Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sensory Detection Thresholds for Full Black Defects in Beverage Coffee

Julien Delarue

jdelarue@ucdavis.edu

UC Davis https://orcid.org/0000-0001-8077-6076

Daniel Schoonbrood

University of California, Davis

Willian Klippel-Huber

UC Davis

Irwin Donis-Gonzalez

UC Davis

William Ristenpart

UC Davis

Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11807774/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Sensory Detection Thresholds for Full Black Defects in Beverage Coffee

Daniel Schoonbrood[1,2], Willian Klippel-Huber[1,3], Irwin R. Donis-Gonzalez[1,3], William D. Ristenpart[1,4], Julien Delarue[1,2*]

[1]UC Davis Coffee Center, University of California Davis, Davis, California, USA

[2]Department of Food Science and Technology, University of California Davis, Davis, California, USA

[3]Department of Biological and Agricultural Engineering, University of California Davis, Davis, California, USA

[4]Department of Chemical Engineering, University of California Davis, Davis, California, USA

[*]Corresponding author: Dr. Julien Delarue, jdelarue@ucdavis.edu

Abstract

Under current coffee grading standards, a single ‘full defect equivalent’ in a 350-gram sample disqualifies a batch from the highest-valued specialty grade. Despite the economic significance of these standards, no research has established sensory thresholds for individual defect types. Here, coffee consumers (n = 69) and experts (n = 16) evaluated concentrations of full black defects in Coffea arabica using the ascending forced-choice method of limits. The consumer threshold was measured as 127±29 full black beans per 350 grams, while the expert threshold ranged from 8 to 22 per 350 grams. Further testing with experts using a standard cupping protocol revealed that overall quality was not significantly impacted at or below 27 full blacks per 350 grams. Our results indicate that industry standards overestimate the sensory impact of full black defects by one to two orders of magnitude, suggesting that coffee producers are economically penalized by unnecessarily restrictive standards.

Keywords

Coffee Defects, Experts, Consumers, Quality, Grading, Psychophysics

Introduction

Coffee quality and its corresponding pricing have long been assessed by certified practitioners counting visibly defective beans in samples of unroasted green coffee[1]. Early nineteenth-century sources reference the impact of these defects[2,3], but the first formal system for quantifying coffee quality via defect counts was introduced by the New York Coffee Exchange in the early 1900's[4,5]. In this system, the various types of defects that arise throughout coffee production were expressed in terms of 'full black equivalents'; for example, three broken beans are equivalent to one full black bean. Despite the early objections of coffee producers, this system imposed by coffee buyers became entrenched within the coffee industry[6]. Under the Specialty Coffee Association (SCA) guidelines, a lot cannot reach the highest-valued 'specialty' grade if a representative (350-gram) sample contains a single Category 1 'full defect equivalent' (FDE) (~1 in 3,500 beans)[7]. This limit corresponds to 1 black bean out of approximately 3,500 beans (0.03% w/w).

Although the presence of that single black bean imposes a significant economic penalty on the coffee producer, its sensory impact on the resulting beverage is unclear. Most research on green coffee defects research has focused on their chemical characteristics (e.g., pH, sucrose, and volatile profiles) rather than how they are perceived by humans[8]. Some practitioners within the specialty coffee community have argued that current FDE standards are overly restrictive[9]. To our knowledge, the only study that has investigated the sensory impact of coffee defects determined that samples spiked with equal parts black, immature, and sour defective beans were detectable at 10.25% (w/w) significantly higher than the current specialty standard of 0.03% (w/w)[10]. However, this estimate reflects a mixture and cannot be attributed to any single defect type.

To address this gap, we sought to establish a sensory detection threshold for full black bean defects in brewed *Coffea arabica*, chosen because of their long-standing significance in grading standards. While the consumer may be the ultimate judge of coffee quality, experts (i.e., coffee graders) effectively serve as gatekeepers to coffee entering the supply chain. Differences in training and experience may influence their sensitivity to coffee defects[11–14], making it important to consider both groups when evaluating the sensory basis of grading standards. We thus aimed to determine the detection thresholds for coffee consumers (Experiment 1) and coffee experts (Experiment 2) using the psychometric ascending forced-choice method of limits paired with an unspecified tetrad test[15,16]. We additionally evaluated how sensitive the traditional SCA cupping form, the industry standard for determining coffee quality beyond defect counts, is to changes in full black concentration (Experiment 3). The main goal was to assess whether current coffee grading standards, and their corresponding economic implications, align with actual perception.

Results

Consumer Detection Threshold (Experiment 1)

At the lowest tested defect concentration (1.33%), 27 of the 69 consumers (39%) correctly identified the sample pairs in the tetrad test, only slightly better than chance (Figure 3). As the defect concentration increased, the number of participants selecting the correct pair dipped slightly at 2% defect concentration, then increased to 37 of 69 participants (53.8%) at a defect concentration of 3%. Further increases yielded a slight uptick in correct responses at 4.5% defect concentration, and then finally 38 of 69 (55%) correct selections at the highest tested concentration of 6.75% defects. We emphasize that at 6.75% black defects by mass (~236 full

blacks per 350 grams), only around half of the participants could correctly discriminate the defective samples from the clean.

Estimated detection probability increased with defect concentration in the consumer panel (β_1 = 0.10, SE = 0.04, z = 2.55, p = 0.011; Figure 4a). Predicted performance at zero defect concentration was 0.35 (β_0 = -0.39, SE = 0.15). Based on the fitted population-level psychometric curve, the consumer detection threshold for full black, green coffee defects was estimated to be 3.64% w/w at d' = 1, with a standard error of SE = 0.82% (Figure 4a).

The random slope standard deviation (σ = 0.10) indicated substantial variation in individual defect sensitivity. Individual threshold estimates ranged from 1.98% to 45.26% w/w (median = 4.15%, IQR: 2.61 - 5.53%). Estimated thresholds exceeded the highest tested concentration (6.75% w/w) for 15 of the 69 participants (22%; green bars in Figure 4b).

Expert Detection Threshold (Experiment 2)

High-Range Series

Threshold testing indicated that experts are much more sensitive to the black defects than consumers (Figure 5). The number of experts correctly identifying the matching pairs in both replicates rose from 8 of 15 at 1.33% to 13 of 15 at 4.5%, then to 11 of 13 at 6.75% (Figure 5a). Estimated detection probability increased with defect concentration in the expert panel (β_1 = 0.48, SE = 0.18, z = 2.74, p = 0.006; Figure 5b). Predicted performance at zero defect concentration was 0.38 (β_0= -0.32, SE = 0.32). Based on the fitted population-level psychometric curve, the expert detection threshold for full black, green coffee defects was estimated to be 0.63% w/w (SE = 0.51%) at d' = 1.

Similar to the results obtained with the consumers, the random slope standard deviation (σ = 0.30) indicated substantial variation in individual defect sensitivity. Individual threshold estimates ranged from 0.42 to 3.74% w/w (median = 0.63%, IQR: 0.42 - 1.44%). Only four experts had estimated thresholds within the tested concentration range, while the remaining estimates fell below the lowest concentration tested. These latter estimates should therefore be interpreted with caution, as they are based on extrapolation beyond the experimental range.

Low-Range Series

The low-range series (0.16% - 2%) more accurately captures the experts' full black defect threshold, since the high-range estimate fell below the lowest tested concentration (1.33%). The number of experts correctly identifying the matching pairs in all three replicates rose from 2 of 16 at 0.16% w/w to 6 of 16 at 2% (Figure 5c). In contrast to the High-Range Series, most experts did not achieve correct responses across all three replicates at any tested concentration, indicating that the lower concentration range captured the ascending portion of the psychometric function rather than its ceiling (Figure 5d). Estimated detection probability tended to increase with defect concentration, though the effect did not reach statistical significance (β_1 = 0.51, SE = 0.29, z = 1.75, p = 0.080; Figure 5d). Predicted performance at zero defect concentration was 0.45 (β_0= -0.13, SE = 0.15). Based on the psychometric curve, the expert detection threshold for full black, green coffee defects was estimated to be 0.23% (SE = 0.25%) at d' = 1.

The random slope standard deviation (σ = 0.76) exceeded the fixed slope, indicating substantial heterogeneity in expert sensitivity at low concentrations. Individual threshold estimates ranged from 0.09 to 1.61% w/w (median = 0.24%, IQR: 0.10 - 0.69%). Negative individual slopes were estimated for several participants (n = 4) and are an artifact of estimation caused by the intercept being fixed in the model. Twelve estimated thresholds fell below the

highest tested concentration of 2%, six of which fell below the lowest tested concentration (0.16%).

Analysis of the shared 2% defect concentration across the two series indicated no difference in the proportion of correct responses (High-Range Series: 66.7%, Low-Range Series: 62.5%; β_1 = -0.14, SE = 0.32, z = -0.45, p = 0.654), supporting the consistency of expert detection across sessions.

Expert Grading Assessment (Experiment 3)

We now turn to the impact of black defects on the traditional SCA cupping form as assessed by the panel of certified Q-graders. The three independently presented defect-free controls did not differ in Cupping Score ($p > 0.05$) and were pooled for subsequent analyses. Neither 1 defect bean per 350-gram sample (0.03%; 82.08 ± 0.56, p = 0.51) nor the 1 defect bean per 13-gram sample (0.77%; 81.32 ± 0.58, p = 0.79) affected the Cupping Score relative to the control (80.79 ± 0.56), with scores remaining above the 80-point specialty coffee cutoff in both cases (Figure 6, Table S6). Cupping Scores began to decline from the control at 1.33% defect (75.27 ± 2.40, p = 0.007) and remained lower than the control through 6.75% (64.54 ± 3.40, $p < 0.001$). Notably, the defect concentration at which scores fell below the 80-point specialty coffee cutoff (1.33%, 47 beans per 350 g) was two to six times higher than the expert detection thresholds reported in Section 3.2 (0.23 - 0.63%, 8 - 22 beans per 350 g).

Increases in full black defect concentration had varied effects on individual attribute scores. As with Cupping Score, 0.03% and 0.77% defect samples did not differ from the control) for any of the measured attributes ($p > 0.05$; Table S6 & S7). At 1.33%, Number of Defect Cups (1.27 ± 0.33, $p < 0.001$) and Defect Intensity (1.47 ± 0.36, $p < 0.001$) increased (Figure S4a, b),

alongside a decrease in Clean Cup (8.14 ± 0.82, p = 0.007; Figure S5c). At 2% defect, Flavor (7.02 ± 0.21, p = 0.004), Acidity (7.11 ± 0.18, p = 0.006), and Aftertaste (6.91 ± 0.20, p = 0.010) also decreased from the control (Figure S5a, b, d). All six attributes remained significantly different from the control through 6.75% (Table S6 & S7). Fragrance/Aroma, Body, Uniformity, and Balance (Figure S6a-d) displayed less consistent patterns, with scores increasing and decreasing in a more stochastic manner rather than following a clear trend. The Clean Cup attribute (Figure S5c) showed the largest overall change of any attribute, while Fragrance/Aroma (Figure S6a) was affected the least.

Between-subject variability was highest for Uniformity (ICC = 0.43), suggesting substantial disagreement among experts (Table S7). In contrast, Defect Cups (ICC = 0.10) displayed the lowest between-subject variance.

Discussion

Under the current SCA guidelines, one full black bean in a 350g sample (~1 in 3,500 beans, or ~0.03% w/w) is sufficient to disqualify a lot from achieving 'specialty' grade[1]. Our results suggest that this limit is far stricter than what consumers and experts can detect in brewed coffee. Coffee consumers could not reliably detect a difference between a clean cup of coffee and one that contained full black defects until their concentration reached 3.64% (~127 beans per sample). Experts performed better, with an estimated detection threshold of 0.63% (~22 beans per sample) for the High-Range Series and 0.23% (~8 beans per sample) for the Low-Range Series. These results are broadly consistent with prior work by Carvalho et al., who reported a detection threshold of 10.25% for coffee spiked with a 1:1:1 mix of black, green, and sour defects amongst Brazilian consumers[10]. Both findings place human detection well above the SCA limit.

Substantial heterogeneity was present in both groups, with individual detection thresholds spanning more than an order of magnitude both within the consumers and the experts. The underlying reason for the heterogeneity is unclear. Although all consumers met the same consumption-frequency criterion, their habits differed: some added adjuncts (dairy, sweeteners, syrups, etc.), while others drank only black coffee. Expert consumption habits were similarly diverse, but unlike consumers, experts are commonly expected to have overcome such biases through training. Nonetheless, expert thresholds varied widely, consistent with research showing considerable differences among experienced assessors within the same domain[17,18].

Coffee graders evaluate coffee on a daily basis. They are trained to recognize defects in both green and brewed coffee, and those working in the industry frequently encounter these defects. In contrast, consumers drinking coffee for pleasure, habit, or utility are unlikely to have experienced defects in isolation or learned to identify them. Given this experiential gap, the difference in sensitivity between groups (experts: 0.23 - 0.63% w/w, consumers: 3.64% w/w) is unsurprising and consistent with previous findings suggesting that domain experts have more acute perception than consumers[11,19,20]. This difference is best understood as a function of training and repeated exposure rather than any underlying perceptual advantage[21]. The more a person experiences a stimulus and codes it with a specific meaning (e.g., tasting a black defect and having themselves of their peers identify it as such), the more fluent they become at recognizing it[13,22]. In the present study, this perceptual fluency may have allowed experts to effectively treat the unspecified tetrad as a specified task, searching for known attributes rather than judging overall similarity, thereby improving discrimination[23].

The gap in full black defect sensitivity between consumers and experts reported in this study should raise questions about whether expert evaluations should be weighted as heavily as

they currently are in the coffee industry, and if they truly represent consumer coffee experience. Similar tensions exist in other domains (e.g., audio engineering, radiology, textiles) where those highly familiar with a stimulus can detect differences where untrained individuals cannot[24–26]. Wine researchers have faced a similar issues with the cork taint defect, where expert thresholds have overestimated its impact on consumer perception, prompting the use of consumer rejection thresholds as a more relevant benchmark[27].

The SCA cupping form was sensitive to full black defects at concentrations well above expert thresholds determined using the ascending forced-choice method of limits. Cupping Score was unaffected by full black defect concentration of 0.77% and lower, only falling below the 80-point 'specialty' cutoff for defect concentrations of 1.33% and above. Defects did not affect all SCA cupping form attributes equally. Number of Defect Cups, Defect Intensity, and Clean Cup were the most sensitive attributes, differing significantly from the control starting at 1.33%. Flavor, Acidity, and Aftertaste were slightly less sensitive to defect concentrations, with significant deviations compared to the control at 2% and above. In contrast, Fragrance/Aroma, Body, Uniformity, and Balance responses showed no consistent pattern when compared to the defect-free reference. Additionally, substantial between-subject variability for these items suggests that graders may interpret the meaning of attributes differently. Handling factors inherent to cupping, such as crust breaking, spoon technique, and cooling over the course of a session, may additionally contribute to measurement noise. The gap between expert detection thresholds (0.23 - 0.63% w/w) and the concentration at which Cupping Scores were first found to significantly differ from the control (1.33% w/w) reflects a difference in task demands. The tetrad is a forced-choice discrimination task in which participants compare control and spiked samples side by side and focus their attention on finding differences and similarities between

them. The SCA cupping protocol, by contrast, asks graders to rate each sample on several quality attributes, so to be penalized, the sensory attributes brought upon by the defect presence must be recognized as a fault.

Several limitations should be considered when interpreting these results. First, the sample size of experts (n = 16) in Experiment 2 fell below the minimum required for 80% power using the tetrad test (n = 6). However, because the experts who participated in this study represented nearly 3% of certified Q-graders in the United States, we considered Experiment 2 informative of this unique population despite an elevated risk of Type II error. To mitigate this, each expert completed the threshold determination task multiple times (two replications for the High-Range Series and three for the Low-Range Series), increasing the number of observations per subject and improving the precision of estimates. Additionally, Experiments 1 and 2 utilized the ascending forced-choice method of limits with defect concentrations presented in increasing order, which limits carryover effects. Although consistent with ASTM standards for threshold testing, this approach may introduce order effects. For instance, participants may learn to recognize the defect as concentration gradually increases, which would lower estimated thresholds. Conversely, sensory fatigue or adaptation over the session would raise them. That being said, fatigue is unlikely to be a major concern in this study because sample volume was relatively small (15 mL).

Beyond the specific detection thresholds identified in this study, many questions remain about how full black sensitivity might shift under real-world consumption conditions. Many coffee consumers add ingredients like cream and sugar to mask coffee's less pleasant flavors or consume coffee as a minor component of a more elaborate drink (e.g., cappuccino, flat white, pumpkin spice latte)[28]. In either scenario, we might expect defect sensitivity to decrease

substantially. Results may also depend on the nature of the coffee itself, as origin, roast level, and other brewing parameters which strongly affect the sensory attributes of coffee[29–31] may also alter the sensory characteristics of defects.

In summary, this study reveals a substantial gap between the SCA's zero tolerance policy for full black defects and consumer sensitivity (127 ± 29 beans per 350 g) and expert sensitivity (8–22 beans per 350 g). Consistent with these estimates, traditional SCA cupping form evaluations performed by Q-graders showed no significant reduction in Cupping Scores when a single defect bean was present per 350-gram sample or even when a single defect bean was present per 13-gram cupping dose. Together, these findings indicate that the current grading standard for full black defects does not correspond with human sensory perception and raise the possibility that other defects may not withstand similar scrutiny. Given that adherence to overly strict standards impede coffee producers' access to specialty markets, revising them based on evidence-based sensory thresholds could reduce unnecessary downgrading without compromising cup quality. We therefore recommend that the coffee industry critically reassess green coffee grading standards.

2. Materials and Methods

The overall experimental methodology is sketched in Figure 1.

2.1 Sample Preparation

2.1.1 Green Coffee Collection and Defect Classification

Washed green Guatemalan *Coffea arabica* donated by Peet's Coffee and Tea, Inc. (Emeryville, CA, USA) served as the clean experimental material to which defects were added (Figure 2). The baseline control coffee was classified as Specialty Grade, characterized by a

physical grading of fewer than five full defect equivalents (FDE) per 350 grams, a predominant sieve retention size of 16 to 18, and a Specialty Coffee Association (SCA) Cupping Score above 80 points (SCA Cupping Score was further corroborated in Experiment 3).

Defective green *Coffea arabica* beans, designated as the defective material stock, were sourced from Guatemala and collected at various stages of quality-inspection separation processes at both wet and dry milling facilities. These samples included beans identified as substandard during quality-control procedures. After collection, we stored the samples in vacuum-sealed plastic bags for shipping to the UC Davis Coffee Center. Full black coffee beans were isolated from the defective material stock using the visual sorting protocol outlined by the SCA.

For long-term storage, samples were transferred to environmental chambers (Model 7000-25-1; Caron Inc., Marietta, OH, USA) with atmospheric conditions set to simulate industry warehouse storage parameters (20°C, 60% relative humidity). Coffee beans reached a wet-basis moisture content of 10.0 ± 0.5% following a 10-day equilibration period, as described by Anokye-Bempah et al.[32].

2.1.2 Defect Concentration Series

A preliminary discrimination test was conducted with lab personnel to estimate the concentration at which full black defects in coffee could be identified 50% of the time[33,34]. This test identified 3% as a relevant center point for the following series. A five-point series (High-Range Series: 1.33%, 2%, 3%, 4.5%, and 6.75% w/w) was geometrically spaced around this center point, yielding equal log-unit steps consistent with ASTM E679-19[15]. The High-Range

Series was evaluated by the consumer panel in Experiment 1 and by the expert panel in Experiment 2.

Unexpectedly, expert participants in Experiment 2 were able to identify the spiked samples better than expected across the High-Range Series. Therefore, to better characterize expert detection sensitivity a second concentration series spanning a lower range (Low-Range Series: 0.16%, 0.33%, 0.67%, and 2% w/w) was introduced in Experiment 2. The 2% concentration was retained from the original series to provide an overlapping data point to measure consistency between the High-Range and Low-Range Series.

Following Experiment 2, the same experts performed a grading task (Experiment 3) using the SCA cupping protocol[7]. Defect concentrations in Experiment 3 consisted of the High-Range Series plus two additional levels: a sample containing one full black bean per 350 g (0.03% w/w), corresponding to the maximum allowable defect count for SCA specialty-grade classification, and a sample containing one full black bean in a standard 13 g cupping dose (0.77% w/w), simulating incidental inclusion of a single defect bean in a grading sample. In total, Experiment 3 included the following defect concentrations: 0.03% (1 bean/350 g), 0.77% (1 bean/13 g), 1.33%, 2%, 3%, 4.5%, and 6.75% w/w.

Because defects are typically reported in the coffee industry as counts per 350 grams rather than as a mass percentage, to provide context we converted our defect mass percentages to an approximate number of unroasted full black beans per 350 grams. For the black beans used here, individual mass measurements of 160 randomly selected full black beans yielded a mean mass of 0.086 ± 0.028 grams. Accordingly, throughout this work we approximate the mass of a single full black bean as 0.1 grams, rounded slightly for ease of calculation. We emphasize that

the black bean counts reported here are thus approximations intended to provide context on the scale customary within the coffee industry, while the reported mass percentages are exact.

2.1.3 Roasting

Before roasting, defective coffee beans were blended with clean beans to prepare spiked coffee samples at the target concentrations, as described in Section 2.1.2. All samples were roasted on a drum sample roaster (Model PIII P01e; Probat GmbH, Emmerich am Rhein, Germany). We adjusted green coffee batch sizes across trials to meet volumetric requirements; for spiked samples, we used 450 g in Experiment 1 and 350 g in Experiments 2 and 3, and we roasted clean control coffee at 450 g in Experiment 1 and 700 g in Experiments 2 and 3. Because subdividing roasted spiked whole-bean batches introduces sampling error and heterogeneity in defect distribution, each spiked lot was roasted as a single unit tailored to the working volume of the subsequent ground coffee homogenization. Description of the homogenization equipment and downstream sample partitioning protocols are provided in Section 2.1.4.

Experiment 1 samples were roasted one day before the first day of testing, which spanned three days. Samples for Experiments 2 and 3 were roasted five days before the first day of testing, which spanned two days. All roasted whole bean lots were stored at a temperature of 21 ± 1 °C in heat-sealed, multi-barrier laminate foil pouches with one-way degassing valves (Item #1001-080-230-V, Sav-on Bags, San Jose, CA, USA).

All batches were roasted under a similar roast profile to a final drop temperature of 207.40 ± 0.21 °C. Thermodynamic curve data and roasting parameters were logged using Cropster Roasting Intelligence software (version 5.11.3; Cropster Inc., Sacramento, CA, USA), and the roast profiles are illustrated in Figure S3. Color measurements of the ground roasted

coffee using a spectrophotometer (ColorFlex EZ, Hunter Associates Laboratory, Inc., Reston, VA, USA) exhibited a total color difference ($\Delta E = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2}$) of less than 2.50 relative to the pooled clean control reference centroid (L^* = 26.30, a^* = 10.25, b^* = 12.65). Individual clean control batches exhibited a dispersion of 0.81 ± 0.09, while defect blends averaged ΔE = 1.60 ± 0.14. Defect loading produced no major differences in overall roast color development (Figure S4). CIELAB color coordinates measurements are provided in Tables S3 and S4.

2.1.4 Grinding and Homogenization

Control sample roasts were homogenized prior to being ground into two distinct baseline matrices: Experiment 1 control roasts were pooled independently, whereas control roasts for Experiments 2 and 3 were combined into a single composite lot to ensure matrix uniformity across these concurrently executed trials.

Following roasting, samples were ground using a commercial coffee grinder (Model KF 804; Ditting Maschinen AG, Bachenbülach, Switzerland) set to standard paper filter size. Particle size distribution across grinding sessions and defect concentrations was determined via laser diffraction equipped with a dry powder disperser unit (Mastersizer 3000+ Aero Series, Malvern Panalytical Ltd., Malvern, U.K.). Representative session median particle sizes (D_{50}) were 994 µm (Experiment 1) and 1050 µm (Experiments 2 and 3). Across all defect loadings (N = 22 measured aliquots), the median particle size remained uniform (D_{50} = 1002 ± 25 µm, compared to 994 ± 13 µm for clean controls), with no major differences in particle size distribution resulting from the inclusion of full black beans (Figure S3). Comprehensive volume percentiles (D_{10}, D_{50}, D_{90}) and relative spans are reported in Supplementary Table S5.

Ground coffee was homogenized using a 2-liter V-type powder mixer (Model 2L V-Type; Lüliang, Shanxi, China) in 350-gram batches within a 304 stainless steel drum at a fixed rotational speed of 24 rpm for 8 minutes[35].

Homogenized samples were portioned into either 13 or 46-gram doses, vacuum-sealed in plastic bags, and stored at -40 °C until brewing. The three lowest Low-Range Series concentrations (0.16%, 0.33%, and 0.67% w/w) were prepared by spiking clean control coffee with the ground homogenized 1.33% w/w samples from the High-Range Series. The 2% w/w Low-Range Series sample was drawn from the original High-Range Series material.

2.1.5 Brewing

All coffees for Experiments 1 and 2 were brewed using a commercial batch brewer (Model 1200.0102; Bunn-O-Matic Co., Springfield, IL, USA) in accordance with SCA Protocols and Best Practices [7]. For each batch, 46 g of ground coffee was brewed with water at a temperature of 94.0 ± 2 °C, using a coffee-to-water ratio of 1:18, yielding approximately 0.73 L of brewed coffee. Water was filtered via reverse osmosis and remineralized with a BWT Bestmin M Premium cartridge, targeting a total dissolved solids concentration of 100 to 150 mg/L, in line with SCA water quality standards[36].

Brewing began with a 5-second pre-infusion (96.6 mL of water) followed by a 30-second pause, then seven pulse-brew cycles of 10 seconds (193.3 mL of water per cycle) separated by 15 second pauses. This profile delivered a total of 830 mL of water over a total brew time of 3 minutes and 27 seconds, yielding 0.73 L of brewed coffee.

Samples brewed for Experiment 1 were transferred to heated carafes (Model 27850.0200; Bunn-O-Matic Co., Springfield, IL, USA) and held at approximately 75 °C on warmers (Model

27875.0200; Bunn-O-Matic Co., Springfield, IL, USA) for no longer than 3 hours. Heated servers and warmers were unavailable for Experiment 2, as such, brewed coffee was instead held in 0.6 L vacuum-insulated stainless-steel carafes (Model FN369; Thermos LLC, Schaumburg, IL, USA) for no longer than 40 minutes.

Samples for Experiment 3 were prepared in accordance with the SCA cupping protocol [7].Water from the source described above was heated to 93 °C and poured over 13g of ground coffee to the rim of a 237 mL (8 fl oz) cup (1:18 brew ratio). Grounds steeped for 3 to 5 minutes before the crust was broken and skimmed.

2.2 Participants and Recruitment

Coffee consumers (n = 69) in Experiment 1 were recruited from Davis, California. All consumers met inclusion criteria requiring consumption of drip, espresso, or immersion-brewed coffee several times per week or more. Expert participants (n = 16) in Experiments 2 and 3 were recruited from across the United States, 15 of whom held current Q-grader certification[37]. All experts were professionals working in the coffee industry. One expert was unavailable for the High-Range Series of Experiment 2 (High-Range Series, n = 15; Low-Range Series, n = 16). Experiment 3 was restricted to certified Q-graders (n = 15), as the task required application of the SCA cupping protocol. Further demographic information is provided in Table S2.

A power analysis was conducted targeting an effect size of d' = 1 with the unspecified tetrad test (Power = 0.80; $\alpha = 0.05$), yielding a minimum required sample size of n = 61. This target was impractical for the expert group, given that only 512 credentialed Q-graders are located in the United States[38]. To reduce the impact of this constraint, experts repeated the

ascending series: two replications for the High-Range Series and three replications for the Low-Range Series.

All participants provided informed consent. Experimental protocol was approved by the Institutional Review Board at the University of California, Davis (IRB # 2371841-1).

2.3 Testing Procedure

Experiments 1 and 2

Experiments 1 and 2 employed the ascending forced-choice method of limits procedure, where successive comparisons involved progressively higher concentrations of defect beans (Section 2.1.2). In Experiment 1, the ascending series was completed once per subject, while in Experiment 2, it was conducted twice for the High-Range Series and three times for the Low-Range Series. Sample discrimination was evaluated using the unspecified tetrad test, which involved presenting two defect-spiked samples alongside two clean coffee control samples[33]. Participants were instructed to taste all samples and identify the two matching pairs. Sample presentation order was balanced across participants.

Samples (15 mL) were dispensed from carafes immediately before presentation and served in 50 mL double-walled borosilicate glass espresso cups at a temperature of 75°C. No rest period was enforced between sample sets. A cup was provided in case participants wished to spit out the samples. Responses were recorded on paper ballots. Sessions lasted approximately 20 minutes.

Experiment 3

In Experiment 3, expert participants evaluated each defect concentration using the SCA cupping protocol[7]. Three cups were presented per unique sample. In addition to the seven spiked samples outlined in Section 2.1.2 three sets of clean samples were added to the sample set. Sample presentation order was randomized, and the sample contents were labeled with blinding codes. Experts tasted samples using cupping spoons, prior to scoring each sample in accordance with the complete SCA cupping form, evaluating all standard attributes: Fragrance/Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity (similarity of triplicate samples), Balance, Clean Cup, Sweetness, Overall, Number of Defect Cups (present in triplicate sample), and Defect Intensity (if present).

2.4 Statistical Analysis

Experiments 1 and 2

All statistical analyses were conducted using R (version 4.6.1; R Core Team, 2026). Participant responses from individual tetrad tests were coded as binary outcomes (correct = 1, incorrect = 0) and fitted using a generalized linear mixed-effects model (GLMM) with a probit link (*glmer* function, lme4 package, version 2.0-6; Bates et al., 2015). Defect concentration was included as a fixed effect, and individual differences in defect sensitivity were modeled with random slopes by subject. The intercept was shared across participants (fixed) and estimated freely rather than set to the tetrad's chance level of 1/3, allowing baseline performance at zero defect concentration to be estimated from the data. The model was specified as follows:

$$\Phi^{-1}\left(P_{\mathrm{c},ij}\right) = \beta_0 + (\beta_1 + b_{1j})C_{\mathrm{def},ij} \quad (1)$$

where Φ^{-1} is the is the probit link, $P_{\mathrm{c},ij}$ is the probability of a correct response for observation i by subject j, $\mathrm{C}_{\mathrm{def},ij}$ is the defect concentration (% w/w) of observation i by subject j, β_0 is the

probability of getting a correct response when $C_{\mathrm{def},ij} = 0$, β_1 is the average effect of defect concentration on the probability of a correct response, and b_{1j} is the subject $j's$ deviation from β_1 and is assumed to be normally distributed, $b_{1j} \sim \mathrm{N}\left(0, \sigma_{b_1}^2\right)$. Models were fitted separately for each experiment.

Expert responses for the overlapping 2% w/w concentration shared between the High- and Low-Range Series were compared using a GLMM with a probit link, with concentration series as a fixed effect and a random intercept for subject. The model was specified as follows:

$$\Phi^{-1}\left(P_{\mathrm{c},ij}\right) = \beta_0 + \beta_1 S_{ij} + b_{0j} \quad (2)$$

where Φ^{-1} is the probit link, $P_{\mathrm{c},ij}$ is the probability of a correct response for observation i by subject j, β_0 is average performance on for the High-Range Series (reference), β_1 is the difference between the Low- and High-Range Series, S_{ij} is a dummy variable (0 = High-Range Series, reference; 1 = Low-Range Series), and b_{0j} is subject $j's$ deviation from β_0 and is assumed to be normally distributed, $b_{0j} \sim \mathrm{N}\left(0, \sigma_{b_0}^2\right)$.

For each model, fixed-effect coefficients are reported alongside their standard errors (SE) Wald z-statistics, and p-values. Random effects are reported as their standard deviations (σ). Statistical significance was set at $\alpha = 0.05$ for all analyses.

Population detection threshold was defined as the concentration at which d′ = 1, corresponding to a probability of a correct response (P_{c}) of 0.49 for the unspecified tetrad (*psyfun* function, sensR package, version 1.5-3, Christensen & Brockhoff, 2008). The population detection threshold ($\hat{T}$) was then estimated by solving the fitted model for the defect concentration at this value:

$$\hat{T} = \frac{\Phi^{-1}(P_c) - \hat{\beta}_0}{\hat{\beta}_1} \quad (3)$$

where Φ^{-1} is the inverse normal probit function, P_c is the corresponding probability of a correct response, $\hat{\beta}_0$ is the estimated intercept, and $\hat{\beta}_1$ is the estimated slope for defect concentration. A standard error (SE) for the estimated population threshold was obtained with the delta method [42] (*deltaMethod* function, car package, version 3.1-5; Fox & Weisberg, 2019).

Individual estimated detection thresholds ($\hat{T}_j$) were estimated by substituting the fixed slope ($\hat{\beta}_1$) in Equation 3 with each individual's conditional slope:

$$\hat{T}_j = \frac{\Phi^{-1}(P_c) - \hat{\beta}_0}{(\hat{\beta}_1 + \hat{b}_{1j})} \quad (4)$$

where $\hat{b}_{1j}$ is subject $j's$ estimated deviation from the fixed slope. The distribution of these individual thresholds was summarized using a histogram (bin width of 0.50% w/w) to visualize individual differences in defect sensitivity.

Experiment 3

Item responses from the modified SCA cupping protocol (Section 2.3) were analyzed separately for each of the twelve scored cupping attributes (Fragrance/Aroma, Flavor, Aftertaste, Acidity, Body, Uniformity, Balance, Clean Cup, Sweetness, Overall, Defect Intensity, and Number of Defect Cups). The composite Cupping Score was calculated as the sum of all attribute scores minus the product of the number of Defect Cups and Defect Intensity. For each response, a linear mixed-effects model was fitted with defect concentration as a fixed effect and a random intercept for subject (*lmer* function, lme4 package, version 2.0-6, Bates et al., 2015):

$$y_{ij} = \beta_0 + \sum_{k=1}^{7} \beta_k \mathrm{D}_{kij} + b_{0j} + e_{ij} \quad (5)$$

where y_{ij} is the response score from subject j on observation i, β_0 is the defect-free control sample mean, β_k is the difference between concentration k and the control, D_{kij} is a dummy variable equal to 1 if the sample was at concentration k (k = 1, …, 7) and 0 otherwise, $b_{0j} \sim N(0, \sigma_{b_0}^2)$ is subject $j's$ random intercept, and $e_{ij} \sim N(0, \sigma_e^2)$ is the residual error. Significance was assessed with Satterthwaite-approximated degrees of freedom (lmerTest package, version 3.2-1, Kuznetsova et al., 2017), Some cups were left unscored by individual Q-graders, as such the number of observations varied slightly by attribute (Table S7). Missing scores were treated as missing values, and models were fitted to all available data.

The three independently presented defect-free control samples were compared with respect to each scored cupping form attribute using a linear mixed model with sample as a fixed effect and a random intercept for participants:

$$y_{ij} = \beta_0 + \beta_1 \text{Control}_{1ij} + \beta_2 \text{Control}_{2ij} + b_{0j} + e_{ij} \quad (6)$$

where y_{ij} is the response score from subject j on observation i, β_0 grand mean across the three defect-free control samples, β_1 and β_2 are the differences between control sample 1 and control 2 from the grand mean. Control_{1ij} and Control_{2ij} are effects-coded variables equal to 1 for Control 1 or Control 2, respectively, and -1 for Control 3. $b_{0j} \sim N(0, \sigma_{b_0}^2)$ is subject $j's$ random intercept, and $e_{ij} \sim N(0, \sigma_e^2)$ is the residual error. Significance was assessed with Satterthwaite-approximated degrees of freedom (lmerTest package, version 3.2-1, Kuznetsova et al., 2017). Analyses showed that control samples did not differ significantly for any scored attribute ($p>0.05$) and as such were pooled as a single reference level.

Intraclass correlation coefficients (ICCs) were determined by calculating the proportion of variance explained between participants σ_{b0}^2 with respect to the total variance within the dataset $\sigma_{b0}^2 + \sigma_e^2$ (*icc* function, performance package, version 0.17.1; Lüdecke et al., 2021).

$$\mathrm{ICC} = \frac{\sigma_{b0}^2}{\sigma_{b0}^2 + \sigma_e^2} \quad (7)$$

Funding

This work was supported by Keurig Dr Pepper Inc., the Folgers Coffee Company (a subsidiary of The J.M. Smucker Company), and seed funding from the Specialty Coffee Association. The funders had no role in study design, data collection and analysis, manuscript preparation, or decisions to publish.

Acknowledgements

We thank the student assistants at the UC Davis Coffee Center and the Delarue Lab for sorting defects and administering the psychophysical tests; ANACAFE, Roberto Soto, and coffee producers in Guatemala for procuring clean and defective beans; Timothy Styczynski for roasting the samples; and Kyle Sala and Hannah Zukswert for assisting with the expert sensory experiments.

Author Contributions

D.S. designed the study, administered psychophysical assessments, analyzed the data, and wrote the original draft. W.K. managed all sample preparation and performed all physical measurements. I.R.D.G. procured the samples from Guatemala. W.D.R. procured funding. J.D. supervised study design and overall execution. All authors contributed to the conception of this study and reviewed and approved the final manuscript.

References

1. Specialty Coffee Association. *Washed Arabica Green Coffee Defect Guide*. (2013).
2. Syers, R. *The Coffee Guide*. (Liverpool : D. Marples, 1832).
3. Thurber, F. B. *Coffee; from Plantation to Cup. A Brief History of Coffee Production and Consumption. With an Appendix Containing Letters Written during a Trip to the Coffee Plantations of the East and through the Coffee Consuming Countries of Europe*. (New York, American grocer publishing association, 1881).
4. United States. Department of Commerce and Labor. Bureau of Manufactures. Monthly Consular and Trade Reports, June 1907, No. 321. (1907).
5. Ukers, W. H. *All about Coffee*. (New York The Tea and Coffee Trade Journal Co, 1922).
6. Caspersen, B. A. DO YOU USE THE SCAA'S GREEN ARABICA COFFEE CLASSIFICATION SYSTEM? *Roast Magazine* 25–40 (2016).
7. Specialty Coffee Association. *Protocols and Best Practices*. https://www.scith.coffee/wp-content/uploads/2021/03/SCA-Protocols-_-Best-Practices.pdf (2018).
8. Barrios Rodriguez, Y., Salas Calderon, K. & Hernández, J. Comparison of sensory attributes and chemical markers of the infrared spectrum between defective and non-defective Colombian coffee samples. *Coffee Science* **15**, 1–10 (2020).
9. Khalifé, C. *Tasting the Standards: A Comprehensive View of Green Coffee Defects*. (YouTube, 2024).

10. Carvalho, C. de M. *et al.* Does the presence of defective beans in Arabica healthy coffee beans affect the sensory perception and acceptance of coffee beverages? *Journal of Sensory Studies* **38**, e12873 (2023).
11. Hopfer, H. & Heymann, H. Judging wine quality: Do we need experts, consumers or trained panelists? *Food Quality and Preference* **32**, 221–233 (2014).
12. Pickering, G. J., Karthik, A., Inglis, D., Sears, M. & Ker, K. Determination of Ortho- and Retronasal Detection Thresholds for 2-Isopropyl-3-Methoxypyrazine in Wine. *Journal of Food Science* **72**, S468–S472 (2007).
13. Tempere, S., de Revel, G. & Sicard, G. Impact of learning and training on wine expertise: a review. *Current Opinion in Food Science* **27**, 98–103 (2019).
14. Tempere, S. *et al.* "Brett character" in wine: Is there a consensus among professional assessors? A perceptual and conceptual approach. *Food Quality and Preference* **34**, 29–36 (2014).
15. ASTM. Standard Practice for Determination of Odor and Taste Thresholds By a Forced-Choice Ascending Concentration Series Method of Limits. (2019) doi:10.1520/E0679-19.
16. Ennis, J. M. & Jesionka, V. The Power of Sensory Discrimination Methods Revisited. *Journal of Sensory Studies* **26**, 371–382 (2011).
17. Hodgson, R. T. An Examination of Judge Reliability at a major U.S. Wine Competition. *Journal of Wine Economics* **3**, 105–113 (2008).
18. Tempere, S. *et al.* The Training Level of Experts Influences their Detection Thresholds for Key Wine Compounds. *Chem. Percept.* **4**, 99–115 (2011).

19. Ballester, J., Patris, B., Symoneaux, R. & Valentin, D. Conceptual vs. perceptual wine spaces: Does expertise matter? *Food Quality and Preference* **19**, 267–276 (2008).
20. Borgogno, M., Favotto, S., Corazzin, M., Cardello, A. V. & Piasentier, E. The role of product familiarity and consumer involvement on liking and perceptions of fresh meat. *Food Quality and Preference* **44**, 139–147 (2015).
21. Ishii, R., Kawaguchi, H., O'Mahony, M. & Rousseau, B. Relating consumer and trained panels' discriminative sensitivities using vanilla flavored ice cream as a medium. *Food Quality and Preference* **18**, 89–96 (2007).
22. Stevenson, R. J. & Boakes, R. A. A mnemonic theory of odor perception. *Psychol Rev* **110**, 340–364 (2003).
23. O'Mahony, M., Masuoka, S. & Ishii, R. A Theoretical Note on Difference Tests: Models, Paradoxes and Cognitive Strategies. *Journal of Sensory Studies* **9**, 247–272 (1994).
24. Hussain, Z. An expert advantage in detecting unfamiliar visual signals in noise. *Proceedings of the National Academy of Sciences* **117**, 25935–25941 (2020).
25. Kanai, H. *et al*. Identification of principal factors of fabric aesthetics by the evaluation from experts on textiles and from untrained consumers. *Textile Research Journal* **81**, 1216–1225 (2011).
26. Schinkel-Bielefeld, N., Lotze, N. & Nagel, F. Audio quality evaluation by experienced and inexperienced listeners. *Proc. Mtgs. Acoust.* **19**, 060016 (2013).
27. Prescott, J., Norris, L., Kunst, M. & Kim, S. Estimating a "consumer rejection threshold" for cork taint in white wine. *Food Quality and Preference* **16**, 345–349 (2005).
28. YouGov. Coffee | YouGov Poll: September 21 - 25, 2022. (2022).

29. Bhumiratana, N., Adhikari, K. & Chambers, E. Evolution of sensory aroma attributes from coffee beans to brewed coffee. *LWT - Food Science and Technology* **44**, 2185–2192 (2011).
30. Frost, S. C., Ristenpart, W. D. & Guinard, J.-X. Effects of brew strength, brew yield, and roast on the sensory quality of drip brewed coffee. *Journal of Food Science* **85**, 2530–2543 (2020).
31. Guinard, J.-X. *et al.* A new Coffee Brewing Control Chart relating sensory properties and consumer liking to brew strength, extraction yield, and brew ratio. *Journal of Food Science* **88**, 2168–2177 (2023).
32. Anokye-Bempah, L., Han, J., Kornbluth, K., Ristenpart, W. & Donis-González, I. R. The use of desiccants for proper moisture preservation in green coffee during storage and transportation. *Journal of Agriculture and Food Research* **11**, 100478 (2023).
33. Rousseau, B. Sensory discrimination testing and consumer relevance. *Food Quality and Preference* **43**, 122–125 (2015).
34. Tanner Jr., W. P. & Swets, J. A. A decision-making theory of visual detection. *Psychological Review* **61**, 401–409 (1954).
35. Ludwig, I., Gerassimenko, A. & Imgrund, P. Investigation of Metal Powder Blending for PBF-LB/M Using Particle Tracing with Ti-6Al-4V. *Journal of Manufacturing and Materials Processing* **8**, 151 (2024).
36. Specialty Coffee Association. *CVA Protocol - A System to Assess Coffee Value*. (2024).
37. Specialty Coffee Association. Education - Q Grader. *Specialty Coffee Association* https://sca.coffee/education/q-grader (2026).

38. Coffee Quality Institute. Arabica Graders. https://database.coffeeinstitute.org/users/graders/arabica (2026).
39. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2026).
40. Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software* **67**, 1–48 (2015).
41. Christensen, R. H. B. & Brockhoff, P. B. sensR: Thurstonian Models for Sensory Discrimination. 1.5-3 https://doi.org/10.32614/CRAN.package.sensR (2008).
42. Moscatelli, A., Mezzetti, M. & Lacquaniti, F. Modeling psychophysical data at the population-level: The generalized linear mixed model. *Journal of Vision* **12**, 26 (2012).
43. Fox, J. & Weisberg, S. *An R Companion to Applied Regression*. (SAGE, Los Angeles, 2019).
44. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. **lmerTest** Package: Tests in Linear Mixed Effects Models. *J. Stat. Soft.* **82**, (2017).
45. Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P. & Makowski, D. performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *JOSS* **6**, 3139 (2021).

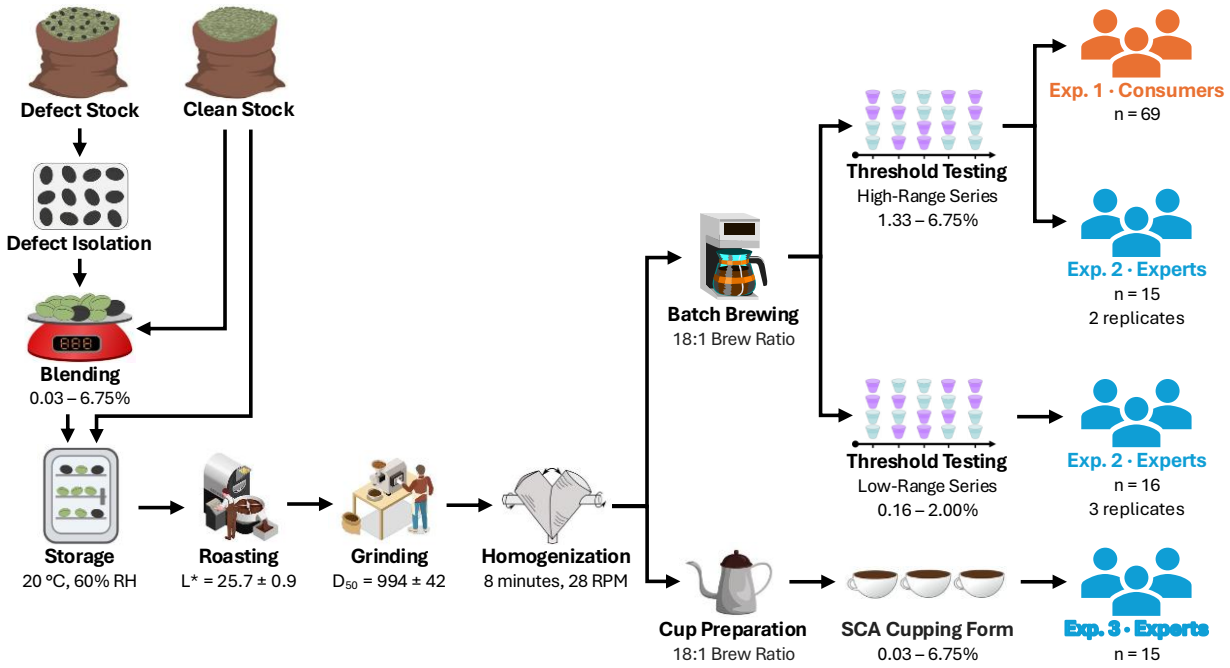


Figure 1. Schematic of experimental workflow. Defective green coffee was blended with clean stock at target concentrations, roasted, ground, homogenized, and prepared for threshold testing (Experiments 1 and 2) and evaluation using the traditional SCA cupping form (Experiment 3).

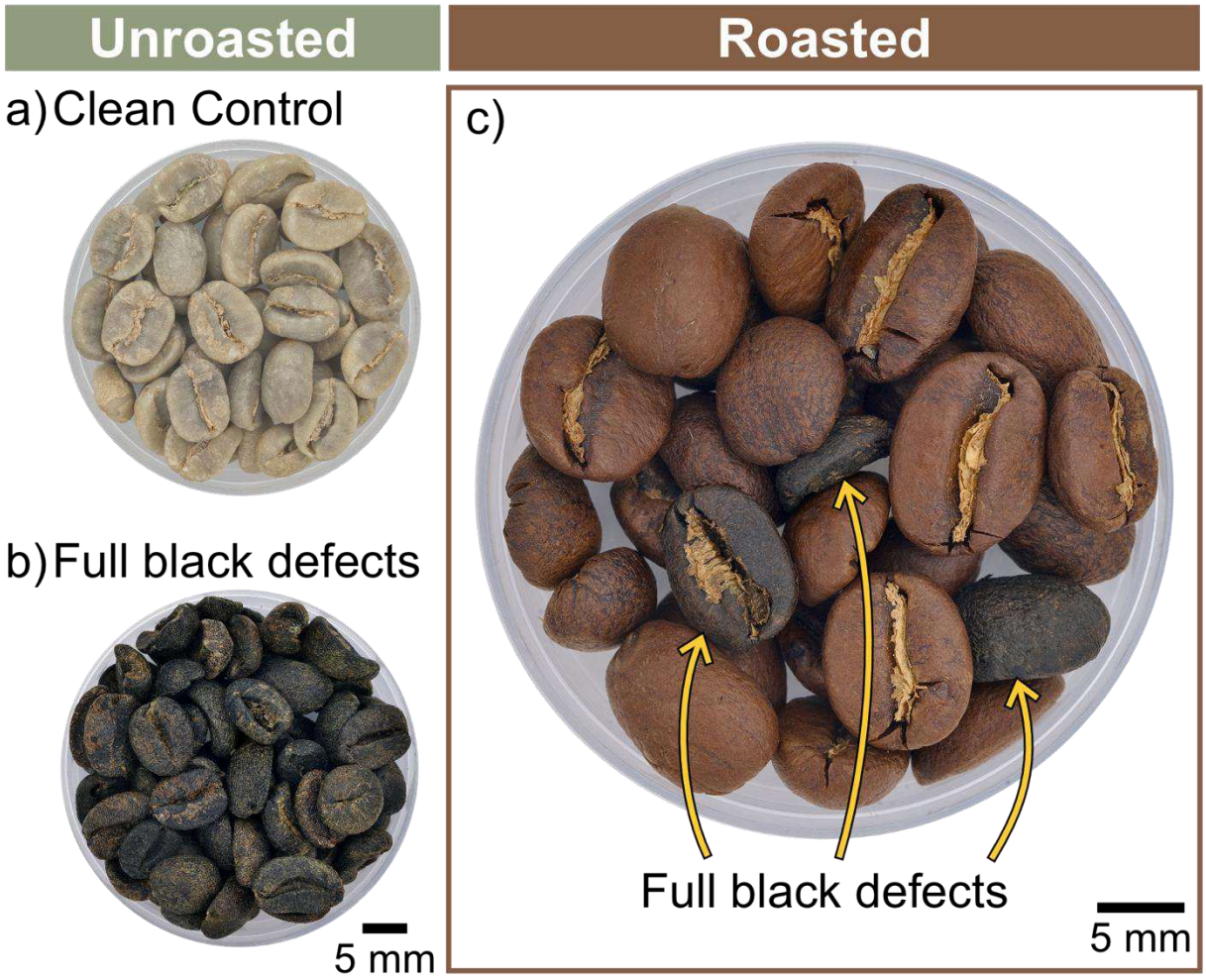
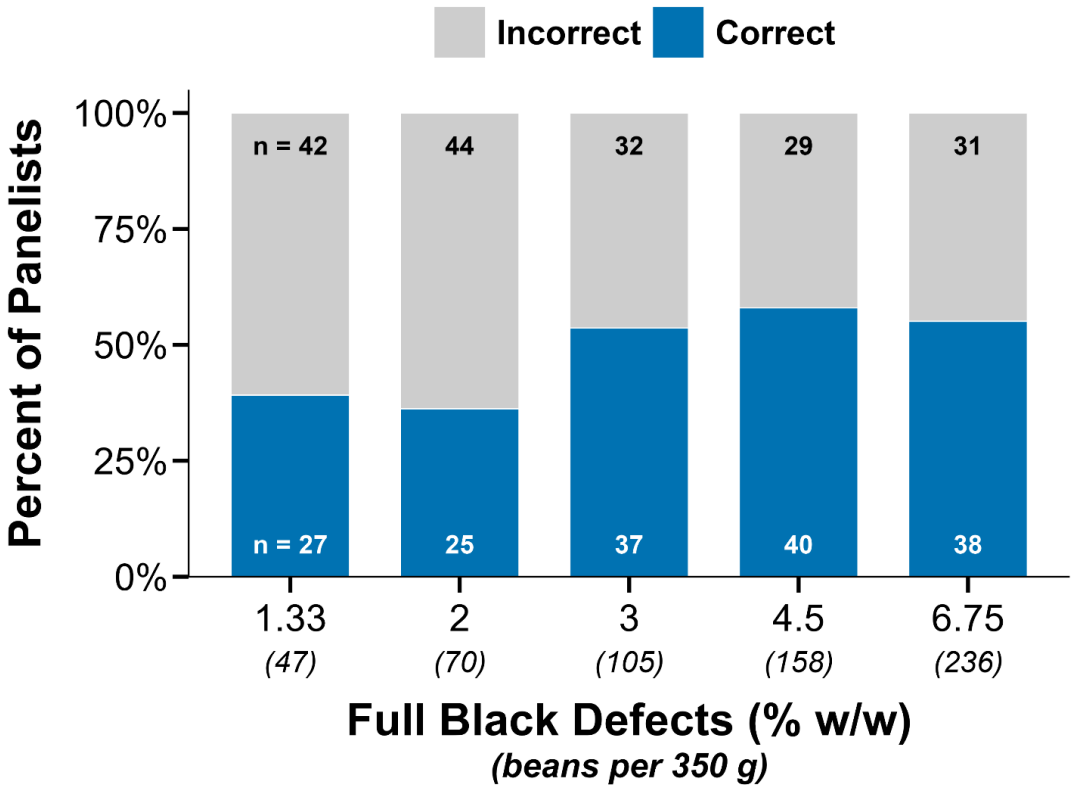


Figure 2. Green and roasted coffee samples. (a) Unroasted clean control beans. (b) Unroasted full black defect beans. (c) Roasted blend of clean beans and full black defect beans (yellow arrows). Scale bars: 5 mm.



644

645 **Figure 3.** Raw outcomes of tetrad testing across full black defect concentrations (Experiment 1,

646 n = 69). Numbers inside the stacked bars represent the number of incorrect (grey) and correct

647 (blue) responses.

648

649

650

651

652

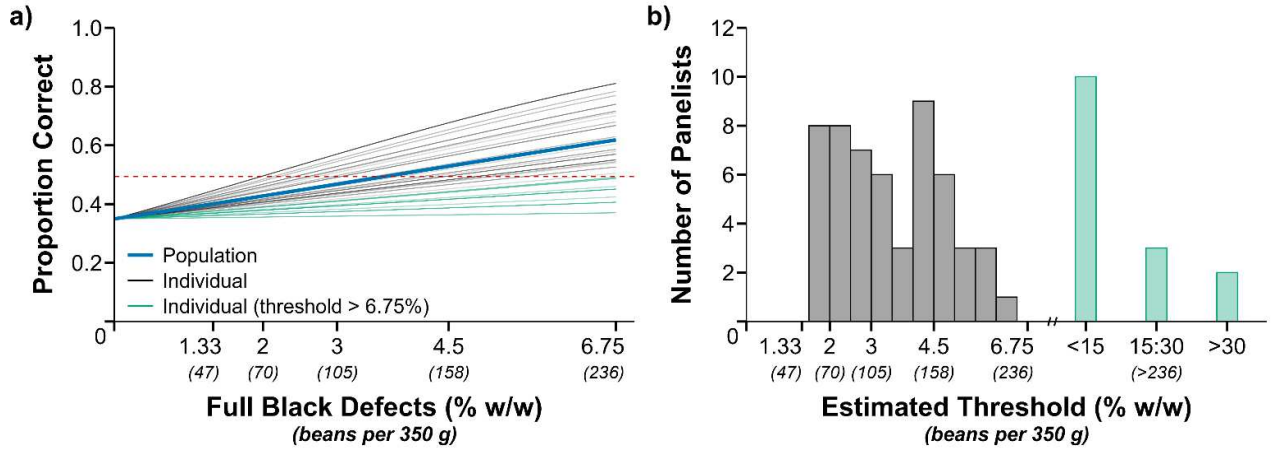


Figure 4. (a) Population and individual psychometric functions for detection of full black defects in brewed coffee (Experiment 1, n = 69). The blue line represents the fitted probit GLMM population-level function relating proportion correct to full black defect concentration. Black lines represent subject-level psychometric functions whose estimated thresholds fall within the tested range (1.33 - 6.75%). Green lines represent subject-level psychometric functions whose estimated thresholds are beyond the highest tested concentration (6.75%). The red dashed line indicates the detection threshold criterion at d′ = 1 (P_c = 0.49). **(b)** Distribution of estimated subject-level full black defect detection thresholds (Experiment 1, n = 69). Thresholds are binned at 0.5% intervals within the tested range (1.33 - 6.75%). Estimates falling above this range are binned into 6.75–15%, 15–30%, and >30% categories, separated by a break in the x-axis.

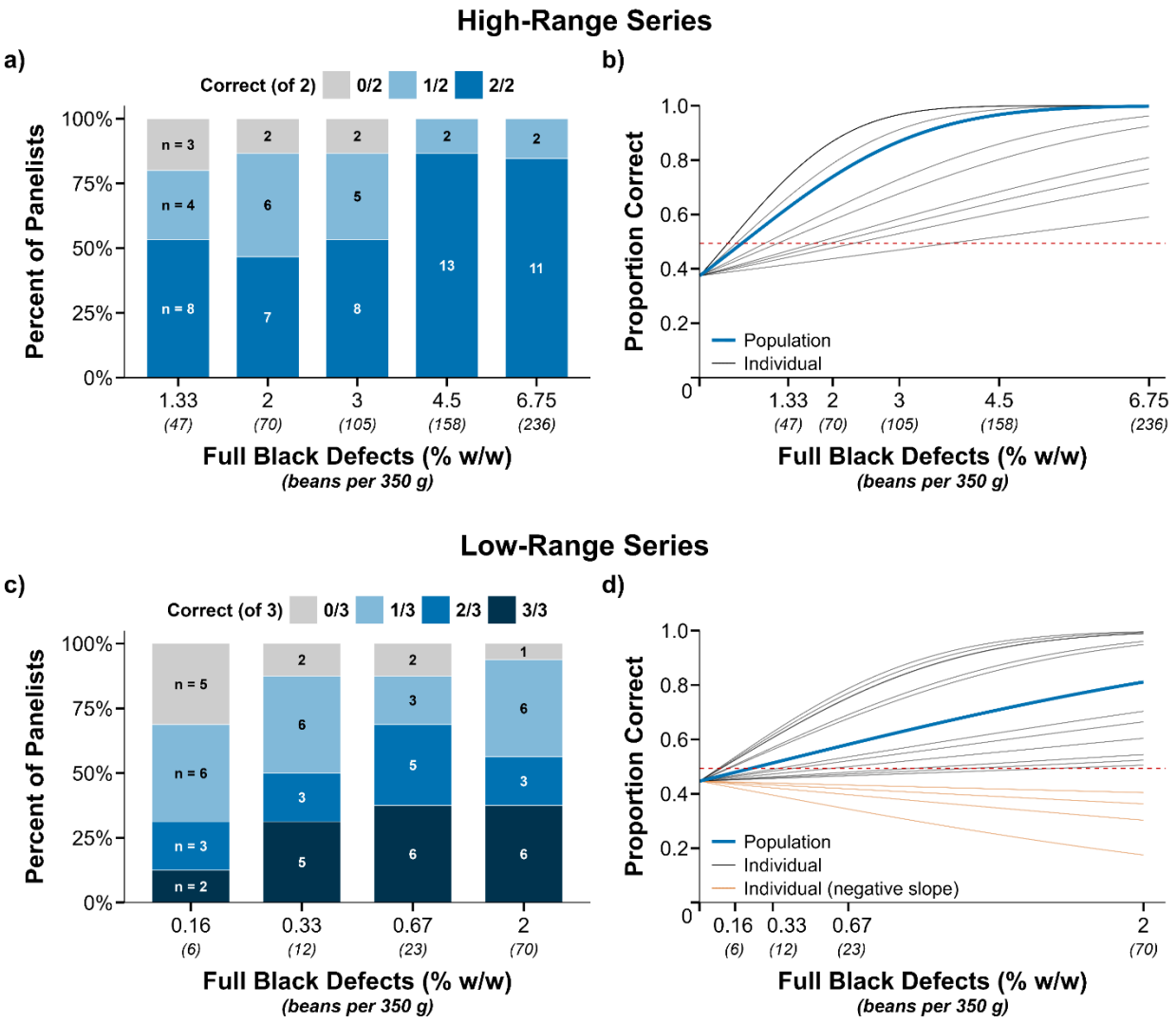
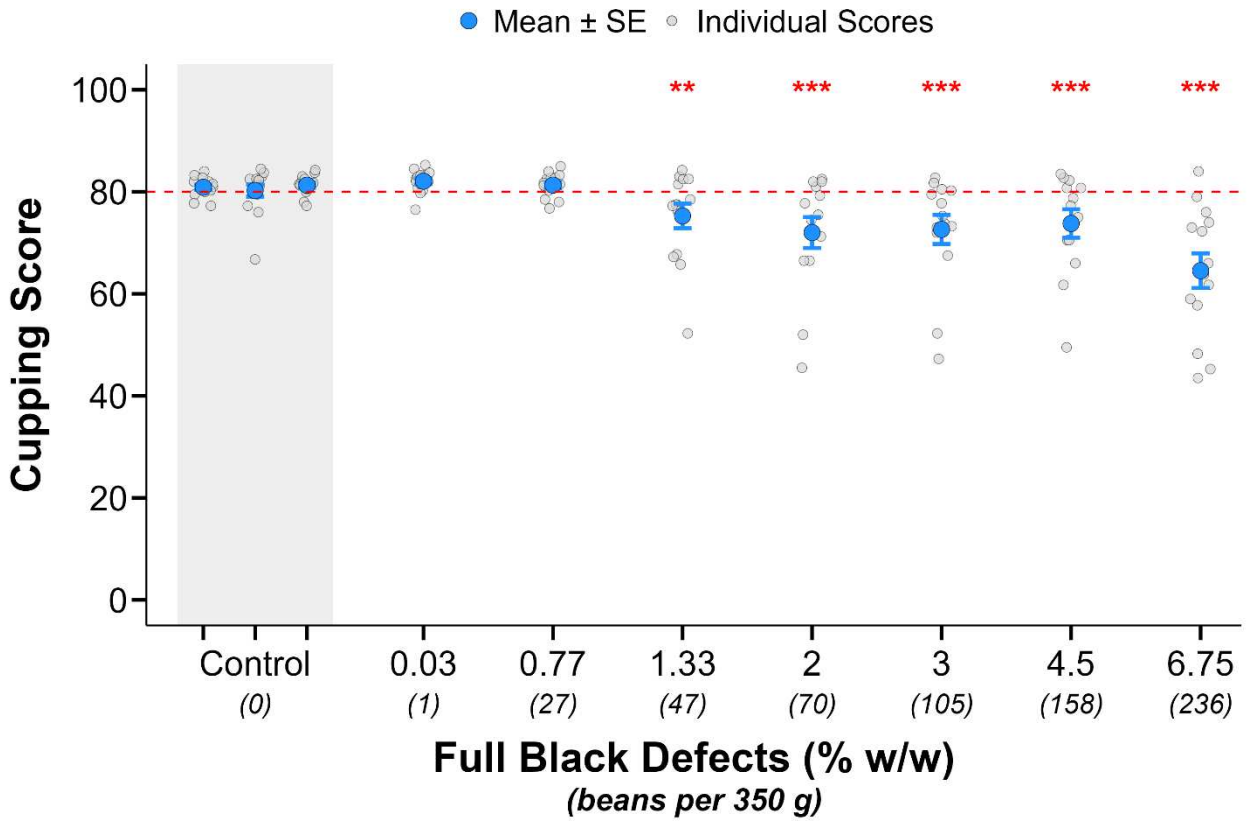


Figure 5. (a) Raw performance outcomes of tetrad testing across full black defect concentrations (Experiment 2, High-Range Series, n = 15). Bars show the number of participants scoring 0, 1, or 2 correct of 2 replicated tetrad tests at each concentration. Numbers located within each segment give the subject count. Participants (n = 2) with incomplete concentration replications for 6.75% were excluded from the plot. **(b)** Population and individual psychometric functions for detection of the full black defect in brewed coffee (Experiment 2, High-Range Series, n = 15). The blue line represents the fitted probit GLMM population-level function relating proportion correct to defect concentration. Black lines represent subject-level psychometric functions. The red dashed

line indicates the detection threshold criterion at d′ = 1 (P_c = 0.49). **(c)** Raw performance outcomes of tetrad testing across full black defect concentrations (Experiment 2, Low-Range Series, n = 16). Bars show the number of participants scoring 0, 1, 2, or 3 correct of 3 replicated tetrad tests at each concentration. Numbers located within each segment give the subject count. Participants with incomplete concentration replications (n = 1) were excluded from the plot. **(d)** Population and individual psychometric functions for detection of the full black defect in brewed coffee (Experiment 2, Low-Range Series, n = 16). The blue line represents the fitted probit GLMM population-level function relating proportion correct to defect concentration. Black lines represent subject-level psychometric functions. Orange lines represent subject-level psychometric functions with negative estimated slopes; this is an artifact of the model's fixed intercept rather than an indication of declining detection with increasing defect concentration. The red dashed line indicates the detection threshold criterion at d′ = 1 (P_c = 0.49).



687

688 **Figure 6.** Cupping Score from traditional SCA cupping form evaluations across full black defect
689 concentrations (Experiment 3, n = 15). Blue points represent mean Cupping Scores ± standard
690 error. Grey points represent individual panelist scores. Individual ticks labeled 'Control'
691 represent independently presented replicates of defect-free samples. Asterisks indicate significant
692 differences from the pooled control (0%) reference. The red dashed line indicates the SCA
693 specialty-grade threshold (80 points). Exact means and standard errors (SEs) are reported in
694 Table S6; p-values in Table S7.

695

696

697

698

699

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.docx
- SupplementaryFigures.docx