International Journal of Research in Marketing 42 (2025) 298–315

Contents lists available at ScienceDirect

International Journal of Research in Marketing

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

A method for measuring consumer confusion due to lookalike labels

Martin Schoemann [a,*,1], Piet van de Mosselaar [b,1], Sonja Perkovic [b], Jacob L. Orquin [b,c]

[a] Department of Psychology, Technische Universität Dresden, Germany
[b] Department of Management, Aarhus University, Denmark
[c] Centre for Research in Marketing and Consumer Psychology, Reykjavik University, Iceland

ARTICLE INFO

Article history:
First received on 4 July 2022 and was under review for 7 months
Available online 29 August 2024

Area Editor: Ralf van der Lans
Accepting Editor: David A. Schweidel

Keywords:
Consumer confusion
Copycatting
Mouse-tracking
Packaging
Third-party certification
Greenwashing

ABSTRACT

We propose that some products carry labels that mimic the features of certified health and sustainability labels and that such lookalike labels can confuse consumers into believing that a product has specific, desirable attributes. To address this, we develop a mouse-tracking method for measuring consumer confusion about product attributes. In Study 1, we show that lookalike labels often mislead consumers into believing a product includes a certified label, and that mouse cursor movements provide insights into confusion levels. By applying signal-detection theory to mouse cursor movements, we develop a novel metric that quantifies product attribute confusion and accurately flags products as either "attribute confusion suspect" or "attribute confusion safe". In Study 2, we replicate our findings and show that attribute confusion is associated with a higher willingness-to-pay. In Study 3, we test the robustness of the metric under different exposure times. The novel attribute confusion metric provides marketers, policymakers, and consumer advocacy groups with a tool to design less confusing labels and can serve as evidence in cases where a product is suspected of misleading consumers or copycatting certified or trademarked labels.

1. Introduction

Brand copycatting is a well-known marketing phenomenon where consumers are misled into believing they are purchasing the original brand (Miaoulis & D'Amato, 1978). We propose that companies can also mislead consumers into believing a product has specific desirable attributes, a concept we term *attribute confusion*. This strategy can benefit marketers if consumers mistakenly perceive a product as more healthful or sustainable. Marketers can cause attribute confusion by using *lookalike labels* that mimic third-party certified labels. These labels are trusted by consumers as guarantees of strict standards, such as sustainable farming or responsible sourcing (Wu et al., 2021). This practice may undermine the value of certified labels by diverting purchases to products with lookalike labels. Although existing methods measure consumer confusion due to copycat brands, none specifically address attribute confusion. To fill this gap, we developed a novel attribute confusion metric using mouse cursor tracking. By integrating signal-detection theory with mouse cursor movements, our metric quantifies product attribute confusion at the individual level, effectively flagging products as either "attribute confusion suspect" (i.e., they create attribute confusion) or "attribute confusion safe" (i.e., they do not). Compared to other

* Corresponding author.
E-mail address: martin.schoemann@tu-dresden.de (M. Schoemann).
[1] Joint first authors.

https://doi.org/10.1016/j.ijresmar.2024.08.010
0167-8116/

metrics based on cursor, choice, or response-time data, the attribute confusion metric offers higher accuracy and more detailed individual-level information. Unlike choice data, which averages across groups, the attribute confusion metric captures individual differences, revealing that attribute confusion can increase willingness-to-pay for uncertified products by about 5 % between the least and most confused consumers.

The attribute confusion metric can help marketers, policymakers, and consumer advocacy groups design product labels that minimize confusion with existing labels, especially when introducing new ones. It also offers valuable insights for identifying deceptive labeling practices and detecting greenwashing, where companies create confusion about a product's sustainability through misleading images, claims, or labels.

1.1. Why lookalike labels may cause attribute confusion

Certified labels signal premium product attributes like sustainability or healthfulness through their presence on a product (Newman et al., 2014). Consumers generally recognize and trust these labels, which enhances a product's perceived value (Wu et al., 2021). For example, products with certified nutrition labels, such the Nordic Keyhole, obtain a price premium of about 14 % (Edenbrandt et al., 2018), while organic labels can add a 30 % premium, depending on the category (Crowder & Reganold, 2015; van Doorn & Verhoef, 2011). However, certification might be challenging or require substantial investments and license fees (e.g., the Fairtrade label; FLOCERT, 2021b). In the absence of regulation, marketers can design their own labels that mimic elements of certified labels, such as color, shape, or size. These lookalike labels could yield similar benefits to certified labels if consumers confuse them with certified labels (for examples of such labels, see Fig. 1).

Consumers are likely to confuse lookalike labels because they do not process product information in detail. On average, they spend only 0.36–0.55 s attending to each product in a supermarket, which is insufficient for a detailed visual inspection of labels or packaging attributes (Hurley et al., 2013). Research shows consumers can quickly and accurately detect certified labels using their peripheral vision that is, without direct gaze (Perkovic et al., 2023). Despite the lower resolution and color sensitivity of peripheral vision (Strasburger et al., 2011), people can still recognize colors far from the point of gaze (Webster et al., 2010). Additionally, they can identify shapes and sizes in their peripheral vision, though this perception is often supplemented by information gathered through central vision (Eymond et al., 2020). Consequently, consumers may confuse a lookalike label with a certified one if the colors are similar, and they expect seeing the certified label. Even when directly gazing at a label, consumers may still be influenced by their expectations and familiarity with certified labels, leading to confusion if the lookalike label's colors, shape, and size closely resemble the certified one (Summerfield & Egner, 2009).

1.2. Measuring attribute confusion

Lookalike labels may harm consumer well-being if they lead to false beliefs about a product's desirable attributes, such as sustainability or health benefits. Additionally, they may undermine competition by diluting the equity of certified labels (Morrin & Jacoby, 2000). Despite these risks, lookalike labels have been overlooked in both public discourse and scientific research. A key barrier is the absence of a method to measure product attribute confusion. Existing research on consumer confusion identifies three sources of confusion: *information similarity*, *overload*, and *ambiguity* (Mitchell & Papavassiliou, 1999). Overload and ambiguity can delay or hinder consumer choice due to motivational issues or cognitive limitations. In contrast,

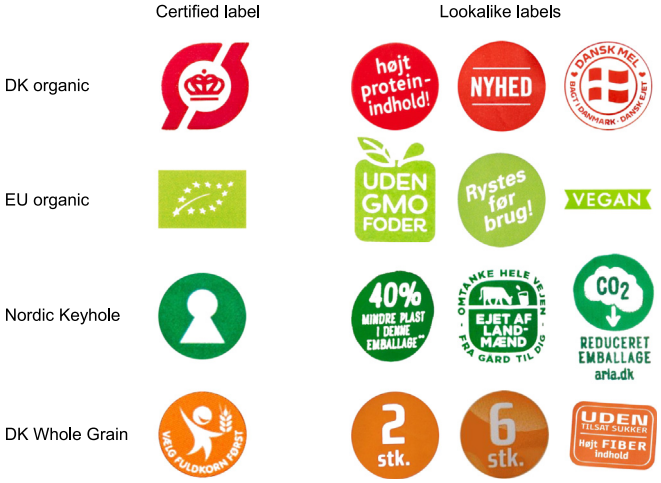

Fig. 1. Four certified labels and examples of lookalike labels. While the lookalike labels make valid claims (e.g., high protein content), they may confuse consumers due to their visual resemblance to the certified labels. The Danish organic and EU organic labels signal that a product is organic, the Nordic Keyhole label signals that a product is among the most healthful in its category, and the Danish Whole Grain label signals that a product is made from whole grains and meets some requirements shared with the Nordic Keyhole label.

similarity confusion occurs when consumers falsely believe different products are functionally similar due to their visual resemblance (Foxman et al., 1992), increasing the risk of unintentional purchases that lack the desired attributes (Walsh & Mitchell, 2005). While similarity confusion refers to false beliefs about an entire product, we define attribute confusion as a specific type of similarity confusion, where consumers mistakenly credit a specific desirable attribute to a product.

Different methods have been proposed to measure similarity confusion due to copycat brands (for an overview, see Mitchell & Kearney, 2002). The earliest study, by Friedman (1966) used a shopping task where participants selected the most economical package in a simulated supermarket, measuring confusion based on their selection errors. Later studies often relied on perceptual tasks to estimate the likelihood of consumers confusing different products. For instance, Loken et al. (1986) used a perceptual rating task where participants, along with providing a confidence rating, indicated whether two products displayed for about 7 s were made by the same or different companies. Confusion was measured using an index derived from their choices and confidence ratings. More recent studies have used self-report methods like questionnaires (e.g., Walsh et al., 2007). However, a key limitation of these methods is the extended exposure to brand packaging, lasting several seconds or minutes, which does not accurately reflect typical consumer behavior (Hurley et al., 2013).

As an early attempt to address this issue, Kapferer (1995) adapted the *tachistoscopic method*, where participants were asked to name brands of products shown at varying exposure times between 0.1 and 1 s. Kapferer created a brand confusion index by dividing the probability of identifying the original brand when a copycat was shown by the probability of identifying the original brand itself. A more recent approach, the *triangle test* by Satomura et al. (2014) exposes consumers to three products for 0.3 s—two identical original products and one visually similar copycat brand—and requires them to identify the odd one out. Satomura et al. (2014) used a drift–diffusion model to analyze both perceptual decisions and response times, leveraging the fact that more difficult decisions, where consumers are uncertain about the odd product, take longer. Consequently, even when consumers correctly identify the copycat brand as different, their longer response times may indicate higher confusion. To accurately measure attribute confusion due to elements like lookalike labels, brief exposure times, as used in the tachistoscopic method (Kapferer, 1995) or the triangle test (Satomura et al., 2014), may be necessary. However, while the triangle test measures confusion between two products, attribute confusion concerns beliefs about a single product. The tachistoscopic method could, in principle, measure this, but it has a key limitation: It requires responses from many consumers, assuming uniform decision-making across individuals, which may obscure important differences. To overcome this, attribute confusion must be measured at the individual level for each specific product.

Mouse cursor tracking is an underutilized method well-suited for studying attribute confusion at the individual consumer and product level. In this approach, participants typically view a stimulus (e.g., a product) briefly ($0.1 - 0.5$ s) and respond quickly by moving the cursor to one of two choice options (e.g., certified label: yes or no) at the top corners of the computer screen (Schoemann et al., 2021). Beyond binary choice data, cursor movements reveal unconscious decision-making processes as they unfold (Scherbaum & Dshemuchadse, 2019). Even when consumers correctly identify that a lookalike-labeled product is not certified, cursor movements can indicate whether the decision was easy (straight trajectory, low confusion) or difficult (curved trajectory, high confusion) (Stillman et al., 2020). Mouse-tracking research has identified five trajectory prototypes (Wulff et al., 2019), including straight trajectories, smoothly curved trajectories, and those with one or more sudden changes in direction. These trajectory prototypes are considered signatures of specific decision processes, and we expect them to reflect attribute confusion.

1.3. Outline

This article presents three preregistered studies using mouse cursor tracking to measure attribute confusion about whether products carry certified labels. In Study 1, we show that attribute confusion is reflected in both choices and cursor movements.

Using signal-detection theory, we develop a novel metric that quantifies attribute confusion and accurately flags products as either "attribute confusion suspect" or "attribute confusion safe". In Study 2, we test this attribute confusion metric on a new sample, replicate results from Study 1, and demonstrate its ability to reveal associations between attribute confusion and other marketing constructs like willingness-to-pay. In Study 3, we replicate Study 1 with varying product exposure times to test the metric's robustness. Finally, we discuss our findings, implications, and future research directions.

2. Study 1: A mouse-tracking method for measuring attribute confusion

In this initial study, we tested whether consumers confuse products with lookalike labels for products with certified labels using mouse cursor tracking. Previous research has shown that participants can discern visually dissimilar brands within just 0.3 s, but they often confuse visually similar ones (Satomura et al., 2014). Accordingly, we expected that participants would confuse lookalike labels with the certified labels they mimic. As noted, cursor movements during decision-making can provide insights into the cognitive process (Schoemann et al., 2021) and help distinguish between easy and difficult decisions (Stillman et al., 2020). A key measure of decision difficulty is the degree to which the cursor trajectory deviates from the direct path to a correct response, with larger deviations typically indicating more challenging decisions. Therefore, we expected that participants' cursor movements would exhibit larger deviation from the direct path when responding correctly to lookalike-labeled products compared to those with neither lookalike nor certified labels. However, larger deviations may also indicate qualitatively different decision processes (Wulff et al., 2019), highlighting the importance

of distinct patterns in cursor trajectories. We propose that consumers facing lookalike labels undergo two sequential processes: an automatic process, likely driven by covert attention (Perkovic et al., 2023), which may bias them toward attribute confusion, followed by a more deliberate process that may correct this bias (Stillman et al., 2018). This is evident in so-called change-of-mind trajectories (Atiya et al., 2020), where the cursor initially moves toward an incorrect response before abruptly shifting to the correct one. Therefore, we expected a higher proportion of change-of-mind trajectories when participants encounter lookalike-labeled products compared to those with neither lookalike nor certified labels.

2.1. Method

2.1.1. Open practices

The study protocol, hypotheses, and analyses were preregistered and executed as planned without deviations.[1] An a priori *safeguard power analysis* (Perugini et al., 2014), based on an effect estimate from 14 pilot participants with d = 0.99, 90% CI = [0.42, 1.52], statistical power of 95%, and a type I error of 5 %, indicated that a sample of 63 participants was required. The study received ethical approval from the IRB at Aarhus University (2021–05).

2.1.2. Participants and design

We recruited 89 participants (63 females, M = 24.69 years, SD = 2.81 years, range = 20 – 33 years) from the university's participant pool for an online study, accounting for potential exclusions as specified in our preregistration. To ensure familiarity with the chosen products and certified labels, participants were native Danish speakers residing in Denmark. Participants were required to use a computer mouse to ensure the quality of the cursor data. They also passed a color perception test prior before participating. Participants were compensated 50 DKK (approx. 8 USD) for their participation in the 15-min study.

The study used a one-factorial, within-subjects design. Participants were shown images of 96 unique food products obtained from Danish online supermarkets. Each product image was presented for 0.3 s, and participants were asked to quickly and accurately judge whether the product carried a predefined certified label (see Fig. 1 for an overview of labels included in the study and Fig. 2 for an illustration of the procedure).

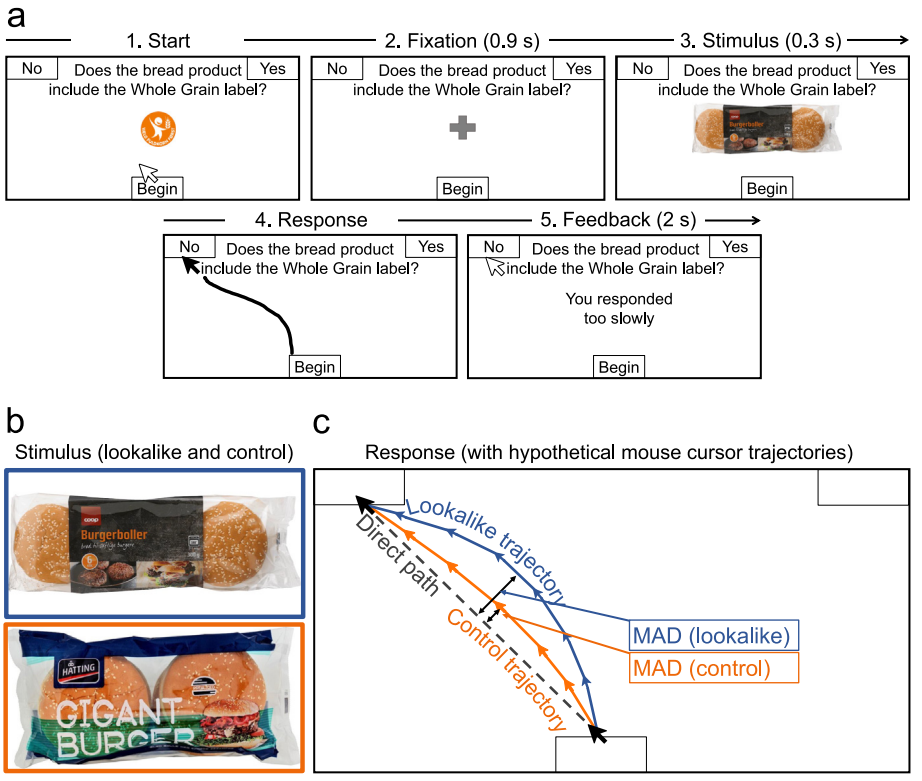

Fig. 2. a) Example of the mouse-tracking procedure. The certified label was displayed only during the initial trial of each block. b) Examples of lookalike-labeled (blue) and control products (orange). c) Hypothetical examples of cursor trajectories for lookalike trials (blue), control trials (orange), and their maximum absolute deviations (MADs) from the direct path (dashed line).

[1] https://osf.io/ym8pg/.

2.1.3. Stimuli

96 unique suitable food products were manually selected from a pool of 1,326 products in the bakery and dairy categories, sourced from the online assortments of Denmark's two leading retail chains. The 96 selected products were organized into 24 sets, each containing four closely matched products from the same category. These 24 sets were divided into four blocks corresponding to the four certified labels (i.e., DK organic, EU organic, Nordic Keyhole, DK Whole Grain; Fig. 1). The order of the blocks and the products within each block was randomized. Within each of the 24 sets, products were carefully selected so that one carried the certified label of interest, one carried a relevant lookalike label, and two control products carried neither certified nor lookalike labels relevant to that block and condition (Fig. 1b). For more details on the stimuli, please see Web Appendix A.

The study was designed in PsychoJS (i.e., the online version of PsychoPy 2020.2; Peirce et al., 2019) and hosted on Pavlovia.org. The study used a standard mouse-tracking setup, with stimuli (900 px × 900 px) presented at the center of the screen, a start button at the bottom center, and two response buttons located in the upper corners (Schoemann et al., 2021, Fig. 2c,). The allocation of the response options, *yes* and *no*, to the response buttons was randomized across participants.

Cursor movements were tracked from the beginning of the stimulus stage until the end of the response stage, with samples taken at each screen refresh. Due to the online setup, cursor speed information was unavailable (cf. Grage et al., 2019).

2.1.4. Procedure

Each trial comprised five consecutive stages (Fig. 2a): start, fixation, stimulus, response, and feedback. In the start stage, participants clicked the start button after a 0.5-s delay (cf. Schoemann et al., 2019). The predefined certified label was displayed only during the initial trial of each block at this stage. In the fixation and stimulus stages, participants were instructed to refrain from moving the mouse cursor while the fixation cross (0.9 s) and the product (0.3 s) were displayed sequentially, with the mouse cursor remaining invisible. The response stage required participants to initiate a response within 0.6 s (i.e., movement initiation deadline) by moving the cursor upward from the start button and to complete the response within 2.5 s (i.e., response deadline) by clicking one of the response buttons (cf. Schoemann et al., 2021). The feedback stage, lasting for 2 s, provided feedback on participants' timing or showed a blank screen if timing thresholds were met. After completing all 96 trials, participants assessed their familiarity with and the importance of each of the four certified labels using slider scales. A questionnaire was administered following the mouse-tracking task to gather insights into participants' observations and their understanding of the study's aim.

2.1.5. Data processing

Data (pre-)processing and statistical analyses were conducted in R (R Core Team, 2020). We used the *lme4* package (Bates et al., 2015) for the (generalized) linear mixed-effects modeling and the *mousetrap* package (Kieslich et al., 2019) for handling the cursor data. We excluded 22 participants who failed to meet the movement initiation deadline or the response deadline in more than one-third of all trials, or who responded incorrectly in more than one-third of the certified trials. This left 67 participants (48 females, M = 24.46 years, SD = 2.60 years, range = 20 − 33 years) for all analyses. The familiarity ratings and importance ratings for the four certified labels among all remaining participant are reported in Web Appendix D.

For the analysis of participants' choice data, we included all available trials (n = 6,432), including those not completed in full-screen or where participants failed to meet one of the timing thresholds. Conversely, for the analyses of participants' cursor data, we included only trials with correct choices and excluded those not completed in full-screen (n = 84) and those with timing inconsistencies (n = 866), yielding a total of 5490 trials. To standardize the cursor data, we mapped the *no* response option to the upper left response button and the *yes* response option to the upper right response button. Consequently, cursor trajectories for correct choices ended in the top-left (top-right) corner for lookalike and control (certified) trials (Fig. 2c). We aligned the trajectories' start and end positions, ignoring the initiation phase (when the cursor was static at the onset) and click phase (when the cursor had landed on the final response button; Kieslich et al., 2019). We operationalized the cursor trajectories' deviation as the maximum absolute deviation (*MAD*) from an ideal straight line between respective start and end positions (Fig. 2c; Kieslich et al., 2019). To analyze whether cursor trajectories deviate more from a direct path when dealing with lookalike-labeled products, we temporally normalized cursor trajectories to 101 points (Spivey et al., 2005). To move beyond aggregate cursor trajectories and assess whether there is an increase in change-of-mind trajectories when a product carries a lookalike label, we spatially normalized the trajectories to 20 points and mapped them to a set of predefined trajectory prototypes based on minimizing the Euclidean distance between measured and prototypical trajectories (Wulff et al., 2019). This set included five prototypes: straight, curved, continuous-change-of-mind (cCoM), discrete-change-of-mind (dCoM), and discrete-double-change-of-mind (dCoM2) trajectories (see Fig. 3c, right panel). These prototypes can be ordered (from bottom to top, Fig. 3c, right panel) according to the degree of response conflict they represent, offering more detailed information into the underlying decision processes that average trajectories may obscure (Wulff et al., 2019).

2.2. Results

First, we examined whether participants misjudged lookalike-labeled products. Our confirmatory mixed-effects regression analysis, using a binomial model with a logit link and a random intercept and slope for participants and sets, supports this notion (Fig. 3a). Compared to lookalike trials, the probability of responding "yes" that a product includes a predefined

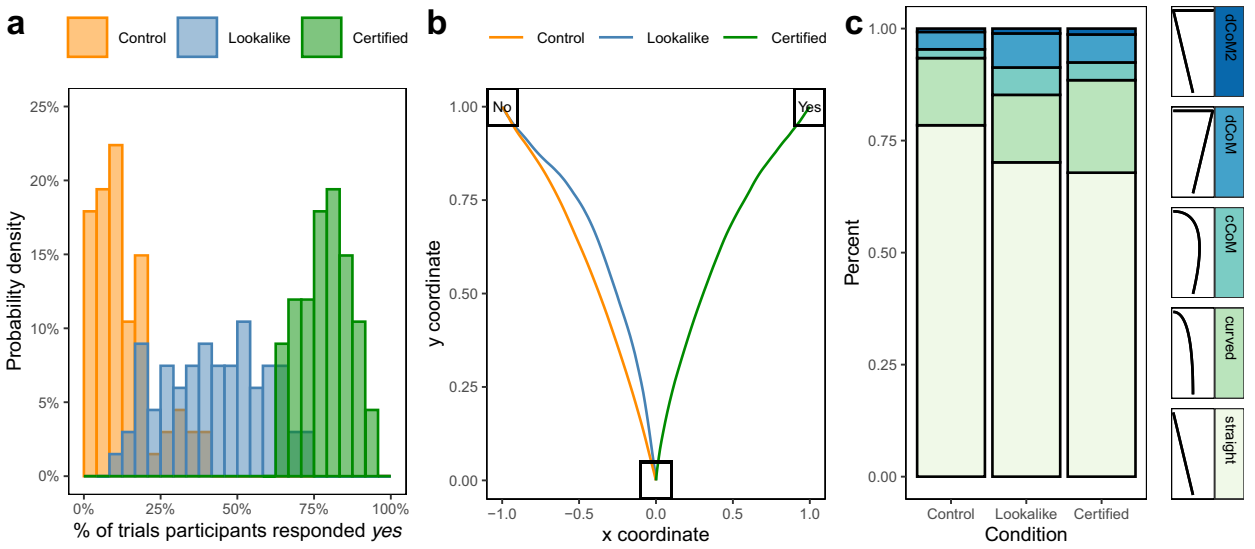

Fig. 3. a) Histograms depict the proportion of trials in which participants responded that the product carries the predefined certified label for each condition (color). b) Lines represent the average cursor trajectories for each condition (color). c) Stacked bar charts illustrate the cumulative proportions of different trajectory types (color; right panel) for each condition (left panel).

certified label was lower in control trials, $b = -1.99$, $SE = 0.26$, $p < 0.001$, and higher in certified trials, $b = 1.90$, $SE = 0.26$, $p < 0.001$. Specifically, the estimated probabilities for "yes" responses were 12.03% in control trials, 43.62% in lookalike trials, and 86.95% in certified trials.[2]

Second, we examined whether participants' cursor trajectories deviated more toward an incorrect response in lookalike trials compared to control trials when they ultimately made a correct response. As expected, our confirmatory mixed-effects regression analysis, using a linear model with a random intercept and slope for participants and sets, shows that the deviation toward the incorrect response was significantly larger in lookalike trials than in control trials, $b = -0.13$, $SE = 0.04$, $p = 0.001$ (Fig. 3b). A directed paired-samples t-test on z-scored (within participants) MAD, mirroring our a priori power analysis, further confirms this result. Average cursor trajectories deviated more from the direct path in lookalike trials, $M = 0.29$, $SD = 0.41$, than in control trials, $M = -0.09$, $SD = 0.15$, $t(66) = 5.94$, $p < 0.001$, $d_z = 0.73$, 90% CI = [0.52, 0.96]. This effect size falls within the 90% confidence interval determined from our pilot participants, which was CI = [0.42, 1.52].[3]

Third, we examined whether participants exhibited more changes-of-mind from an initially incorrect response to a correct response in lookalike trials than in control trials. As expected, a confirmatory X^2-test shows significant differences in the proportion of trajectory types between lookalike and control trials, $X^2(4) = 62.28$, $p < 0.001$ (Fig. 3c).[4] Indeed, lookalike trials presented fewer straight trajectories, $\text{std.res}_{\text{straight}} = -6.58$, and more change-of-mind trajectories, $\text{std.res}_{\text{cCoM}} = -4.25$, $\text{std.res}_{\text{dCoM}} = 5.73$) than control trials, particularly for the discrete-change-of-mind trajectories (dCoM, second from the top, Fig. 3c right panel). Hence, the larger deviation in average cursor trajectories during lookalike trials is due to differences in the proportion of trajectory types, supporting the notion of a two-system process where participants first complete a biased automatic process, followed by a correcting deliberate process.

2.3. Discussion

In Study 1, we use mouse cursor tracking to examine whether products with lookalike labels cause attribute confusion. Our results show that after just 0.3 s of exposure, participants accurately detected certified labels. This aligns with previous research showing that people can perceive the gist of an ad in a single fixation (Pieters et al., 2012) or distinguish between dissimilar brands in 0.3 s (Satomura et al., 2014). In contrast, the presence of a lookalike label often led participants to incor-

[2] An exploratory analysis, using the same exclusion parameters at the trial level as in the mouse-tracking analyses (i.e., excluding trials not completed in full screen and those with timing inconsistencies), yielded similar results: Compared to lookalike trials, the probability of "yes" responses was lower in control trials, $b = -2.18$, $SE = 0.28$, $p < 0.001$, and higher in certified trials, $b = 1.93$, $SE = 0.30$, $p < 0.001$ The estimated probabilities for "yes" responses were 10.20% in control trials, 43.95% in lookalike trials, and 87.28% in certified trials.

[3] An exploratory mixed-effects regression analysis revealed no significant difference in deviation toward incorrect responses between certified and lookalike trials, $b = -0.05$, $SE = 0.04$, $p = 0.228$.

[4] An exploratory ordinal mixed-effect regression with random intercept and slope for participants and sets corroborated the X^2-test result, showing that the trajectory types produced during control trials exhibited lower levels of response conflict compared to those produced during lookalike trials, $b = -0.82$, $SE = 0.19$, $p < 0.001$. This occurs when the five predefined trajectory prototypes are arranged according to the level of response conflict they depict. (from bottom to top, Fig. 3c right panel; Wulff et al., 2019).

rectly respond that the product carried a certified label (Fig. 3a). Furthermore, even when participants did not misjudge the labels, our mouse-tracking analysis revealed more pronounced curvatures toward the incorrect response in lookalike trials, suggesting increased decision-making difficulty. This difference in curvature (Fig. 3b) was primarily driven by a higher prevalence of changes-of-mind trajectories in lookalike trials (Fig. 3c). This suggests that participants often initially confused a lookalike-labeled product with a certified one but were sometimes able to correct their judgments. In the post-experimental questionnaire, about 10 out of 89 participants reported noticing visual similarities between lookalike-labeled and certified-labeled products. For instance, one participant reported confusion between labels featuring the Danish flag and the DK organic label, while another mistook product quantity descriptions for the DK Whole Grain label (Fig. 1).

3. Developing an attribute confusion metric

Study 1 provides evidence that lookalike labels confuse consumers into believing that products carry certified labels. In this section, we propose a metric to quantify levels of attribute confusion and identify products that cause such confusion due to lookalike labels.[5] We propose that consumer attribute confusion caused by lookalike labels can be formalized as a signal-detection problem (Green & Swets, 1966), where consumers receive noisy perceptual information about product attributes (presence of certified label: yes vs. no) and must determine whether the product carries a certified label (response: yes vs. no). In signal-detection terminology, when a product carries a certified label and the consumer correctly identifies it (i.e., responds "yes"), this is referred to as a *hit*. If the consumer fails to recognize the certified label (i.e., responds "no"), this is called a *miss*. Conversely, when a product does not carry a certified label (i.e., lookalike and control trials) and the consumer incorrectly identifies it as carrying one (i.e., responds "yes"), this is called a *false alarm*. A *correct rejection* occurs when the consumer correctly identifies the absence of a certified label (i.e., responds "no"). The probabilities of hits and false alarms, known as the *hit rate* and *false-alarm rate*, respectively, fully describe decision performance (Fig. 4a; Stanislaw & Todorov, 1999).

Kapferer (1995) applied this framework to develop an early brand confusion index by calculating the ratio of the false-alarm rate (i.e., mistakenly identifying a copycat brand as the original brand) to the hit rate (i.e., correctly identifying the

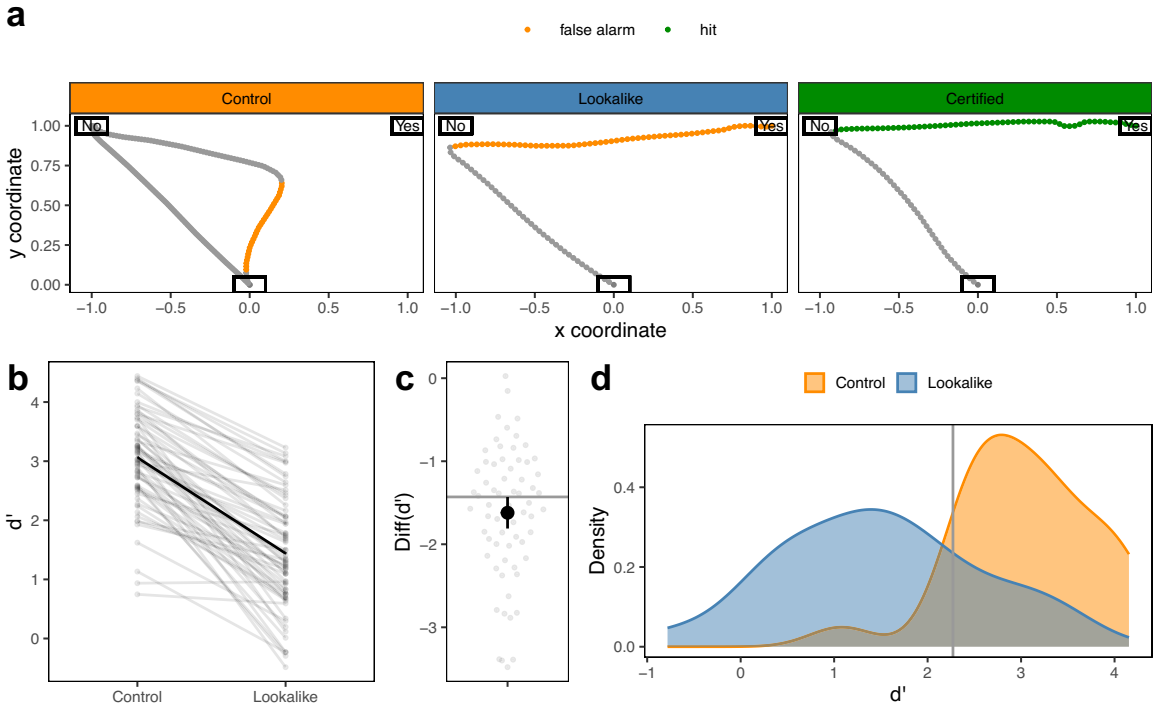

Fig. 4. a) Example of how cursor trajectories are segmented into 100 discrete steps, each classified as hits or false alarms (color). b) Within-subjects effect on sensitivity (d') averaged within participants and across sets (grey; black = grand mean) for control and lookalike-labeled products. c) Within-subjects effect on sensitivity (d') as the difference between the comparison conditions. The error bar depicts the 95 % CI of the mean, and the horizontal line depicts its upper boundary, which is used in Study 2 and Study 3 to evaluate replication success (see Lakens et al., 2018). d) Density distribution of sensitivity (d') averaged within sets and across participants for control and lookalike-labeled products (color). The vertical line depicts the classification threshold that maximizes classification accuracy.

[5] The following analyses were not part of our preregistration.

original brand). However, Kapferer's confusion index has two caveats. First, the ratio of the false-alarm and hit rate inherently reflect both *response bias* (i.e., a consumer's general tendency to respond "yes" or "no") and *sensitivity* (i.e., a consumer's ability to discriminate between copycats and originals). Additionally, the ratio varies on a logarithmic scale, making it highly sensitive to extreme rates (i.e., values close to zero or one). Second, calculating false-alarm and hit rates for this index requires a substantial number of within- or between-subjects responses to the same stimuli, limiting the feasibility of single observation designs. Single observation designs are often necessary in marketing research when marketers need to assess the impact of an individual advertisement (Cradit et al., 1994) or product (Bloch, 1995), requiring aggregation of responses across consumers. This aggregation can be problematic, as it assumes a uniform decision-making process across consumers and may overlook important individual differences.

To address these caveats, we propose an attribute confusion metric by computing two signal-detection parameters, sensitivity and response bias, based on cursor movements. We address the first caveat by estimating these signal detection parameters, allowing us to determine whether systematic differences in false-alarm rates between lookalike and control trials are due to variations in sensitivity, response bias, or both (Stanislaw & Todorov, 1999). Moreover, these parameters vary on an equal-interval scale, ensuring that changes in scores are directly proportional to changes in the underlying psychological process (Wixted, 1990), making differences between conditions easier to interpret (Loftus, 1978).

The second caveat concerns the need for a substantial number of responses. In experimental psychology, signal detection parameters have traditionally been calculated by aggregating discrimination performance across stimuli to study individuals. In contrast, marketers typically study stimuli by aggregating responses across individuals. This difference in focus leads to differences in the available data. Calculating false-alarm and hit rates requires multiple responses per individual, which is usually feasible in experimental psychology because psychologically equivalent stimuli can be presented multiple times. However, marketing research often focuses on a unique stimulus and must rely on a single response per individual and stimulus, thus frequently adhering to single observation designs. The meaningful calculation of signal detection parameters depends on the assumption that all data (i.e., the false-alarm and hit rates) result from a consistent decision rule. However, individual consumers are likely to differ in both sensitivity and response bias parameters. Therefore, these parameters should be estimated for individual consumers across stimuli, rather than for individual stimuli across consumers (Macmillan & Kaplan, 1985). Even if it were feasible to estimate the parameters for individual stimuli across consumers under certain conditions (Cradit et al., 1994), this would require observations from at least 100 consumers (Kadlec, 1999). We address the second caveat by using cursor data, which enables the estimation of signal detection parameters from a single response per individual and unique stimulus. As demonstrated in our confirmatory analyses in Study 1, cursor data provide more observations for each response and offers additional insights into the decision-making process. For instance, a change-of-mind trajectory reveals that a consumer initially leaned toward one response but then revised the decision.

We leverage this information about the decision-making process by segmenting the cursor trajectories into 100 discrete steps. Each step is classified as a "yes" or a "no" based on the direction of the cursor trajectory at that step, that is, whether the trajectory points toward the "yes" or "no" response. For example, a change-of-mind trajectory may be segmented into 40 steps toward the "yes" response and 60 steps toward the "no" response. Conversely, a straight trajectory would produce 100 steps exclusively toward either the "yes" or "no" response. In certified trials, this means that the change-of-mind trajectory would produce 40% hits and 60% misses, while the straight trajectory would produce 100% hits or misses (Fig. 4a, right panel). In lookalike or control trials, the change-of-mind trajectory would produce 40% false alarms and 60% correct rejections, whereas the straight trajectory would produce 100% false alarms or correct rejections (Fig. 4a, left and middle panel).

In the following section, we describe the computation of these trajectory values and their application in computing the Kapferer index as well as the signal-detection parameters. Finally, we evaluate and compare how different metrics can be used to identify products with lookalike labels that cause attribute confusion ("attribute confusion suspects") versus products that do not cause confusion ("attribute confusion safe").

3.1. Confusion metrics setup

To begin, we computed the Kapferer index, k' (notation introduced by the authors; Kapferer, 1995), along with two signal-detection parameters: response bias, β, and sensitivity, d' (Green & Swets, 1966). The computation of k' is based on the false alarm rate (F) and the hit rate (H) (Kapferer, 1995), with an additional logarithmic transformation applied (introduced by the authors) to bring it onto an equal-interval scale:

$$k' = \ln \frac{F}{H}$$

The computations for both β and d' are derived from the z-scores (Φ^{-1}) of the hit and false-alarm rates (Stanislaw & Todorov, 1999):

$$\ln \beta = \frac{\left[\Phi^{-1}(F)\right]^2 - \left[\Phi^{-1}(H)\right]^2}{2}$$

$$d' = \Phi^{-1}(H) - \Phi^{-1}(F)$$

The hit rates (H) and false-alarm rates (F) were obtained from length-normalized cursor trajectories. Each trajectory was divided into 100 discrete steps, and for each step, we determined whether the trajectory approached the "yes" or the "no" response based on the change in Euclidean distance toward the two response buttons. In certified trials, we classified each step as a hit or miss depending on whether the cursor moved toward the "yes" or "no" response, respectively (Fig. 4a, right panel). In lookalike and control trials, each step was classified as a false alarm or correct rejection depending on whether the cursor moved toward the "yes" or "no" response, respectively (Fig. 4a, left and middle panels). The hit-rate was computed as the sum of hits divided by the sum of hits and misses, and the false-alarm rate was computed as the sum of false-alarms divided by the sum of false alarms and correct rejections.

Since the data comprised 24 product sets, each including a certified-label, a lookalike-labeled, and two control products with neither label, we calculated k', ln β, and d' for each participant and product set by pairing a certified-labeled product with either a lookalike-labeled product or both control products. For more details on the calculation, see Web Appendix B.

Such pairings (i.e., matched comparisons) require that each set include at least one valid trial for each condition. Therefore, we excluded any set for a participant that did not meet this requirement, leaving a total of 4,611 trials in this exploratory analysis.

3.2. Results

First, we examined which of the calculated metrics systematically varied between comparison conditions. Separate mixed-effects regression analyses, using linear models with random intercepts and slopes for participants and sets, revealed that both the Kapferer index (k') and sensitivity (d') varied between comparison conditions, whereas the log-transformed response bias (ln β) did not, $b = 0.02$, $SE = 0.12$, $p = 0.874$. The Kapferer index (k') was significantly larger when comparing certified-labeled with lookalike-labeled products than when comparing certified-labeled to control products, $b = 1.61$, $SE = 0.25$, $p < 0.001$. Sensitivity (d') was significantly lower when comparing certified-labeled with lookalike-labeled products than when comparing certified-labeled to control products, $b = -1.57$, $SE = 0.21$, $p < 0.001$ (Fig. 4b & c). The difference in effect direction is due to the logic of the two metrics: the Kapferer index increases with a higher false-alarm rate, whereas sensitivity decreases. However, the effect size is similar, as shown by comparing the standardized coefficients for the Kapferer index, $b = 0.26$, 95% CI = [0.17, 0.34], and sensitivity, $b = -0.29$, 95% CI = [−0.370.21]. Thus, lookalike-labeled products influence both the Kapferer index and sensitivity, but not the response bias, suggesting that k' and d' are useful for measuring attribute confusion and identifying lookalike labels.

Next, we examined if the Kapferer index and sensitivity could distinguish products that cause attribute confusion due to lookalike labels ("attribute confusion suspects") from control products that do not cause attribute confusion ("attribute confusion safe"). To do so, we switched the criterion and predictor variables from the previous mixed-effects regression analyses and used simple logistic regression (due to singularity issues with both participant and set as random effects). Separate logistic regression analyses revealed that both the Kapferer index, $b = 0.18$, $SE = 0.01$, $p < 0.001$, and sensitivity, $b = -0.24$, $SE = 0.02$, $p < 0.001$, significantly predicted the probability that a matched comparison included a lookalike-labeled product. To evaluate the performance of these confusion metrics compared to simpler process data, we fitted logistic regression models based on differences in response times (RT) and mean absolute deviation (MAD) between certified trials and control or lookalike trials, respectively, within each set. We then used these models to classify whether a product pair included a lookalike-labeled product and ran 100x6 fold cross-validations for each model, assessing classification performance (mean and standard deviation in classification accuracy). The results showed that the Kapferer index, $M = 61.39\%$, $SD = 0.07\%$, and sensitivity, $M = 63.49\%$, $SD = 0.08\%$, had an accuracy above chance level (50% due to the experimental design), with sensitivity performing best. The results also revealed that both confusion metrics outperformed the simpler process data (RT, $M = 51.07\%$, $SD = 0.13\%$, and MAD, $M = 52.56\%$, SD = 0.16%), which showed accuracies close to chance level. These findings indicate that sensitivity (d') provides a classification accuracy of about 64%, even with only two mouse-tracking responses measured (one for the certified-labeled and one for the lookalike-labeled product).

Third, we explored ways to improve the classification accuracy beyond 64%. To achieve this, we averaged the confusion metrics within the 24 sets and the two comparison conditions across the 67 participants. Averaging across participants allowed us to compare metrics computed from cursor data to those derived from participants' choices. Although computing confusion metrics across participants' choices is theoretically flawed, as earlier explained, it provides insight into whether the additional effort in mouse cursor tracking is justified by improved classification performance. We again fitted logistic regression models, used them to classify whether a product pair included a lookalike-labeled product, and ran 100x6-fold cross-validations for each model. The results showed that averaging confusion metrics and simpler process data substantially improved classification performance. Sensitivity, $M = 84.29\%$, $SD = 2.12\%$, showed the highest accuracy, followed by the Kapferer index, $M = 75.25\%$, $SD = 1.52\%$, MAD, $M = 73.60\%$, $SD = 0.98\%$, and RT, $M = 62.17\%$, $SD = 0.97$ %. Computing the confusion metrics across participants' final responses also improved classification performance, with comparable accuracy for both sensitivity, $M = 84.04\%$, $SD = 1.40\%$, and the Kapferer index, $M = 82.12\%$, $SD = 1.12\%$.

In sum, sensitivity (d') performed best, leading us to examine the optimal classification threshold for d' and the exact classifications of the product pairs. Using the averaged d', we conducted a grid search to find the value that maximized classification accuracy. The grid search yielded a $d' = 2.27$ with an accuracy of 87.50% (Fig. 4d). Of the 24 lookalike-labeled products, 19 were classified as "attribute confusion suspects". Of the 24 sets of control products, 23 were classified as "at-

tribute confusion safe" (for detailed classification data, see Table C1 in Web Appendix C). One set of control products was classified as "attribute confusion suspect" presumably due to the low hit rate in that set (Set 24).

3.3. Discussion

In this section, we proposed an attribute confusion metric for measuring attribute confusion and identifying "attribute confusion suspects", referring to products that cause attribute confusion due to lookalike labels. We benchmarked the metric against variations of more established metrics. Using mouse-tracking data in a label classification task, our metric computes a well-established signal-detection parameter. It outperforms comparable metrics in terms of classification accuracy and provides an easy-to-apply classification threshold. The metric indicates that 19 of the 24 products carrying lookalike labels are "attribute confusion suspects", while 23 of the 24 sets of control products carrying neither lookalike nor certified labels are "attribute confusion safe". In Set 24, both the product with a lookalike label ($d' = -0.78$) and the control products ($d' = 1.08$) had low sensitivity values, and both were categorized as "attribute confusion suspects" (for detailed images of these products, see Figure C1 in Web Appendix C). A plausible explanation is that participants found it difficult to identify the certified label on the product with the certified label in Set 24, as indicated by the low hit rate (Table C1 in Web Appendix C). Consequently, the d' values from comparisons between the lookalike-labeled and certified-labeled product, as well as between the control and certified-labeled products are likely to fall below the d' threshold that maximizes classification accuracy across 24 sets on average.

4. Study 2: Replication and association

In Study 2, we aimed to replicate the findings from Study 1 and use the attribute confusion metric's information advantage to test the product-level association between attribute confusion and willingness-to-pay. We expected the results to replicate and that attribute confusion would be associated with willingness-to-pay – specifically, that consumers would be willing to pay more for a product without a certified label when they are more confused about whether it carries the certified label.

4.1. Method

4.1.1. Open practices

The study protocol and analyses were preregistered and executed as planned, with one minor deviation in data processing, described below.[6] The sample size was determined by an a priori power analysis based on a correlation test for analyzing the correlation between paired samples of the attribute confusion metric and measures of willingness-to-pay for lookalike-labeled products. Given a directed *t*-test of a correlation, we aimed to detect a small-to-medium effect ($r = 0.30$) with 80% power and a type I error of 5%, which required 67 participants. This sample size roughly matches the sample size of Study 1, ensuring sufficient power to replicate both the confirmatory and exploratory analyses presented in Study 1. The study received ethical approval from the IRB at Aarhus University (2021–05).

4.1.2. Participants and design

We recruited 94 participants (55 females, $M = 26.12$ years, $SD = 4.51$ years, range = 21 – 48 years) from the university's participant pool, accounting for potential exclusions as defined in our preregistration. Exclusion criteria, informed consent, and compensation were identical to those in Study 1. The study employed the same task using the same one-factorial, within-subjects design as described in Study 1 (see Fig. 2 for details).

4.1.3. Stimuli, procedure, and data processing

The stimuli, procedure, and data processing were very similar to those in Study 1. The experimental task, as depicted in Fig. 2a, and the subsequent questionnaire remained the same. However, an additional rating task was included at the beginning. In this rating task, each trial comprised three stages: fixation, stimulus, and rating. During the fixation and stimulus stages, the fixation cross (0.9 s) and the product (0.6 s) were displayed sequentially, and participants were instructed to carefully process the product. In the rating stage, participants were asked how much they would be willing to pay for the displayed product. For data processing, we excluded 25 participants who failed to meet the movement initiation deadline or the response deadline in more than one-third of all trials, or who responded incorrectly in more than one-third of the certified trials. This left 69 participants (37 females, $M = 25.75$ years, $SD = 4.07$ years, range = 21 – 48 years). The familiarity and importance ratings of the four certified labels for all remaining participant are reported in Web Appendix D. For all analyses, trials that were not completed in full-screen mode ($n = 73$) and those with imperfect timing ($n = 903$) were excluded. The exclusion of trials according to the predefined criteria may have resulted in incomplete sets – sets that did not include at least one valid trial in each condition. Since complete sets are required to compute the confusion metrics, trials belonging to incomplete sets ($n = 1{,}019$) were also excluded, which is a deviation from the preregistration. This process left 4,640 out of 6,624 trials.

[6] https://osf.io/b7dyx.

4.2. Results

4.2.1. Replication

First, we examined whether we replicated the confirmatory results from Study 1 using the same statistical models. Regarding participants' choices, the probability of responding "yes" that a product includes a predefined certified label was, relative to lookalike trials, lower in control trials, $b = -1.71$, $SE = 0.25$, $p < 0.001$, and higher in certified trials, $b = 2.11$, $SE = 0.27$, $p < 0.001$. The estimated probabilities for "yes" responses were 15.38% in control trials, 42.46% in lookalike trials, and 89.16% in certified trials. Regarding participants' decision-making processes during trials in which they responded correctly, the deviation from the direct path was significantly larger in lookalike trials than in control trials, $b = -0.15$, $SE = 0.03$, $p < 0.001$. The trajectory types produced during control trials exhibited lower levels of response conflict compared to those produced during lookalike trials, $b = -0.74$, $SE = 0.15$, $p < 0.001$. In sum, we successfully replicated all confirmatory results from Study 1.

Second, we examined whether we replicated the exploratory results from Study 1 using the same statistical models and procedures regarding the measurement of attribute confusion and identification of "attribute confusion suspects". Mixed-effects regression analyses revealed that both the Kapferer index (k'), $b = 1.23$, $SE = 0.21$, $p < 0.001$, and sensitivity (d'), $b = -1.30$, $SE = 0.18$, $p < 0.001$ (Fig. 5a & b), varied between comparison conditions, whereas the log-transformed response bias ($\ln \beta$) did not, $b = 0.21$, $SE = 0.12$, $p = 0.101$. The effect sizes, as indicated by standardized coefficients for the Kapferer index, $b = 0.20$, 95% CI = [0.13, 0.27], and sensitivity, $b = -0.24$, 95% CI = [−0.31, −0.18], were very similar. Regarding the classification of "attribute confusion suspects", logistic regression analyses revealed that both the Kapferer index, $b = 0.13$, $SE = 0.01$, $p < 0.001$, and sensitivity, $b = -0.19$, $SE = 0.02$, $p < 0.001$, significantly predicted the probability that a matched comparison included a lookalike-labeled product. 100x6-fold cross-validations for each model revealed that the Kapferer index, $M = 57.65$ %, SD=0.06%, and sensitivity, $M = 61.71\%$, $SD = 0.05\%$, showed classification accuracy above chance level, with sensitivity showing the best performance and outperforming the simpler process data (RT, $M = 51.49\%$, $SD = 0.08$ %, and MAD, $M = 50.89\%$, $SD = 0.13\%$), which showed classification accuracy close to chance level. To improve the classification accuracy of around 62%, we performed 100x6-fold cross-validations with logistic regressions based on averaged data across participants. Averaging substantially improved classification performance for all metrics, with sensitivity, $M = 82.56\%$, $SD = 1.72\%$, showing the highest accuracy, followed by the Kapferer index, $M = 75.04\%$, $SD = 1.54\%$, MAD, $M = 69.06\%$, $SD = 1.20\%$, and RT, $M = 69.60\%$, $SD = 1.60\%$. The results also revealed that computing the confusion metrics across participants' choices improved classification performance, with comparable accuracy for both sensitivity, $M = 71.65\%$, $SD = 1.18\%$, and the Kapferer index, $M = 79.65\%$, $SD = 1.56\%$.

Regarding the optimal classification threshold, a grid search for the sensitivity value that maximizes classification accuracy yielded a $d' = 2.11$ with an accuracy of 85.42% (Fig. 5c). For detailed classifications of the product pairs, please refer to Table C2 in Web Appendix C, which shows that 20 out of 24 products with manually identified lookalike labels were classified as "attribute confusion suspects", and 22 out of 24 sets of control products were classified as "attribute confusion safe". Two sets of control products were classified as "attribute confusion suspects" (Set 4 & Set 24).

4.2.2. Association between attribute confusion and willingness-to-pay

Next, we examined whether our attribute confusion metric is associated with willingness-to-pay (in DKK; all products: $M = 17.30$, $SD = 8.99$; certified-labeled products: $M = 17.61$, $SD = 8.53$; lookalike-labeled products: $M = 17.60$, $SD = 9.52$; control products: $M = 16.98$, $SD = 8.93$) through correlational analyses. A mixed-effects regression analysis, using a linear model with

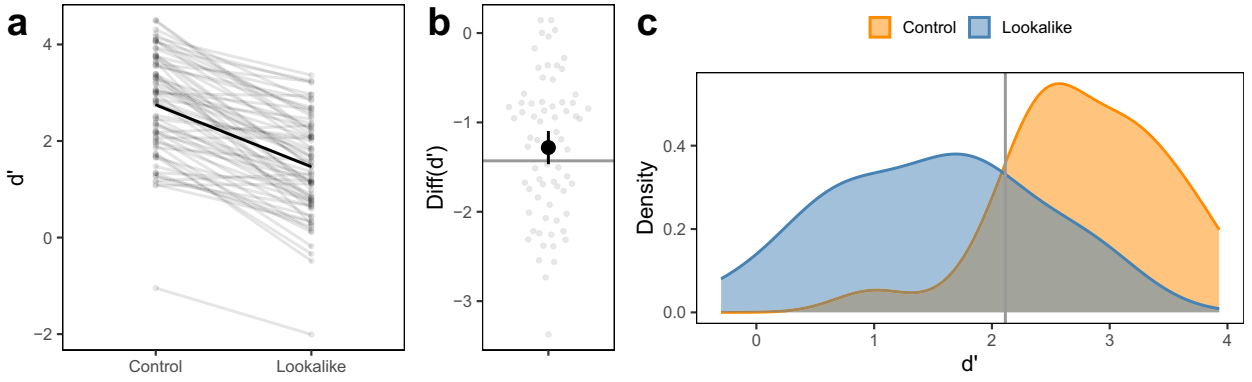

Fig. 5. a) Within-subjects effect on sensitivity (d'), averaged within participants and across sets (grey; black = grand mean), for control and lookalike-labeled products. b). Within-subjects effect on sensitivity (d') as the difference between comparison conditions. The error bar depicts the 95% CI of the mean, with the horizontal line representing the upper boundary from Study 1 for comparison. c) Density distribution of sensitivity (d'), averaged within sets and across participants, for control and lookalike-labeled products (color). The vertical line indicates the classification threshold that maximizes classification accuracy.

random intercepts for participants and sets, revealed that both comparison condition, $b = 0.49$, $SE = 0.22$, $p = 0.030$, and sensitivity (d'), $b = -0.13$, SE=0.05, $p = 0.004$, significantly predicted willingness-to-pay for the respective product, without any meaningful interaction. Thus, on average, participants were willing to pay 3% (0.50 DKK or $0.07) more for products featuring a lookalike label than for control products. Additionally, highly confused participants (i.e., $d' = 0$) were willing to pay 5% (0.80 DKK or $0.12) more for control or lookalike-labeled products compared to participants who were not confused (i.e., $d' = 6$).

4.3. Discussion

In Study 2, we replicated the findings from Study 1 and confirmed the product-level association between the attribute confusion metric and willingness-to-pay. The metric again outperformed comparable metrics in terms of classification accuracy. It identified 20 of the 24 lookalike-labeled products as "attribute confusion suspects" and 22 of the 24 set of control products as "attribute confusion safe". In Set 4 and Set 24, the control products were categorized as "attribute confusion suspects", likely due to low hit rates for the products with certified labels (for details, see Table C2 in Web Appendix C). Beyond replicating the findings from Study 1, we demonstrated the additional value of the attribute confusion metric in predicting other consumer behavior measures. Specifically, we showed that the attribute confusion metric predicts willingness-to-pay: The more confused a participant is about a control or lookalike-labeled product, the more they are willing to pay for that product. It would not be possible to demonstrate a similar association using choice data, as this would require data aggregation across participants.

5. Study 3: Effect of exposure time on consumer confusion

In Study 3, we aimed to once more replicate the main findings from Study 1 regarding our attribute confusion metric and test its robustness under varying exposure times. We expected the results to be consistent across different exposure times. However, with sufficient exposure time, consumers may be better equipped to accurately judge whether a product carries a certified label, potentially causing the metric to fail in detecting confusion due to lookalike labels. We replicated the task from Study 1, with the exception that we varied the exposure time for product images (0.3 s, 0.6 s, 0.9 s, or 1.2 s) between participants.

5.1. Method

5.1.1. Open practices

The study protocol and analyses were preregistered and executed as planned, with one minor deviation in data processing, which is described below.[7] This deviation altered some numbers in the power analyses but did not affect the overall rationale; the updated numbers are presented below. The sample size was determined by an a priori safeguard power analysis (Perugini et al., 2014), based on an inferiority test (Lakens et al., 2018), to analyze the difference in sensitivity (d') between lookalike and control trials compared to the effect found in Study 1, $M_{Diff} = -1.62$, $SD = 0.79$, $SE = 0.10$, $d_z = -2.06$, 95% CI = [−2.50, −1.78] (Fig. 4c). The inferiority test seeks to reject the null hypothesis that, within each level of exposure time, the difference is equal to or lower than $M_{Diff} = -1.62 + 1.96 * 0.10 = -1.43$. This implies that the alternative hypothesis assumes the effect magnitude to be smaller than $d_z = -1.78$. Given a t-test against a constant, we aimed to detect a medium effect, $d_z = 0.50$, with 80 % power and a conventional type I error of 5%, which required 27 participants within each level of exposure time. The study received ethical approval from the IRB at Aarhus University (2021–05).

5.1.2. Participants and design

We recruited 165 participants (111 females, $M = 25.06$ years, $SD = 7.69$ years, range = 18 − 67 years) from the university's participant pool, accounting for potential exclusions as defined in our preregistration. Exclusion criteria, informed consent, and compensation were identical to those in Study 1. The study employed the same task as described in Study 1 but with a two-factorial, mixed-subjects design. The exposure time varied between participants and was set at 0.3 s, 0.6 s, 0.9 s, and 1.2 s (see Fig. 2 for details).

5.1.3. Stimuli, procedure, and data processing

The stimuli, procedure, and data processing were identical to those in Study 1. We excluded 30 participants who failed to meet the movement initiation deadline or the response deadline in more than one-third of all trials or who responded incorrectly in more than one-third of the certified trials. This left 135 participants (90 females, $M = 24.76$ years, $SD = 7.05$ years, range = 18 − 67 years) distributed across each level of exposure time as follows: 37 (0.3 s), 39 (0.6 s), 31 (0.9 s), and 28 (1.2 s). The familiarity and importance ratings of the four certified labels for all remaining participant are reported in Web Appendix D. For all analyses, trials that were not completed in full-screen mode ($n = 514$) and imperfectly timed trials ($n = 1{,}996$) were excluded. The exclusion of trials based on the predefined criteria may have resulted in incomplete sets, that is, sets that did not include at least one valid trial in each condition. Since complete sets are required to compute the confusion metrics, trials

[7] https://osf.io/sjhb3.

belonging to incomplete sets (n = 2,199) were also excluded, which is a deviation from the preregistration. The exclusion of trials left 8,312 out of 12,960 trials.

5.2. Results

We examined whether we replicated the main results regarding our attribute confusion metric from Study 1 using the same statistical models and procedures across different exposure times (0.3 s, 0.6 s, 0.9 s, and 1.2 s). Regarding the measurement of attribute confusion, separate mixed-effects regression analyses for each exposure time revealed that sensitivity (d') varied between comparison conditions at all exposure times, with the effect size decreasing as exposure time increased (Fig. 6a): 0.3 s, b = −1.59, SE = 0.21, p < 0.001, 0.6 s, b = −0.80, SE = 0.17, p < 0.001, 0.9 s, b = −0.36, SE = 0.12, p = 0.008, and 1.2 s, b = −0.24, SE = 0.12, p = 0.060. Thus, we successfully replicated the effect for the 0.3 s exposure time, but the effect was significantly smaller for 0.6 s, 0.9, and 1.2 s.

Regarding the identification of "attribute confusion suspects", separate logistic regression analyses for each exposure time revealed that sensitivity (d') significantly predicted the probability that a matched comparison included a product carrying a lookalike label at all exposure times, with the effect size decreasing as exposure time increased: 0.3 s, b = − 0.24, SE = 0.02, p < 0.001, 0.6 s, b = −0.15, SE = 0.03, p < 0.001, 0.9 s, b = −0.09, SE = 0.03, p = 0.006, and 1.2 s, b = −0.07, SE = 0.04, p = 0.048. We then used these models to classify "attribute confusion suspects" using 100x6-fold cross-validations for each model at the respective exposure time. The results revealed that sensitivity showed a meaningful accuracy above chance level for exposure times of 0.3 s, M = 64.25%, SD = 0.08%, and 0.6 s, M = 57.19%, SD = 0.10%, while accuracy was around chance level for exposure times of 0.9 s, M = 51.53%, SD = 0.11%, and 1.2 s, M = 50.73%, SD = 0.37%. Again, 100x6- fold cross-validations of logistic regressions based on data averaged across participants revealed that classification performance improved substantially at all exposure times with a similar slope as exposure time increased: 0.3 s, M = 75.23%, SD = 0.93%, 0.6 s, M = 76.31%, SD = 1.24%, 0.9 s, M = 55.15%, SD = 1.84%, and 1.2 s, M = 56.52%, SD = 1.77 %.

In terms of the optimal classification threshold, separate grid searches for the sensitivity that maximizes classification accuracy for each exposure time yielded increasing sensitivities and decreasing accuracies as exposure time increased: 0.3 s, d' = 1.91 with 83.33%, 0.6 s, d' = 3.52 with 77.08%, 0.9 s, d' = 4.89 with 77.08%, 1.2 s, d' = 4.15 with 60.42% (Fig. 6b).

5.3. Discussion

In Study 3, we explored how exposure time affects attribute confusion as detected by our attribute confusion metric. We manipulated the exposure times for product images (0.3 s, 0.6 s, 0.9 s, and 1.2 s) to determine whether longer exposure times would help consumers better distinguish between certified-labeled and lookalike-labeled products. Our results indicate that longer exposure times reduced attribute confusion, as demonstrated by a decline in the difference in sensitivity (d') values. We replicated the results from Study 1 for an exposure time of 0.3 s. However, for exposure times of 0.6 s, 0.9 s, and 1.2 s, the degree of attribute confusion was significantly reduced. In hindsight, it is clear that longer exposure times naturally allow participants more opportunity to inspect the labels carefully, leading to decreased confusion. Despite this decline, mixed-effects regression analyses demonstrated that the sensitivity values remained significantly lower for lookalike-labeled compared to control products across all exposure times. This indicates that lookalike labels still cause attribute confusion regardless of exposure time.

Regarding optimal classification thresholds, separate grid searches showed that longer exposure times require higher thresholds. The findings show that while longer exposure times help consumers more accurately differentiate between certified and lookalike labels, lookalike labels continue to create confusion even at the longest exposure. Thus, our proposed

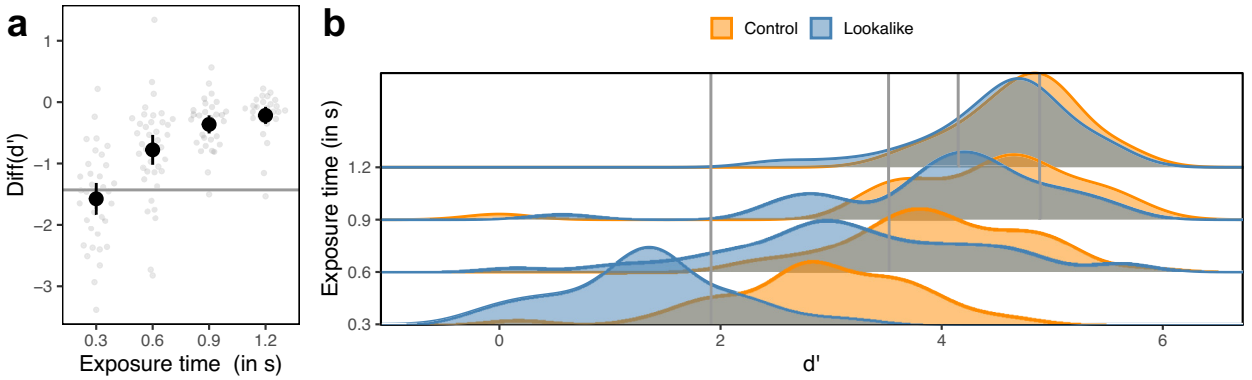

Fig. 6. a) Within-subjects effect on sensitivity (d′) as the difference between the comparison conditions for each level of exposure time. The error bars depict the 95% CI of the mean, with the horizontal line depicting the upper boundary from Study 1 for comparison. b) Density distributions of sensitivity (d′) averaged within sets and across participants for the control and lookalike products (color) for each level of exposure time. The vertical lines depict the classification thresholds that maximize classification accuracy for each level of exposure time.

attribute confusion metric is robust to variations in exposure times, though its effectiveness diminishes with longer exposure times. Based on this, we recommend that other researchers use a 0.3 s exposure time.

6. General discussion

Copycatting pays off by confusing consumers about products. While typically studied in branding, we propose that companies can also create *attribute confusion*, causing consumers to mistakenly believe a product has desirable attributes like healthfulness or sustainability. We investigated whether lookalike labels mimicking third-party certifications can cause this confusion. To address this, we developed a novel attribute confusion metric using mouse cursor tracking and applied it in three preregistered studies. Here, we summarize the main findings, discuss implications, address limitations, suggest future research, and provide a final conclusion.

6.1. Summary of main findings

In Study 1, we showed that participants confused products with lookalike labels for those with certified labels. Their cursor movements deviated more from the direct path when correctly identifying lookalike-labeled products compared to those with neither lookalike nor certified labels. Based on this, we developed an attribute confusion metric to quantify attribute confusion for each participant using signal-detection theory applied to the mouse-tracking data. Values close to zero indicate near-perfect confusion, while higher values indicate less confusion. Through binary classification, the metric reliably flagged products as either "attribute confusion suspect" or "attribute confusion safe", identifying 19 of 24 lookalike-labeled products as "attribute confusion suspect".

In Study 2, we replicated findings from Study 1 and demonstrated the unique advantage of the attribute confusion metric over alternative metrics. Unlike choice data, which aggregates information across groups, the attribute confusion metric quantifies attribute confusion for individual consumers. Using this advantage, we showed that attribute confusion predicts willingness-to-pay for uncertified products, with a premium of approximately 5% between the least and most confused consumers.

In Study 3, we tested the robustness of the attribute confusion metric across different product exposure times. While consumers remained confused by lookalike labels at all exposure times, confusion decreased as exposure time increased. At longer exposure times, the attribute confusion metric was less effective at flagging lookalike-labeled products as "attribute confusion suspect", suggesting optimal identification at 0.3 s, with some effectiveness at 0.6 s. Overall, our results replicate well, showing that consumers remained confused by lookalike labels even at longer exposure times (1.2 s), which exceeds the typical time consumers spend attending to products (i.e., 0.36 − 0.55 s, Hurley et al., 2013).

6.2. Practical implications

The concept and measurement of consumer attribute confusion have important practical implications for marketers, policymakers, and consumer advocacy groups. Our findings suggest that lookalike labels may exploit the equity of certified or trademarked product labels. For example, consumers are willing to pay up to 30 % more for products with certified organic labels (Crowder & Reganold, 2015) and up to 14 % more for products with certified nutrition labels (Edenbrandt et al., 2018). We found that highly confused consumers are willing to pay 5 % more for products with lookalike labels, indicating that attribute confusion could negatively impact products with certified labels if some sales shift to lookalike-labeled products (Mitchell & Papavassiliou, 1999). Furthermore, lookalike labels may dilute the label equity of certified labels even without direct confusion, through *blurring* (where the distinctiveness of certified labels is reduced due to redundant claims) or *tarnishment* (where lookalike labels create unwarranted associations), leading consumers to view certified labels less favorably (Morrin & Jacoby, 2000).

Marketers using certified labels can leverage our method in two key ways. First, they can monitor lookalike labels in their product categories and use the attribute confusion metric to detect if competitors are successfully confusing consumers. In response, marketers can communicate premium product features through alternative means, such as distinctive packaging design or advertising, or they might pursue legal action for trademark infringement or misleading advertising. Second, marketers can use the attribute confusion metric to design product labels that minimize confusion with existing labels. For example, when choosing green for a CO_2 label due to its environmental associations, it is crucial to ensure it is distinct from similar hues used in other labels. The attribute confusion metric provides a systematic approach to measure and adjust these similarities, ensuring new labels are distinctive and less likely to cause confusion.

From a policy and consumer advocacy perspective, our findings suggest that lookalike labels may mislead consumers into purchasing less sustainable or healthful products, posing a threat to environmental goals and consumer well-being. To ensure consumers can easily make informed purchasing decisions, regulations should prohibit marketers from mimicking certified labels. Policymakers and consumer advocacy groups could use the attribute confusion metric as evidence in legal actions against products with lookalike labels, arguing that this constitutes misleading advertising (cf. Mitchell & Kearney, 2002). Legal actions might also be warranted due to the dilution of certified label equity, which can reduce incentives for producers to improve their products (Vyth et al., 2010), thereby harming competition and undermining regulatory efforts.

Moreover, the metric is versatile and can be applied to any potentially misleading labels, images, symbols, or claims suspected of causing attribute confusion concerning brand identity, healthfulness, ethical production, or sustainability. This is particularly relevant for detecting subtler forms of greenwashing (Carlson et al., 1993), where companies use colors or symbols to falsely imply sustainability. The only practical caveat is that the metric requires two products: one lacking the attribute in question and a comparison product that features it. The comparison product must not only have the attribute but also reliably convey it to consumers. Hence, finding a suitable comparison product may be challenging in some categories or for certain attributes.

6.3. Theoretical implications

Mouse cursor tracking reveals new insights into attribute confusion due to lookalike labels. Our findings suggest that participants are initially biased toward perceiving lookalike labels as certified ones, a response they sometimes manage to correct (Stone et al., 2022). The decision trajectories, as indicated by cursor movements, imply a dual-system process through which lookalike labels confuse consumer decision-making (Evans, 2003): an automatic quick identification of relevant product features, followed by a more deliberate process to distinguish lookalike from certified labels. This initial bias may be exacerbated by consumers often relying on covert attention (Perkovic et al., 2023)—using peripheral vision without direct gaze—which is prone to errors due to lower resolution (Strasburger et al., 2011) and a reliance on expectations (Eymond et al., 2020). Interestingly, even when using foveal vision, expectations guide the detailed processing of visual information, making consumers fill in gaps based on familiar labels (Summerfield & de Lange, 2014). As a result, it is unnecessary to repeatedly attend to familiar objects in detail (Summerfield & Egner, 2009). If a lookalike label initially resembles a familiar certified label, consumers may not scrutinize it, instead filling in details based on expectations of the certified label (Summerfield & Egner, 2009). Expectations are shaped by contextual factors (Summerfield & Egner, 2009; Summerfield & de Lange, 2014), which lookalike labels can exploit. For instance, since the DK Whole Grain label is common on bread but not dairy products, consumers are more likely to mistake a DK Whole Grain lookalike label as the real one on bread. Other factors influencing expectations include price (e.g., organic products tend to be more expensive; Juhl et al., 2017), shelf position (e.g., premium products are usually on top shelves; Valenzuela et al., 2013), product category (e.g., certain certified labels are limited to specific categories; Danish Veterinary and Food Administration, 2019; FLOCERT, 2021a), and label position (e.g., certified labels are placed in consistent locations on products; Orquin et al., 2020). In sum, our findings suggest that consumer expectations about the presence of certified labels, combined with fast, automatic visual detection, create an initial bias toward mistaking lookalike labels for certified ones, a bias that is sometimes corrected through deliberate processing.

6.4. Methodological implications

Our method is the first to directly address attribute confusion. We compared our attribute confusion metric with alternative metrics derived from choice data, response times, mouse-tracking data, and an adapted version of the Kapferer index. Although not explored here, attribute confusion could potentially also be measured using survey methods. Of these, our metric is likely the most complex in terms of methodology and computation. This complexity is a disadvantage compared to simpler methods like surveys, where participants could rate how confusing a product is (e.g., Walsh et al., 2007). However, our metric has key advantages, such as better mimicking the fast decision-making processes typical of consumer behavior, thanks to limited exposure time (Hurley et al., 2013). While the triangle test by Satomura et al. (2014) also uses limited exposure, it does not specifically address attribute confusion. The Kapferer index is the most practical alternative, being simpler and offering comparable accuracy in detecting products with lookalike labels. However, our metric has the crucial advantage of providing detailed information on confusion levels for each individual consumer and product, enabling new insights, such as predicting willingness-to-pay as shown in Study 2. In contrast, drawing similar conclusions from choice data would require regressing willingness-to-pay on confusion levels across multiple products, which could overlook the hierarchical nature of the data and be confounded by Simpson's paradox (Simpson, 1951), leading to incorrect conclusions about the relationship between confusion and willingness-to-pay.

We recommend that practitioners who intend to use the attribute confusion metric closely follow our established procedure, particularly in setting up the mouse-tracking method and exposure time. To classify products as "attribute confusion safe" or "attribute confusion suspect", practitioners should use a threshold in the range proposed d' thresholds (2.11, 2.27). While the number of participants and products may vary, we suggest maintaining similar sample sizes to ensure external validity. Practitioners may also opt for on-site data collection to gain more control over the mouse cursor tracking, increasing data quality and reducing the need to exclude participants and trials.

6.5. Limitations and directions for future research

This article introduces a new research topic by demonstrating that consumers can be confused by product labels mimicking third-party certified labels. We identify several limitations and propose directions for future research.

A key limitation is the absence of purchase data for lookalike-labeled products. Future choice or field studies are needed to determine whether attribute confusion leads to consumer purchases based on false beliefs. These studies could also exam-

ine the relationship between label similarity and confusion levels, addressing another limitation: identifying which specific label features and their degree of similarity contribute to consumer confusion.

Another limitation is the lack of clarity on the mechanisms driving attribute confusion due to lookalike labels. Eye-tracking studies are needed to determine whether this confusion is primarily driven by covert or overt attention (Perkovic et al., 2023), or an alternative mechanism, such as implicit learning of statistical associations between labels and brands (Perkovic & Orquin, 2018). Understanding these mechanisms could help marketers and policymakers design product labels or other methods of communicating premium product features that minimize confusion with certified labels or other attributes that contribute to brand equity. It would also aid in identifying factors that moderate attribute confusion, such as the prevalence of certified labels, shelf position, brand, and price.

A third limitation of our research is its exclusive focus on lookalike labels. While we have suggested that the metric can be applied to any product feature, future studies could specifically test its application to features like brand names, images, symbols, or claims, and evaluate their potential to create attribute confusion. Additionally, research could investigate adapting our metric for trade dress imitation, where a copycat product mimics the overall theme of an original brand. These studies could also examine whether attribute confusion extends to other product attributes, for instance, through the halo effect (Huber & McCann, 1982).

7. Conclusion

This article introduced a novel metric using mouse cursor tracking to measure consumer attribute confusion caused by lookalike product labels. By combining signal-detection theory with mouse cursor movements, the metric quantifies confusion and categorizes products as either "attribute confusion suspect" or "attribute confusion safe". In three preregistered studies, we found that lookalike labels often mislead consumers into believing that a product carries a certified label, and that this confusion increases willingness-to-pay. Attribute confusion is most pronounced at short exposure times, reflecting typical consumer behavior at the point of purchase. The attribute confusion metric equips marketers with a tool to design less confusing labels and provides concrete evidence when products are suspected of misleading consumers.

Author note

The authors thank Sascha Steinmann, Klaus G. Grunert, and Valdimar Sigurðsson for helpful comments on an earlier manuscript version. This research was supported by the Independent Research Fund Denmark (grant 8046-00014A). The funders had no role in conceptualization, data analysis, decision to publish, or preparation of the manuscript; the authors declare no competing interests. All datasets and analyses scripts are openly available at https://osf.io/gn3qd/. We report how we determined our sample size, all data exclusions (if any), all manipulations, all measures in the studies, and all relevant preregistrations. All studies were preregistered (please find the links in the respective sections). Martin Schoemann and Piet van de Mosselaar contributed equally to this article and share first-authorship. Correspondence concerning this article should be addressed to Martin Schoemann, Department of Psychology, Technische Universität Dresden, Zellescher Weg 17, 01,069 Dresden – Germany. E-mail: martin.schoemann@tu-dresden.de.

CRediT authorship contribution statement

Martin Schoemann: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Piet van de Mosselaar:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sonja Perkovic:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Jacob L. Orquin:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Data availability

All datasets and analysis scripts are openly available (https://osf.io/gn3qd/)

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijresmar.2024.08.010

References

Atiya, N. A. A., Zgonnikov, A., O'Hora, D., Schoemann, M., Scherbaum, S., & Wong-Lin, K. (2020). Changes-of-mind in the absence of new post-decision evidence. *PLOS Computational Biology, 16*(2), e1007149. https://doi.org/10.1371/journal.pcbi.1007149.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bloch, P. H. (1995). Seeking the ideal form: Product design and consumer response. *Journal of Marketing, 59*(3), 16–29. https://doi.org/10.1177/002224299505900302.

Carlson, L., Grove, S. J., & Kangun, N. (1993). A content analysis of environmental advertising claims: A matrix method approach. *Journal of Advertising, 22*(3), 27–39. https://doi.org/10.1080/00913367.1993.10673409.

Cradit, J. D., Tashchian, A., & Hofacker, C. F. (1994). Signal detection theory and single observation designs: Methods and indices for advertising recognition testing. *Journal of Marketing Research, 31*(1), 117–127. https://doi.org/10.1177/002224379403100110.

Crowder, D. W., & Reganold, J. P. (2015). Financial competitiveness of organic agriculture on a global scale. *Proceedings of the National Academy of Sciences, 112*(24), 7611. https://doi.org/10.1073/pnas.1423674112.

Danish Veterinary and Food Administration. (2019). *Vejledning om anvendelse af nøglehulsmærket på fødevarer m.v.* Retrieved November 22, 2021, from https://www.retsinformation.dk/eli/retsinfo/2019/9665.

Edenbrandt, A. K., Smed, S., & Jansen, L. (2018). A hedonic analysis of nutrition labels across product types and countries. *European Review of Agricultural Economics, 45*(1), 101–120. https://doi.org/10.1093/erae/jbx025.

Evans, J. S. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012.

Eymond, C., Seidel Malkinson, T., & Naccache, L. (2020). Learning to see the ebbinghaus illusion in the periphery reveals a top-down stabilization of size perception across the visual field. *Scientific Reports, 10*(1), 12622. https://doi.org/10.1038/s41598-020-69329-9.

Flocert (2021a). Fairtrade products Retrieved December 16, 2021, from https://www.fairtrade.net/product.

Flocert (2021b). Glossary: Licence fee Retrieved November 4, 2021, from https://www.flocert.net/glossary/licence-fee/.

Foxman, E. R., Berger, P. W., & Cote, J. A. (1992). Consumer brand confusion: A conceptual framework. *Psychology and Marketing, 9*(2), 123–141. https://doi.org/10.1002/mar.4220090204.

Friedman, M. P. (1966). Consumer confusion in the selection of supermarket products. *Journal of Applied Psychology, 50*(6), 529–534. https://doi.org/10.1037/h0024048.

Grage, T., Schoemann, M., Kieslich, P. J., & Scherbaum, S. (2019). Lost to translation: How design factors of the mouse-tracking procedure impact the inference from action to cognition. *Attention, Perception, & Psychophysics, 81*(7), 2538–2557. https://doi.org/10.3758/s13414-019-01889-z.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Huber, J., & McCann, J. (1982). The impact of inferential beliefs on product evaluations. *Journal of Marketing Research, 19*(3), 324–333. https://doi.org/10.1177/002224378201900305.

Hurley, R. A., Ouzts, A., Fischer, J., & Gomes, T. (2013). Effects of private and public label packaging on consumer purchase patterns. *Packaging Technology and Science, 26*(7), 399–412. https://doi.org/10.1002/pts.2012.

Juhl, H. J., Fenger, M. H. J., & Thøgersen, J. (2017). Will the consistent organic food consumer step forward? an empirical analysis. *Journal of Consumer Research, 44*(3), 519–535. https://doi.org/10.1093/jcr/ucx052.

Kadlec, H. (1999). Statistical properties of d' and β estimates of signal detection theory. *Psychological Methods, 4*(1), 22–43. https://doi.org/10.1037/1082-989X.4.1.22.

Kapferer, J.-N. (1995). Brand confusion: Empirical study of a legal concept. *Psychology & Marketing, 12*(6), 551–568. https://doi.org/10.1002/mar.4220120607.

Kieslich, P. J., Henninger, F., Wulff, D. U., Haslbeck, J. M. B., & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: A practical guide to implementation and analysis. In M. Schulte-Mecklenbeck, A. Kühberger, & J. G. Johnson (Eds.), *A handbook of process tracing methods*. Routledge.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. https://doi.org/10.1177/2515245918770963.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition, 6*(3), 312–319. https://doi.org/10.3758/BF03197461.

Loken, B., Ross, I., & Hinkle, R. L. (1986). Consumer "confusion" of origin and brand similarity perceptions. *Journal of Public Policy & Marketing, 5*(1), 195–211. https://doi.org/10.1177/074391568600500114.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98*(1), 185–199. https://doi.org/10.1037/0033-2909.98.1.185.

Miaoulis, G., & D'Amato, N. (1978). Consumer confusion & trademark infringement. *Journal of Marketing, 42*(2), 48–55. https://doi.org/10.1177/002224297804200208.

Mitchell, V.-W., & Kearney, Í. (2002). A critique of legal measures of brand confusion. *Journal of Product & Brand Management, 11*(6), 357–379. https://doi.org/10.1108/10610420210445497.

Mitchell, V.-W., & Papavassiliou, V. (1999). Marketing causes and implications of consumer confusion. *Journal of Product & Brand Management, 8*(4), 319–342. https://doi.org/10.1108/10610429910284300.

Morrin, M., & Jacoby, J. (2000). Trademark dilution: Empirical measures for an elusive concept. *Journal of Public Policy & Marketing, 19*(2), 265–276. https://doi.org/10.1509/jppm.19.2.265.17137.

Newman, C. L., Howlett, E., & Burton, S. (2014). Shopper response to front-of-package nutrition labeling programs: Potential consumer and retail store benefits. *Journal of Retailing, 90*(1), 13–26. https://doi.org/10.1016/j.jretai.2013.11.001.

Orquin, J. L., Bagger, M. P., Lahm, E. S., Grunert, K. G., & Scholderer, J. (2020). The visual ecology of product packaging and its effects on consumer attention. *Journal of Business Research, 111*, 187–195. https://doi.org/10.1016/j.jbusres.2019.01.043.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Hochenberger, R., Sogo, H., Kastman, E., & Lindelov, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y.

Perkovic, S., & Orquin, J. L. (2018). Implicit statistical learning in real-world environments leads to ecologically rational decision making. *Psychological Science, 29*(1), 34–44. https://doi.org/10.1177/0956797617733831.

Perkovic, S., Schoemann, M., Lagerkvist, C.-J., & Orquin, J. L. (2023). Covert attention leads to fast and accurate decision-making. *Journal of Experimental Psychology: Applied, 29*(1), 78–94. https://doi.org/10.1037/xap0000425.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9*(3), 319–332. https://doi.org/10.1177/1745691614528519.

Pieters, R., Wedel, M., & Smith, R. H. (2012). Ad gist: Ad communication in a single eye fixation. *Marketing Science, 31*(1), 59–73. https://doi.org/10.1287/mksc.1110.0673.

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Satomura, T., Wedel, M., & Pieters, R. (2014). Copy alert: A method and metric to detect visual copycat brands. *Journal of Marketing Research, 51*, 1–13. https://doi.org/10.1509/jmr.11.0467.

Scherbaum, S., & Dshemuchadse, M. (2019). Psychometrics of the continuous mind: Measuring cognitive sub-processes via mouse tracking. *Memory & Cognition, 48*(3), 436–454. https://doi.org/10.3758/s13421-019-00981-x.

Schoemann, M., Lüken, M., Grage, T., Kieslich, P. J., & Scherbaum, S. (2019). Validating mouse-tracking: How design factors influence action dynamics in intertemporal decision making. *Behavior Research Methods, 51*(5), 2356–2377. https://doi.org/10.3758/s13428-018-1179-4.

Schoemann, M., O'Hora, D., Dale, R., & Scherbaum, S. (2021). Using mouse cursor tracking to investigate online cognition: Preserving methodological ingenuity while moving toward reproducible science. *Psychonomic Bulletin & Review, 28*(3), 766–787. https://doi.org/10.3758/s13423-020-01851-3.
Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 13*(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x.
Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences, 102*(29), 10393–10398. https://doi.org/10.1073/pnas.0503903102.
Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137–149. https://doi.org/10.3758/BF03207704.
Stillman, P. E., Krajbich, I., & Ferguson, M. J. (2020). Using dynamic monitoring of choices to predict and understand risk preferences. *Proceedings of the National Academy of Sciences, 117*(50), 31738. https://doi.org/10.1073/pnas.2010056117.
Stillman, P. E., Shen, X., & Ferguson, M. J. (2018). How mouse-tracking can advance social cognitive theory. *Trends in Cognitive Sciences, 22*(6), 531–543. https://doi.org/10.1016/j.tics.2018.03.012.
Stone, C., Mattingley, J. B., & Rangelov, D. (2022). On second thoughts: Changes of mind in decision-making. *Trends in Cognitive Sciences, 26*(5), 419–431. https://doi.org/10.1016/j.tics.2022.02.004.
Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision, 11*(5), 13. https://doi.org/10.1167/11.5.13.
Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences, 13*(9), 403–409. https://doi.org/10.1016/j.tics.2009.06.003.
Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience, 15*(11), 745–756. https://doi.org/10.1038/nrn3838.
Valenzuela, A., Raghubir, P., & Mitakakis, C. (2013). Shelf space schemas: Myth or reality? *Journal of Business Research, 66*(7), 881–888. https://doi.org/10.1016/j.jbusres.2011.12.006.
van Doorn, J., & Verhoef, P. C. (2011). Willingness to pay for organic products: Differences between virtue and vice foods. *International Journal of Research in Marketing, 28*(3), 167–180. https://doi.org/10.1016/j.ijresmar.2011.02.005.
Vyth, E. L., Steenhuis, I. H. M., Roodenburg, A. J. C., Brug, J., & Seidell, J. C. (2010). Front-of-pack nutrition label stimulates healthier product development: A quantitative analysis. *International Journal of Behavioral Nutrition and Physical Activity, 7*(1), 65. https://doi.org/10.1186/1479-5868-7-65.
Walsh, G., Hennig-Thurau, T., & Mitchell, V.-W. (2007). Consumer confusion proneness: Scale development, validation, and application. *Journal of Marketing Management, 23*(7–8), 697–721. https://doi.org/10.1362/026725707X230009.
Walsh, G., & Mitchell, V.-W. (2005). Consumer vulnerability to perceived product similarity problems: Scale development and identification. *Journal of Macromarketing, 25*(2), 140–152. https://doi.org/10.1177/0276146705280613.
Webster, M. A., Halen, K., Meyers, A. J., Winkler, P., & Werner, J. S. (2010). Colour appearance and compensation in the near periphery. *Proceedings of the Royal Society B: Biological Sciences, 277*(1689), 1817–1825. https://doi.org/10.1098/rspb.2009.1832.
Wixted, J. T. (1990). Analyzing the empirical course of forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(5), 927–935. https://doi.org/10.1037/0278-7393.16.5.927.
Wu, W., Zhang, A., van Klinken, R. D., Schrobback, P., & Muller, J. M. (2021). Consumer trust in food and the food system: A critical review. *Foods, 10*(10), 2490. https://doi.org/10.3390/foods10102490.
Wulff, D., Haslbeck, J., Kieslich, P., Henninger, F., & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: Detecting types in movement trajectories. In M. Schulte-Mecklenbeck, A. Kühberger, & J. G. Johnson (Eds.), *A handbook of process tracing methods*. Routledge.