Learning to Cup in Hawaii: Part 1

I would guess that when most people in specialty coffee think about places to learn about coffee and cupping, Hawaii does not come readily to mind. Much of the Hawaiian coffee that we see lands right around 80 points. It tends to be simple and nutty, have citric acidity, maybe some lemon flavor, and in some cases perhaps a bit of an herbal aftertaste.

Hawaii: the official coffee of 10% blends.

The nicer Hawaiian coffees that we've brought in for Cafe Imports have scored up to around 85 points, with a notable standout coffee from K'au that landed at 88 points. We know that there is good quality in Hawaii, perhaps excellent coffee, but somewhere between the competition and the cost at these score levels we seem to lose track of the nicer Hawaiian offerings. These are not cheap coffees, at any score—and this is important. We use score to justify price. This makes sense a simple, meritocracy sort of way. Of course score is hardly the only merit, and this underscores the reality that meritocracy is far from reality.

What if a coffee reminds you of your honeymoon and rekindles a flame that seemed to have gone out? What if it brings you back to that week on the beach, when things felt right, carefree, and you experienced child-level happiness again? Does it matter that it's only an 80-point coffee? Does it have merit? Maybe we need to talk about the value of scores. Nevertheless, these are expensive coffees, and if you're really just talking about Primes and basic Extra Primes, then the dollar per point ratio is probably one of the highest in coffee, outside of top auction lots.

Hawaii: the official coffee of “Hey Jamaica, hold my beer.”

Hawaii: the official coffee of slowly extracting the fingernails from Adam Smith’s invisible hand.

For many of us, we had our introduction to “nice” coffee when some local shop or other put out their holiday Kona, charged an arm and a leg for a cup and told us that this was the good stuff. We mostly grew out of it, introduced to lighter roasts and denser beans. Growing out of it, at least for myself, I didn’t look back, I put it behind myself, moved on and gave it no more thought. In the odd moments where Hawaiian coffee came back into my life, it happened to play to my expectations and so the opinion was reinforced.

Hawaiian coffee is like family, or old friend groups. No matter how much we learn and grow, after a couple of drinks and by the end of the reunion or as the pie comes out at Thanksgiving, we find ourselves reanimating our old roles. Bring up Hawaiian coffee and we immediately fall back into our old habits and associations: the dark roasts, herbal aromas and bitter finishes.

Hawaii: the official coffee of your aunt and uncle adamantly proclaiming they’re sharing with you the best coffee they’ve ever had and that you’ll just love while scooping tabs of pre-ground ash from a gold foil bag with 1993 label graphics despite the fact that you bring them AA Kenyas, G1 Yirgacheffes and Honey coffees from Costa Rica that were hand-picked, processed and meticulously dried by a guy who produces like four bags per year, every time you visit.

But maybe we have it wrong. Maybe Hawaiian coffee is not only the Hawaiian coffee that we’re willing to buy because just enough people will gift it if we price-point it with a cheap enough blender around the holidays. Maybe there’s more to Hawaiian coffee than that. Maybe we’ve become so habituated to Amazon and Walmart that we’re not able to see beyond the first page of a lowest-price filter and we don’t know the first thing about what Hawaiian coffee is.

In the last two or three years, I’ve had the opportunity to relieve myself of some of these previously held opinions regarding Hawaiian coffee. I’ve learned a lot in the process. Traveling to Kona for two very different cupping competitions and also getting to work closely with the folks running those competitions proved very instructive. In one we explored new avenues for the organization of coffee competitions, significantly emphasizing transparency and objectivity in scoring. In the other we worked with a panel of Q graders using the SCA-Q scorecard and protocol, allowing me to re-investigate the industry standard practice first hand. In both cases the coffees presented were significantly better and more diverse than I imagined they would be.

The Kona Cultural Coffee Festival

In 2016 I launched into the project of developing and introducing a new cupping form for our sensory program at Cafe Imports. Its creation was informed by my introductory studies in sensory science and sought to remedy items in our protocols that were at odds with those studies. Having used that form now for a couple of years we have learned a few things. After cupping on older, more qualitative and subjective forms, it’s a relief to use such a straightforward system. It’s a relief to set aside subjective qualification and simply use the form to describe what is there. That said, this scorecard and its supporting protocols are imperfect. Changes to both are in the works, though those details will be presented elsewhere.

In 2017 I was approached by Shawn Steiman PhD, of Daylight Mind Coffee Company and Coffea Consulting in Kona, Hawaii. He was given the opportunity to take the reins of a green coffee competition as part of an annual festival called The Kona Cultural Coffee Festival. He wanted to do something different. And with good reason: Shawn is a scientist and his background in biology, chemistry, and sensory science, coupled with his many years in specialty coffee, put him squarely in the middle of that sliver of Venn diagram that lies between Sensory Science and Specialty Coffee.

Shawn reached out in part because he had heard about our having introduced a new form and was curious about the challenges that we encountered, pushback, and how we dealt with it. After e-mailing back and forth about our process, as well as his own goals and vision, he eventually invited me to participate as a judge. There is much to report on the topic of the contest and the coffees themselves, though for now I will stick to the matter at hand. Shawn wanted to create a contest scoring system that would minimize panel bias and increase criteria transparency for those entering their coffees. This is worth thinking about. Shawn received some flack for his system, and in two important ways (that I can think of) it can be validly critiqued. At the same time it is easily the most transparent contest system that I have seen or used. Further, it was designed as a contest system and it functions with great precision in that role.

Here's how it worked:

The judges developed two target profiles before the competition began, one called the Heritage profile and the other called the Modern profile. The profiles were designed to recognize two distinct streams in Kona coffee. The first is more traditional Kona coffee. This harks back to the days of homogenization and regional blending, when you might choose between a Colombia Excelso or a Guatemala SHB and have a reasonable expectation for the difference between them. By the same token, you would not expect any significant difference between two SHB Guats. Kona coffee is no different and really came into its own as a distinct, quality offering in that milieu. Regional isolation, uniform processing, and blending, and majority Kona Typica coffee trees created a specific, pleasant and sought-after cup profile.

The Modern profile looked to the more recent developments whereby some Kona producers have brought in new coffee varieties and begun using different processing methods, along with lot separation. These coffees can be a significant departure from the classic profile and mirror much of what we see as the leading work throughout Latin America.

Shawn and the competition judges held discussions to try and determine the scorecard profile for both of these. It is noteworthy that this is not an exercise that most cuppers have participated in and yet it is a basic component of setting up a sensory analysis panel. Specifically, this part of the project allowed us the opportunity to communicate and calibrate our experience and begin with some rudimentary lexicon building. Contrary to popular usage, lexicons are more appropriately local than universal (we may seek to create a universal lexicon, but we will always calibrate and apply it locally, and the validity of remote-universal application is seriously questionable). A good experimental design will allocate at least some time to establishing and calibrating the terms of usage (the lexicon at play) for the experiment (competition).

The goal of the competition was not to have the most or best (highest-scoring) acidity. The goal was to have the acidity level that most closely matched the profile target. For example, the Modern profile target score for acidity was 7 out of 10, with 0 being absent and 10 being the most acidity there could be while still being pleasant. To help with calibration before the competition we used water-taste/flavor solutions to dial in and define the attribute parameters that we were using.

In addition to acidity, we also assessed sweetness, body, floralness (how floral is the coffee) and coffeeness (how coffee is the coffee). Scores were tabulated by their squared differences from the target. If one sample scored an 8 on acidity and another scored a 6 (with a target 7), they were tied with a squared difference from the target of 1. Descriptors that were not handled by the attribute categories, as well as defects, were listed to the side in boxes called Complexity and Defects, respectively. If multiple judges recorded equivalent descriptors (lemon and lime) or defects (dirt and earth), the coffee would receive a bonus (or negative) point.

The bases of transparency and critique are related in this system. In terms of transparency, this system is unparallelled. It details exactly what attribute scores are required to win and publishes those targets in advance. In other competitions, the winning profile is indeterminate until the contest has ended and the scores have been tabulated. As an example, we have participated in and are as likely to expect the bidding up of a coffee ranking from 5th to 7th in a Cup of Excellence competition over and above the bidding afforded higher ranking lots.

We have found that frequently the top ranking coffees are simply the most intense, despite the Cup of Excellence placing emphasis on Sweet and Clean. Depending on the panel and its preferences, any number of profiles might win. If I am processing coffee for a competition, how do I know what to submit? Do I leave the washing process loose or move into Honey or Natural processing to try and create a fruitier cup? Or do I try to make a pristinely Washed lot? There's no way to know because the winning profile is indeterminate until the judges arrive and complete their scoring.

Because Shawn's system states the target attribute levels, there is greater transparency. I can look at the targets and see that fairly acidic and floral coffees will be closer to the target than fruity ones (or however the target is presented). This brings in the first critique which is that for the Modern profile, Shawn's system limits the potential paths to victory to just one. This makes more sense for the Heritage profile. Modern profiles are more diverse, frequently using different combinations of growing, variety and processing techniques. Further, in competitions like the Cup of Excellence it is always a possibility that a coffee will show up with a completely unique or new profile. Of course, this is really a double edged blade because the thing that makes Shawn's system so transparent is also what I see as it's primary limitation.

In a strange twist, Shawn's system is actually fairly conservative, despite being so forward-looking from a sensory perspective. It can really only be set up to target known profiles, as opposed to less transparent and more indeterminate systems, which can allow more readily for the unknown to emerge. You've heard the old adage about building: cheap, fast, and good. Pick two. This applies to sensory tests also, in a way. If you want to emphasize a particular component of experimental design (such as transparency, efficiency, quality discovery, repeatability, accuracy, etc) in a series of sensory tests, then you will have to pay for it with the loss of some other component.

Shawn's system is an excellent response to requests for transparency and objectivity. These are both fairly "expensive" components in the testing protocol parts bin. With efficiency being dictated by budget, the unconstrained, open-ended, and indeterminate nature of a discovery protocol makes the best sense to use as the payment for the requested transparency and objectivity. Ultimately it's not a question of liking or disliking Shawn's system. It's a question of what you value in testing. If you want transparency, objectivity, and capacity for quality discovery, then you'll need to be prepared to arrange and fund a very intensive and expensive testing protocol.

The secondary critique is with numerical output. This again has two sides. In Shawn's system points are tabulated in terms of how far you are from the target. The closer your coffee ends up to 0, the better. Thus, an entrant might win the competition with a score of 2. A question arises here about experimental design and the point of running these "cupping tests." Of course you design and set up a "cupping test" (or series thereof) for a particular reason. As an importer we often cup for quality discovery, for example. Most cupping would probably be generically described vaguely as seeking quality assessment, or expert grading. Meaning that most cupping tests are run in search of a final score and notes. Specialty coffee is most often talked about with scores out of 100. We assume that we all know what we all mean when we all say that a coffee is an 86.

This is where it gets interesting. We honestly have only a very vague idea of what anybody who we don't cup with regularly means when they give a score. Further, in a contest setting, by

definition the purpose and so the desired output is to determine the winner. The most transparent and direct way to support that goal is not necessarily going to be the same thing as setting up testing for quality discovery, or broad product assessment or description. One of the reasons that Shawn's system is so good is that it dispenses with extraneous goals and so can be highly focused and efficient.

The downside is that people still want to talk in cup scores. They don't want to talk in terms of having gotten a 2 or a 3. In my view, this is an issue of communication and understanding the purpose of tests (not a strong suit in spec coffee). But it is an issue. For better or for worse, I think that as long as we remain tied to a 100-point system (which, if we're honest we really only use maybe 14 points of), we will have to build translation functions into every attempt that we make to create new scoring systems. When someone asks, "What does that mean, what's a 6? Is this an 85, or what?," we need to be able to answer, "Yes/No."

To make things worse, it falls on us as designers to do this without bogging down our assessors with extra tasks by complicating the assessment procedure. We correctly seek to remove as much as possible outdated and unhelpful practices from the mechanical processes of the cupping test as well as from the mental and psychological space of our assessors. At the same time, we are tasked with building-in to our designs an avenue for digestible communication with those who will consume our results.

I don't know what the final answer/s will be. I do hope to see Shawn's system or something like it in practice again. The system is very much worth working on (and reassessing if you are in the Hawaiian coffee world) and learning to utilize. Competitions are great. Creating and communicating a clear statement of purpose prior to competitions beginning—Why are we having this competition? What are we trying to determine and how is it measured? What is of primary importance and what is less important?—will help in making them even better. We can design tests in accordance with our audiences' expectations. At the same time, our audiences must be willing to align their expectations with the realities of good test design.

In part 2 we'll catch up on some much more recent history with the 2019 Hawaiian Coffee Association Cupping Competition. This contest was built around the SCA cupping form and protocol and was paneled by Q graders from four of Hawaii's growing regions. We'll take a new look at a component of the SCA-Q cupping form and protocol that I had not previously appreciated (or even noticed) and dig in just a little with Signal Detection Theory.

Learning to Cup in Hawaii: Part 2

The HCA Competition

This year I was invited back to Kona. South Kona, to a town called Captain Cook to be precise, where I was invited to be the head judge for the Hawaiian Coffee Association Cupping Competition. The HCA competition is a statewide contest with entrants from nearly all of Hawaii's coffee-growing regions. This invitation was personally interesting to me due to my current work in the development of Cafe Imports' new cupping protocols. The HCA competition was to be run on the SCA standard cupping form, and judged by a panel of Q graders. (Note: I am no longer a Q grader, though was previously and am familiar with the SCA-Q methodology.)

This contest afforded me the opportunity to work with and observe first hand the SCA-Q system with a group of people who were certified (and up-to-date) to use it. This was a unique chance to work in a closed setting with a select group of cuppers. Of course, having been introduced to the breadth and depth of Kona coffee at the Kona Cultural Coffee Festival a couple of years ago, my curiosity was piqued to see what would be on the tables.

The contest was hosted by Pacific Coffee Research in their facility in Captain Cook. To begin, I have to point out that the running of the competition, from what I was able to see, was flawless. I have no doubt that there were at least a couple of "OMG" moments behind the scenes. There always are. The trick is in how you handle them, and Brit and Brian, my hosts at Pacific Coffee Research, made the entire operation seamless. But the coffees were always well within specification (and in fact within a very tight tolerance) for cupping roast, the tables were always cleaned and reset efficiently, the water was always temped and poured precisely, and the timing of it all was well thought out and brisk, without being either rushed or too lax. Brit and Brian were both clearly invested in not only the integrity of the contest that they were hosting, but also in their work in general.

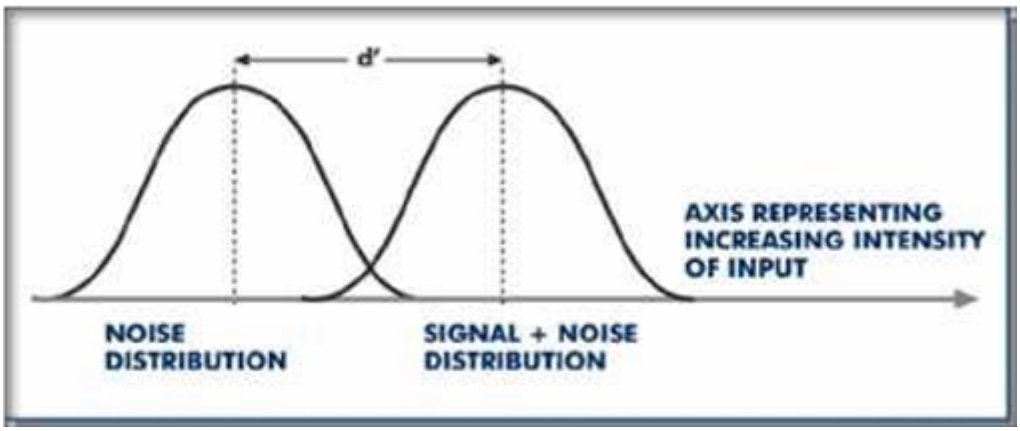
Frankly, the SCA-Q form and protocol are not my favorite. They rely heavily on the expertise of the assessors both with focussing on assessing the quality of an attribute (rather than its intensity) and also with the inclusion of attributes that require a high level of expertise to assess (balance, clean cup, uniformity, overall). That being said, I did learn quite a bit more about what the SCA-Q methodology does well. It should be emphasized that all cupping forms and protocols have strengths and weaknesses.

The SCA-Q form and protocol utilize the concept of the criterion from signal detection theory in a really interesting way. It is possible to critique the SCA-Q method for assessing sweetness (for example) as a cup-by-cup Yes/No, with disproportionate scoring assignments for each response. Other attributes, like flavor, are scored from 6–10 in 0.25-point increments while sweetness is scored Yes or No for each of 5 cups with each Yes being worth 2 points. On other forms, sweetness is scored on a scale just like other attributes in recognition that a coffee can be more or less sweet and that this is an important descriptive metric.

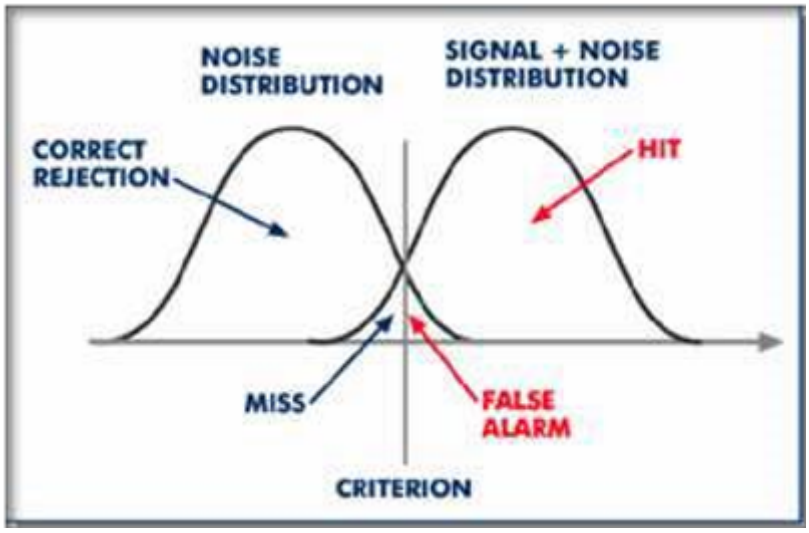
With the SCA-Q form there is no difference noted in sweetness between a base-level specialty coffee that is, relative to commercial coffee, technically sweet, and a very high-end microlot that is, relative to all coffee, very sweet. This can be frustrating. However, the crux (the criterion) here is in the term “relative to.” The SCA-Q form correctly (in my eyes) identifies Sweetness, Cleanliness, and Uniformity as the primary indicators of specialty coffee. This is of course a value statement, a standard. The impressive part of the design of this form is the way in which this value statement (Sweetness, Cleanliness and Uniformity are the *sine qua non* of specialty coffee) is imbedded in the form. First, a quick detour...

Signal detection theory is a decision model that was first developed in the early 1950s. We have been working with SDT at Cafe Imports for the last year and will go into more depth on it and our projects involving it elsewhere. Here, we will limit ourselves to some basic points. Whenever we seek to assess something (this acidity is good/bad, this acidity is strong/weak, etc) there is another function that must be performed. We must first determine whether we are assessing acidity or not.

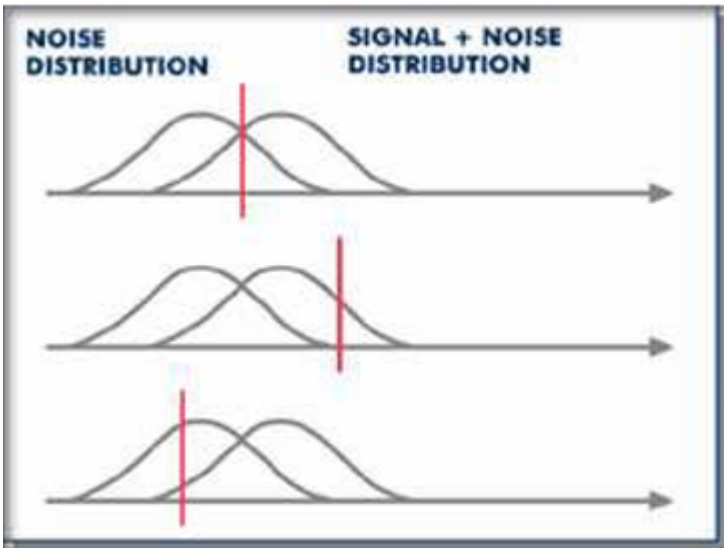
For example, a person reading a radar may need to determine if a blip is an enemy warship or perhaps just a whale. Because both are possible, the radar reader is trained to distinguish between their radar signatures. Imagine two distributions, one for instances of detecting a whale (noise) and the other for instances of detecting an enemy submarine (signal), both along an axis of “intensity of input.” The following graphics from UC Davis will help.



If the two distributions are separate, that means that they are easy to tell apart. If the two overlap, then that means that there is a decision to be made as to whether a given blip is a whale or a sub. There is an area where the submarine signal is weak enough to be confused with that of a whale. We can call the overlap between our two distributions the decision space. Our responses in this space, and the space itself, will be influenced by our momentary sensitivity and also by our bias. We draw a vertical line through our distributions. This is the *Criterion*. To the right of this line we will call everything "enemy submarine," and to the left of it we will call everything "whale."



We can move this line back and forth along the x axis depending on our needs. For example, calling an enemy submarine a whale (miss) is much worse than the alternative of calling a whale an enemy submarine (false alarm). To account for that disparity, we can move our radar reading criterion to the left, deeper into the Noise (whale) distribution. This amounts to saying that when we are in doubt, we will treat inputs as though they are submarines. We will accept an increase in False Alarms in order to secure a reduction or elimination of Misses.



How does all this relate to the SCA-Q form? With the selection of and treatment given to Sweetness, Uniformity, and Clean Cup, the SCA-Q form attempts to build in a three-part criterion for specialty coffee. Similar to the submarine vs whale example, the SCA-Q assessor is tasked with determining whether a coffee [cup] is Sweet, Uniform, and Clean. Again, similarly to the radar reader, an SCA-Q cupper must be trained to identify what counts as sweet in coffee and what does not, what counts as uniform in coffee and what does not, and what counts as clean in coffee and what does not. The SCA-Q cupper must decide for each of five cups presented for every sample whether they fall to the right or the left of the specialty coffee criteria for Sweetness, Clean Cup and Uniformity. Any coffee that fails to rise to signal level (this is sweet, clean, and/or uniform) will be effectively eliminated from specialty consideration.

This particular device is clever for a couple of reasons. In the first place, one might use an SDT model to simply ask, “Is this specialty or not?” Such a question is very difficult to train, let alone audit. However, it would be valid, in particular given a large enough sampling pool. A strength of SDT is that it can effectively use both subjectivity and variability as data. However, by reducing the question, “Is this specialty or not?” to the attributes of Sweetness, Uniformity, and Clean Cup, the SCA-Q form accomplishes something important. It builds in a trainable baseline standard that assessors have no choice but to follow.

Ultimately it is of little importance whether a cupper understands the foundational values of a cupping form, as long as the form deals with those values directly. If your form is vague (as in asking, “Is this specialty or not?,” or, “Is this quality or not?”), then your cuppers will need to understand and actively consider these values. This opens an entirely new can of worms, which we will heartily dig into elsewhere. For the time being, suffice it to say that particularly when cupping a series of coffees in which the question, “Is this specialty or not?” is likely to be asked, the SCA-Q form and protocol can be effective for doing so. Further, it more than other forms ensures that coffee assessments are built on a foundation of the hallmarks of specialty coffee-sweetness, cleanliness and uniformity.

We’ll continue to dig into signal detection theory over the coming months, as well as into the SCA and alternative cupping protocols. Since beginning to investigate and work with SDT in 2018 we’ve come to increasingly see it as foundational to cupping. Finding an aspect of it applied in the SCA form, whether intentionally put there or not, has been a good impetus for reflection on both.

I hope as well that we’ll have the opportunity to continue digging into Hawaiian coffee. My experiences with Dr. Shawn Steiman at Coffea and with Brian and Brit at PCR, and of course with beginning to see more of what Hawaiian coffee has to offer, have been wonderful and eye-opening. Cuppers should always be willing to learn to cup. In the end, our task in cupping boils down to calling it how we see it. The trouble arises when we start seeing it how we call it.