38
	Bored	1.07	0.30	2.90	0.06	2.68	0.07
	Content	12.5	**< 0.01**	9.07	**< 0.01**	0.44	0.65
	Excited	22.1	**< 0.01**	2.15	0.12	1.55	0.21
	Nostalgic	8.51	**<0.01**	0.19	0.83	0.78	0.49
	Disconfirmed	1.55	0.21	0.99	0.37	1.92	0.15
	Disgusted	6.27	**0.01**	0.41	0.66	2.09	0.12
	Tame/Safe	1.89	0.17	0.15	0.86	1.08	0.34
	Underwhelmed	0.24	0.63	4.90	**<0.01**	0.47	0.63
	Curious	4.97	**0.03**	6.08	**<0.01**	0.41	0.66
	Liking	0.40	0.53	0.29	0.75	0.94	0.39

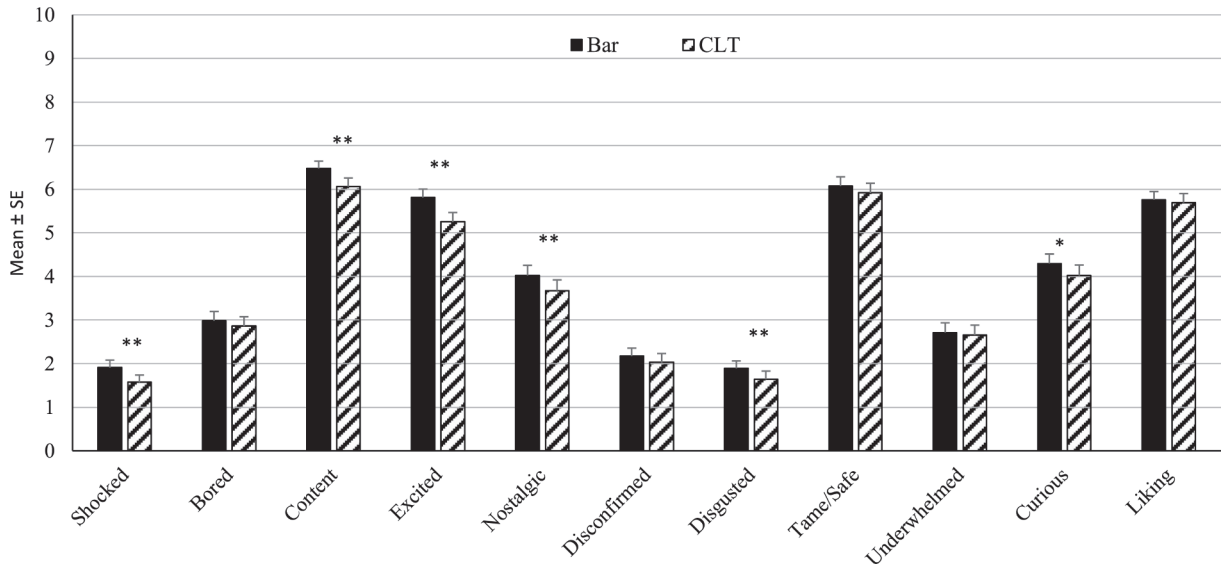

Fig. 2. Mean scores and standard errors for emotion categories and liking, when evaluated by 97 participants in the Bar and Central location Test (CLT). All data was collected on a continuous line scale from 0 to 10. * and ** indicate a significant difference between de Bar and CLT at $p < 0.05$ and $p < 0.01$ respectively.

Table 2
Intra-Class Correlation coefficients between test and retest data collected from 62 participants at three stages of consumption (after a sip, half a glass and an entire glass) in the Bar and CLT. Cells are shaded so that darker cells highlight higher coefficients.

		Bar			CLT		
		1 sip	Half	Glass	1 sip	Half	Glass
Emotional response	Shocked	0.637	0.690	0.837	0.634	0.646	0.682
	Bored	0.452	0.629	0.640	0.605	0.473	0.765
	Content	0.596	0.554	0.676	0.364	0.180	0.532
	Excited	0.659	0.613	0.529	0.446	0.231	0.765
	Nostalgic	0.679	0.743	0.768	0.610	0.631	0.651
	Disconfirmed	0.467	0.601	0.691	0.562	0.675	0.639
	Disgusted	0.622	0.661	0.676	0.590	0.445	0.536
	Tame/safe	0.505	0.826	0.635	0.610	0.777	0.796
	Underwhelmed	0.786	0.809	0.827	0.673	0.717	0.554
	Curious	0.666	0.738	0.742	0.769	0.506	0.692
	Liking	0.613	0.580	0.621	0.691	0.646	0.756

test–retest reliability of each emotion category and liking, evaluated per consumed amount in the Bar and CLT contexts by 62 participants. ICC ranged from 0.180 to 0.837 and most indicated a moderate (0.5 to 0.75) to good (0.75 to 0.9) test–retest reliability (Koo & Li, 2016).

Three-way ANOVA found an overall significant effect for context ($F(1) = 6.8$, $p = 0.017$), consumed amount ($F(1) = 5.4$, $p = 0.013$) and consumer responses ($F(1) = 4.1$, $p = 0.004$) on test–retest reliability. For context effect, in general, ICC was significantly higher in the Bar (0.66 ± 0.1) than the CLT (0.6 ± 0.15). However, when exploring further, this is actually due to an interaction. As an approaching significant interaction for Context*Consumed amount was found ($F(1) = 3.2$, $p = 0.06$), where in the CLT context there appeared to be a drop in reliability after consumption of half the glass of beer (Fig. 4), which was observed for emotion categories *Bored*, *Content*, *Excited*, *Disgusted* and *Curious* (Table 2).

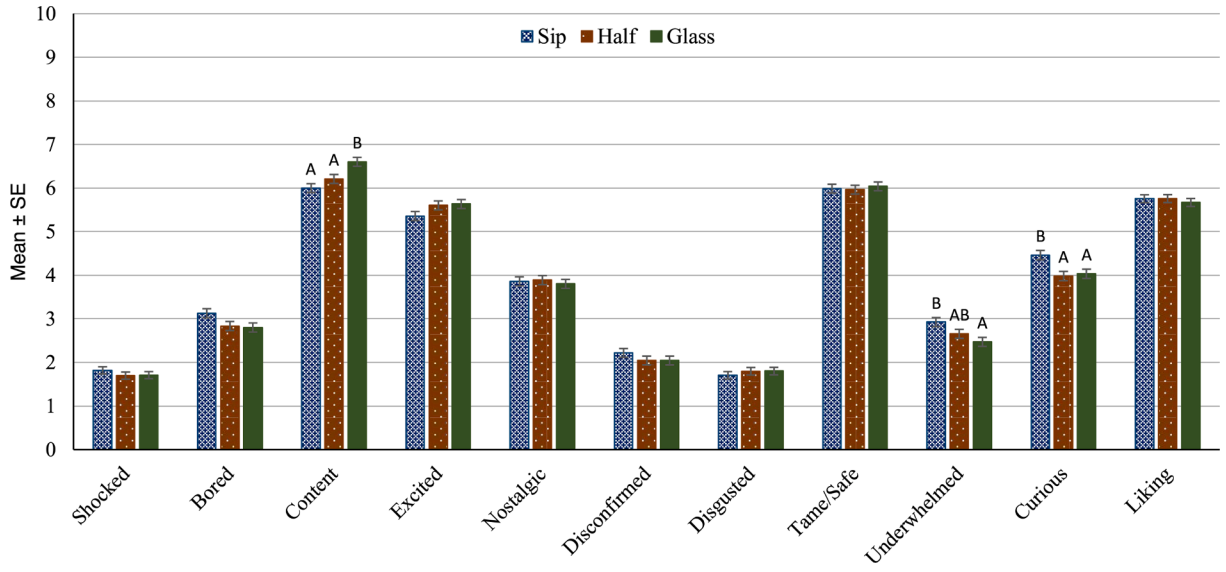

Fig. 3. Mean scores and standard errors for emotion categories and liking, when evaluated by 97 participants after one sip (Sip), after consuming half of the beer in the glass (half) and after consumption of an entire glass of beer (Glass). A and B indicate Tukey's post hoc groupings at $p < 0.05$.

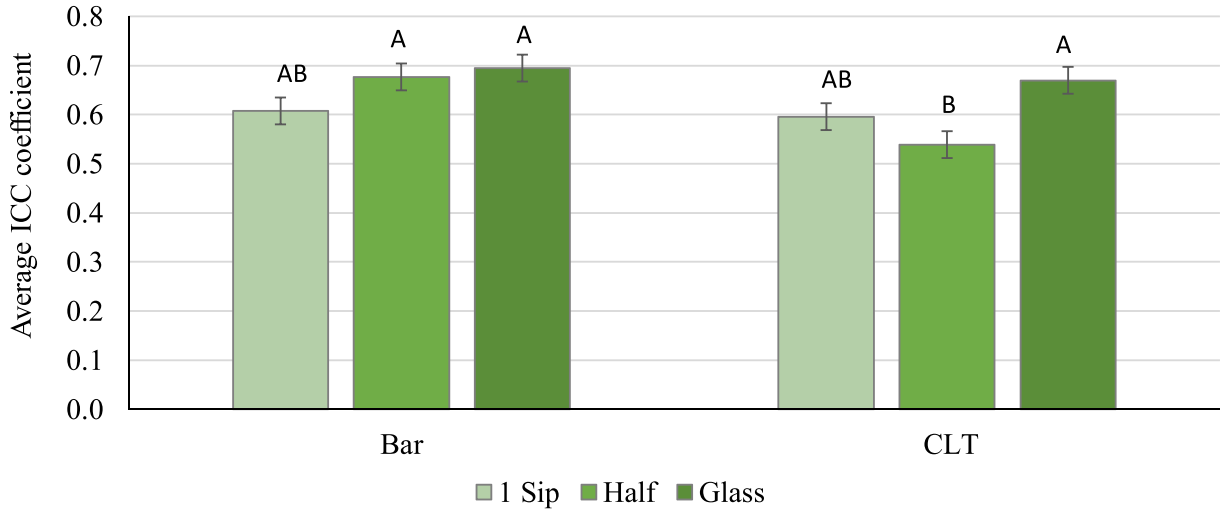

Fig. 4. Average ICC over all variables (10 emotion categories and liking) per consumed amount (one sip, half a glass and full glass) of beer and context (Bar and CLT) with Tukey's post hoc groupings at $p < 0.05$ comparing amounts within each context.

Table 3
Mean scores (averaged over 62 participants for three consumed amounts) and standard errors (SE) for emotional response and liking collected during the first (dummy), second and third (test), fourth and fifth (retest) sessions. All data was collected on a continuous line scale from 0 to 10. A and B indicate Tukey's HSD post hoc groupings at $p < 0.05$.

		Session Means ($n = 62$)										ANOVA	
		1st		*2nd*		*3rd*		*4th*		*5th*		Session	
		M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*F*	*p*
Emotional response	Shocked	1.57	0.12	1.84	0.13	1.79	0.12	1.81	0.12	1.75	0.13	0.60	0.67
	Bored	2.97	0.17	3.23	0.16	2.89	0.16	2.67	0.14	2.79	0.16	1.51	0.2
	Content	6.81A	0.12	6.30AB	0.13	6.24B	0.13	6.29AB	0.13	6.11B	0.13	3.60	0.007
	Excited	5.92A	0.14	5.40AB	0.15	5.63AB	0.14	5.67AB	0.14	5.27B	0.15	2.53	0.04
	Nostalgic	4.10	0.20	4.08	0.17	4.32	0.18	4.00	0.17	4.12	0.18	0.42	0.79
	Disconfirmed	2.04	0.15	2.22	0.15	2.13	0.14	2.06	0.14	2.06	0.16	0.22	0.93
	Disgusted	1.65	0.14	1.89	0.14	2.06	0.15	1.88	0.13	1.78	0.13	1.01	0.40
	Tame/Safe	6.08	0.15	6.00	0.15	5.80	0.16	5.76	0.17	5.82	0.18	0.82	0.52
	Underwhelmed	2.84	0.17	2.80	0.16	2.62	0.17	2.38	0.16	2.40	0.16	1.72	0.15
	Curious	4.60	0.17	4.06	0.17	4.21	0.18	4.16	0.17	4.37	0.19	1.32	0.26
	Liking	5.89	0.14	5.69	0.15	6.03	0.14	6.12	0.13	6.07	0.16	1.11	0.35

For most emotion categories and liking, higher ICC values were found for ratings given after a glass of beer than after a single sip (Table 2). When pooling all data together, a significant higher ICC after a glass of beer (0.68 ± 0.1) was found compared to the ICC for half glass (0.61 ± 0.16) and 1 sip (0.6 ± 0.1). It is also worth noting that the ICC values between context and consumed amount are between 0.6 and 0.68, indicating moderate reliability.

Interestingly, for consumer responses, *Content* (0.48 ± 0.18) had significantly lower ICC value than *Underwhelmed* (0.73 ± 0.1), *Tame/safe* (0.69 ± 0.13), *Shocked* (0.69 ± 0.08), *Curious* (0.69 ± 0.09), and *Nostalgic* (0.68 ± 0.06), indicating *Content* is the least reliable emotion.

3.3. Effect of repeated exposure on consumer responses

Emotional response and liking scores were compared across session orders for participants that completed all five sensory sessions ($n = 62$). Average scores per session and F and p-values from the one-factor ANOVA comparing the sessions (Table 3) revealed that only two emotion categories (*Content* and *Excited*) were affected by repeated exposure. Post-Hoc groupings showed that both *Content* and *Excited* were scored significantly higher during the dummy session (1st Session) than the last session (5th Session) (Table 3).

4. Discussion

4.1. Testing with context and representative amounts of beer

Drinking beer in the bar evoked a significantly different emotional response for *Shocked*, *Content, Excited, Nostalgic*, *Disgusted* and *Curious* emotional response than drinking beer in the CLT. Although the effects are modest, they suggest drinking beer in the bar could evoke higher positive and negative emotional response than evaluating beer in the CLT. However, further research would be needed to understand how different evoked emotions would affect consumer choice. One limitation of this study is that only one beer sample was explored, it would be interesting to explore context effect on product differentiation.

When considering the effect of the consumed amount of beer, no significant differences were found for 7 out of 10 emotion categories and liking, indicating that the consumption volume has a limited impact on emotional and hedonic responses. Participants only felt relatively more *Content* after drinking the full glass of beer and felt more *Curious* and *Underwhelmed* after one sip which could indicate different emotions at different consumption stages but the differences were small and should be interpreted with caution. However, Worch et al. (2020) examined sample order effect on emotional responses by using Check-All-That-Apply (CATA), and found that some emotions (e.g. satisfied, friendly and refreshed) are generally have higher citations when evaluated first or last in the design, indicating order effect could appear in emotional response evaluation. In the current study, whilst consumed amount didn't affect overall liking data, it is worth noting that emotional response was collected prior to overall liking which may have affected responses resulting in less discrimination (Popper, Rosenstock, Schraidt, & Kroll, 2004), therefore asking emotion before liking could potentially influence the overall liking ratings.

In the current study test–retest reliability coefficients (ICC) ranged between 0.36 and 0.79 for a single sip and 0.53 and 0.84 for full glass of beer. This was comparable to earlier findings on reliability of emotional response to food consumption by Cardello et al. (2012) who found test–retest correlations ranged between 0.37 and 0.70. Interestingly, different emotion items had different reliability scores. *Underwhelmed*, *Tame/safe*, *Shocked*, *Curious* and *Nostalgic* emotions had moderate to high reliabilities (0.68–0.73). However, *Content* had the lowest reliability coefficient (0.48), highlighting caution needs to be taken when interpreting the data for *Content* emotion. When looking at interactions, relatively low reliability coefficients for some of the emotion categories were found for measurement after consumption of half a glass of beer in the CLT context. Especially for emotion categories *Content* (0.18) and *Excited* (0.23) which also appeared to differ more between contexts. It is unclear what might have caused this lower reliability after half a glass of beer in comparison to one sip and a full glass in the CLT for these specific emotions, but it could be due to the fact that in CLT, consumers might concentrate more on the task at the beginning (during one sip) and the end (after a full glass). Since the half glass was the midpoint of the experiment, resulting in lower reliability for emotions that may be less relevant for the CLT context (e.g. *Excited* and *Content*). Whereas for the Bar context, this represents a real consumption scenario, and consumers might chat with their friends between answering the questionnaire, resulting in greater focus during the actual task. However, this hypothesis requires a further exploration. Emotion can vary across individuals as well as within a person across instances (Feldman Barrett, 2016). Although a slightly higher ICC scores in the bar (0.66) was found in comparison to CLT (0.6), the difference is very small, and caused by an interaction, therefore, caution needs to be taken when interpreting the test–retest reliability data between the two contexts. Interestingly, test–retest reliability of emotional response was found to be somewhat higher for results collected after consumption of full glass of beer than those collected after a single sip, in both the Bar and CLT. This suggests that testing with representative volumes of beer led to more repeatable emotion measurements, regardless of context.

Liking was stable across the two contexts and between a single sip, half glass and full glass of beer. These findings are in line with earlier research that found no differences between first and last liking scores from a temporal liking procedure for beer samples that yielded dynamic emotional responses over time and number of sips (Silva et al., 2019). Whist higher ICC scores were found for liking in the CLT after consuming a glass compared to one sip and half a glass the difference was very modest, especially when comparing it to the emotion categories. The sample size for the test–retest reliability was 62 participants, thus caution needs to be taken when interpreting the data due to the lower sample size. Future studies with much larger sample size are recommended to confirm the findings.

4.2. Stability of response with repeated exposure

Considering the high number of factors that were not controlled in the Bar environment, such as music and the presence of other customers, it is surprising to see ICC scores in the Bar were comparable as in the CLT context. This result showed that testing in a controlled environment may not necessarily provide more stable data compared to testing in a real-life consumption environment.

For liking data the use of test–retest reliability has been criticised, since hedonic measurements are likely to change over time. Köster, Couronne, Léon, Lévy, and Marcelino (2003) have argued that repeatability is not a good criteria to measure the effectiveness of a method since first preferences seem to be a poor indication of later food choices. Since liking can change as a result of repeated exposure, it can be questioned whether emotional responses change over time and with repeated number of exposures. Repeated exposures to the same product have been suggested to increase boredom and decrease curiosity (Köster et al., 2003), but no such findings were found in the current study. Instead, in the current study, repeated exposure for both over different sessions and within a session (consumed volume) did not affect liking and self-reported emotional response, apart from *Content* for both different sessions and within a session, and *Excited* emotion for different sessions only. However, low reliability was found for *Content* emotion, cautions would need to be taken for interpretation. For *Excited* emotion, there is a tendency that the first session (dummy session) evoked a higher *Excited* emotion, which declined as repeating the experiment, and became significantly less *Excited* for the last session (5th Session). This finding is supported by a previous research where Nijman et al. (2019) found emotional response to beer was generally more positive in the first session than subsequent sessions, however, this study only found a

difference between first session and last session (5th session), suggesting the first session order effect is modest.

It should be noted that the repetitiveness of the questionnaire and study design might have led to the reported stability of responses. The nature of the task can have an effect on ecological validity (Jaeger & Porcherot, 2017) and emotional response may be impacted by self-report measurement (Kaneko et al., 2018). In the current study participants might have experienced fatigue from repeatedly completing the same questionnaire, despite the use of a reduced emotion lexicon in an attempt to minimise the required time and effort. Therefore, further research is required to determine the effect of repeated exposure on product differentiation using emotional response measures, in order to ascertain that reliability does not come at the cost of detecting product differences.

5. Conclusion

The current study investigated the reliability of emotional response and liking for beer across different consumed amounts in two different contexts. Test-retest reliability was comparable between a real-life environment (Bar) and controlled setting (CLT), suggesting no adverse effects of conducting consumer testing in realistic environments on the stability of consumer responses. Interestingly, the *Content* emotion had the lowest reliability when repeating the experiment. No significant repeated exposure effect was observed for most emotional response categories and liking, apart from *Content* and *Excited, for* which responses from the first (dummy) session were significantly different to responses in the last session (5th). This suggests that the dummy session didn't produce a significant first order effect for most emotional response in the current study, indicating that a dummy session may not need to be considered in future studies. Further research is needed to determine the effects of consumed amount and repeated exposure on product differentiation as this was not investigated here.

CRediT authorship contribution statement

Marit Nijman: Writing – original draft, Investigation, Formal analysis, Visualization, Methodology. **Qian Yang:** Supervision, Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Claire Hidrio:** Project administration, Funding acquisition. **Rebecca Ford:** Funding acquisition, Supervision, Project administration, Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Cardello, A. V., Meiselman, H. L., Schutz, H. G., Craig, C., Given, Z., Lesher, L. L., et al. (2012). Measuring emotional responses to foods and food names using questionnaires. *Food Quality and Preference, 24*(2), 243–250.

Dalenberg, J. R., Gutjar, S., Ter Horst, G. J., de Graaf, K., Renken, R. J., & Jager, G. (2014). Evoked emotions predict food choice. *PLoS One, 9*(12), 1.

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.

Dorado, R., Pérez-Hugalde, C., Picard, A., & Chaya, C. (2016). Influence of first position effect on emotional response. *Food Quality and Preference, 49*, 189–196.

Eaton, C., Chaya, C., Smart, K. A., & Hort, J. (2019). Comparing a full and reduced version of a consumer-led lexicon to measure emotional response to beer. *Journal of Sensory Studies, 34*(2), Article e12481.

Feldman Barrett, L. (2016). Navigating the Science of Emotion. In H. L. Meiselman (Ed.), *Emotion*. Measurement: Woodhead Publishing.

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

Jaeger, S. R., & Porcherot, C. (2017). Consumption context in consumer research: Methodological perspectives. *Current Opinion in Food Science, 15*, 30–37.

Kaneko, D., Toet, A., Brouwer, A.-M., Kallen, V., & van Erp, J. B. F. (2018). Methods for evaluating emotions evoked by food experiences: A literature review. *Frontiers in Psychology, 9*(911).

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

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

Mora, M., Giussani, B., Pagliarini, E., & Chaya, C. (2019). Improvement of an emotional lexicon for the evaluation of beers. *Food Quality and Preference, 71*, 158–162.

Nijman, M., James, S., Dehrmann, F., Smart, K., Ford, R., & Hort, J. (2019). The effect of consumption context on consumer hedonics, emotional response and beer choice. *Food Quality and Preference, 74*, 59–71.

Popper, R., Rosenstock, W., Schraidt, M., & Kroll, B. J. (2004). The effect of attribute questions on overall liking ratings. *Food Quality and Preference, 15*(7–8), 853–858.

Silva, A. P., Voss, H.-P., Van Zyl, H., Hogg, T., De Graaf, C., Pintado, M., et al. (2019). Effect of adding hop aroma in beer analysed by temporal dominance of sensations and emotions coupled with temporal liking. *Food Quality and Preference, 75*, 54–63.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55.

Worch, T., Sinesio, F., Moneta, E., Abbà, S., Dreyfuss, L., McEwan, J. A., et al. (2020). Influence of different test conditions on the emotional responses elicited by beers. *Food Quality and Preference, 83*, Article 103895.

Yang, Q., Dorado, R., Chaya, C., & Hort, J. (2018). The impact of PROP and thermal taster status on the emotional response to beer. *Food Quality and Preference*.

Yen, M., & Lo, L. H. (2002). Examining test-retest reliability: An intra-class correlation approach. *Nursing Research, 51*(1), 59–62.