Received: 13 January 2023 | Revised: 28 February 2023 | Accepted: 1 April 2023

DOI: 10.1111/joss.12833

ORIGINAL ARTICLE

Journal of Sensory Studies | WILEY

How do CATA questions work? Relationship between likelihood of selecting a term and perceived attribute intensity

Sara R. Jaeger[1,2] | Sok L. Chheang[1] | David Jin[1] | Grace S. Ryan[1] | Gastón Ares[3]

[1]The New Zealand Institute for Plant & Food Research (PFR) Limited, Mt Albert Research Centre, Auckland, New Zealand

[2]Vescor Research, Copenhagen, Denmark

[3]Sensometrics & Consumer Science, Facultad de Química, Universidad de la República, Canelones, Uruguay

Correspondence
Sara R. Jaeger, The New Zealand Institute for Plant & Food Research (PFR) Limited, Mt Albert Research Centre, Private Bag 92169, Victoria Street West, Auckland 1142, New Zealand.
Email: sara.r.jaeger@gmail.com

Funding information
Ministry for Business Innovation and Employment; New Zealand Institute for Plant and Food Research Limited

Abstract

The present research contributed to a better understanding of how check-all-that-apply (CATA) questions work by examining the relationship between likelihood of selecting a term and perceived attribute intensity. Seven consumer studies were conducted (147–157 people per study) using within-subjects experimental designs where participants twice evaluated the same set of stimuli on the same set of terms (or attributes), respectively with CATA questions and intensity scaling (7-point category scale; 1 = "not at all," 7 = "extremely"). As a function of perceived intensity, the average CATA citation frequency tended to follow a sigmoidal-like relationship where likelihood of selecting a CATA term increased more slowly at the extreme ends of the intensity scale (1–2 and 6–7) and linearly otherwise. This illuminates why for a given term, CATA questions are less suited for discriminating between samples that are of similar "low" or "high" intensity.

Practical Applications

CATA questions are popular for sensory product characterization tasks with consumers. Despite their simplicity, they accurately discriminate among samples, and term citation frequency is a proxy for perceived intensity, albeit not a direct measure hereof. Versatility and applicability of CATA questions to characterize diverse stimuli using diverse types of terms/attributes was demonstrated. By showing that likelihood of CATA term selection typically increases with perceived intensity according to a sigmoidal-like shape, the present research shows that CATA terms best discriminate between samples when these vary in intensity rather than being of similar "low" or "high" perceived intensity.

1 | INTRODUCTION

Check-all-that-apply (CATA) questions are frequently used for consumer-derived sensory product characterizations (Jaeger & Ares, 2022). In such tasks, consumers are presented with a list of terms (or attributes) and select those that apply to the presented sample, and term citation frequency (or frequency of term selection) is calculated across participants to derive profiling information. While studies using CATA questions for applied research are plentiful, a central question has remained somewhat neglected: what do CATA term citation frequencies measure?

Despite their simplicity, CATA questions have been reported to accurately discriminate among samples (Jaeger & Ares, 2022). Several studies have reported that citation frequency is a proxy for perceived intensity, albeit not a direct measure hereof (Bruzzone et al., 2012; Choi & Lee, 2019; Jaeger, Chheang, et al., 2020; Oliveira et al., 2018;

Reinbach et al., 2014; Vidal et al., 2021). If for a given term (X), the citation frequency for two samples (A and B) are X_A and X_B, then the average intensity of X can be inferred to be higher in A than B if $X_A > X_B$. The ability of CATA questions to discriminate among samples in a given sensory attribute can be explained considering that assessors do not always select all the CATA terms they perceive in a specific sample, but only those that exceed a person-, category-, and attribute-specific threshold (Jaeger, Beresford, et al., 2020). Thus, when CATA question responses are pooled across many consumers, relative differences emerge between samples can be identified based on term citation frequency.

In this context, the aim of the present research was to continue investigations into how CATA questions work. The study aimed at exploring the relationship between the likelihood of selecting a term in a CATA question and perceived attribute intensity. For this purpose, direct comparisons of consumers' responses to CATA questions and intensity scales when evaluating identical samples using identical terms were performed. Based on past findings, it was expected that the likelihood of CATA term selection will increase with perceived intensity. Moreover, drawing on the relationship between stimulus strength and perceived intensity often established in psychophysics (e.g., Lawless & Heymann, 2010; Mather, 2022), it was probable that the relationship be sigmoidal (or S-shaped). That is, the likelihood of selecting a CATA term was expected to slowly increase for low and high perceived intensities, whereas a linear relationship was expected in the middle of the scale. Results are expected to contribute to the interpretation of results from CATA questions, contributing to the development of guidelines for best practice.

Since previous related studies have used tasted stimuli and been conducted in CLT settings, the present research deliberately used visual stimuli (images of foods and beverages) and were largely conducted as online surveys. The growing reliance on online surveys in consumer research (e.g., Jaeger & Cardello, 2022; Menon & Muraleedharan, 2020) supports this, as does the increasing dominance of electronic media and exposure to food and beverage stimuli herein. Motivated by the "beyond liking" paradigm, where food-related consumer research increasingly extends beyond the sensory and affective domain into emotions, conceptualizations, and situational appraisals (e.g., Meiselman et al., 2022), the present research also included such "non-sensory" terms. Collectively, these extensions relative to past research increase the ability to generalize findings, and in turn, further support the popularity of CATA questions for consumer-driven product characterization (Meiselman et al., 2022).

2 | MATERIALS AND METHODS

A total of seven studies were conducted (Table 1) and apart from using different product categories (and accompanying terms), the studies were highly similar. Within-subjects experimental designs were used in all studies, and participants completed both the CATA and intensity scaling tasks in a single research session.

2.1 | Participants

Seven studies were conducted, each with 147–157 participants (48%–51% female, 18–69 years old) from New Zealand (see Data S1 for further details). Consumers from Auckland who were registered on a database maintained by a recruitment agency took part in Studies 1 and 2 (CLTs). National coverage was achieved for participants in Studies 3–7 (online surveys). Participants in these studies had self-

TABLE 1 Overview of studies included in the research. All studies were conducted in New Zealand as CLTs (Studies 1 and 2) or online surveys (Studies 3–7).

Study	Stimuli	Stimuli description	Location (*N* consumers)
1	8 fresh fruit	Photographs of real fruit. "Large" stimuli differences (e.g., banana, cherries, pineapple)	CLT (147)
2	9 yellow-fleshed kiwifruit	Artist renditions of fruit. "Medium" stimuli differences relating to skin color, core size, seeds, and fruit flesh	CLT (157)
3	8 fresh fruit	Identical to Study 1	Online (152)
4	9 yellow-fleshed kiwifruit	Identical to Study 2	Online (152)
5	9 green-fleshed kiwifruit	Artist renditions of fruit. "Medium" stimuli differences relating to skin color, core size, seeds, and fruit flesh	Online (152)
6	7 milk (dairy, plant-based)	Photographs of small cups of dairy milk and plant-based alternatives. "Small" stimuli differences relating to milk color based on ingredient (e.g., soy, cashew, oat)	Online (153)
7	7 red wine	Photographs of red wine in glasses. "Small" stimuli differences relating to color based on quality, vineyard, and country	Online (152)

Note: Within-subjects experimental designs were used in all studies, with participants evaluating the stimuli (product images in color) once using CATA questions and once using 7-point category intensity scales (1 = "not at all," 7 = "extremely"). The terms used in each study are listed in Table 2. Figure 1 has exemplar product images (Data S1 has all images).
Abbreviations: CATA, check-all-that-apply; CLT, central location test.

registered on a database managed by a web panel provider with ISO 20252:2019 accreditation (International Organization for Standardization, 2019).

2.1.1 | Human participants ethics statement

All studies were covered by a general approval for sensory and consumer research obtained from the Human Research Ethics Committee at the New Zealand Institute for Plant and Food Research (PFR). Informed consent was obtained, and financial compensation given. Participants were aware that their identities would remain anonymous and that they were free to leave the study at any time.

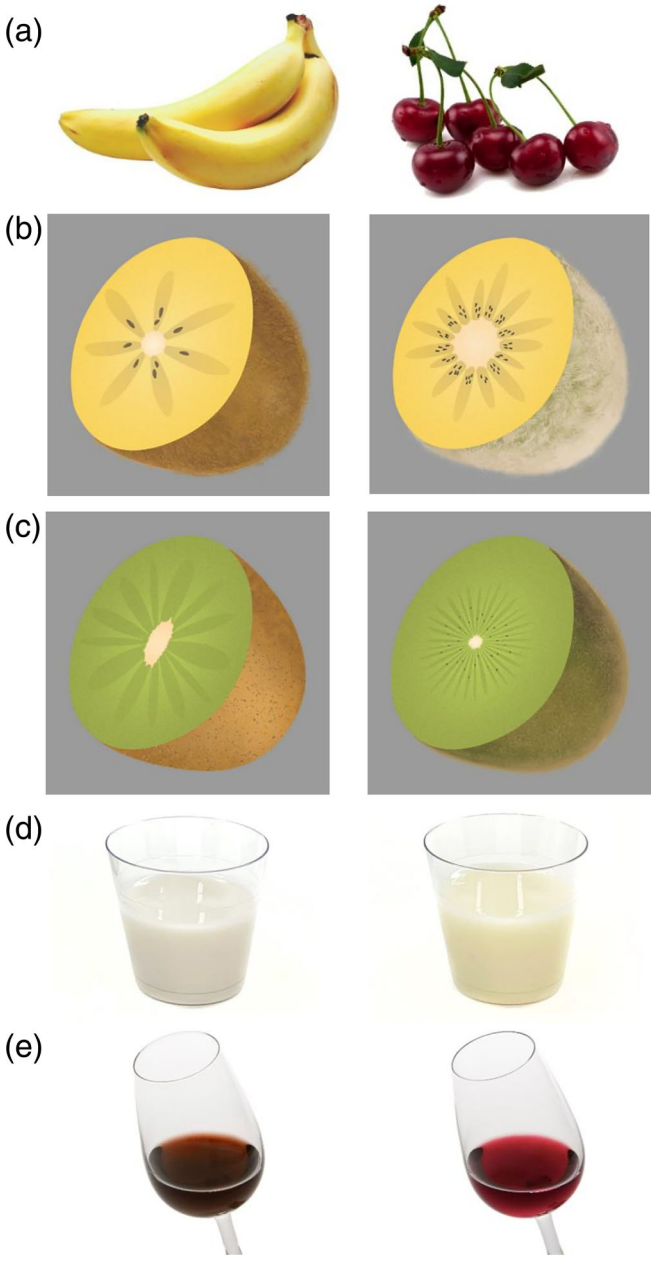

FIGURE 1 Examples of images used as product stimuli in the research, shown by study: (a) Study 1 and 3, (b) Study 2 and 4, (c) Study 5, (d) Study 6, and (e) Study 7.

2.2 | Stimuli and terms

The same stimuli and terms were used in both experimental conditions. Images of foods and beverages were used as stimuli (Figure 1). The studies included 7, 8, or 9 stimuli (Table 1) and the differences between stimuli varied between studies from "large" (fresh fruit) to "medium" (kiwifruit) to "small" (milk, wine) (Data S1 shows all stimuli). Identical samples were used in Studies 1 and 3 (fresh fruit), CLT and online and Studies 2 and 4 (yellow kiwi, CLT, and online). The images were photos or artist impressions and were without copyright restrictions or used with prior approval (Data S1 has full details).

Table 2 lists the terms (or attributes) used in each study ($n = 12$ or 13) and their classification as "sensory" ($n = 5$–9) or "non-sensory" ($n = 3$–7). This classification did not impact on data collection and the

TABLE 2 Product characterization terms used in each of the seven studies from Table 1 (S1–S7).

Product images[a]	Type of term[b]	Terms
S1, S3: Fresh fruit	Sensory (6)	Hard, Juicy, Seeds, Soft, Sweet, Tart/acidic
	Non-sensory (6)	Boring, Convenient, Exciting, Familiar, Inconvenient, Novel
S2, S4: Yellow-fleshed kiwifruit	Sensory (10)	Dry, Hard, Juicy, Kiwifruit flavor, Large core, Seeds, Soft, Sour/acidic, Sweet, Tropical flavor
	Non-sensory (3)	Exciting, Sophisticated, Unfamiliar
S5: Green-fleshed kiwifruit	Sensory (9)	Dry, Hard, Juicy, Kiwifruit flavor, Large core, Seeds, Soft, Sour/acidic, Sweet
	Non-sensory	Exciting, Sophisticated, Unfamiliar
S6: Milk (cow, plant-based)	Sensory (9)	Creamy appearance, Golden color, Gray color, Nutty taste, Oat/grain taste, Sweet taste, Thin appearance, Watery appearance, White color
	Non-sensory (3)	Healthy, Natural, Refreshing
S7: Red wine	Sensory (5)	Brown color/hue, Complex flavor, Pink color/hue, Purple color/hue, Simple flavor
	Non-sensory (7)	Cheap, Classy, Feminine, Goes well with many foods, Masculine, Modern, Traditional

Note: The terms are listed (alphabetical order) by group classification (sensory or non-sensory).
[a]The same terms were used in Studies 1 and 3 (fresh fruit) and Studies 2 and 4 (yellow-fleshed kiwifruit).
[b]The numbers of terms are shown between brackets.

two types of terms were not presented in separate CATA questions/blocks of rated attributes. This decision was justified by the recommendation to not use CATA questions with <10–12 terms (Jaeger & Ares, 2022). The terms were based on the authors' product-relevant experience and extant literature (e.g., Cardello et al., 2022; Jaeger et al., 2019; Jaeger, Roigard, et al., 2020; Longo et al., 2020; Thomson, 2016).

2.3 | Empirical procedures

Within studies, one-half of participants completed the CATA task before the rating task, while the other half of participants completed the rating task before the CATA task. The average time to complete the two product evaluation tasks varied by study and ranged between 6 and 8.5 min. Completion of the two tasks was separated by filler tasks which lasted 5–8 min. These filler tasks and other additional tasks also completed in the same sessions are not considered further for lack of relevance to the current research.

For each stimulus, participants were instructed to look at the product image (according to Table 1). When answering CATA questions, the question was "Which of the following attributes do you expect to describe this sample? Please select all the attributes that apply." For intensity rating, the elicitation question was "At what intensity do you expect to perceive the following attributes?" When responding using paper ballots (CLT conditions, Studies 1 and 2) the supplementary instruction was to select one answer for each attribute, while in online surveys (Studies 3 to 7), the supplementary instruction was "please adjust the marker to indicate your response on the scale." Category scales were used, with seven response options and end-point anchors only (1 = "not at all" and 7 = "extremely"). In the CLT studies, the verbal anchors were positioned to the left and right of the scale while in the online surveys the anchors were positioned above the scale and vertically aligned with the two extreme points of the scale. In the online studies, the visual layout of the scale partially resembled a line scale since the boxes were connected with a line which changed color when a response was made to show degree of perceptual intensity (Data S1 contains visual examples of the data collection and ballot variations).

Studies 1 and 2 were conducted under CLT conditions using a stapled booklet of pen-and-paper ballots with one stimulus per page. The stimuli were evaluated monadically according to designs based on Williams' Latin Squares. The order of terms varied across and within participants (Ares et al., 2015). Participants were seated in standard sensory testing booths (white lighting, positive air flow, 20°C–22°C).

Studies 3–7 were conducted as online surveys. The stimuli were presented in randomized order, as the order of terms was randomized across and within participants. Once participants had completed evaluation of one stimulus they clicked on the "next" button to progress the task. Participants completed the studies in a private location of their choosing. Data S1 contains a data quality statement in accordance with recommendations by Jaeger and Cardello et al. (2022) regarding online survey research. This statement describes criteria used for exclusion of participants who completed the surveys based on indices that are linked to inferior data quality.

2.4 | Data analysis

The data from each study were analyzed separately, using the same set of procedures. All data analyses were run on R software version 4.2.0 (R Core Team, 2022).

Because within-subjects experimental designs were used, it was possible to designate some participants as "bad responders." These were defined a priori according to the logic that perceived intensity should, on average, be higher when CATA terms are selected than when they are not. For each participant this criterion was implemented by (1) calculating across all samples and all terms the average intensity when CATA terms were selected ($M_{CATA=1}$) and the average when CATA terms were not selected ($M_{CATA=0}$), (2) calculating the difference between these values as $M_{diff} = M_{CATA=1} - M_{CATA=0}$, and (3) determining if $M_{diff} > 0$. Only participants who satisfied this criterion were retained for further analysis (i.e., participants were excluded if $M_{diff} \leq 0$).

For descriptive purposes, range of average values (CATA citation frequency and intensity) was calculated. Across retained participants, the average CATA citation frequency was calculated for each scale point on the 7-point intensity scale. The relationship between the two sets of values was shown in tabular format or visualized using line plots. Besides performing these analyses across all terms, analyses were also performed across terms identified as sensory or non-sensory (Table 2). Bootstrapped 95% confidence intervals based on percentiles (5% and 95%) around citation proportions were calculated considering 500 iterations.

3 | RESULTS

The criterion for excluding participants from analysis ($M_{diff} \leq 0$) varied by study and was for S1–S7, respectively: 0%, 0.01%, 10.5%, 17.8%, 13.2%, 8.5%, and 11.2%. A difference between participant exclusion in CLT and online studies (Studies 1–2 vs. 3–7) was, thus, apparent. Cumulative histograms of the distribution of M_{diff} (see Data S1) showed study differences in the means and standard deviations for M_{diff} in support of person-to-person differences in the perceived intensity at which CATA terms are selected.

Table 3 shows the range of CATA citation frequency and perceived intensity across sensory attributes and samples for the seven studies. It fitted expectations that the study where stimuli differences were largest (fresh fruit) was also where the greatest range in average responses were seen (e.g., when a study includes very different types of fruit, it is expected that some are not very sweet while other are very sweet). There was a difference between the two types of terms (sensory and non-sensory) where the latter spanned less of the scale range (on a study-by-study basis). Finally, the online studies (Studies

3–7) were generally characterized by lower response values than studies conducted in CLT settings (Studies 1 and 2).

Also in Table 3, it was seen that for most terms the likelihood of selecting a CATA term when perceived intensity corresponded to the maximum score largely differed from 100%. This corresponded with Ares et al. (2014) who used eye-tracking technology to show that consumers do not fixate their gaze on all the terms of a CATA question for all samples.

Table 4 contains the key results relative to the stated research aim. As predicted, CATA citation proportions increased with increased perceived intensity following, more or less a sigmoidal curve. Figure 2a–g show this, and the accompanying confidence intervals for each scale point are in Data S1 (not included in Figure 2 to retain visual clarity). In all the studies, the average CATA citation frequency for a given intensity level (scale point) was always higher than the value at the previous intensity level. However, the difference in citation proportions between consecutive intensity scores differed across the scale, and differences were generally smaller in the extreme of the scales compared with the middle rage (Table 4).

The CATA citation proportions corresponding to a certain perceived intensity varied across studies and terms. Notably, the citation proportion that corresponded to an intensity score was higher for sensory terms than non-sensory terms (Table 4 and Figure 2a–g). It was also observed that the line plots for sensory and non-sensory terms evolved more or less in parallel in some studies (e.g., Figure 2a: fresh fruit), but not in others (e.g., Figure 2e: green kiwi). Finally, differences between the lines for sensory and non-sensory terms was study-dependent and smaller in Study 6 (Figure 2f: milk) and Study 7 (Figure 2g: red wine) than the two studies with yellow kiwifruit (Figure 2b: Study 2 and Figure 2d: Study 4).

TABLE 3 Results for the seven studies included in the research (S1–S7), showing range of responses for the different experimental conditions according to term type ("sensory" or "non-sensory").

Study	Type of term	Range: CATA[a]	Range: Rating[b]
S1	Sensory	1.4–99.3	1.3–6.6
S2	Sensory	1.3–76.4	1.4–5.4
S3	Sensory	0.7–82.2	2.1–5.9
S4	Sensory	4.6–55.3	2.7–4.8
S5	Sensory	5.3–64.5	2.6–5.5
S6	Sensory	0.7–83.7	2.1–5.5
S7	Sensory	2.0–58.5	2.1–5.0
S1	Non-sensory	0.7–89.1	1.6–6.7
S2	Non-sensory	8.3–35.0	2.6–4.5
S3	Non-sensory	1.3–69.1	2.1–6.1
S4	Non-sensory	6.6–19.1	3.1–3.8
S5	Non-sensory	3.9–38.3	2.5–3.6
S6	Non-sensory	3.9–36.6	2.8–5.0
S7	Non-sensory	3.3–44.7	2.8–4.5

[a]Citation frequency from CATA (check-all-that-apply) questions can range between 0% and 100%.
[b]Average intensity ratings from 7-point category scales (1 = "not at all"; 7 = "extremely").

4 | DISCUSSION AND CONCLUSIONS

The present research aimed to contribute to the literature by providing additional insights into how CATA questions work when used by consumers. Across multiple sets of data with images of foods and beverages as stimuli, the relationship between the likelihood of selecting a term and perceived attribute intensity was assessed.[i] The findings inform recommendations regarding suitability of using CATA questions when ability to discriminate between a given set of stimuli is paramount to achieving stated research objectives.

Fitting expectations, the relationship between perceived intensity and CATA citation frequencies resembled a sigmoid curve. Specifically, if perceived intensity corresponded to the lowest range of the scale (1–2), participants were unlikely to select a CATA term to describe a sample. This suggested that CATA questions may not the best methodological choice when for the discrimination of samples with similar and low attribute intensity. In such situations, the use of two-sample directed difference tests (2AFC; Lawless & Heymann, 2010) which would require focus on a single attribute, may

TABLE 4 Average citation frequency from CATA questions (0% to 100%) at different intensity scores (1 = "not at all" to 7 = "extremely") by study (S1–S7), calculated across all samples and terms.

Intensity score	S1	S2	S3	S4	S5	S6	S7
1	3.1 [2.0–4.1]	8.6 [6.7–10.7]	1.7 [1.0–2.8]	5.1 [3.5–6.5]	7.8 [4.8–12.9]	6.7 [5.4–7.9]	4.8 [3.4–6.6]
2	6.8 [5.3–8.5]	15.9 [13.7–18.0]	5.8 [4.2–7.5]	11.2 [9.0–13.6]	10.6 [8.5–12.9]	9.8 [7.9–11.6]	9.2 [7.2–11.2]
3	14.9 [12.1–18.0]	22.8 [20.7–25.0]	11.3 [9.1–13.9]	16.4 [13.3–20.8]	19.7 [16.9–22.7]	17.7 [14.6–20.8]	14.9 [12.7–17.6]
4	27.1 [24.0–30.9]	29.7 [27.0–32.2]	21.8 [18.5–25.3]	25.9 [21.9–30.3]	28.9 [24.6–33.5]	24.0 [20.9–27.2]	20.2 [18.0–22.8]
5	49.0 [45.7–52.1]	47.1 [44.0–49.9]	41.6 [37.9–45.6]	43.5 [38.0–49.2]	42.6 [37.4–48.0]	36.4 [31.9–40.8]	35.7 [31.9–39.5]
6	70.3 [67.0–73.8]	61.8 [57.6–65.8]	56.3 [49.6–62.5]	44.4 [34.8–55.0]	52.0 [44.1–59.3]	39.5 [33.3–46.9]	45.0 [37.0–54.5]
7	82.9 [79.7–85.8]	73.2 [68.3–77.7]	63.2 [55.4–70.1]	56.9 [44.6–70.8]	56.5 [46.0–67.0]	43.2 [35.0–52.6]	46.5 [38.0–56.0]

Note: Bootstrapped 95% confidence intervals are shown between square brackets.

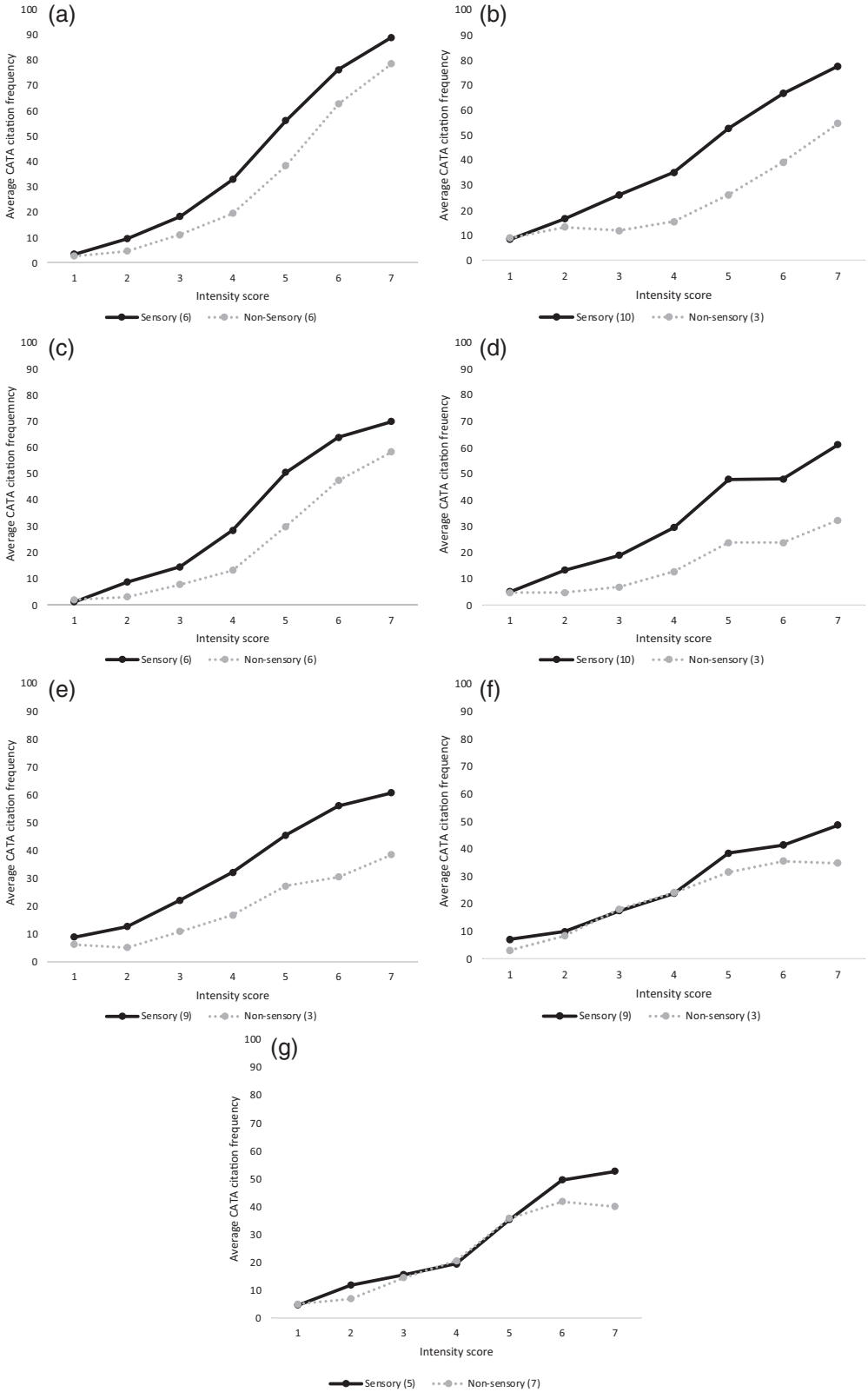

FIGURE 2 Line plots of average CATA citation frequency (0%–100%) by intensity score (1 = "not at all" to 7 = "extremely") shown for each study in the research with terms grouped as sensory (black line) and non-sensory (dotted line) according to Table 1. In alphabetical order the plots are: (a) Study 1 (fresh fruit images, CLT), (b) Study 2 (yellow kiwi images, CLT), (c) Study 3 (fresh fruit images, online), (d) Study 4 (yellow kiwi images, online), (e) Study 5 (green kiwi images, online), (f) Study 6 (milk images, online), and (g) Study 7 (red wine images, online).

be recommended. However, 2AFC requires repeated pairwise tests when applied to more than two samples and several terms/attributes.

As perceived terms intensity increased, likelihood of selecting a CATA term tended to linearly increase. This matched results from previous research by, for example, Bruzzone et al. (2012), Reinbach et al. (2014), Oliveira et al. (2018), Jaeger, Chheang, et al. (2020), Choi and Lee (2019), and Vidal et al. (2021).

At the end range of the intensity scale (6–7), the linear relationship between intensity and likelihood of selecting a CATA term tended to no longer hold. Instead, the likelihood of selecting a CATA

term increased slowly in the upper range of the intensity scale. This suggested that CATA questions may be unlikely to discriminate among samples with similar and high intensity of a given term/attribute. In such situations, the uses of RATA (rate-all-that-apply) questions may be necessary, as recommended by Vidal et al. (2018, 2021). Briefly, in RATA questions, participants select if a given term applies or not to describe a given sample, and if it does, they rate the perceived intensity on an intensity scale. For completeness, it is noted that RATA questions can also improve discrimination (relative to CATA questions) among samples with similar and low attribute intensity (Jaeger & Ares, 2022).

Besides contributing to a better understanding of how CATA questions work, the present research also confirmed large individual differences in when CATA terms are selected. Fitting expectations, the average intensity score when CATA terms were selected was higher than the average score when CATA terms were not selected for most participants. Inspection of difference scores (M_{diff}) revealed considerable heterogeneity where $M_{diff} \sim 0$ for some participants where and $M_{diff} \sim 5$ for other participants. This was observed in all studies (Data S1 shows cumulative histograms for M_{diff} by study) and fitted with the earlier finding by Jaeger, Beresford, et al. (2020) suggesting person-specific thresholds exist for CATA term selection. This suggests considerable individual variability in the perceived intensity at which people selects a term to describe a sample using CATA questions, and a key insight from these M_{diff} distributions is that there can exist no universal "conversion" to/from average CATA citation frequency and perceived intensity.

Following the recommendation of Vidal et al. (2021), the present research extended beyond the use of CATA questions for sensory product characterization. This was warranted because CATA questions are also being used to characterize other aspects of eating and drinking experiences notably emotional associations, conceptualizations, and situational appropriateness (e.g., Cardello et al., 2022; Jaeger & Ares, 2022). The key findings replicated for sensory and non-sensory attributes. However, it was also apparent that the two sets of results were not identical since the increase in average CATA citation frequency with increasing intensity scale score tended to be reduced among non-sensory terms relative to sensory terms. From a measurement perspective, there is no apparent reason why this would be the case. However, it does make sense that sensory terms which are a class of descriptors primarily used to characterize foods and beverages can be perceived as closer to maximum value (7 = "extremely"). Conversely, emotions and conceptualizations apply to a much broader range of stimuli and experiences in life and the variation experience within the food/beverage domain is likely to be limited relative to other life experiences. For example, it is possible to imagine much more intense/extreme feelings of "happy" than those linked to consuming fresh fruit or plant-based milk (say, graduating from university, getting married, becoming a parent, etc.), and, likewise, it is possible to imagine objects that would be regarded as more unfamiliar than a white-haired kiwifruit without seeds (say, a UFO, a dancing robot, etc.). This difference does not diminish the present results but regards them through an appropriate interpretative lens, giving focus to the important relationship with is that of an increase in CATA citation frequency as perceived intensity rises. While this result could suggest lower sample discrimination may be obtained using non-sensory attributes compared to sensory attributes using CATA questions, it must be considered that non-sensory differences between the focal stimuli were likely smaller than those for sensory attributes. Thus, before drawing such a conclusion, research with stimuli selected to span widely across non-sensory features should be performed. We speculate that stimuli which vary across brand, packaging, claim, and price variables may lead to results that are more similar to those seen for sensory attributes in the present research.

4.1 | Limitations and suggestions for future research

A central assumption underpinning this research is that the perceived intensity ratings provided by consumers are "correct." The interpretation of CATA citation frequencies is made against the intensity scores, and it is assumed that participants can accurately scale perceived intensity and that they use the 7-point intensity scale correctly. The latter includes the assumption that the intensity scale has interval-level measurement properties (seen visually in Figure 2 by spacing the interval scale points evenly apart on the horizontal axes). It is beyond the scope of the present research to evaluate these assumptions and doing so would make the paper into one that is focused more on consumers' ability to perform intensity ratings than being a paper that focuses on understanding how CATA questions work. Nonetheless, if the intensity ratings were to possess ordinal-level measurement properties rather than interval-level measurement properties, rank ordering would be maintained to reflect increased probability of CATA term citation. The relationship may look different in terms of the shape of the sigmoidal-like curve, which, for example, could be very flat at high intensity while rising more steeply at low intensity and remaining linear at moderate intensity.

Conducting the research with different product categories and in different test "locations" increased the generalizability of findings. This was a strength, but the limitation of this research strategy lay in not being able satisfactorily explain between-study differences, which would have required more systematic study-to-study variation. Other researchers may have preferred this, but the present strategy increased applied relevance, and this was deemed more important considering the popularity of CATA questions in these settings. For this reason, it may also be relevant for future research to compare rate-all-that-apply (RATA) questions to intensity rating in a similar manner to that done here and by previous authors.

A systematic difference between studies conducted online and in CLT conditions was apparent based on participant exclusion according to M_{diff} scores, where $M_{diff} \leq 0$ was interpreted as low/poor attention to the task since it is logically inconsistent that average intensity be higher when CATA terms are not selected compared with when they are selected. Exclusions in CLT studies were very near absent, while between 8% and 20% of participants were excluded in the online studies. This was despite having already implemented a "speedster"

criterion in the online studies which excluded participants who performed the full survey faster than 1/3 of the median time (see Data S1 for details). Although the difference between CLT and online studies can likely be attributed to reduced quality of the data from online survey (e.g., Brühlmann et al., 2020; Jaeger & Cardello, 2022), differences in the implementation of the intensity scale could have also contributed to the lower quality of the data collected in the online studies. In the CLT studies, the verbal scale anchors were placed to the left/right of the extreme scale points, while in the CLT studies they were placed above "1" and "7." Tentatively, this difference in layout "compressed" the scale range in participants' minds. While performance of ratings scales and layout variations in online research has been investigated (e.g., Lenzner & Höhne, 2022; Menold, 2020) there are, to our knowledge, no directly relevant past studies. Considering the increasing interest in online surveys, additional methodological research is needed to develop best practice recommendations for the implementation of intensity scales and other data collection tools.

Because the research used within-subjects experimental designs, there is a risk that participants purposefully tried to recall previous answers and make the two sets of answers more similar or that other biases due to repeating the "same" task twice could have influenced the results. This cannot be ruled out, but we regard it as unlikely due to the studies being embedded in larger session sessions/surveys where other tasks were completed before, between, and after the data collection focal to the present research. There were no instructions that drew participants attention to this particular set of tasks or to work slowly/deliberately that may have increased the cognitive effort and increased ability to recall previous answers.

The present research was limited to visual stimuli. Further research is needed to extend the sigmoidal-like relationship between perceived attribute intensity and citation proportions with experience-based sensory ratings of smelled or tasted stimuli. Based on Jaeger, Beresford, et al. (2020), who reported the existence of person-, attribute-, and category-specific thresholds for selecting a term in a CATA question, it can be hypothesized that results from the present research would be replicated.

AUTHOR CONTRIBUTIONS

Sara R. Jaeger: Conceptualization, Methodology, Formal analysis, Visualization, Writing – Original draft, Writing – Editing & Reviewing. Sok L. Chheang, David Jin, and Grace S. Ryan: Resources, Investigation. Gastón Ares: Conceptualization, Methodology, Formal analysis, Writing – Original draft, Writing – Editing & Reviewing.

ACKNOWLEDGMENTS

Dr Leticia Vidal from Universidad de la República (Uruguay) is thanked for help in bootstrapping to determine confidence intervals for likelihood of CATA term selection at different levels of perceived intensity. Staff at the Sensory & Consumer Science Team at Plant and Food Research (New Zealand) are thanked for help in pilot work and data curation. Open access publishing facilitated by New Zealand Institute for Plant and Food Research Ltd, as part of the Wiley - New Zealand Institute for Plant and Food Research Ltd agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST STATEMENT

All authors declare no conflicts of interest. The funding sources had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Sara R. Jaeger https://orcid.org/0000-0002-4960-5233

ENDNOTE

[i] Readers who are also interested in how average intensity changes as a function of CATA citation frequency are referred to Parts 8 and 9 of Data S1 which contains the relevant figures.

REFERENCES

Ares, G., Etchemendy, E., Antunez, L., Vidal, L., Gimenez, A., & Jaeger, S. R. (2014). Visual attention by consumers to check-all-that-apply questions: Insights to support methodological development. *Food Quality and Preference*, *32*, 210–220.

Ares, G., Reis, F., Oliveira, D., Antúnez, L., Vidal, L., Giménez, A., Chheang, S. L., Hunter, D. C., Kam, K., Roigard, C. M., Paisley, A. G., Beresford, M. K., Jin, D., & Jaeger, S. R. (2015). Recommendations for use of balanced presentation order of terms in CATA questions for sensory product characterization. *Food Quality and Preference*, *46*, 137–141.

Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, *2*, 100022.

Bruzzone, F., Ares, G., & Gimenez, A. (2012). Consumers' texture perception of milk desserts. II – Comparison with trained assessors' data. *Journal of Texture Studies*, *43*, 214–226.

Cardello, A. V., Llobell, F., Giacalone, D., Roigard, C. M., & Jaeger, S. R. (2022). Plant-based alternatives vs. dairy milk: Consumer segments and their sensory, emotional, cognitive and situational use responses to tasted products. *Food Quality and Preference*, *100*, 104599.

Choi, Y., & Lee, J. (2019). The effect of extrinsic cues on consumer perception: A study using milk tea products. *Food Quality and Preference*, *71*, 343–353.

International Organization for Standardization. (2019). Market, opinion and social research, including insights and data analytics — Vocabulary and service requirements (ISO standard No. 20252). https://www.iso.org/obp/ui/#iso:std:iso:20252:ed-3:v1:en

Jaeger, S. R., & Ares, G. (2022). Using check-all-that-apply (CATA) questions in emotion questionnaires. In M. Bensafi (Ed.), *Basic protocols on emotions, senses, and foods*. Springer Nature.

Jaeger, S. R., Beresford, M. K., Lo, K. L., Hunter, D. C., Chheang, S. L., & Ares, G. (2020). What does it mean to check-all-that-apply? Four case studies with beverages. *Food Quality and Preference*, *80*, e103794.

Jaeger, S. R., & Cardello, A. V. (2022). Factors affecting data quality of online questionnaires: Issues and metrics for sensory and consumer research. *Food Quality and Preference*, *2022*, 104676.

Jaeger, S. R., Chheang, S. L., Jin, D., Roigard, C. M., & Ares, G. (2020). Check-all-that-apply (CATA) questions: Sensory term citation

frequency reflects rated term intensity and applicability. *Food Quality and Preference*, *86*, 103986.

Jaeger, S. R., Hunter, D. C., Vidal, L., Chheang, S. L., Ares, G., & Harker, F. R. (2019). Sensory product characterization by consumers using check-all-that-apply questions: Investigations linked to term development using kiwifruit as a case study. *Journal of Sensory Studies*, *34*(3), 12490.

Jaeger, S. R., Roigard, C. M., Jin, D., Xia, Y., Zhong, F., & Hedderley, D. I. (2020). A single-response emotion word questionnaire for measuring product-related emotional associations inspired by a circumplex model of core affect: Method characterisation with an applied focus. *Food Quality and Preference*, *83*, 103805.

Lawless, H., & Heymann, H. (2010). *Sensory evaluation of food science principles and practices* (2nd ed.). Springer Verlag.

Lenzner, T., & Höhne, J. K. (2022). Measuring subjective social stratification: How does the graphical layout of rating scales affect response distributions, response effort, and criterion validity in web surveys? *International Journal of Social Research Methodology*, *25*(2), 269–275.

Longo, R., Pearson, W., Merry, A., Solomon, M., Nicolotti, L., Westmore, H., Dambergs, R., & Kerslake, F. (2020). Preliminary study of Australian pinot noir wines by colour and volatile analyses, and the Pivot© profile method using wine professionals. *Food*, *9*(9), 1142.

Mather, G. (2022). *Foundations of sensation and perception* (4th ed.). Psychology Press. https://doi.org/10.4324/9781003335481

Meiselman, H. L., Jaeger, S. R., Carr, B. T., & Churchill, A. (2022). Approaching 100 years of sensory and consumer science: Developments and ongoing issues. *Food Quality and Preference*, *100*, 104614.

Menold, N. (2020). Rating-scale labeling in online surveys: An experimental comparison of verbal and numeric rating scales with respect to measurement quality and respondents' cognitive processes. *Sociological Methods & Research*, *49*, 107–179.

Menon, V., & Muraleedharan, A. (2020). Internet-based surveys: Relevance, methodological considerations and troubleshooting strategies. *General Psychiatry*, *33*(5), e100264.

Oliveira, D., Galhardo, J., Ares, G., Cunha, L. M., & Deliza, R. (2018). Sugar reduction in fruit nectars: Impact on consumers' sensory and hedonic perception. *Food Research International*, *107*, 371–377.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reinbach, H. C., Giacalone, D., Ribeiro, L. M., Bredie, W. L. P., & Frøst, M. B. (2014). Comparison of three sensory profiling methods based on consumer perception: CATA, CATA with intensity and napping®. *Food Quality and Preference*, *32*(2), 160–166.

Thomson, D. M. H. (2016). Conceptual profiling. In H. Meiselman (Ed.), *Emotion measurement* (pp. 239–272). Woodhead Publishing.

Vidal, L., Ares, G., Hedderley, D. I., Meyners, M., & Jaeger, S. R. (2018). Comparison of rate-all-that-apply (RATA) and check-all-that-apply (CATA) questions across seven consumer studies. *Food Quality and Preference*, *67*, 49–58.

Vidal, L., Ares, G., & Jaeger, S. R. (2021). The perceived intensity of sensory attributes can be indirectly obtained using CATA questions with a group of respondents. *Journal of Sensory Studies*, *36*(5), 12695.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jaeger, S. R., Chheang, S. L., Jin, D., Ryan, G. S., & Ares, G. (2023). How do CATA questions work? Relationship between likelihood of selecting a term and perceived attribute intensity. *Journal of Sensory Studies*, e12833. https://doi.org/10.1111/joss.12833