Received: 9 December 2019 | Revised: 13 April 2020 | Accepted: 21 June 2020

DOI: 10.1111/joss.12596

SHORT CONTRIBUTIONS

Journal of Sensory Studies

WILEY

Inter-rater reliability of “clean cup” scores by coffee experts

Davide Giacalone[1] | Ida Steen[2] | Jesper Alstrup[2] | Morten Münchow[2]

[1]Department of Technology and Innovation, University of Southern Denmark, Odense, Denmark

[2]CoffeeMind ApS, Valby, Denmark

Correspondence
Davide Giacalone, Department of Technology and Innovation, University of Southern Denmark, Campusvej 55, DK-5230 Odense, Denmark.
Email: dg@iti.sdu.dk

Abstract

Cupping scores from experts are extensively used in the coffee industry for a variety of applications, from quality control to judging coffee competitions. In this paper, we examined inter-rater reliability (IRR) of “clean cup” ratings by coffee experts (“cuppers”) in two studies. In both studies, IRR reliability was found to be low, denoting a lack of concept alignment among experts. Remarkably, however, within-assessor reproducibility was high, suggesting that expert cuppers have their own individual understanding of “clean cup.”

Practical applications

The results presented suggested that “clean cup” scores have a fundamentally subjective nature. Since cupping scores are routinely used to drive business decisions (particularly in the context of quality control), it would be advisable that such attributes be anchored in a precise definition (in the case of “clean cup” of what constitute a defect from a sensory point of view) developed based on properly conducted sensory studies.

1 | INTRODUCTION

Coffee is one of the most important beverage commodities traded in world markets (Chambers IV et al., 2016). From a sensory point of view, coffee is a complex matrix with a wide range of flavor active components, making sensory analysis, a very important tool for the evaluation of its quality (Chambers IV et al., 2016; Giacalone et al., 2019; Giacalone, Fosgaard, Steen, & Münchow, 2016; Spencer, Sage, Velez, & Guinard, 2016; Worku, Duchateau, & Boeckx, 2016).

The most common form of sensory evaluation in the coffee industry is “cupping” with coffee experts (“cuppers”). The term “experts” in this context does not correspond to trained assessors in sensory descriptive analysis, but rather denotes individuals who have a long-standing experience with a product and whose expertise is called upon for quality control and product development purposes (Lawless & Heymann, 2010). Most quality grading systems currently used for coffee, for instance cupping protocols by the International Standard Organization (ISO) or The Specialty Coffee Association (SCA), rely on such experts (di Donfrancesco, Gutierrez Guzman, & Chambers IV, 2014; Feria-Morales, 2002). A “cupping” session generally consists of tasting at least five cups of the same coffee, prepared according to brewing conditions standardized with regard to temperature, contact time, water to coffee ratio, water quality, and brewing method (ISO, 2008; SCA, 2019). A cupper scores each cup on a predefined set of flavor attributes for coffee (in the current version of the SCA protocol, for instance, these are “fragrance/aroma,” “flavor,” “aftertaste,” “acidity,” “body,” “balance,” “uniformity,” “clean cup,” “sweetness,” “defects,” and “overall”), ranging from 0 to 10. It should be noted here that, unlike assessors in sensory descriptive analysis, cuppers do not rate attribute *intensity* per se but rather give a *subjective appraisal* of the individual attributes. For instance, a high grade in “acidity” would indicate how well the “quality” of the sourness of the coffee fits within the expectations for that particular coffee, regardless of absolute intensity (Giacalone et al., 2019). This blend of hedonic and analytical assessment marks a very important difference with scientific sensory analysis.

Another major difference between the two is that in cupping there is usually no consensus regarding the sensory vocabulary or the use of particular scales. While recent contributions from sensory scientists (Chambers IV et al., 2016; Spencer et al., 2016) promise to improve cupping protocols, at present there is still a substantial degree of variation across countries and even individual companies performing the cupping. Worku et al. provide a telling example in a paper showing that cupping scores provided by importers and exporters are uncorrelated and thus not interchangeable (Worku et al., 2016).

https://doi.org/10.1111/joss.12596

A cupping attribute that really embodies the two problems highlighted here—mixture of analytical and holistic aspects and lack of concept alignments—is "clean cup." "Clean cup" is a concept used by coffee experts to denote absence of defects, and is widely used in the industry for quality grading (particularly for green coffee) and also to evaluate the quality of the brews in coffee competitions, and so forth. The holistic character of "clean cup" means that those ratings rely substantially more on the experts' product knowledge and expectations regarding what is desirable in a coffee (similar to typicality judgments for wine), rather than on clearly defined sensory properties. Moreover, the concept blends hedonic and analytical elements, and not surprisingly previous research has found that these ratings do not correlate well with sensory descriptive analysis ratings (Di Donfrancesco et al., 2014; Feria-Morales, 2002). Although coffee experts purportedly have a shared understanding of its meaning, it is unclear to which extent these ratings vary from one assessor to the other. This question has some practical implications given that typically a very small number of cuppers (often only one person) are in charge for "clean cup" grading. While the importance of this attribute in quality control (particularly of green coffee) has been noted in the literature (Feria-Morales, 2002), focused investigations are lacking. To fill this gap of knowledge, in this work we focused on the inter-rater reliability (IRR) of "clean cup" ratings by coffee experts, drawing on results from two "cupping" studies involving a large (relative to current practices) number of coffee experts.

2 | MATERIALS AND METHODS

2.1 | Assessors

Assessors consisted of coffee professionals participating in the annual Nordic Roaster Forum event (www.nordicroasterforum.com), in two consecutive years (2016, in Copenhagen, Denmark, and in 2017 in Oslo, Norway), which we refer to as "Study 1" and "Study 2" (N = 49 and N = 45, respectively) in the remainder of the paper.

2.2 | Samples

Assessors evaluated three coffee samples in Study 1, and four coffee samples in Study 2. All samples came from a single origin washed *Arabica* bean from Colombia (from the Horizontes wet mill) whose roasting conditions were varied in order to obtain different sensory profiles. Green beans were roasted using a Probat drum roaster (model "Probatino," 1 kg Probat-Werke, Germany) in both studies. The range of roasting conditions for all samples is given in Table 1.

TABLE 1 Roasting conditions (time in min) used to obtain the different samples in the two studies. Start temperature was 200°C in all cases. The last two columns report, respectively, the Agtron Gourmet value (a spectrophotometric measure indicating the color of the roast), and the starting temperature for each sample (air temperature when the beans enter the roaster)

Study	Sample	First crack	End	Color
1	1	9:19	11:15	75
	2	10:07	14:07	76
	3	9:33	11:53	78
2	1	8:52	10:22	77
	2	9:08	11:31	77
	3	9:49	14:21	75
	4	9:31	16:01	76

Sample preparation for the sensory evaluation followed procedures used in our previous research (Giacalone et al., 2019). The coffee beans for each sample were ground to a slightly coarse particle size in a Mahlkönig (Hamburg, Germany) grinder. Grinding was performed within 1 hr prior to brewing to ensure freshness. The coffee was brewed by adding 50 g of coarse coffee to a French press brewer (Bodum Chambord French Press, coffee maker) and adding 900 ml of hot water (92°C). After a 4 min extraction time, the mixture was gently stirred 10 times with a spoon, foam was removed, and the plunger was pressed to the bottom. The coffee was poured into thermo bottles before being poured into 200 ml cupping bowls and served to the sensory panel at a temperature of 55°C.

2.3 | Procedures

The cupping evaluation took place under central location test conditions in a tasting room provided by the event organizers during the event. The latter consisted in a conference room adapted for the purpose; assessors sat at approximately 50 cm from one another and were instructed not to interact during the evaluations. Evaluation took place under normal lighting at a room temperature of approximately 22°C.

Assessors evaluated the samples by the cupping method, where the coffee is aspirated into the mouth from a spoon (SCA, 2019). Specifically, assessors took a 10 ml sample in the mouth with the spoon and swirled the coffee slowly in the mouth before spitting it out. Samples were evaluated in blind and presented to the assessors with a 3-digit code label. The serving order was fully randomized between assessors. The coffee samples were evaluated in three replicates in a single session lasting approximately 45 min, meaning that the assessors evaluated nine samples in Study 1, and 12 in Study 2. This is a slightly higher number of samples than what is common for sensory studies on coffee (generally 6–10 samples in a single session), especially in Study 2. To reduce the risk of habituation and fatigue, assessors were given a recuperation time of approximately 3 min between each sample.

Assessors evaluated the samples on several attributes including, three basic tastes (bitterness, acidity, sweetness); two flavor attributes (roasted bread, fruity); two mouthfeel attributes (body and aftertaste); and two overall attributes: "clean cup" and "balance" (the latter only in Study 1). All attributes were rated on each assessor's smartphone using Google analytics survey software with the anchor points "absent" to "a lot." In this paper, we will focus on results concerning "clean cup," though results from other attributes are given in a separate publication (Münchow, Alstrup, Steen, & Giacalone, 2020) and we may refer to them occasionally to aid the interpretation.

We should note that the use of a quantitative scale departs somewhat on the way this attribute is scored in some cupping protocols.[1] However, since there exists significant country-to-country differences in "clean cup" scoring (International Coffee Organization, 2018), the choice of using a line scale was taken to avoid mixing different scale types in the same ballot (thereby reducing the burden for the assessors), as well as for the additional data analytical advantages this offers compared to a non-quantitative scale.

TABLE 2 ANOVA tables showing main and interaction effects for replicate and assessors on "clean cup" ratings in the two studies

	df	SS	MS	F	*p*
Study 1					
Replicate	2	32.6	16.3	1.6	.198
Assessor	48	2,439.6	50.8	5.1	<.001
Replicate:Assessor	96	776.4	8.1	0.8	.892
Residuals	294	2,948	10		
Study 2					
Replicate	2	3.6	1.8	0.2	.789
Assessor	44	2,651.2	60.2	7.9	<.001
Replicate:Assessor	88	390.2	4.4	0.6	.999
Residuals	405	3,068	7.6		

Abbreviation: ANOVA, analysis of variance.

2.4 | Data analysis

Data were rendered in a matrix crossing samples and assessors. "Clean cup" ratings were analyzed using two-way analysis of variance (ANOVA) to evaluate the main effects of assessors and replicates, as well as possible interactions. Secondly, to quantify inter-assessor reliability we computed a single intraclass correlation coefficient (ICC, Shrout & Fleiss, 1979) using a two-way model, meaning that both samples and assessors are considered as randomly chosen from a bigger pool. Finally, interassessor agreement was also quantified and visualized by computing Pearson's correlation coefficients between each pair of assessors.

All analyses were done in R (R Development Core Team, 2018) using native functions and functions from the IRR (Gamer, Lemon, & Singh, 2017) and corrplot (Wei et al., 2017) packages.

3 | RESULTS AND DISCUSSION

ANOVA results are presented in Table 2. In both studies, there was a significant assessor effect, whereas the effect of replicate and the

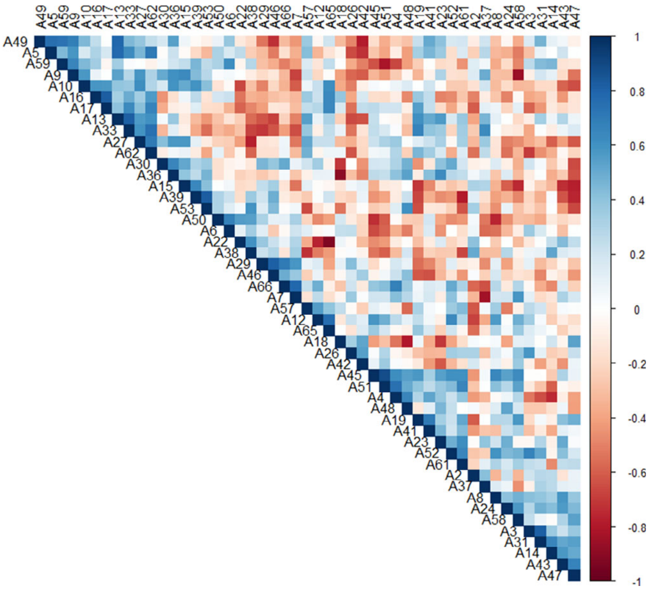

FIGURE 1 Correlogram showing correlation coefficients between individual assessors in Study 1. Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients

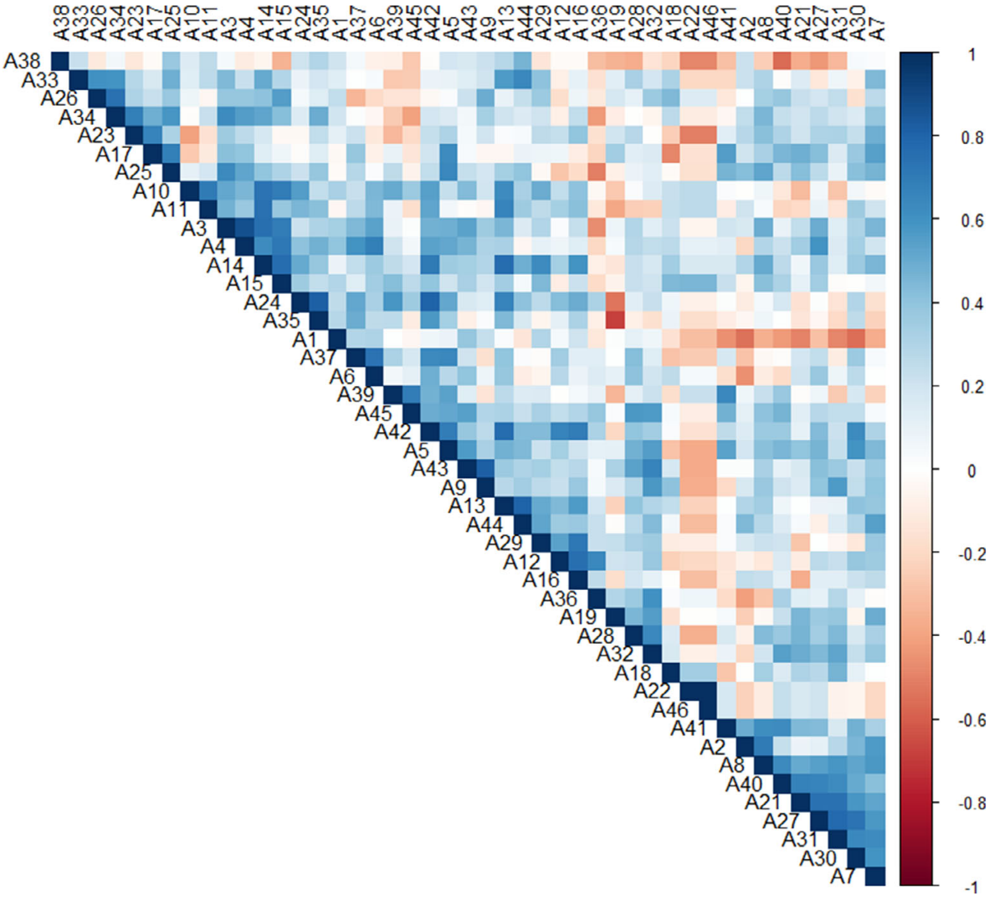

FIGURE 2 Correlogram showing correlation coefficients between individual assessors in Study 2. Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients

TABLE 3 ICC values for all attributes in both studies and associated significance level

	Clean cup	Acidity	Bitterness	Sweetness	Body	Balance	Roasted bread	Fruity	Aftertaste
Study 1	.01$^{\text{n.s.}}$	.03**	.01$^{\text{n.s.}}$	.03**	0$^{\text{n.s.}}$	0$^{\text{n.s.}}$	.01$^{\text{n.s.}}$	.04**	0$^{\text{n.s.}}$
Study 2	.13***	.20***	.12***	.12***	.01$^{\text{n.s.}}$	—	.19***	.17***	—

Note: $^{\text{n.s.}}$ $p > .05$.
***$p < .001$.
**$p < .01$.

interaction term were not significant. The lack of a significant replicate and interaction term is remarkable and show that the "clean cup" ratings were highly reproducible *within* individual assessors. By contrast, the highly significant assessor effect indicates a poorer performance in terms of reproducibility *between* assessors. Accordingly, the ICC coefficients values were 0.01 in Study 1 ($F_{(8,384)}$ = 1.6, p = .11), and 0.13 ($F_{(11,484)}$ = 7.7, p < .001) in Study 2—indicating a very low degree of IRR (McGraw & Wong, 1996). These results on "clean cup" ratings are consistent with previous work on overall cupping scores (particularly from Worku et al. (2016)), who reported a higher variability between than within expert cuppers.

Further insights into the results are presented in Figures 1 and 2, which show correlations coefficients between each pair of assessors. Visual assessments of the two figures already suggest that, while the majority of these correlations are blue (positive), a sizeable minority are either white or red (indicating a lack of correlation or a negative correlation, respectively). More precisely, in Study, 1 53.1% of the correlation coefficients were positive, while 46.5% were negative. However, most assessors' pairs were uncorrelated or rather weakly so (only 11% of the correlation coefficients were r > .5, and only 8% were <.05). Reliability between assessors was slightly better in Study 2, where 65.2% of the correlation coefficients were positive (15.2% >0.5) and 33.4% were negative (2.3% <0.5). The fact that several correlations are negative indicates that the assessor effect is not merely due to differences in scale usage, but rather that the actual ranking of the samples was different and in some cases opposite.

Taken collectively, these results indicate that, on the one hand, expert cuppers have their own individual understanding of "clean cup," which they can reliably draw on (as we have seen, the effect of replicate was not significant). On the other hand, the ICC and correlation results indicate that their understanding varies from one individual to another, denoting a lack of concept alignment among the assessors.

To complement this interpretation, we also looked at whether result for other attributes in the ballot was similar or dissimilar. Compared to them, "clean cup" had a more intrinsically subjective character, so one could expect that experts' agreement for the other attributes would be higher. However, it is not clear that was the case: as Table 3 shows, ICC values for all attributes were all relatively low and close to that of "clean cup" (range$_{(\text{Study } 1)}$: 0 < ICC-0.04; range$_{(\text{Study } 2)}$: 0.01 < ICC < 0.2). This would suggest that, save for a few exceptions such as "acidity" and "roasted bread" in Study 2, most attributes may have been used idiosyncratically by expert cuppers. It is, however, interesting to notice that the lowest ICC values were observed for the attributes "body" and "balance" which, like "clean cup," lack a clear definition and have a distinct subjective component.

4 | CONCLUSIONS

We examined IRR of "clean cup" ratings by coffee experts ("cuppers") in two studies. In both studies, IRR reliability was found to be low (ICC > 0.13), indicating poor agreement between experts, likely due to a different understanding of this attribute's meaning. Remarkably, however, within-assessor reproducibility was high, suggesting that expert cuppers have their own individual understanding of "clean cup." These results more generally reflect the fundamental difference between expert grading of food quality and sensory descriptive analysis, where the former are based on individual expertise and the latter on concept alignment achieved through training and the use of reference.

These results should be relevant to practical implications considering that cupping scores from experts are extensively used in the coffee industry for a variety of applications, from quality control to judging coffee competitions. The results presented suggested that "clean cup" scores have a fundamentally subjective nature. Since business decisions routinely depend on these evaluations, it seems advisable that attributes such as "clean cup" be anchored in a precise definition of what constitute a defect in coffee from a sensory point of view (e.g., see the paper by Giacalone et al. (2019) documenting chemical and sensory markers associated with the roasting process). While such sensory markers are necessarily specific to each variety, country of origin, and so forth, it is important that they are developed based on carefully designed experiment and properly conducted sensory studies.

ORCID

Davide Giacalone https://orcid.org/0000-0003-2498-0632

ENDNOTE

[1] For instance, the SCA protocol (SCA, 2019) uses a checklist format where the cupper has to tick whether a sample has "clean cup" or not. If the "clean cup" is not ticked the cupper can then, in a separate section of the ballot, whether defects are present with two intensity levels ("Taint" and "Fault").

REFERENCES

Chambers, E., IV, Sanchez, K., Phan, U. X., Miller, R., Civille, G. V., & di Donfrancesco, B. (2016). Development of a "living" lexicon for descriptive sensory analysis of brewed coffee. *Journal of Sensory Studies*, *31*, 465–480.

di Donfrancesco, B., Gutierrez Guzman, N., & Chambers, E., IV. (2014). Comparison of results from cupping and descriptive sensory analysis of Colombian brewed coffee. *Journal of Sensory Studies*, *29*, 301–311.

Feria-Morales, A. M. (2002). Examining the case of green coffee to illustrate the limitations of grading systems/expert tasters in sensory evaluation for quality control. *Food Quality and Preference*, *13*, 355–367.

Gamer, M., Lemon, J., & Singh, I. F. P. (2017). 'irr': Various coefficients of interrater reliability and agreement. 2012. R package version 0.84. Retrieved from https://cran.r-project.org/web/packages/irr/index.html

Giacalone, D., Degn, T. K., Yang, N., Liu, C., Fisk, I., & Münchow, M. (2019). Common roasting defects in coffee: Aroma composition, sensory characterization and consumer perception. *Food Quality and Preference*, *71*, 463–474.

Giacalone, D., Fosgaard, T. R., Steen, I., & Münchow, M. (2016). "Quality does not sell itself" divergence between "objective" product quality and preference for coffee in naïve consumers. *British Food Journal*, *118*, 2462–2474.

International Coffee Organization (2018). National Quality Standards. Retrieved from http://www.ico.org/documents/cy2017-18/icc-122-12e-national-quality-standards.pdf.

ISO 6668. (2008). *Green coffee—Preparation of samples for use in sensory analysis*. Genève, Switzerland: ISO.

Lawless, H. T., & Heymann, H. (2010). *Sensory Evaluation of Food: Principles and Practices*, New York, NY: Springer Science & Business Media.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Münchow, M., Alstrup, J., Steen, I., & Giacalone, D. (2020). Roasting conditions and coffee flavor: A multi-study empirical investigation. *Beverages*, *6*(2), 29.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

SCA. (2019). SCA protocols and best practises. Cupping protocols. Retrieved from http://www.scaa.org/?page=resources&d=cupping-protocols.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Spencer, M., Sage, E., Velez, M., & Guinard, J. X. (2016). Using single free sorting and multivariate exploratory methods to design a new coffee taster's flavor wheel. *Journal of Food Science*, *81*, 2997–3005.

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., & Zemla, J. (2017). 'corrplot': Visualization of a correlation matrix. R package version 0.84. Retrieved from https://cran.r-project.org/web/packages/corrplot/.

Worku, M., Duchateau, L., & Boeckx, P. (2016). Reproducibility of coffee quality cupping scores delivered by cupping centers in Ethiopia. *Journal of Sensory Studies*, *31*, 423–429.

How to cite this article: Giacalone D, Steen I, Alstrup J, Münchow M. Inter-rater reliability of "clean cup" scores by coffee experts. *J Sens Stud*. 2020;35:e12596. https://doi.org/10.1111/joss.12596