The Cafe Imports Coffee Rose: Rich Content CATA for Specialty Coffee Cupping

Ian Fretheim

Cafe Imports

March 21, 2022

More technically, unlike linear equations (the type most prevalent in science), nonlinear ones are very difficult to solve *analytically*, and demand the use of detailed numerical simulations carried out with the help of digital machines. This limitation of analytical tools for the study of nonlinear dynamics becomes even more constraining in the case of nonlinear combinatorics. In this case, certain combinations will display *emergent properties*, that is, properties of the combination as a whole which are more than the sum of its individual parts. These emergent (or "synergistic") properties belong to the *interactions between* parts, so it follows that a top-down analytical approach that begins with the whole and dissects it into its constituent parts (an ecosystem into species, a society into institutions, [*a coffee into attributes*]), is bound to miss precisely those properties. In other words, analyzing a whole into parts and then attempting to model it by *adding up* the components will fail to capture any property that emerged from complex interactions, since the effect of the latter may be multiplicative (e.g. mutual-enhancement) and not just additive.

Of course, analytical tools cannot simply be dismissed due to this inherent limitation. Rather, a top-down approach to the study of complex entities needs to be *complemented* with a bottom-up approach: analysis needs to go hand in hand with synthesis.

- Manuel De Landa, p.17-18, A Thousand Years of Nonlinear History, Swerve Editions 1997

Contents

Specialty Coffee and Sensory Science

This white paper describes the process and results of a project that has spanned the last three years for us in the Cafe Imports cupping lab: the development of a novel cupping form. By the end of this report I will have introduced our new cupping form, as well as its scoring engine and lexicon. Along the way, I'll explore some of the paths that we followed to get to this intersection of the fields of specialty coffee and sensory science. With this extended write up, I hope to contribute a useful voice to our industry's conversation about the art and science of cupping.

How it Started

Early in 2018, the cupping lab at Cafe Imports began a deep dive into exploring sensory science. I had come to suspect that the science of sensory analysis, a well-established, technical discipline with numerous branches, methods, and applications, was largely absent from what most of us knew as "cupping" in specialty coffee. I liked the scientific approach to the study of sensory and wanted to understand its potential application within our own cupping program. I set out with a set of questions:

- Why aren't the scientific, replicable, and transparent aspects of sensory studies more integrated into specialty coffee, and is there a good reason they aren't?
- If I was correct in my assessment that they are removed from one another, just how far removed is specialty coffee cupping from sensory science?
- Is there viable ground where the two can overlap?
- Are there aspects of sensory science that we can apply to our own program, to steer ourselves toward the most objective, replicable, and transparent cupping approach possible?

What is Sensory Science?

Sensory evaluation can be defined as "a scientific method used to evoke, measure, analyze and interpret those responses to products as perceived through the senses of sight, smell, touch, and hearing" (Lawless, 1999, p.2).[1] To this end, sensory science draws on and synthesizes an array of technical, "hard science" fields (e.g. physics, chemistry, physiology, psychology, statistics, etc). As I scratched the surface more on sensory science, I saw that the diversity of approaches and applications in the well-established scientific field, combined with what appears to be an obvious overlap in scope with coffee cupping, suggested that I would indeed be able to find common ground between the two. It made me hopeful that we would be able to employ much more sensory science in our cupping protocols.

[1] Lawless H., Heymann H. (1999). Sensory Evaluation of Food: Principles and Practices. *Springer Science & Business Media*

According to the authors of the textbook “Sensory Evaluation Techniques”:

> Dependable sensory analysis is based on the skill of the sensory analyst in optimizing the four factors, which we all recognize because they are the ones that govern any measurement (Pfenninger1979).
>
> 1. Definition of the problem: what is to be measured must be precisely defined; important as this is in “hard” science, it is much more so with senses and feelings.
> 2. Test design: not only must the design leave no room for subjectivity and take into account the known sources of bias, but it also must minimize the amount of testing required to produce the desired accuracy of results.
> 3. Instrumentation: the test subjects must be selected and trained to give a reproducible verdict; the analyst must work with them until he/she knows their sensitivity and bias in the given situation.
> 4. Interpretation of results: using statistics, the analyst chooses the correct null hypothesis and the correct alternative hypothesis, and draws only those conclusions that are warranted by the results (Meilgaard, 2007, p2).”[2]

I’ll be returning to touch on each of these four points over the course of this work, but can set the tone by jumping right in with an exploration of the last one because the degree to which statistics are used in sensory science is a characteristic point of interest and contrast with specialty coffee cupping programs. A couple of things are notable just with this one observation. One very practical item is that most relevant statistical models require input levels (sample sizes) that are not realistic for many cupping programs. Another is simply that in specialty coffee cupping, the basic paradigm of our approach is very different from one that employs a heavy reliance on statistics.

When people think of cupping, they often think of experts slurping coffees, in places called “labs,” maybe even in far flung and exotic corners of the globe. These cuppers then relay their expert experiences of those slurps to the rest of us in the form of specific, personal, and occasionally esoteric descriptions and scores. The cupping forms that they use have fairly complex and formal appearances, being composed of numerous anchored scales and various fields and boxes. Some cuppers are even capable of generating high decibel whistles when they slurp. It can all be very intimidating and one might be tempted to think that sensory science is essentially similar to cupping, just with more scales, more boxes, and louder whistles. One would be wrong.

Sensory science is a decidedly boring, methodical, and logical discipline. In a good sensory test or program, much of the hard work is done outside of the tasting itself. This takes the form of arranging the test to control variables and target critical information, and also of analyzing results after the fact. In the eyes of sensory science, the act of cupping is much less the vaunted expert art form that we know and enjoy and much more the staid activity of carefully collecting sensory data. In specialty coffee cupping, cuppers are humans doing their best to relay their

[2] Meilgaard M., Civille G., Carr B. (2007). Sensory Evaluation Techniques. *CRC Press*

impressions and opinions of their coffee tasting experiences. In sensory science, and in particular in descriptive analysis, humans are considered to be lab instruments used by test administrators to measure the intensities of target attributes.

Rather than glorifying one and vilifying the other, we will be better served to come to an understanding of 1) what each discipline offers and 2) what each requires. We may then be able to form a better picture of what specialty coffee analysis could stand to improve on and what we might adopt from the world of sensory science (this sort of borrowing and adaptation is a move that is well suited to both specialty coffee *and* sensory science).

A common use case for sensory science might look like this:

A company makes a product that is well liked in its market. Unfortunately, the product is expensive to produce and as a result, its margins are unsustainably thin. Rather than scrap the product, sensory science might be used here to help relieve some of that pricing pressure by doing a reformulation study. The idea would be to devise a reformulation of the product with less expensive ingredients in such a way that only an allowable percentage of users would be *likely* to *notice* the change. Of those that notice the change, a percentage may find it to be unacceptable to the point that they reduce or even stop purchasing it. The goal is to increase profitability by decreasing the cost of production without alienating too many customers. This is done by estimating a likely and acceptable attrition for a given formulation change against the savings generated by that cheaper formulation or process.

For many of us in specialty coffee, this use case may be at odds with our core values, focused as they are on relationships, honesty, and transparency in buying, selling, and roasting the best possible coffees. Nevertheless, it is important to at least sketch some rough outlines of the fields that we will be discussing.

The example of an uneven coin illustrates another aspect of sensory science’s statistical paradigm that will lead us to a familiar corollary in specialty coffee. If you believe a coin is weighted such that it unfairly flips tails, one way to demonstrate your case might be to flip the coin. Note that flipping the coin and demonstrating a certain behavior does not prove that the coin is out of balance. To do that, you would need some way to directly measure the balance of the coin. A suitable procedure may not come readily to mind, or be practical, or even available. In many cases it will be more practical to demonstrate the behavior and compare it to a standard, or the expected behavior of a known reference. The statistician and the sensory scientist will be interested in the degree to which the target object fails to meet the expected standard or behavior.

To begin, you would need to flip this coin more than once in order to demonstrate uneven flipping behavior. An even coin has a 50% chance of landing heads and 50% chance of landing tails. A single flip landing tails is well within the probable outcome for flipping an even coin. Over numerous flips we would expect an even coin to yield a balance of around 50/50 heads and tails. If the claim is that a trick coin is uneven toward landing tails, we would expect the trick coin to deviate from that balance toward tails. With enough tests and enough deviation, we would eventually be comfortable saying that the coin is unlikely to be even.

Note that the language “coin is unlikely to be even” is intentional. What we do with probability is establish a model behavior or population of expected results. When we flip (or model the flipping) a standard coin 1 million times, we develop a population of results against which we can test our real coin. We then flip the alleged trick coin some number of times, which generates a sample of observations. Our probability question is “How likely is it for our sample observations to come from the larger expected population?”

How many tests are enough? While the answer to that question will vary depending on how certain you need to be, probability supplies us with both the logic and the math to get there. For this we employ something called hypothesis testing. The hypothesis test has three primary parts: the effect (or claim), the null hypothesis, and the alternative hypothesis. The claim in our example is that we have a coin which flips tails unevenly. The next step is to formulate what is called the null hypothesis. The null hypothesis represents the normal state that is assumed to be true and would correspond to the standard population concept just discussed. It is the hypothesis that we must nullify in order to support our claim. In this case, the null hypothesis is that our coin is a normal (even or fair) coin. Finally, we state the alternative hypothesis. The alternative hypothesis formulates our original claim and is mutually exclusive to our null hypothesis. In this case the alternative hypothesis is that our coin is in fact uneven, or more specifically it could be that our coin is uneven *toward* tails.

Over the course of testing we will presumably collect evidence in the form of observations that the coin is uneven. In other words, we will flip *significantly* more tails than heads. At some point we would reject the null hypothesis (that the coin is even) based on that evidence and the improbability that our observations would arise from flipping a fair coin. Functionally, the coin is observed to behave far enough outside the parameters of a fair coin’s known (assumed) behavior to support this rejection (i.e. given the evidence, it is unlikely to X degree that this coin is fair). On the other hand, if our coin did flip fairly, we would fail to reject the null hypothesis.

Starting with the null hypothesis has some advantages. Presumably, we have access to many examples of even coins. We can easily define and model the behavior of what an even coin should do. The chances of getting tails for 1 flip of an even coin is 50% and all subsequent flips are entirely independent of the outcome of any prior flips. This gives us the parameters for what the behavior of an even coin should be.

The chances of flipping two tails in a row is 50% * 50%, 25%. The chances of flipping three in a row is 50% * 50% * 50%, 12.5%. And so on. Of course, we don’t need to get 100% consecutive flips to demonstrate an unfair coin. We would more likely look at the probability of flipping at least Y tails out of X attempts. While the math for consecutive probability and that for “at least Y out of X” probability is different, in general this is how independent probability works, and this is the basic process underlying the comparison of our tests to what would happen if our null hypothesis were in fact true, which forms the basis of the hypothesis test.

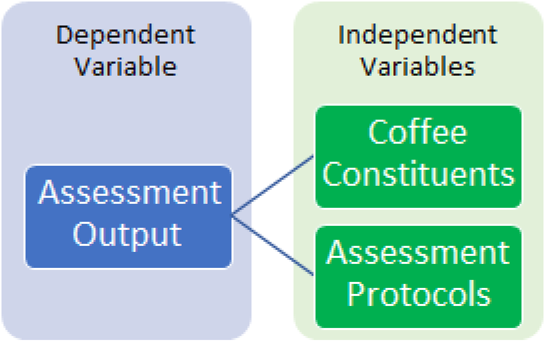
The probability of flipping *at least* nine tails in 10 tries is around 1%. That of flipping *at least* eight tails is around 5.5%. This gets clever as we ramp up the number of flips. The probability of flipping at least 16 tails in 20 tries is around 0.6%. Note that the proportion of successes to attempts is the same as with flipping at least eight out of 10, while the probability declines dramatically. Outperforming is increasingly unsustainable as the number of tests increases. Remember, we're not trying to *prove* that the coin is uneven. We're just trying to reject the hypothesis that it's even. It's up to stakeholders to determine at what level of probability the null hypothesis can be rejected, though often the 5% level is used.

There is a certain elegance here. There is also a bit of brute force. The more tests we do, the harder it becomes for the null hypothesis to keep up if it, in fact, should be rejected. Bringing this back to coffee: triangle tests test for difference on the basis of the same principle. The null hypothesis in a triangle test is that there is not a discernible difference between two treatments presented in a blind group of three samples. This hypothesis means that any selection made by an assessor can only be made at random and that any correct answer is due to chance. For each set in a triangle test (similar to what we in coffee call "triangulations"), there is a 33.3% chance of correctly selecting the odd sample. Each subsequent set is independent of the prior and over enough sets it becomes increasingly unlikely that an assessor will outperform the chance success rate if they are unable to discern the difference between the treatments.

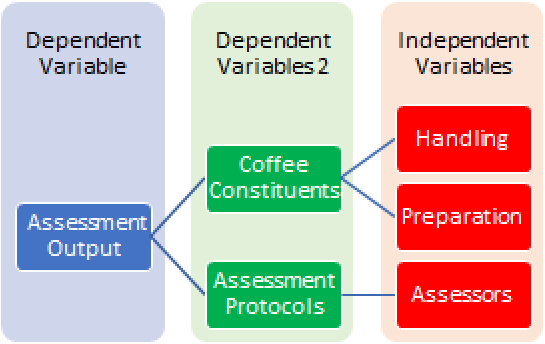
For clarity, the usual application of a triangle test (a type of difference test in sensory science) is to determine the likelihood of a difference existing or being detectable between samples, processes, or treatments, as opposed to the more gamified "triangulation" that we are familiar with in specialty coffee. In a triangle test, we would repeat the same two samples/treatments over many placements in order to establish a meaningful test sample size. Appropriate applications for this in specialty coffee might be in checking the consistency of repeated roasts or of a replacement component in an established blend.

These are just examples to get us warmed up, aimed at illustrating a basic difference in paradigm between specialty coffee cupping and sensory science. A cupper might approach the question of difference between two coffees by tasting them and comparing the notation and scores. From a sensory science perspective, the question: "are they different?" calls for some form of difference test (triangle or otherwise). The question "How are they different?" is a separate question and unique tools will be preferred. For this question, descriptive analysis would be favored. This is a common pattern for these two fields, and ultimately we'll see if we can Venn ourselves some ground between both.

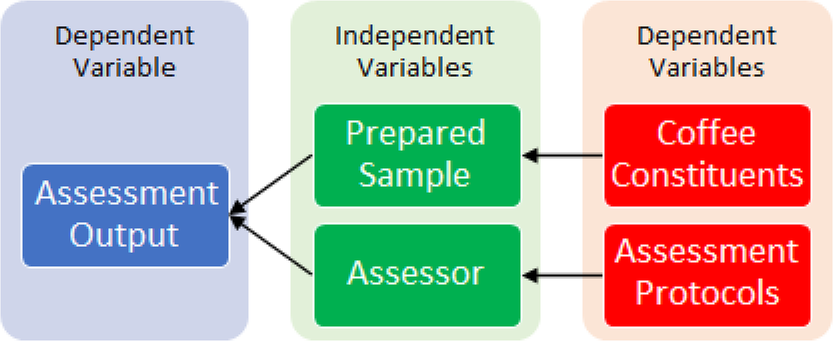
In distinction to sensory science's (descriptive analysis') focus on the object, cupping in specialty coffee is about the cupper. This point will become clear as we progress, and is crucial for understanding not only the critique, but also the potential benefits of bringing some sensory science into specialty coffee. Structurally, we might consider coffee assessment output (description and score) to be our dependent variable (output/effect), while the coffee itself (perceivable constituents) and the applied assessment protocols are our primary independent variables (inputs/causes).

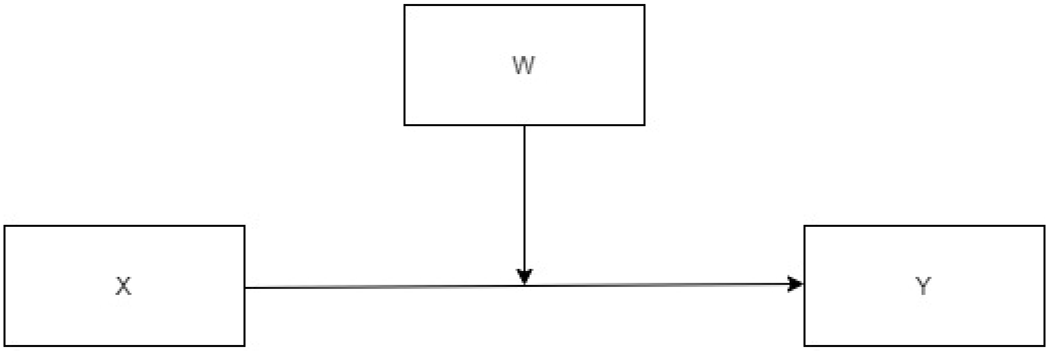


Because we do not assess coffee in a raw state, the coffee itself is also a dependent variable subject to independent variables like handling and preparation. Similarly, assessment protocols are dependent on the assessors that apply them. This is even more the case in cupping systems that do not have explicit standards and that ask the cupper whether attributes are good or bad, liked or disliked. While we tend to talk about cupping results with reference to the underlying (green or "ideal") coffee, as if it were an independent variable itself free of significant dependencies, it is important to note that this is not the case. Realistically, any cupping experiment model is more properly describing the relationship between a specific *prepared coffee sample*, a group of assessors applying a set of assessment protocols, and their assessment output than it is that between the coffee itself, the protocols, and that description.



This is important to understand. In the same way that we do not have access to the coffee itself because we must modify it via preparation, we also do not have independent assessment protocols, in particular when we use affective metrics and leave their determination up to cuppers in the moment of cupping. Descriptive analysis techniques are designed to get as close as possible to the original model, minimizing the modifying impact of the assessors and preparation. Assessment output for these tests is in fact aimed at describing the coffee constituents while recognizing the limits to that aim. Affective testing techniques, on the other hand, are designed to focus on the assessors. Assessment output for affective tests is aimed at describing the assessors' response to an input. I'll circle back to this distinction later. By moving from a hierarchy to a process view we can see that this model mirrors a simple moderation model diagram:





Conceptual diagram of a simple moderation model in which the effect of the focal antecedent (X) on the outcome (Y) is influenced or dependent on a moderator (W).

The crux of the dilemma is this: Our primary interest in cupping is to understand and characterize something about the coffee itself. This is well aligned with descriptive analysis. However, cupping as commonly practiced is affective in nature and so we end up extrapolating results that we obtain *about ourselves* to the coffees we think we're assessing, while ignoring or underestimating the impact of moderation. The failure to recognize moderation effects in test design, coupled with the use of affective testing for descriptive-analytic ends reveals a fundamental flaw at the root of many cupping programs that undermines their accuracy and transparency.

While there are no alternatives to preparing coffees for assessment by assessors, we can work to limit the variability (and moderation) at play in our preparation and assessments. As an industry we have become relatively good at limiting the variability in our sample preparation (handling, storage, roasting, grinding, extraction, etc). However, we have largely ignored the variability and moderation resulting from our assessments and their protocols. In particular, in standardless and preference based cupping systems, the cupper tends to be a poorly controlled moderator variable relative to an assessment output that is inappropriately assumed to describe a set of coffee constituents.

In preference based cupping systems (score the quality, score the distinctiveness), the assessment output does not describe coffee constituents or prepared sample constituents as commonly assumed, but rather it describes the assessor and their opinion about those constituents. Such a scorecard does not tell you that the flavor of a coffee is good. It tells you that *the cupper thinks* the flavor is good. It tells you that *the cupper likes* the flavor of a coffee. This is not coffee information, it is cupper information. This type of information is useful for upstream stakeholders to gauge consumer sentiment but should not be considered reliable or suitable for quality determinations or attribute descriptions.

In sensory science, the assessor is recognized as a variable that we want to control and not report results on (unless we explicitly are testing for assessors). Tests are specifically designed to control for interaction, bias, and other errors and effects that can occur between the assessor and coffee variables. Tests are designed differently depending on whether we want information about coffees or about coffee tasters, and again, depending on what specific information about each that we want. The triangle test is designed specifically to isolate the dependent variable (difference between samples) and also to undermine the sustainability of an assessor succeeding by means other than identification of that variable.

This brings me back to that original, very practical point about statistics in sensory science. A common trait across most methods of sensory science is the use of larger sample sizes than many of us have realistic access to in our regular workflows in specialty coffee. Remember the jump in reliability in the coin flip example above: there's a 5.5% chance of flipping eight out of 10 tails, but only a 0.6% chance of flipping 16 out of 20.

For many of us, regularly achieving a sample of 10 tests for coffees is fairly out of reach. Sample sizes can be increased either by increasing the number of assessors participating in a study, or by increasing the number of times each assessor sees a given sample. The former is expensive in time, training, and payroll, while the latter can quickly overwhelm an assessor's palate while also slowing down response times to buying and selling partners. In general, sensory science testing occurs with a finite set of samples within a specified period of time. Sensory tests, or studies, are often discrete and designed to gather information that will remain applicable within a given product space. Of course, in an importer's sample room, we receive new samples every day and rarely have the luxury of holding exhaustive, closed studies for the analysis of those samples[3, 4].

Importantly, lower sample volume cupping operations have an opportunity here to mitigate the potential downside of seeing less coffee by creating more robust sampling replication and assessment protocols than most exporters, importers, or other high volume, constant flow samplers have the bandwidth for. We should always look for the potential strengths inherent to

[3] At Cafe Imports we have instituted a system aimed at targeted increases in the number of individual, independent cupping tests performed on green samples. Our system is structured to allow us to adjust the number of tests performed on coffee samples on the basis of their priority and urgency without impacting our blinding protocol. As of March 2022, we averaged 3 tests per green sample and regularly see test counts as low as 1 and as high as 6. We have been increasing the number of independent tests per sample and anticipate averaging 4 - 5 in 2023, with high counts for individual samples pushing closer to 10.

[4] It's worth noting that the formula for standard error is sample standard deviation divided by the square root of sample size. *As sample size increases, nominal error is reduced.*

our specific situations. This discussion points back to the second item in the aforementioned Meilgaard's list: "Test design: not only must the design leave no room for subjectivity and take into account the known sources of bias, but it also must minimize the amount of testing required to produce the desired accuracy of results." Minimization here has to do with efficiency, but always in service to the level of accuracy that we desire.

Carl Staub of Agtron once told me and another participant in a seminar that we should strive to be scientists and experimenters. The point that he was making was to encourage us to try new things, to keep records, to engage our curiosity and intelligence, and to work actively and logically on solving problems. It seems impressive to say that we cup some relatively large number of samples each year. There are certain advantages that come with that exposure. At the same time, cupping half as many *could* mean having twice the bandwidth to focus on those fewer coffees.

I believe that the point Staub was making is the same that can be found running through much of sensory science. Each of us is attempting to answer questions arising in our real, day-to-day environments. The caricature of science here would be to insist on unrealistic and pedantic solutions, as opposed to searching for realistic and practical ones. It would be in setting greater importance on the claims one could make than on the progress one did make. It would be in replacing the Cupping Bae with a Sensory Science Bae, beautiful and terrible as they might be.

To be clear, there are much better scientists in coffee than myself. While they obviously bring a crucially needed foundation of scientific method to the table and to their projects, they also bring an inquisitive nature and cultivated creativity to problem solving. Ever onboard and answer operational questions for a scientist? Oof. May we all be so lucky… I wouldn't wish it on anyone. Many of us do not have formal training in the scientific method. Be that as it may, we can do our best with method and learning while bringing forth and nurturing our inquisitive natures and creativities, neither of which are in short supply in specialty coffee. While there are certainly differences in approach and orientation between these disciplines, there is also some compelling common ground.

There may well be practical limits to what a specialty coffee cupping program can or should do with regard to sensory science. It is not a holy grail or promised land. That's fine. Sensory science has already demonstrated that there are limits to what specialty coffee cupping programs can do without it. Our aim then is to test the former, while keeping in mind their practicality and trying to improve upon the latter. Rather than scrap the idea of integrating a stronger sensory science foundation into our cupping program on the basis of observed incongruities and challenges, we've focused our work on baseline principles and more generalizable guidelines like sources of error, controls for those errors, and focus in design. Additionally, we've sought to develop a complementary synthesis approach to sensory science's analysis.

Sensory science is neither an incantation nor a speech act whereby merely talking about it imbues one's cupping system with its principles. Implementation requires critical investigation of one’s protocols and a willingness to adapt them. For our part, we looked at the strategies and logic employed in various sensory testing procedures to see what might be applicable to our own. We ran trials to demonstrate and validate various of the common physiological and psychological errors identified by sensory science. We evaluated and reevaluated what our needs actually were.

While I will aim to keep myself more closely on topic throughout this report than I’ve done in this initial essay, it’s important to know that this project and this paper flow out of a larger investigation of an enormous field of study. Sometimes you get taken in by the flowers, and this investigation has provided much more than we can hope to present in this space.

Quality and Standards

Cupping is one of the most interesting and engaging activities available to specialty coffee professionals. Unfortunately, it is also one of the most gate-kept, poorly defined, and arguably antiquated activities. Sensory science has come a long way since its inception in the 1940's. Nevertheless the general approach to cupping in specialty coffee has not kept pace, and is itself much the same as when it was conceptually introduced in the mid 1980's and certainly as when it was refined in the late 90's and early 2000's. While there are real limits to the suitable application of sensory science tools in many of the common use cases associated with assessing specialty coffee, there are also notable shortcomings in the common approach to cupping. The recently published SCA Coffee Sensory and Cupping Handbook (and supporting work) provides an excellent and concise primer, even if I do not think that it sufficiently prioritizes the application of its contents to advancing the outdated and unreliable components of the common cupping model. As an independent company, we at Cafe Imports have much more freedom to innovate (I do not envy anyone charged with herding the collective cats of us, the specialty coffee industry) and have sought to uncover, implement, and integrate applicable sensory science principles and protocols in our cupping program.

I recognize that many of us compete with one another in the coffee business space. Nevertheless, we all share a concern to see the entire specialty coffee sector flourish. Many of us further share the goals of improving transparency and equity along coffee supply chains such that people who historically have been treated in an opaque and primarily extractive way can have an increased stake and insight into how the coffees they produce make their way to roasters and cafes around the world. One of my aims here is to add to the larger conversation about how specialty coffee functions, in this case through its cupping discipline, how it could function differently, and how those things may impact the larger goals for equity that are commonly shared in our industry.

Prior to sensory science emerging as a distinct discipline, the norm for acquiring sensory or quality information about a product was to solicit expert opinions. This has been found by researchers to be problematic because most people, even heavy product users, are generally not experts and are not therefore connected with the experience and opinions of experts. We also know that even for experts, opinions of good and bad are conditional, contextual, subject to influence and to physiological and psychological errors. In diverse fields, in particular when the question is opinion based, experts have been shown to be both remarkably noisy and biased.

Despite this, cupping as commonly presented in specialty coffee is designed for and accessible primarily to experts (our industry has made great efforts to define and elevate expert status, but few to corral even its most obvious flaws). Resources are spent attempting to train expertise and opinion alignment, while many of the now commonplace techniques for limiting sensory system error are overlooked or ignored. The basic metric for quality in most cases remains the momentary opinion of the assessor. As a result, cupping outputs tell us very little about the

coffees that they ostensibly describe. Rather, they tell us about the person or group that cupped the coffees.
The SCA's Coffee Sensory and Cupping Handbook explains: "Affective testing differs from the other major groups of sensory tests in a major way: in discriminative and descriptive testing, the aim is to evaluate a product or a person objectively, but in affective testing, **the aim is to document the subjective experience of a person** or group of people. For this reason, more than any other sensory tests, affective testing embraces the discipline of psychology and combines a psychological understanding of human experience with the tools used in sensory analysis. **Ideas about quality**, acceptability, preference, value, and purity are profoundly influenced by cultural norms and an individual's psychology, and so the affective approach to sensory analysis deeply respects an individual's subjective experience and seeks to understand it. While other sensory analysis techniques intentionally minimize idiosyncratic experience and personal bias, **affective testing deliberately focuses on these**. In an affective text, **subjective experience is the thing that is to be measured; it is the point of the test**…It is to be remembered, however, that these concepts are subjective in nature (or, collectively subjective in the case of a cultural norm), because they are based on human experience and are measurable in no other way (Fernández-Alduenda, 83)."[5]

Emphasis mine.

As I've been discussing, most current cupping protocols do not measure coffee attributes, they measure a cupper's liking or acceptance of those attributes. "Quality of Flavor" is not a measurable *coffee* attribute. It is a [measurable] human one. Same with balance, quality of acidity, etc. Quality based assessment metrics measure coffee cuppers' momentary hedonic opinions about coffees and coffee attributes, in particular in the absence of explicit standards for what quality is and is not. There definitely is a place for preference and liking measures in both sensory science and specialty coffee. It is my view, however, that this place is not when we are working to determine and measure coffee quality.

Standard protocol in the specialty coffee industry is to score attributes on their *quality*, as opposed to their *quantity*. While it may seem logical to score on perceived quality when you're trying to determine quality, there are issues with this. For one, quality is not a measurable coffee attribute. It is a judgment that is made given an analysis of attributes and an application of stakeholder decisions (opinions), ideally that have been codified into standards. A coffee assessor can tell you directly if a coffee's acidity is citric or malic, and again relatively how intense that acidity is. However, determining the *quality* of those acids at given intensities or in various contexts is a different order of procedure that requires active value judgment and decision making. Humans don't decide how much or what kind of acidity a coffee has. We do decide how we value those things, or how distinctive we think they are when we observe them. It is important to note that such a decision (the quality or valuation decision) will always be made somewhere in the assessment process. The issue that I'll dig into is in leaving it to the assessor, whereas the alternative is in building it into the protocol via form design and valuation standards.

[5] Fernández-Alduenda M., Giuliano P. (2021). Coffee Sensory and Cupping Handbook, edition no. 01. *sca.coffee*

With this distinction in mind it should become clear that we need definitions and standards in order to determine quality at any scope beyond personal preference. These can either be explicitly stated, or as is more common they can be tacit, implied, and vague. In this latter mode, when cuppers score an attribute with a quality score, the best case scenario is that they are noting observable characteristics and then referring to some internally held standard for how to value those observations. More realistically, they are simply giving voice to their liking and even worse their estimates of others' liking, in particular when prompted to "observe" unobservable characteristics like quality and balance.

Confounding the situation is that there is no definition for what quality in coffee means, for what constitutes quality, and for which high scores should be awarded. It is a tautology and not a definition to say that "quality" flavor is flavor that is "excellent" or "distinctive." What is the measure of excellence? What makes an attribute distinctive? Without a stated standard, these are no less personal, malleable, and unobservable than the generic term "quality."

In fact, the concept of quality *requires* a standard and suggests a comparison against that standard. As provided by the Oxford English Dictionary:

noun: quality; plural noun: qualities

1. the standard of something as measured against other things of a similar kind; the degree of excellence of something.
2. a distinctive attribute or characteristic possessed by someone or something.

Even if we would seek to avoid the challenge that comes with stating a standard and the vulnerability that comes with making measurements against it, the second entry for quality discussing distinction in attributes still requires a standard for what is and is not distinctive in order for a statement of distinction to be at all meaningful. How else will cuppers know what are distinct versus common attributes? What meaning or understanding can a third party possibly take away from conflicting personal opinions of the distinctiveness of a coffee or attribute?

Surely we are not proposing to just shrug off the entire coffee assessment paradigm to whoever makes up the most creative words or whoever has the least coffee exposure (and for whom everything novel is distinct). Surely the immediately distinct, and in much of specialty coffee, rarely tasted attributes of Monsooned Malabars and Vietnamese Robustas are not intended to rate in parity with boutique and meticulously processed Arabicas, nor even with the well processed examples that have become so common as to already be undervalued in today's specialty coffee environment. Surely the goal is not to undermine the work for coffee sector equity by defining quality in terms that can only be meaningfully applied with exclusivity and that are inevitably personal in scope.

Surely these are not the case. And yet, here we are. Just as there is no document stating that malic and citric acids are equivalent, better, or worse than one another, there is none stating that either is equivalent, more, or less distinctive than the other. Ultimately this may be for the best. At the industry leadership level, it is likely more appropriate to introduce ideas, tools, guidelines, and guardrails than hard standards. Imposed standards are rarely engaged enthusiastically, are thankless to develop, and have proven impractical to apply across such a broad and decentralized space as specialty coffee, in particular with such highly personal metrics as “good” and “distinctive”. Importantly, the promotion and use of standards does not require industry level standardization. Nor does the removal of good and bad from cupping forms remove it from coffee assessment.

Standards should be taken up as a local responsibility, minimally in the form of a statement explaining “this is how we cup, and this is what our scores mean,” while using terms that avoid the tautologies noted above. Such statements allow us to consistently explain results and bridge gaps in cupping outcomes, improving communication. At the industry level, we do not currently have an outline for good and bad attribute valuation standards, or even a recommendation for their local creation. The recent proposal of distinctiveness as the measure for quality appears to further erode the situation. While the concept of distinction lends itself well to our current individual expert preference system for valuing coffees, it extends privilege to novelty, supports a sort of flavor arms race, and promotes assessor inexperience and whim as viable paths to outperforming score generation. These are problematic for the communicability, meaningfulness, and reliability of cupping output, as well as for the goals of transparency, equity, and sustainability along our coffee supply chains. Speaking with cuppers, one suspects that most are under the impression that their affective judgements describe coffees and not themselves. Certainly the language of coffee description and reporting reflects a belief in coffee centrism. This is understandable given that common industry cupping protocols are presented as descriptive and analytic when in truth they are affective.

Lexicons: Tools for Communication

Hand in hand with the ideas of quality and standards comes that of lexicons. Loosely, a lexicon is the vocabulary of a language or branch of knowledge. In sensory science, lexicons take on a more technical meaning. “A lexicon is a set of standardized vocabularies developed by highly trained panelists for describing a wide array of sensory attributes present in a product.”[6] Along with this definition come a number of rules for what may or may not be included in a sensory lexicon. For example, “terms listed in the lexicon must be extensive and complete, non-hedonic, singular (not integrated), and non-redundant, and also must capture all product differences.”[7]

Attempts to create, unify, and standardize a specialty coffee lexicon have been made with limited success. Efforts such as those made by World Coffee Research have in fact both generated a lexicon and also modeled an appropriate methodology to generate such a lexicon. It is possible that emphasis has been placed too much on the lexicon itself and not enough on the methodology used to develop it as an exemplar for the development of locally applicable lexicons. In a sensory science descriptive testing environment, lexicon creation is generally one of the foundational tasks that a panel or group of experimenters undertakes. Importantly, the lexicon creation task is performed as an iterative process among stakeholder participants, with the iterations being determined by group concept generation, meaningful clustering and differentiation between concepts as applied in the testing environment, observed concept alignment between assessors, and reliable concept utilization by assessors.

While broad conceptual lexicons can be created centrally for distribution and universal application, practical, used lexicons are often created or *adapted* locally for implementation of organization level standards. For most specialty coffee cuppers the target outcomes of cupping are a combination of quality assessment or classification and flavor or profile description. An underlying goal shared by these cuppers, even when their protocols and expressions vary, is to communicate pertinent coffee information between people. This reinforces the problem underpinning the earlier discussion of affective testing being mistaken for descriptive.

Coffee communication can be facilitated in a myriad of ways, from imposing a central or universal lexicon to allowing a free-for-all in language use. While the former may ensure that we are all using the same words, it may lack meaning for local users and may even foster a false sense of calibration. At the other extreme, descriptive free-for-alls tend to lack meaning outside of their immediate, local, and often specifically personal context. In practice, neither extreme is necessarily as bad as they may seem at first blush. People tend to bend imposed, top down, or commonly accepted lexicons to their local purposes, as we have seen in linguistic evolution throughout our spoken histories. Similarly, apart from perhaps the wildest extremes, descriptive free-for-alls do not necessarily lead to incorrigibly divergent expressions so much as they do

[6] Suwonsichon S. (2019). The Importance of Sensory Lexicons for Research and Development of Food Products. *Foods (Basel, Switzerland)*, *8*(1), 27. https://doi.org/10.3390/foods8010027
[7] Suwonsichon S. (2019).

unreliable and poorly communicable ones.[8] Such free-for-alls can lead people to seek out the meaning of unusual descriptions, potentially fostering communication, even if it is primarily in service of bridging initial gaps.

Our work at the industry level can be beneficially focussed on two things. First, on developing tools that allow participants to create the most locally applicable and robust assessment procedures possible. Second, on devising communication and assessment protocols aimed at bridging the inevitable gaps that will emerge between somewhat novel lexicons and standards, or between novel applications of a universal one. Rather than attempting to avoid or whitewash differences in expression or understanding, we can anticipate them and leverage the work it takes to bridge them into an even deeper understanding of the coffees we are assessing and the partners that we are working with.

In an essay titled Developing General Models and Theories of Addiction, authors Robert West, Simon Christmas, Janna Hastings, and Susan Michie (The Routledge Handbook of Philosophy and Science of Addiction, p. 166) discuss the semantic web as a model for resolving the extremes of imposed vs undefined language.

> In principle the use of namespaces could result in anarchy if no-one agreed about the definitions of anything. But this is no different from what happens with natural language. The difference in the case of the Semantic Web and ontologies is that **disagreements in how terms are used are explicit and defined. They can then be subject to scrutiny, and differences potentially resolved where such resolution is useful**. In practice, many researchers *do* agree on key definitions and can therefore use a single shared namespace to reflect this consensus. This is directly analogous to referencing the definition of a construct in another paper, with the important practical difference being that such links become computable.
>
> The concept of a 'namespace' allows for a diversity of perspectives on the Semantic Web, and **ensures that anyone can say anything about anything. There are, however, practical limitations to diversity**. In order to have debates about the construct of 'addiction' for example, **we do need to agree on the meaning of a set of more fundamental terms and predicates we use in that debate**: terms such as 'is a subclass of' and 'has attribute'.

Bold emphasis mine. The point is not that we need to usher in Web 3.0, but rather that we can aim for creating systems that allow for diversity and disagreement, and that capture and quantify them, making them available for "scrutiny" and "potential resolution", rather than systems that ignore, hide, suppress, or alternately revel in that novelty and disagreement. We can aim to

[8] A new cupper in my lab once described a cellulose-y and "papery" coffee with the term "bumblebee wing." This cupper often used such colorful descriptions and I think to some extent was doubted for doing so. As a rule her explanations of these descriptors were in fact descriptive at least of my own best assessment of the "technical components" and "commonly accepted descriptions" of a given coffee).

develop lexicons that allow freedom of expression by discovering, agreeing upon, and leveraging the more fundamental building blocks and operators of our cupping systems.

In the notably decentralized and perhaps proudly iconoclastic specialty coffee industry, emphasis should be placed on the creation, defense, and engagement of locally meaningful and universally transparent lexicons and standards, or on the adaptation of universal ones to local use. We need to be careful with the imposition of work-product that may not be applicable and for which users may not be engaged, let alone included, as stakeholders. The assessment of coffee is complex. Clear communication is also complex. Rather than pitting creative and precise uses of language against each other, rather than undermining coffee assessment by giving one or the other too much weight, we can use the precise to support, aim, and contextualize the creative. We can use the creative to give life and flavor to the precise. Taken together, *synthesized*, we can aim to raise our cupping language to a new level of descriptive accuracy.

Transparency in Cupping

Transparency is a key concept underlying the development of assessment forms, protocols, standards, and sensory attribute valuation models. However, for many participants, observers, and stakeholders, cupping is essentially a black box where coffee goes in and scores and descriptions come out. While the idea of transparency has seen a resurgence in specialty coffee conversation in recent years as we've turned our attention increasingly to supply chain equity, this uptick has focussed largely on economic and social concerns while ignoring less striking or immediately obvious aspects of the coffee supply chain, including assessment standards and protocols.

Transparency in cupping assessment protocols and standards is paramount because cupping is the medium through which so much of the value that we place on coffees is technically determined. It is the ground on which our conversations about coffees are built, and it is the language around which many of our relationships and contracts are formed and maintained.

In terms of the current industry transparency conversation, we want to know who grew, processed, and produced our coffee, even if we're not yet quite at the point of wanting to know who wrapped it and where the wrap came from, who hauled it on their shoulder, who drove it, who inspected it, or who shipped it. This makes sense. A shipping company may have a fascinating story to tell, but they don't offer quite the romanticism that a connection with a farmer does.

Nevertheless, we should be careful that our push for transparency doesn't inadvertently become a form of opacity, greenwashing things that we find less attractive or interesting, or reducing complex systems to photo ops, slogans, and memes. In a similar vein, transparency should not be one sided, where for example consumers demand transparency around the flow of the money they spend, but then refuse to reciprocate with regard to something as fundamental as the grading and valuation processes behind the spending of that money.

As mentioned, transparency is a foundational concept in sensory analysis. For the conversation here I can highlight two ways that transparency is important:

1) It allows us to communicate and validate our procedures and results.
2) It allows stakeholders, particularly those on the sell side, to understand how their coffee is being assessed and valued.

One of the driving reasons for applying sensory analysis paradigms and procedures in cupping is to reduce uncertainty and doubt with regard to the reliability of assessments. Consider again the example of the triangle test given above. The triangle test not only acknowledges but is built around the problem of uncertainty. It is designed to measure observations against an established baseline of maximum uncertainty (random chance). In doing so it provides us with a quantifiable context for understanding the likelihood and therefore certainty of our results. This is similar for controlling error, bias, and noise.

Sensory science has identified numerous ways in which human assessors can fall into error, as well as strategies for controlling the impact of those errors. How do we know that an assessor's mood, a sample's table placement or presentation, or a sample's preparation isn't unduly impacting our results? Saying "I'm a really good expert" is the "trust me, bro" of sensory analysis. It is much better to say "We do X, Y, and Z in order to mitigate the common errors A, B, and C, that particularly tend to arise in situations E, G, and F." We present samples under blind, random codes in order to mitigate expectation error, which particularly tends to arise when a coffee's origin is known, or beans are seen. We employ blind replication of samples in order to limit physiological errors like adaptation and cross potentiation, which can occur unnoticed in single placement protocols.

When patterns emerge in a robust sensory environment, we can be much more confident that those patterns are meaningful to the coffees under observation, as opposed to their having occurred either by random chance, outside influence, or preconception, etc. The act of developing and following protocols that target and address specific psychological and physiological errors, before we even look at cup scores and coffee descriptions, lays a foundation of transparency for the validation of our eventual cupping results. In this case, transparency involves establishing, following, and making available intentional protocols and standards for sensory assessment. It further means that those protocols and standards should not be reducible to or circumventable by the whims and opinions of assessors. Standards and protocols *can* be changed, but this should be done purposefully at the administrative level rather than individually, in the moment and on the fly.

Transparency in protocol design and implementation is crucial because of the high level of complexity, wide profile variance, and potential for variability in coffee preparation and assessment. Coffee cupping is notably challenging due to both the chemical complexity of coffee and also the ephemerality of its preparations. Arabica coffee boasts a staggering number of identifiable aromatic compounds and the relative perceptibility of those compounds can change dramatically with temperature in very short periods of time.

In order to be assessed, coffee must be prepared by roasting, dosing, grinding and brewing. The number of variables introduced by these processes, the inevitability of introducing them, and the reliance of coffee flavor profiles on these same variables contributes a significant suite of challenges to the reliable assessment and description of coffee. While these latter preparation variables are technically extrinsic to any given coffee, they must be recognized as *intrinsic to the assessment process* because they must occur in order for assessment to be possible. As described earlier, this sets up a series of dependencies in our cupping designs. Of course we control these variables to the full extent that we are able, but the state of the art of sample preparation is still such that in any given cupping program they will generate some degree of variability.

Sensory tests (both consumer/affective and analytical/descriptive) aim to use extrinsic factors like cupping form and test design, scaling and scoring standards, and sample presentation to further limit the impact of the earlier mentioned assessment variables (roast, dose, grind, brew) beyond just dialing in the technical components of preparation. These extrinsic factors are also designed to mitigate common sensory errors so that cuppers can more completely assess the target intrinsic variable of coffee chemical complexity. Rather than mitigating, however, current industry cupping practice actually adds significant procedural complexity to the already challenging task of coffee assessment. In standard cupping practice, cuppers are at minimum required to:

1) determine whether attributes are good or bad
    a) without reference to an assessment standard defining good and bad
    b) as an overall judgment considering possible perceived differences between grouped cups
2) give attributes a score on the basis of that judgment
    a) without reference to a scoring standard for degrees of good and bad
    b) as an overall judgment considering possible perceived differences between grouped cups
3) determine whether the combination of those attributes is good or bad
4) determine whether some attributes are pass or fail
    a) as a localized cup by cup judgment
5) identify whether there are defects
    a) determine and then classify the severity of those defects
    b) calculate a score for the number and severity of those defects
6) mathematically add their attribute and defect scores to generate a final score
7) verify that the totalled attribute scores align with their overall impression and description of the coffee
8) describe both the attributes and the overall coffee in associative, descriptive, and qualifying language.

This is an ambitious program for any product, let alone for one as challenging and complex as coffee. Beyond the many modes of thought and processing that must occur in this model, it is striking that the base metrics of good and bad (distinct, not distinct) are not defined but instead are left to the momentary judgment of the individual. By leaving these assessment metrics to the momentary judgment of the individual, we start from a place of *significant opacity*. Not only do outside observers and stakeholders have no way of knowing how conclusions are reached, but individuals themselves may also have poor insight into the moment by moment expression of their own preferences and so cannot be said to be utilizing a reliably operative metric (i.e. transparent protocol) for determining good and bad.

Remember, the “idiosyncratic experience and personal bias” that we control for in descriptive testing (opacity), are the actual items of interest in affective testing, these are **“**the point of the test.” To the extent that scoring metrics are held in the head (let alone determined at the whim) of the assessor, we cannot know what those metrics of assessment are. Much of the above described system complexity is therefore tied into its opacity, or lack of transparency. Indeed, the system's opacity makes it more complex to navigate, while its complexity makes it more opaque. By not laying out the terms of assessment in advance, such systems occlude outside knowledge (including that of the assessor her or himself) by pushing the development and deployment of those terms onto the individual assessor *at the moment of assessment*.

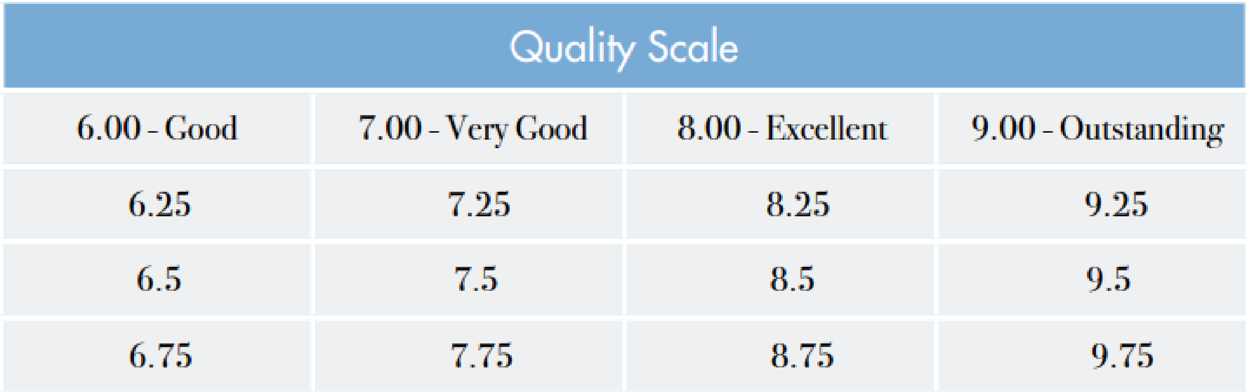
While the specialty coffee industry has recently taken a stronger interest in the ideas of sustainability and equity along production and consumption lines, we have not yet widely addressed the opacity and lack of equity that is persistent in coffee assessment. One of the keys to realizing the equity that we seek is in increasing the transparency of coffee assessment and classification such that the terms of any assessment can be known in advance and scrutinized in retrospect if needed. Unfortunately, the procedures employed and outputs generated by the standard practice of cupping as described above remain doggedly opaque.

Some years ago, Shawn Steiman PhD, of Coffea Consulting, pointed out to me that producers entering their coffee into competitions have no way of knowing how their coffee will perform. I thought, well of course not, it's a competition. If we knew the results in advance, it'd just be an award show. I misunderstood the importance of the point he was making. Entrants cannot know the likely performance of their entries because neither the administrators nor the judges of the competitions themselves know even *what kind* of coffees will perform well. Opacity in cupping is not malevolent, but it is systemic. At the time, I was stuck thinking from a place that assumes that we all know and agree on what's good in coffee. We do not. While we all take for granted that our unified task is to identify the best coffees, we fail to realize that we have very little idea what that will mean for any given group of assessors on any given day.

It bears repeating: Entrants cannot know the likely performance of their entries because neither the administrators nor the judges of the competitions themselves know even *what kind* of coffees will perform well. We ground competitions on the basis of "quality," "what's good," and "finding the best," but we don't actually know how these things are determined or what they mean. Competition and purchasing quality metrics are not transparent even to ourselves and so of course they will not be to the people submitting their coffees.

Scoring in most cupping rooms revolves around individuals determining if attributes are good or bad, liked or disliked, and soon distinct or (presumably) banal. This is often done without stated standards for *what* is good and *what* is bad, or for what a cupper as part of an event, organization, or industry, *should* "like" or "dislike". The rejection of the idea that anyone else can tell us what we should or should not like is almost instinctive, but the resistance or inability to set aside personal preferences in order to perform a specified coffee assessment is, bluntly, unprofessional. The rub from an assessment of coffee perspective is not that some attributes are considered good while others are considered bad, it's that those decisions are made individually by assessors on the spot, when they should be made publicly in advance. Saying in advance that X is good allows assessors to assess whether coffees meet X as a stated criteria while all stakeholders are aware from the beginning that X is a criteria for success.

Entrants to competitions, and generally anyone submitting coffee for assessment, should at least be able to form an idea about how their coffee will perform on the basis of a competition's or organization's stated standards for what the assessed attributes will be and how they will be valued. Instead of this, submitting coffee for assessment can feel more like a crap shoot to many coffee growers and suppliers. "Here's the coffee, I hope you like it." This is completely logical when you consider that the metric for success in quality scaled coffee assessment is "excellent to outstanding" attributes.



Quality Scale			
6.00 - Good	7.00 - Very Good	8.00 - Excellent	9.00 - Outstanding
6.25	7.25	8.25	9.25
6.5	7.5	8.5	9.5
6.75	7.75	8.75	9.75

SCAA Quality Scale

The ambiguity of this scaling leads to opacity in part because there are no explicit standards for the meaning of these terms. There is nothing to tell us what "excellent" or "outstanding" means. Scores are generated as the momentary hedonic opinions of the judges. From year to year or contest to contest, retrospective extrapolation of what standards would appear to have been, given the results, obviously vary. One year we see that tart, fruity coffees are preferred while savory, fermenty coffees are rejected. The following year, fermenty coffees are on trend and more subtly fruity coffees are passed over. There generally are no guidelines (either fixed or dynamic) to tell us if it's a good or bad thing to taste any given attribute (with the exception of major defects) in a cup of coffee, let alone how good or how bad.

This lack of transparency is not just a cupping problem. It is a problem for specialty coffee as a whole, specifically as we attempt to foster equity in our supply chain relationships. As instinctively (and I think, correctly) as we reject the idea that someone else might dictate to us what is good and what is bad, we must also recognize the untenable position of transacting on the basis of good and bad in a market without compass for what is good and what is bad.

Some may want to describe (or, god forbid, dictate) where the money for coffee should go, down to the last penny, but as long as the allocation of that money hinges on the momentary opinions, moods, and whims of coffee assessors, like a one way mirror the transparency of our systems will remain deeply consumer centric on the one side and frustratingly opaque for those standing on the other.

A transparent cupping protocol will include a lexicon in which terms are defined to the extent that they can be. It will further include values or connotations for the items in that lexicon. The final score of a coffee should be clearly related if not directly tied to its description, and the terms of its description should be valued consistently and transparently.

The standards employed and valuations given to coffee descriptors do not need to be identical, imposed, or enforced across our industry. It is often the case that sensory lexicons are created locally, by stakeholder participants, rather than being imported universally. An overarching, universal lexicon could be of use for mediating or starting a conversation between cuppers, but could also become an impediment for engaged assessment and authentic communication. One could imagine even a single roaster having somewhat novel lexicons and valuation metrics in their sourcing and roasting/production capacities. The questions and goals for each are potentially different enough to warrant such a model, and employing novel protocols could allow each to be more finely tuned to their specific tasks.

While they may not need to be universal, lexicons and standards do need to be transparent such that we can communicate and explain our cupping results to one another and to our partners, impartially and consistently. And they must exist. Even if we do not all need to play by the same rules (it is okay for us to value attributes differently), we must formulate, state, and stick to the rules that we ourselves are playing by, at least if we wish to make claims about sensory fidelity, commercial transparency, and supply chain equity.

Context: Behind the Project

At Cafe Imports, our preparation, presentation, and assessment protocols are all designed to limit and prevent common sensory errors. Our cupping forms, the lexicons that they utilize, and the standards that they are built on should take those common sensory errors into account as well. I've sought to remove complex tasks and thinking from the job of coffee assessment, opting to design into our protocols and scorecard as much qualitative information as I could. This leaves our assessors free to report and describe the attributes that they taste, without worrying about determining whether they are good or bad.

Sensory science has developed many techniques and tools over the last several decades. These techniques and tools have different goals, ask and answer different questions, and thus have different applications. Selecting tools appropriate for the task is of course an important step towards achieving desirable outcomes. Even more so, we should consider that the task in question here is not merely "cupping." What we call "cupping," is in reality a suite of different practices, with different questions and different target answers (e.g. cupping for pass/fail, cupping for description, cupping for quality assessment, cupping for roast refinement, cupping for brew suitability, etc.). It makes sense that we should create cupping tools that are purpose built to deal with the specific questions we are asking. We should also be very careful about trying to jam too many functions into a single test.

Today, standard practice cupping fits most closely with the practices that emerged in the early and pre-history of sensory science. Prior to what may be considered the beginning of sensory science in the late 1940s and early 1950s at the Quartermaster Food and Container Institute for the Armed Forces in Chicago, companies within the food industry utilized expert opinion to determine and manage the quality of their products. Early work in sensory science shifted the focus from expert judgment toward trained assessors in an interdisciplinary approach for "both the study of basic taste, olfaction, appetite and hunger, and the development of reliable and valid methods of discriminative and affective measurement tools."[9]

Cupping today looks somewhat like an amalgamation of the expert opinions, judgments, and apprenticeships that characterized the food industry prior to the emergence of sensory science, informed by a particular aspect of an early descriptive analysis technique developed by the Arthur D. Little Company in the late 1940s known as the Flavor Profile Method. The Flavor Profile Method created literal consensus among its panelists, whereas today in cupping we tend to use mean scores that are "within calibration," while actively augmenting that calibration with post flight discussions of results. Most current cupping protocols utilize a discussion period after a flight (a grouping of assessments) has been completed to help panelists form and maintain consensus throughout a larger series of flights and assessments. While the changing or amending of scores and notes is generally not allowed during this phase of a cupping session,

[9] Schutz H (1998). Evolution of the Sensory Science Discipline. Food Technology, 52 (8), 43.

the overall attitude is one of collaboration and is aimed at fostering a sort of meta-consensus, colloquially called “calibration.”

Along with the similarity between the concepts of consensus and calibration, individual evaluation and smaller panel sizes are the primary connections between these two approaches. One important difference between current industry cupping practice and the Flavor Profile Method is that the latter, already in the 1940s, asked assessors for intensities of attributes rather than judgements of their quality.

Since around 2016, we at Cafe Imports have used a simplified form of another sensory method known as Quantitative Descriptive Analysis (QDA). In its most basic form, a QDA test presents independent assessors with a series of target attributes and asks those assessors to indicate the intensity of those attributes, generally by using anchored scales. Our cupping form was designed on a scoring model that was indexed to common specialty coffee industry scoring and compression, while assessing the intensity of attributes rather than perceived quality. We paired this cupping form with the processing based scoring standards that we developed in 2012.

This cupping form and scoring standard combination checked many boxes with regard to bringing Cafe Imports’ cupping program closer to the larger field of sensory science. Of course, it also left much to be desired. When I first presented that cupping form, I summed it up like this:

> “This isn’t the final word in score cards. For me, it’s just the second or third word. Remember: We’re Perfecting an enemy of the Good. Should we find success such that our scorecard is one day itself Good, another enemy will be needed.”

We began work on a new cupping form, and a new scoring model, in early 2019. We had been doing a lot of work with signal detection, sensory errors, and their resolutions starting in 2017 or 2018. A driving motivation behind this work was the desire to bring our cupping program more in line with sensory science. I wanted to develop a model based on the principle that “cuppers assess coffees; standards score assessments.” This meant breaking scoring off from cupping, which in turn meant that we needed to design both a new cupping form and a program to score that form. By the end of 2019, I was looking at establishing a runway for rolling out the new form and scoring engine that I had developed. Then the COVID-19 pandemic hit.

COVID did two things to that project. 1) It put the brakes on it, hard and 2) it created the space for a major breakthrough and redesign of the model, sometime in the fever dream of the early summer of 2020 in Minneapolis. Which brings us to our current project: the Cafe Imports Coffee Rose.

The Cafe Imports Coffee Rose

Today, we introduce the next development in Cafe Imports' ongoing work to improve our coffee sensory assessment. Over the last two years we've been busy building a new suite of tools for coffee assessment. These include a new cupping form, a scoring engine, a lexicon with assessment value standards, a flight building and blinding module, as well as various reporting features.

The Cafe Imports Coffee Rose is a dynamic, rich content CATA cupping form paired with a user independent scoring engine. We'll break down what each of these things are in some detail.

CATA is a sensory science acronym that stands for Check All That Apply. It is an easy to use method for eliciting descriptive data about a product. While CATA forms are frequently given to consumers, they have also been shown to work very well with trained and semi-trained individuals. They are useful for capturing a large amount of qualitative data and also for converting that qualitative data to quantitative data.

A standard CATA form for a product, in our case a coffee sample, presents the user with an array of descriptive terms that may be applicable to the coffee in question. This array will generally be presented in an easy-to-read table, list, or grid format. The assessor simply tastes the coffee sample and checks (endorses) the terms that apply to that sample. Terms that do not apply are ignored.

This format eliminates the need for assessors to both come up with descriptions and also to determine whether those descriptions are good or bad, or liked or disliked. In the context of specialty coffee cupping, it also removes numerical scoring and mathematical processing from the assessor's task. Relative to what many of us are accustomed to in cupping assessment, the CATA methodology marks a significant reduction in task complexity. On the other hand, CATA forms can be alternately cumbersome and restrictive, depending on how much descriptive resolution is provided. We believe that we've dramatically reduced these downsides with the introduction of this tool.

As described earlier, with most specialty coffee cupping forms assessors are given a multitude of nominally related but functionally different tasks. Cuppers are asked to determine whether a series of attributes are good or bad. They must quantify the goodness of each of those attributes on a numerical scale. These ratings are added to generate a final score, which itself must be validated by the cupper to verify that it fits with their overall assessment of the goodness of the coffee. Further, assessors are tasked with generating descriptions of both the attributes and the coffee as a whole. Cuppers must also identify defects, rate their strength, and calculate them into their attribute qualitative scoring.

The Coffee Rose is presented in the familiar format of a flavor wheel. The flavor wheel itself functions as the CATA array, where each of the flavors, aromas, and tastes displayed on the wheel are active buttons that can be used to select or endorse applicable descriptors. We've characterized this CATA form as both “rich content” and “dynamic.” These are descriptive, rather than technical terms and will be the subjects of the following sections of this discussion.

In short, the “rich content” component refers to two primary attributes of the wheel; the first is its tiered structure and the ability to endorse either simple, individual descriptors or to build and endorse more complex descriptor strings. The second is that the wheel's endorsement protocol utilizes an indication for intensity. These two components allow us to pack significantly more descriptive punch into our CATA array than would be possible in a static list or table format. The “dynamic” component refers to the way that the wheel alters its appearance depending on the inputs that it is provided with, expanding the active section as well as visually providing reference to the underlying standard, though without compromising the principle separating the assessor from the final valuation process.

Rich Content CATA

The Coffee Rose is organized in hierarchies or tiers (Categories, Qualifiers, Types, and Specifics) of attributes. At the base level, we have seven broad Categories which correlate in scale, scope, and descriptive power to the attributes on a standard cupping form (e.g. sweetness, bitterness, acidity, etc). Each of these Categories contain specific pathways that allow users to build descriptive strings by selecting from increasingly more specific terminology.

On a standard CATA form, you might have the option of selecting descriptors like “cooked fruit,” “cooked berry,” “raspberry,” or “cooked raspberry.” On the Coffee Rose, the structure of the categories and their hierarchies are such that rather than scanning through an array of prebuilt, fixed terms, you progressively build out the terminology that you wish to use. Assessors begin with the most general Category tier, then if applicable they can move on to select from the Qualifier and Type tiers, and finally the Specific tier. The structure of the Coffee Rose is, from general to specific: Category ⇛ Qualifier ⇛ Type ⇛ Specific. Examples of four tier descriptor strings could be: fruit ⇛ cooked ⇛ berry ⇛ raspberry, or fruit ⇛ dried ⇛ berry ⇛ raspberry.

CATA for broad green assessment of specialty grade coffee can be challenging because of the limitations imposed by presenting a manageable array of terms vs the need for terminology to be descriptive enough to sufficiently cover the breadth and depth of coffee tasting experience. Deconstructing complex terminology and presenting it in hierarchies of building blocks allows for a much more extensive CATA without requiring a correspondingly daunting array.

Further, the Coffee Rose incorporates a measurement functionality as utilized in descriptive analysis. After a user selects a descriptor or descriptor string (e.g. fruit ⇛ cooked ⇛ berry ⇛ raspberry), they enter it into the system by indicating its intensity, relative to a supplied or ideal reference. This reference coffee is what we call a “flat 80.” 80 points on industry standard scoring is the accepted floor for specialty coffee. Most scoring systems are additive, meaning that the final score is generated by adding the values given to many attributes. This in turn means that the same score can be arrived at by very different coffees or attribute scoring profiles. A coffee with very high marks in one attribute and very low marks in all others can achieve the same final score as a coffee with very high marks in a different category than the first, and again with low marks in all others. Further, a coffee with mediocre marks in all categories can also achieve that same score.

Conceptually, this last example is a “flat 80”. It is a coffee that scores around 80 points with generally flat or similar scoring across its attributes. The idea is that the attributes of a flat 80 are reasonable estimates of what the general floor for each of those attributes are in specialty coffee. The characteristics of a flat 80 are quite neutral. In the larger world of coffee, a flat 80 would properly be described as “mild, sweet, and clean,” representing an obvious level of sweetness and a low level of bitterness relative to non-arabicas and poorly prepped arabicas. For specialty coffee people, flat 80s will have no notable or distinctive flavor attributes, and mild acidity and sweetness. There will be no notable defects or overt problems. In short, for specialty coffee people, flat 80s are bland but still specialty grade coffees by technical assessment. For the larger coffee world, flat 80s are more obviously specialty grade by their mildness, relative sweetness, and lack of offending bitter and astringent qualities.

We draw out the description of the flat 80 in detail because the intensity scaling that we employ on the Coffee Rose is designed relative to that concept. We rate attribute intensity at four levels: Reference, Noticeably (More than Reference), Significantly (More than Reference), and Much (More than Reference). Consider the structure and example given earlier:

(category ⇛ qualifier ⇛ type ⇛ specific)

(fruit ⇛ cooked ⇛ berry ⇛ raspberry)

After building this descriptor string, the assessor then enters it into the system by indicating if the coffee that they are assessing tastes Noticeably, Significantly, or Much more like the above string than the Reference does. A complete descriptor string then includes a notation for Intensity. Importantly, only the Category level is required to be entered, meaning a more accurate rendering of the structure would look like this:

(category ⇛ [qualifier] ⇛ [type] ⇛ [specific] ⇛ intensity)

The base entry for a coffee Category that tastes very similar to that of the provided Reference would simply be Reference. For a Category or descriptor that is less intense than the Reference, but for which there is not a negative description (such as sweetness), the entry is still Reference. In essence, the intensity indicators are <= Reference, Noticeably > Reference, Significantly > Reference, and Much > Reference.

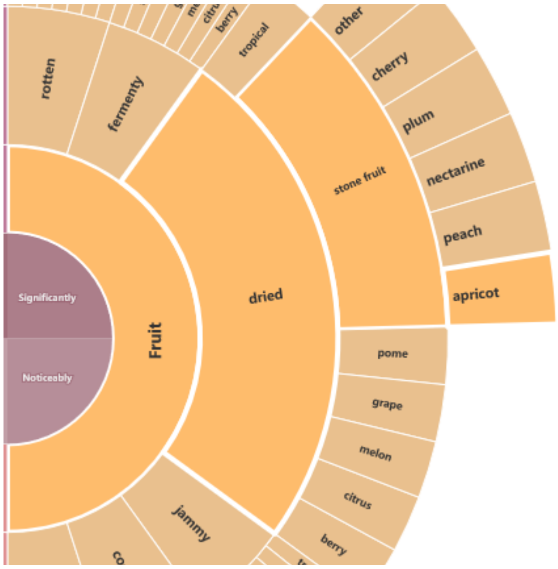
There are a couple of reasons for this move. One is that because we deal almost exclusively with specialty coffee, we’ve decided to focus our available bandwidth on measuring specialty coffee (i.e. coffees > 80). Another is that we’ve observed that it is much easier to anchor a measurement scale and limit it to a single direction (greater than) than it is to float an anchor and try to measure reliably both above and below it.

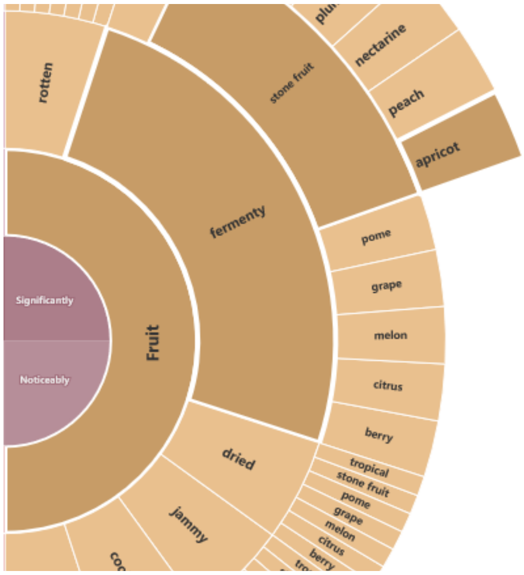
We've also observed that coffees with < specialty attributes frequently have > specialty non-specialty attributes (e.g. coffees that lack sweetness also often have excessive bitterness). Because we're able to logically present both positive and negative attributes on the CATA wheel, we've found it advantageous to limit our scaling to items that *stand out as greater than the reference*, whether good or bad, as opposed to attempting to measure them on bipolar scales around the reference. Remember, on a standard CATA form, the assessor only checks items that apply, ignoring those that do not. The same can be said for our wheel, where the criterion for the question, "Does this apply?" is the flat 80 Reference coffee.

Dynamic CATA

The Coffee Rose is a dynamic CATA because rather than being a static array or list of terms it changes in response to user inputs. The Rose rotates and expands to highlight the category and hierarchy in which a cupper is working. As it reorients itself with each selection, it also limits the selections available for the active descriptor string to those that are applicable, somewhat akin to dynamic surveys. This emphasizes the organizational structure and offers an ease of use compared to static CATA arrays.

The Coffee Rose also alters its appearance in real time according to the scoring engine and the standardized good/bad connotation value associated with the items that a user selects. When building a descriptor string, the items selected in that string are highlighted on the Rose with a thick outline. This helps the assessor keep track of their place and what they're doing. Further, the descriptors themselves change in tint or shade depending on if the current cumulative value underlying a given descriptor string is positive or negative. The lighter tinted selection on the left indicates a positive connotation, while the darker shaded selection on the right indicates a negative.





While we seek to remove valuation, scoring, and mathematics from the cupper’s assignment, we need to remember that cupping from an analytic perspective still requires calibration and concept alignment. These can and should be improved with ongoing training, but as we’ve learned during Covid, in-person work is not always possible. Nevertheless, we want to maintain very low latency feedback loops for our cuppers. This connotative highlighting helps reinforce the training that *has* occurred and guards against people using the same terminology to describe different things or different terminology to describe the same things.

An example of where we might need this functionality is the increasing popularity of anaerobically processed coffees. It is conceivable that the Qualifier for Fruit “Fermenty” could be used as a neutral, negative, or even positive term. We use that term to indicate a negative quality in the cup, and so if a person tried to use it to describe a positive experience, the value output would not match their expectation, even if the description did. By making the standards and valuation ranges employed behind the scenes available for reference, and by indicating active connotation as an assessor makes descriptive selections, we are able to minimize such conceptual misalignments among assessors.

Signal Detection: The Roots of Cupping

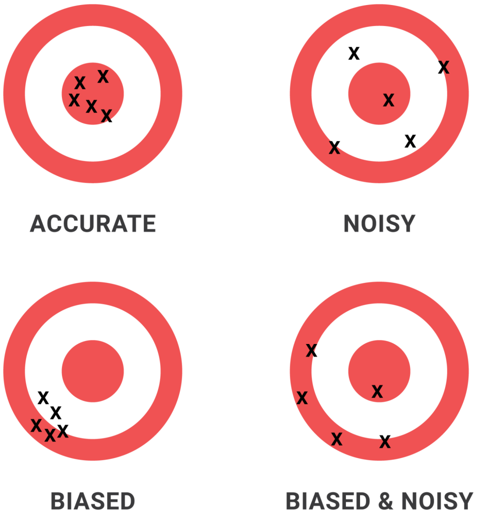
While the nuts and bolts of the Coffee Rose are CATA with a nod from QDA, its root is signal detection. Signal detection theory is a decision model which underpins cupping activity. At bottom, the goal of SDT is:

1) to determine whether or not something has occurred
2) to estimate the probability of incorrect determinations, and
3) to mitigate the impact of incorrect determinations.

In its simplest form SDT reduces the experienced world to two inputs: signal and noise, and two possible responses: SIGNAL and NOISE.

Signal is whatever our target is. A standard cupping form can be viewed as a series of signal detection questions such that each attribute presents the cupper with a target, or signal alert for the underlying coffee. Noise is an umbrella term that encompasses a few concepts. In one sense, noise is *everything* that is not signal. Signal detection is single minded. When we're assessing acidity, sweetness is noise. Further, noise includes interference such as the technical errors that were mentioned earlier, and even your own mental state to the extent that it is distracting and impinging on your ability to reliably detect a target signal. In another sense, noise is the baseline or base state from which we detect (or fail to detect) signals, or from which signals emerge as identifiable from noise.

For clarity, we should mention that noise is also what we call the portion of system error that is not attributable to bias. Daniel Kahneman helps us to visualize this in his book titled "Noise." The first target in the upper left of the following figure is accurate. The shots are clustered near one another and also near to the center of the target. The target in the lower left of the figure shows a biased grouping. Here the shots are still clustered, but they also systematically miss the target. The two targets on the right hand side of the figure show Noisy and Noisy & Biased distributions. The noisy distributions display a reduction or even lack of overall pattern, or a looseness of grouping effect. This use of the word noise is a bit different than what we're dealing with in signal detection. Nevertheless it is useful to know both in order to avoid confusion and also because we use this bias and noise framework to characterize and validate our cupping output.



NOISE: Daniel Kahneman, Olivier Sabony, and Cass R. Sunstein

SDT is a decision model because an assessor ultimately must *decide* whether to say that a signal is present or not present. We cannot say that there is strong acidity or good acidity without first deciding that enough acidity is in fact present for us to say something. Logically, we think that things are there or they are not. This is chemically correct. At the same time, assessment is more subtle, complex, and human than that.

> When sensory analysts study the relationship between a given physical stimulus and the subject's response, the outcome is often regarded as a one-step process. In fact, there are at least three steps in the process….The stimulus hits the sense organ and is converted to a nerve signal that travels to the brain. With previous experiences in memory, the brain then interprets, organizes, and integrates the incoming sensations into perceptions. Finally, a response is formulated based on the subject's perceptions (Schiffman 1996). (Meilgaard, 2007, p3)[10]

The presence of citric acid is not the same as the detectable presence, which again may not be the same as the reportable presence or the certain presence. While sensitivity is variable, calibration among professional cuppers is largely agnostic to personal thresholds of detection. I like to tell my cuppers that in many cases it is no more useful to be a parts per billion person in a room full of parts per million people than the reverse. When we're dealing with calibration of different humans, or even the same human at different times, we employ an error mitigation tool from signal detection called the criterion. More on that below.

[10] Meilgaard, M (2007). Sensory Evaluation Techniques, 3. *CRC Press*.

In its simplest form, SDT presents the assessor with a stimulus that can be either signal or noise. In turn, the assessor can identify that stimulus as either SIGNAL or NOISE. These options combine to generate four possible outcomes:

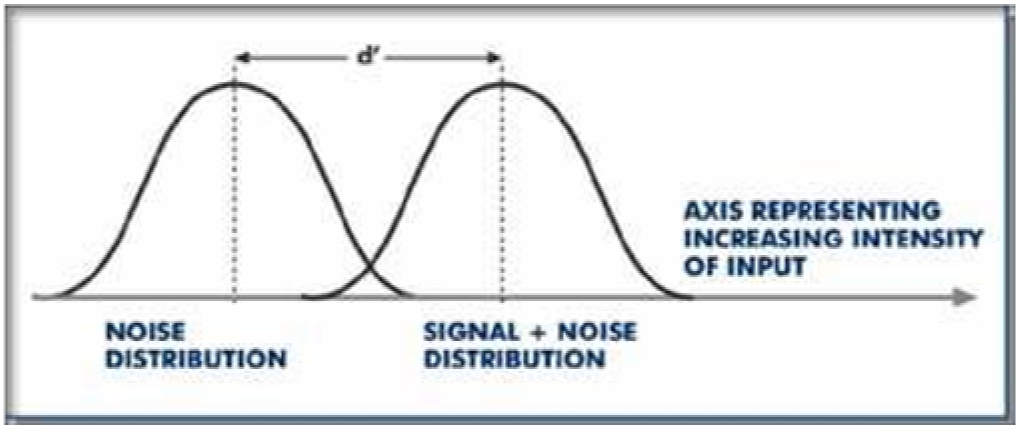
Simple Signal Detection Array		**RESPONSES**	
		SIGNAL	NOISE
INPUTS	signal	Hit	Miss
	noise	False Alarm	Correct Rejection

Presented with signal, the assessor can correctly identify SIGNAL for a hit, or incorrectly identify NOISE for a miss.

Presented with noise, the assessor can incorrectly identify SIGNAL for a false alarm, or correctly identify NOISE for a correct rejection.

Obviously this decision process is not a problem when signal is very different from noise. However, when signal and noise more closely resemble one another we encounter more misses and false alarms. As a part of test design we need to be cognizant of this balance and in some cases decide whether hits or correct rejections are more important, and whether false alarms or misses are more costly.

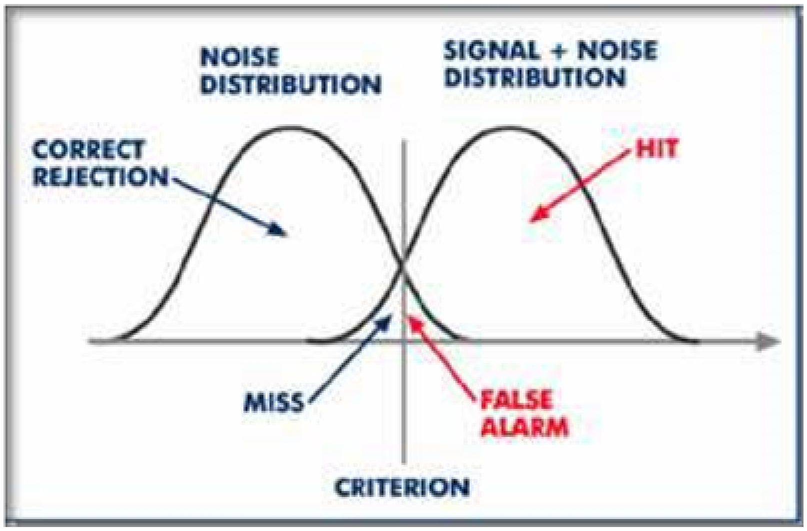
We can visualize this by plotting two distributions on the same axis of input intensity, one distribution for noise and the other for signal. When these two distributions are well separated, the decision between signal and noise is easy. When they overlap, the decision becomes more challenging. University of California, Davis' Applied Sensory and Consumer Science Certificate Program provides three figures to help demonstrate these concepts.


d'
AXIS REPRESENTING INCREASING INTENSITY OF INPUT
NOISE DISTRIBUTION
SIGNAL + NOISE DISTRIBUTION

The first diagram shows the basic layout for signal and noise distributions described above. In terms of coffee, and in terms of our project here with the Coffee Rose, we can think of noise as being an 80 point washed coffee (the flat 80 mentioned earlier) or corresponding attribute. The signal distribution could be for any other coffee (or attribute) that we might assess. If we were assessing the acidity of a coffee from Kenya, or the saltiness of a monsooned Malabar, we would expect the noise and signal distributions to be notably separated. If the assessment target were a standard Central American strictly high grown coffee, however, the distributions may have some overlap.

These distributions show the number of times that observations occur at each intensity. For example, on a 10 point form with SCA math, the flat 80 acidity distribution would probably span from around 6.75 to 7.5, and would most frequently be assessed at around 7.00 or 7.25. The Kenya would maybe span from 8.5 to 9.5 and be very easy to distinguish. On the other hand a series of observations of the acidity of a Central American SHG may span from something like 7.25 to 8.25. This yields an overlap between the two distributions, that is, an area where the SHG acidity is difficult to distinguish from that of the flat 80. There would be instances where the flat 80 acidity was assessed to its upper end of 7.5, and instances where the SHG acidity was assessed near its floor around 7.25.

The next diagram maps the outcome features from the table above (Hit, Miss, Correct Rejection, False Alarm) to the different areas of our two distributions. You'll notice one further new element called the Criterion.



The criterion is where SDT becomes really interesting for professional coffee assessors. The criterion is the decision point below which an assessor will indicate NOISE, and above which they will indicate SIGNAL. In the diagram above, the criterion is placed in the neutral position whereby both hits and correct rejections are maximized relative to one another and misses and false alarms are minimized. This is not the only possible position for the criterion, nor is it necessarily the ideal position.

SDT is a decision model because the criterion is *movable*. The placement of the criterion is the decision. For example, by shifting the criterion left you can achieve a greater hit rate, but only at the cost of increasing false alarms. You could eliminate false alarms, but only at the cost of increasing misses. The neutral position may not always be the ideal location for the criterion because the value of hits, misses, correct rejections, and false alarms may not be equal.

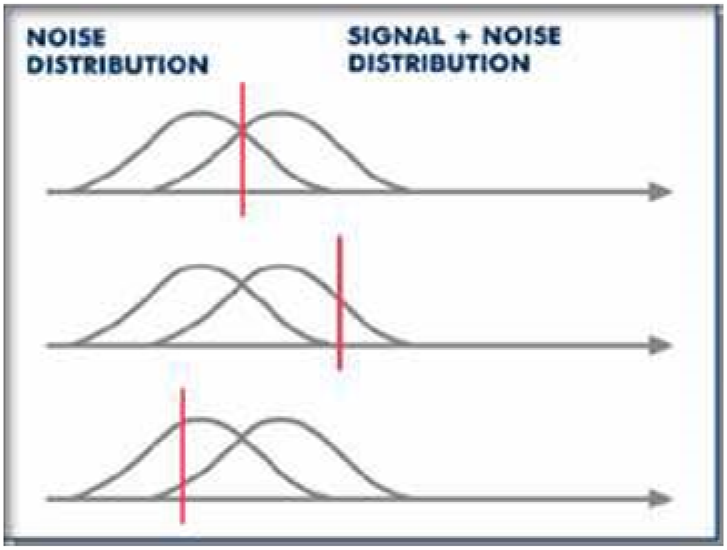
We mentioned that professional calibration is agnostic to personal thresholds of sensitivity or detection. Thinking back to the SDT diagrams above, any two cuppers, given many observations of the same signal and reference coffees, will produce somewhat different distributions for the attributes of each. Cupper A may experience less differentiation between the acidity of the two coffees than cupper B. Cupper A would then experience greater overlap of acidity distributions in an SDT model and therefore a higher proportion of false alarms and misses.

This difference is of course an issue of sensitivity, but it is not resolved by changes in sensitivity. After initial training and validation, sensitivity is not particularly changeable (once we know how to recognize a flavor or compound, we do not become markedly more or less sensitive to it as a matter of further training: we can taste X ppm but not Y ppm). Professional calibration comes into play when Cupper A and Cupper B recognize that they are operating on different distributions with different criterions for the same underlying inputs and then work to adjust their respective criterions to approximate an agreed upon standard.

As mentioned, it’s not necessarily beneficial to be a parts per billion taster in a parts per million world. Obviously neither is the opposite. Cuppers who are very sensitive need to learn how their impression of “strong” aligns with the “moderate” of their peers. Similarly, cuppers who are dull on a given attribute need to learn how their impression of “moderate” aligns, or does not align, with that of their counterparts. Calibration means that cuppers tighten or loosen their criterions according to a determined standard, which may include statements about the acceptable levels of misses and false alarms, or the relative importance of hits and correct rejections.

Consider the example of tasting a coffee with someone. You might say that you found the coffee to be somewhat floral. The other person might reply that they “wouldn’t really call it floral… more herbal maybe.” Assuming you were both well trained, similarly sensitive, and generally calibrated and aligned, this could be an example of overlapping (hard to distinguish) distributions, and or slightly differing placements of the criterion for floral. A divergence in criterion placement is exactly a divergence in calibration, which is to say that when we talk about “calibrating” we’re talking about lining up our criterions for the various coffee attributes that we’re assessing.

The final diagram from UC Davis shows the criterion in different locations. Referencing the prior diagram and its sectors, this last figure demonstrates the cost-benefit of different criterion placements.

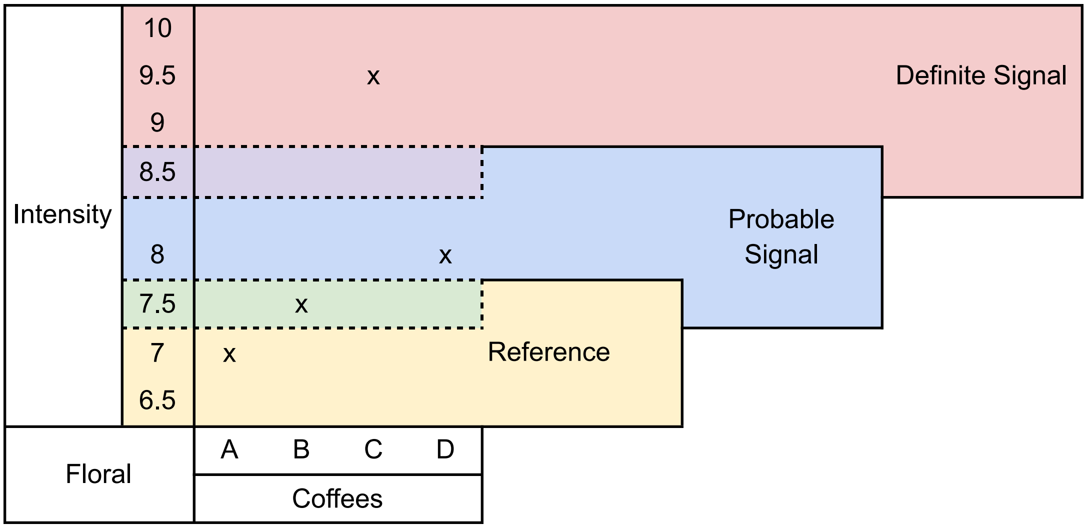


From a SDT perspective, we will frequently speak in terms of the perceived probability of signal presence. Beyond the initial, simple presentation of two possible responses (SIGNAL and NOISE), we might instead allow the use of a scale of probability. An assessor could respond to the same signal or noise input by saying something like: definitely NOISE, probably NOISE, possibly NOISE, possibly SIGNAL, probably SIGNAL, or definitely SIGNAL. This increases the complexity of the analysis, but it also increases the resolution. For example, given a signal input a response of possibly NOISE is more correct than a response of Probably or Definitely NOISE.

This is what the more complex array looks like:

Complex Signal Detection Array		**RESPONSES**					
		Definitely SIGNAL	Probably SIGNAL	Possibly SIGNAL	Possibly NOISE	Probably NOISE	Definitely NOISE
INPUTS	signal	Hit	Hit-ish	Hit-esque	Miss-esque	Miss-ish	Miss
	noise	False Alarm	FA-ish	FA-esque	CR-esque	CR-ish	Correct rejection

In the next figure, this probability is simplified with emphasis placed on two signal zones, the probable and the definite. Attribute scores are associated with the zones along the y axis, while four coffee samples (A, B, C, and D) are given along the x axis. The reference zone is of course our noise input, or flat 80 reference coffee. While these questions (“How intense is it? How probable is it?”) are not the same, we have observed that they display some overlap. More intense attributes will more frequently be identified as more “probably signal.’



To navigate this figure, imagine a flight of four coffees (A, B, C, and D). We are assessing the floralness of each coffee. As per standard protocol, we have a flat 80 reference coffee at the head of our table. The key questions for a floral attribute test would be “Is this floral?” and then “How likely, or how floral, is it?” When we taste coffee A we get nothing. Coffee B it seems like there’s *maybe* a little blip of floral. Coffee C *obviously* pops, and then coffee D seems apparently, or *probably* floral.

Note that there are two areas in between the primary zones where the colors are blended. Green (could be called “possibly signal”) between the reference and probable zones and purple (could be called “likely signal”) between the probable and definite zones. These represent the decision areas where different distributions from our previous examples might overlap. We know from experience that the reference under the correct conditions could rate as high as a 7.5, while a probably floral coffee could rate as low as a 7.5.

When we approach a cupping table with a cupping form that asks us to report the intensity of the floral attribute (or even just the “goodness” of the flavor attribute), we don’t want to go into each cup looking for notes of jasmine, rose, or bergamot. It is very much preferable to taste the coffees initially with a simple openness to whether they “ping” floral or not. It is a waste of time and palate to dig into coffee A and try to tease out a note or two about its potential floralness. That coffee may have definite attributes elsewhere, and if so, identifying and describing those will be a better use of an assessor’s attention.

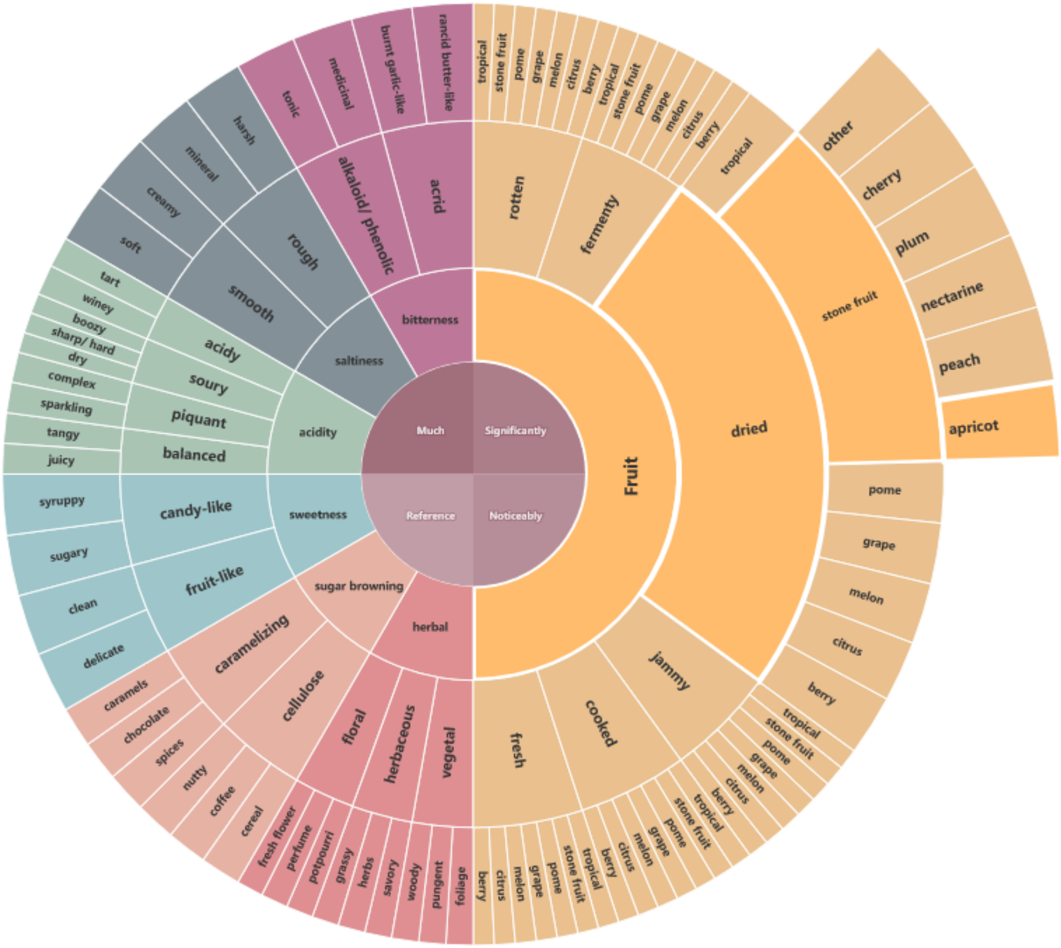
Coffee cuppers need to trust themselves. It is possible that we may miss some things from time to time, but neurotically screening for misses will exhaust our endurance (leading to more misses) and increase our rate of false alarms (finding attributes where they are not, or what I like to call “putting flavors into the cup”). SDT establishes a criterion zone around our flat 80 reference coffee, allowing assessors to reduce their initial test protocol to a decision of signal/noise for each attribute. For attributes determined to be signal, secondary refinements (how signal is it e.g. 8, 8.5, 10? what signal is it e.g. jasmine, generic, rose?) can be focused on with second and third passes by the table. For attributes falling into the uncertain space, they can be compared directly against others that yielded more transparent outputs. Uncertainty is also information, and in most cases it is better (more accurately descriptive) information than a forced or manufactured notation.

Ultimately, most experienced cuppers will process the associative notation simultaneously when the category “pings”. Thus, for coffee C above, I would expect to get a big shot of some specific floral note essentially immediately upon tasting that sample. For a taster working within a well trained associative lexicon on a familiar cupping form, I think that this is the normal experience. For new tasters seeing more experienced cuppers seemingly effortlessly pulling notes out of coffees, there is a risk of overlooking the importance of the effort that has gone into the “effortlessness.” As a result we might see a new cupper get stuck tasting a coffee over and over again, trying to pull notes out whereby they end up pushing notes in. Or in some cases we see what looks like a sort of free association of words that have very low reliability and therefore descriptive power (despite coming off as specifically descriptive).

The base approach to cupping suggested by SDT is quite technical and methodical, and for a beginner either to cupping or to the SDT paradigm, it will definitely feel that way. In time the strictness of the approach loosens up, while the technical and methodical base remains. This approach is likely to help all coffee assessors, on whichever cupping form they utilize. It is a logical, practical, and practicable method for helping to realize the often heard admonition for cuppers to “build their cupping scores.”

The Cafe Imports Coffee Rose is built on this structural, noise to signal, large to small, category to specific building block approach.

The Coffee Rose: Structure



The Coffee Rose is composed of seven attribute categories, four tiers, and an intensity indicator. These combine to create a multi dimensional, descriptive coffee scoring form that is sensitive not only to the broad qualitative categories that we are accustomed to but also to the degree of specificity noted within each, the specific content of notations, and to the quantitative intensity of each impression. The Coffee Rose is sensitive to:

1) Category: attributes on a standard cupping form; these provide context for a description but low descriptive resolution; Acidity is measured separately from Sweetness.
2) Specificity: how general or specific descriptions are; on standard forms specificity is decoupled from scoring; Specific tier descriptions have greater descriptive power and impact than Qualifier tier descriptions.

3) Content: the specific content of descriptors and descriptor strings; on standard forms content is decoupled from scoring; Apricot has greater value in the Cafe Imports system than Coffee Cherry.
4) Intensity: how intense a descriptor impression is, relative to a reference; standard cupping forms tend to be qualitative whereas the Coffee Rose is quantitative; higher intensity descriptors have a greater impact on coffee description and score than lower.

The first tier of the Rose contains seven attribute categories that correspond in resolution and descriptive power to the attributes found on most cupping forms. They are:

1) Fruit
2) Herbal
3) Sugar browning
4) Acidity
5) Sweetness
6) Bitterness
7) Saltiness

These categories are divided into flavors - aromas (1 - 3), and tastes (4 - 7). Since 2016, Cafe Imports has assessed coffee flavor under the categories “Fruit”, “Floral”, and “Caramel” in contrast to the overarching category “Flavor” and the redundant (in practice) category of “Aftertaste.” This is an example of building some qualitative function into the form itself, rather than leaving it for individual assessors to deal with. We’ve structured our form on the assumption that these three broad categories of flavor are *generally* good. Because we’ve also been assessing intensity as in QDA, indicating more caramel flavor will yield a higher score, again without the assessor needing to be responsible for making a value judgment about the goodness of caramel.

With the Coffee Rose, we maintain the basic idea of building qualitative judgments into the form itself, but expand on it and refine it significantly by associating unique values with different permutations of a given category. A weakness of our current form is that it is unable to capture alternate connotations within a category. The Rose has no problem with that at all. The same principle is at play, but is here applied with a much higher degree of resolution.

The second tier of the Rose has what we call Qualifiers. These are descriptors that qualify or further describe their parent categories. For example, a coffee could be described as fruity, and that fruitiness could be described as being like cooked fruit or fresh fruit. “Fruit” is the category, and both “cooked” and “fresh” are potential qualifiers.

The third tier is called Types. These are types of items that fit into the category selected. For example, citrus and berries are both types of fruit. These give greater specificity to a category description. In order to access this tier an assessor must first specify the category and an applicable qualifier for that category. In terms of sensitivity of the scoring engine, all other things being equal, the indication that a coffee is fruity will have a lower impact on the descriptive and valuation output than indicating that it is citrusy, etc.

The fourth tier is called Specifics. Here we find specific examples of a selected type. Raspberries and blueberries are both specific examples of berries. At this level there is also an option for “other” that allows a user to enter anything that they want, but that they do not find on the flavor wheel or in the location where they would like to contextualize it. Perhaps you taste tomato, but mean to indicate it as a vegetal note rather than as a fruity note where it is found on the wheel. You could just click out through the appropriate descriptive pathway and enter tomato at the specific location that best captures your intent.

This function also allows more “free associative” style tasters to utilize any notation that they want. Administratively, the Coffee Rose is still able to capture and score strange, personal, and esoteric notes by requiring assessors to contextualize their novel entries within an existing descriptor string (in order to enter a novel specific descriptor you must first enter the category, qualifier, and type that it best fits with). In fact, this contextualization requirement actually supports the use of those more esoteric or personal descriptors by providing a conceptual pathway for other users or observers to understand their meaning and placement in what is essentially a locally co-created lexicon of coffee communication.

The center of the Coffee Rose contains the intensity indicator, which also functions as the “enter” button for each descriptor string. We rate intensity in comparison to a flat 80 reference coffee and use four degrees for that rating. Descriptor intensity can be:

1) Less than or equal to that found in the reference (labeled as Reference),
2) Noticeably more than the reference (labeled as Noticeably),
3) Significantly more than the reference (labeled as Significantly), or
4) Much more than the reference (labeled as Much).

In practice, the Rose itself spins so that whichever category an assessor is working on is expanded and rotated to fill the right half of the wheel while the rest of the categories are compressed on the opposite half. Further, the specific level descriptors only pop out when a user selects an option in the corresponding type tier (e.g. selecting berry will make the specific berry selections appear). This saves space and clutter in the interface, and also reinforces the assessment process of starting with larger, more general building blocks before moving to smaller, more specific ones.

The Scoring Engine

Behind the Coffee Rose we have what we call the scoring engine. The scoring engine assigns values to individual entries on the Rose, calculates values for descriptor strings, sorts and tallies category and then sample level scores, processes the descriptor strings into natural language forms, and selects from a coffee's generated descriptor pool for top level descriptor output.

On a standard cupping form, scores are generated on the basis of values being assigned to attributes by assessors. For example, I might score the flavor of a coffee at an 8.5 and then elsewhere on the form describe that flavor as chocolate and raspberry. Someone else might describe the flavor as chocolate and raspberry, but only score it at 7.5. Another person might score the flavor at 8.5, but they might describe it as savory and floral. Yet another person might score it at 8.5 without offering any description at all.

In each of the above cases, a score is applied by the assessor to the generic concept of flavor, as opposed to the specific notation that they provide. This is because any descriptions (and of course intensities) that are noted are just supplements to the base impression of quality. Descriptions in most cupping forms are assumed to correlate to scores without having any functional bearing on or causal connection to those scores. The available attribute scoring range on most cupping forms is around four points (6 - 10). Even if we allow that people actually utilize this full range (most do not), an attribute difference of a single point (that between 7.5 and 8.5) amounts to a full 20% of that total available range. This sort of gap can easily occur between two people using the same descriptive language, or between two assessments performed by the same person, even if the described experiences are not materially different.

On the Coffee Rose, attribute scores are generated directly on the basis of descriptive notation and intensity measurement. These scores are therefore sensitive both to how attributes are described qualitatively as well as to the quantitative intensities those descriptions are observed at. While it is still the case that two people can generate different descriptions and arrive at different outcomes for the same coffee, similarity in descriptions reduces differences in scoring when using the Rose. Further, as mentioned in the discussion pertaining to lexicons, the differences that do arise between cuppers are made computable by the Coffee Rose, as opposed to when they are individually generated ad hoc.

The scoring engine has its roots in a project that we did back in 2012. During that time, we began assessing coffees based on scoring standards built around how they were processed. As an example, from a washed-centric perspective, many wet hulled coffees will technically be defective or, at the very least, will tend to score poorly. One cupper might score 84.5 points saying "for a wet hulled coffee, this is very good," while another taster could say, "this coffee is not clean, being vegetal and having flavors of lipid oxidation, and should not score higher than 79 points." Both could be correct. We designed our processing based scoring standards by considering the perspective of what is sought after, expected, and accepted in non-washed process coffees. We then included a process code with our blind table codes so that we knew which standard applied for which coffee.

The Coffee Rose's scoring engine applies these standards automatically. One downside to providing a processing code in a blind presentation is that it impinges on the blinding and triggers expectation error in assessors. While we want our cuppers to be oriented to assessing the particular profile novelties of wet hulled and other coffees, we don't want them to project those novelties onto the samples. It's important to remember that technical sensory errors are human errors. They are very challenging to avoid in the moment, even when you know the risks. The scoring engine allows us to remove the processing code from the sample presentation while still applying those processing standards to the scoring process by assigning different weights and values to descriptors depending on coffee processing. When a cupper describes a "blinded" coffee as being herbaceous and it turns out to be wet hulled, not only can we assign value to that description that is appropriate for the processing, but we can also have a very high level of confidence that the reported profile is true to the coffee as opposed to conforming to the assessor's expectation for that process.

The larger arc from process based scoring standards to this new scoring engine gives a practical example of a design cycle for protocols aimed at reducing specific types of error. Initially, we had noisy *and* biased cupping in instances where coffee processing was apparent and influential, but not standardized for value, and where assessments were preference based. This was problematic even when disagreeing cuppers were both "correct" in their assessments. We reduced these errors significantly by developing process specific scoring standards and by indicating which standard was to be used as part of a blind coding presentation for samples. Unfortunately, this necessarily reduced the "blindness" of our protocol. The net reduction in error, noise, and bias made the trade off worth it. We further refined our process by adopting the QDA style protocol described earlier. The Coffee Rose (and scoring engine) allows us to refine this protocol and reduce error rates even further by combining the strategies of fully blind presentation with processing specific valuation standards.

The scoring engine is, of course, not something that an everyday user directly interacts with. This is by design. However, recalling the discussions on standards, lexicons, and transparency, it is clear that the engine can't be a total black box. In addition to the dynamic tinting and shading connotation features of the Rose itself, users can also look up descriptor values in the glossary. The actual math employed by the scoring engine for descriptor strings is in the form of multipliers. As such, the raw numerical value (multiplier) for any given descriptor is not particularly useful to know because in each case the multiplier is aimed at the range of possible output scores and therefore takes into account the different possible descriptors that come before and after it. It is not useful to compare the multiplier values for raspberry and jasmine because they are not determined relative to one another, but rather are determined by the output value range of the potential descriptor strings that might contain them.

Because of this nuance of the scoring engine, the glossary lists two more meaningful value metrics for each descriptor. The first is the possible output range for a given descriptor (when assessed at intensities greater than reference). These ranges take into account intensity and prior tier options for a given descriptor. For example, a descriptor string like Fruit ⇛ Berry Jammy ⇛ Raspberry ⇛ Much will get close to the maximum score potential for the descriptor Raspberry, whereas Fruit ⇛ Berry ⇛ Fresh ⇛ Raspberry ⇛ Noticeably will yield fairly close to the lowest possible value for Raspberry. The second metric that the glossary provides for descriptor value is a percentile. For each descriptor, this metric locates the percentile amongst the low range of all descriptors and again for the high range. It then reports the mean of those two. In this way, we can see at a glance that apricot is a relatively more highly ranked descriptor (96%) than cranberry (65%). These metrics let us see the actual scoring ranges for each descriptor, as well as their relative rank amongst all other descriptors.

Spinning the Wheel

The Coffee Rose form has four primary components:

1) the wheel itself
2) a table that is generated by your cupping notes as you select them from the wheel
    a) this is in a drawer to the upper right of the form, and is designated by a small spoon icon
3) a list for indicating defects
    a) this is in a drawer to the lower right of the form, and is designated by a small thumbs down icon
4) a panel listing flight information, the extractions (coffees) on the current flight, and a definition box that supplies a definition for each descriptor button that is selected
    a) this is on a panel on the left hand side of the form

The Coffee Rose is arranged in four tiers that wrap around the central intensity selection hub. These tiers are used to select simple descriptors and also to build more complex descriptor strings. The intensity selection hub is used both to report the intensity of impressions and also to endorse descriptors (or descriptor strings) into the system. Beginning in the center and progressing outward, the descriptors in each tier are more specific than the last (e.g. Fruit, Jammy, Berry, Raspberry). Combining items from more than one tier forms a descriptor string.

Some rules for using the Coffee Rose:

1) The first thing that a cupper does when assessing a coffee on the Rose is to select a category. Selecting a Category rotates the Coffee Rose so that the selected category is on the right hand side of the wheel, expanded to one half the total arc of the wheel. This indicates that the selected category is active and ready to be used.

2) One entry is required for each category, even if it’s just to say that a category is equal to or less than the Reference.

a) Defects can be selected from the drawer on the lower right of the form. Cuppers simply select the defect and the number of cups that are impacted. Cuppers can then continue cupping the sample, or if all cups are impacted, can move on to the next sample. Marking a defect will discontinue a sample's scoring from counting toward any total or composite score unless a complete assessment is completed on any clean cups remaining. The Coffee Rose will present a composite profile based on "clean" cups while also reporting the type and number of defects that were found.

3) The Category tier is the minimum required entry. The bare minimum description (barring defects) that this form requires for each coffee is 1x7 Categories + Reference.

4) For each subsequent tier, the prior tier must have a selection made. Thus, in order to select the Qualifier *candy-like* for the Sweet Category, the Sweet Category itself must first be selected.

5) An intensity indication must be made to enter each descriptor or descriptor string that you wish to input. Taking raspberry as an example, the actual steps of entry would be to press the Fruit button, then the Fresh (or other qualifier) button, then the Berry button, then the Raspberry button, and finally the intensity button that best indicates how much stronger the Raspberry flavor in the subject coffee is than in the reference.

6) There is no limit to how many descriptors or descriptor strings a cupper can enter for a given coffee or for a given category attribute.

7) As you select and endorse descriptors and descriptor strings on the Coffee Rose, a table is generated containing each of your selections. This table allows you to see what you've said about a coffee, to see where you're at in terms of completing the necessary selections, and also to delete any entries that may have been made in error or where you've changed your mind. The table can be sorted by column headings for easier navigation.

8) If you wish to delete an entry, you can select the small x next to it on the display table.

9) You'll know that a category has met its minimum entry requirements because the button for that category on the wheel itself will remain highlighted after you navigate away from it.

10) You'll also be able to see in the descriptor table which categories have received entries. You'll know when the minimum entry requirements for a given coffee have been met because the status indicator on the extraction card will change from "incomplete" to "complete".

11) Table navigation is done by selecting from the array of sample codes (on what are called “extraction cards”) located in the panel to the left of the screen. These 3-digit codes can be generated randomly by the system or input manually in the flight builder.

12) As selections are made, the buttons and Category will highlight, indicating whether the current cumulative entry for the category carries a positive or negative scoring connotation.

13) Selecting a descriptor will populate a definition window in the lower left of the form, giving a brief explanation of that descriptor.

14) The scoring engine develops a cumulative value for each Category based on the entered descriptor strings that are furthest from zero. For example, if two positive value descriptors are entered to describe the acidity, the higher value descriptor will be the one used to calculate the score for that category. Similarly if two negative value descriptors are entered then the lower value (more negative, further from zero) will be used. If both positive and negative value descriptors are entered, the furthest from zero negative value descriptor will be subtracted from the furthest from zero positive value descriptor in determining the final Category value. The net value for the category will be indicated in the highlight color of the completed category on the wheel.

Conclusion

One of the driving goals for this project was to bring our cupping program into greater alignment with sensory science, but without forcing a round peg into a square hole. I would like to conclude this paper with one of the “aha!” moments that I had around that idea during the development process of this cupping form. For the longest time, I thought that in order to hew closely to the Sensory Science line I would need to push back against the qualitative, preference, and personal language aspects of cupping in favor of emphasizing the quantitative, numeric, and measurement components. Indeed, some pushback there is warranted. As is some finesse.

Sensory Science is really an interesting field. Included in all the logic, statistics, and best practices is a foundational commitment to a project’s stakeholders and the realities of their business environment, along with a patient and iterative approach to discovering the best possible solution to the stated problem. In the real world of coffee assessment, there is rarely time for the sample sizing and replication common to most sensory science studies, which are highly oriented around statistical validation. Further, much of specialty coffee cupping, in practice, boils down to communication, description, and sharing a human tasting experience that is explicitly not captured by cupping scores or, speaking bluntly, strictly accurate measurement of coffee constituents. Consider that the legendary La Esmeralda Geisha from the Best of Panama competition in 2004 only scored 96 points. This may seem like a high score, but this coffee was described as an “otherworldly beautiful coffee” the likes of which at least one cupper present “had never tasted before”… which begs the question as to where anyone would ever expect to find those last four points and what sort of nectar of the gods level coffee would warrant them!

In many ways, a strict sensory science approach (of the sort that I initially *imagined*) runs the risk of taking the “specialty” out of specialty coffee, at least for the specialists. Do we want that? By the same token, we have to ask ourselves if we’re okay with the “specialty” of specialty coffee meaning **anecdotal**, **imprecise**, and **personal**. As much as we may prefer to deny that it is these things, as much as we may want to project rigorous (and dare I say “scientific”) precision and validity, at bottom I suspect that in large part we are not quite willing to let these things go entirely. And beyond shoring up some fundamental issues, I’m not sure that we should be.

There are no “otherworldly beautiful coffees” in the lab. Nor is there iconoclastic, cynical, and delicious rejection of norms. There are just samples with more and less sweetness, more and less acidity, etc. But coffee is clearly personal. From the merely habitual drinkers who, “can’t start *my* day without *my* coffee,” even if it’s just a grocery store, pre-ground drip, for whom coffee is already on an equivalent linguistic footing (and one suspects existential) as their very day, to the deep cut connoisseur who knows more about their morning (and afternoon, and evening) extraction than most of the people involved in getting it to them. From the coffee that *I* grew and that *I* processed, to the roast that *I* dialed in and perfected, to the flavors that *I* identified, experienced, and described, coffee is clearly a very personal thing for many, many people.

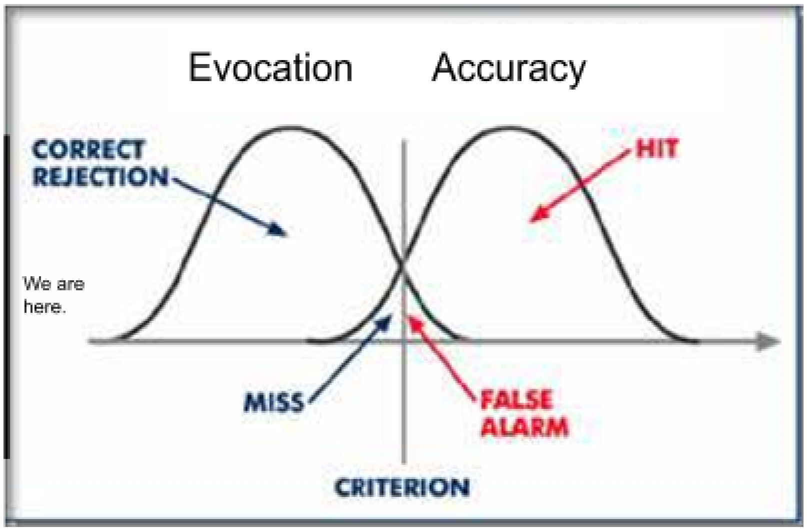
This personal-ness is in large part why we trade so heavily in anecdotes. It's our love language and our mother tongue. Many specialty coffee professionals have a go-to anecdote describing their (generally eye opening, if not outright life changing) introduction and entry into specialty coffee. Some even have more than one. As for precision, for all of our sifting of coffee grounds and weighing of…everything, at the end of the day most of us still just end up talking about what we like and don't like, rather than measuring the components of the coffee solution we just took so much care preparing. While lacking in precision, this is deeply personal and often *is* best expressed through anecdote in an attempt to capture and share (and proselytize and defend) something of the personal novelty not just of the coffee, but of *my experience* of the coffee, of what the coffee did to *me* and how it made *me* feel. Anecdote is the medium of the personal, and that space is naturally imprecise. We may not want to completely eliminate the personal, the anecdotal, and the imprecise, but I think that we need to recognize, balance, and temper those things with practices that reduce their noisiness and bias, making them more reliable and transparent. In doing so, they can become *more* descriptive and expressive, supported by the widely recognized foundations of sensory science.

It took me a while to understand that sensory science's focus on measuring measurable things in as objective a manner as possible is really about describing those things in as meaningful a way as possible, and that in order for a description to be meaningful, it has to be meaningful not only to the object of study but also to an audience. Saying that X coffee, tasted N times, has Y amount of tartaric acid in malic acid equivalents with Z significance is technically much more descriptive (if also much dryer) than saying X coffee has "oodles of soul crushingly beautiful" acidity. However, if the former statement is not meaningful to its audience, then practically speaking it's no more descriptive than the latter.

The goal is no more to be correct, stuffy, and incomprehensible than it is to be a parts per billion person in a parts per million room. The goal is to communicate to an audience or with a peer something meaningful and informative about a coffee, in a way that is reliable and transparent. We can critique the paucity of meaning in hedonic ratings that are dressed up analytically, but that neither means that hedonics has no place in specialty coffee nor that we must swing our assessments to as robotic a place as possible. We can critique the gate-keepy and imprecise predilections of hedonic expertise that our current industry models are steeped in without ignoring the interest and insight afforded by coffee specialism.

Taking a lesson from signal detection, where the criterion is movable and increasing our hit rate comes at the cost of accepting some false alarms, in specialty coffee, we can decide how much error, how much noise and bias is acceptable in our protocols. We can discuss and decide how much accuracy, transparency, and probability (reliability) we need, recognizing that everything comes at a cost. We know that error, bias, and noise increase the more our cupping protocols are built on preference, hedonics, and momentary opinions while being presented and run as analytic rather than consumer tests. Soliciting this information from experts may have some sensory benefit, but that benefit is offset by a lack of transparency, reliability, and validation. We also can observe that the flavor and evocation of our sensory descriptions *improve* when we allow our coffee tasting expertise some room to run and a platform to chime in.

We can reimagine the signal and noise distributions from earlier such that the noise distribution becomes a flavor preference or evocation distribution and the signal distribution becomes an analytic or measurement accuracy distribution. If, as I suggest, the criterion is currently shifted deep into the evocative, flavor preference distribution, and if we decide that we want to correct that, it does not mean that we must blindly impose the strictest possible criterion and shift correspondingly as far into the analytic, measurement accuracy distribution.



Rather, we'll be better served by discussing our needs and then aiming to place our criterion on that basis. As an opinionated and iconoclastic bunch, I think that we'd probably do well to maintain some room for opinion in our mix. As a story, connection, and conversation oriented industry, we want to ensure that we do not snuff the heart out of what we do. At the same time, as an industry seeking transparency in business and equity in trade, we must find and employ strategies that actively and practically support these goals throughout our workflows. As professionals in a far reaching livelihood ecosystem, we must also strive to be as accurate and consistent as we can. We must be willing to show our work. This means reining in those cupping protocols where personal narrative = coffee assessment and updating them with practices that allow cuppers to more transparently profile the coffees that they assess.

Sensory science asks us to clarify our questions before selecting or designing our tests. After talking with folks on both sides of the bean, so to speak, it became clear that some people have a much greater need for scores that track with industry expectations, while others need cupping data to be as communicably descriptive and conversational as possible. Both groups obviously stand to benefit from process improvements that reduce error, increase transparency, and make output more reliable.

In designing the Coffee Rose and its scoring engine, we wanted to make it possible for our cuppers to focus completely on describing the coffees that they tasted. We wanted to relieve them of the tasks of judging good and bad, like and dislike, and of trying to then value those judgements. We wanted to design a system where coffee scores were directly related to their descriptions, and where the core principle was: “cuppers describe coffees; standards value/score descriptions.” We wanted to develop a protocol that accounted for known sensory errors and that controlled for them as much as possible.

We also wanted to maintain focus on the basic question that we’re all asking when we cup: What does this coffee taste like? Pushing cupping to technical extremes that prohibit reasonable access or that restrict comprehension and communication is as off the mark as communicating freely about made up metrics and highly personal associations. The Coffee Rose is an attempt to pull as much as we could from the sensory science toolkit, but in service to the needs of specialty coffee cuppers and communication throughout the coffee livelihood ecosystem.

At bottom, the outputs from our sensory tests should not just be accurate, transparent, and reliable, they should not just be understandable and communicable, they should also foster communication and connection. To the best of my current ability, I think that the Cafe Imports Coffee Rose helps further these goals.

Ian Fretheim
21 March 2022