Article

S|S|S

Social Studies of Science
2016, Vol. 46(3) 461–481

DOI: 10.1177/0306312716651504
sss.sagepub.com
SAGE

The taste machine: Sense, subjectivity, and statistics in the California wine world

Christopher J Phillips
Department of History, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

This article is about mid-20th-century attempts to turn subjective judgments about the quality and composition of wine into objective knowledge. It focuses on the research of Maynard Amerine at the University of California, Davis, and his project to formalize the procedures of sensory evaluation. Using controlled experimental conditions, Amerine and colleagues transcribed judgments about taste into numbers that could then be aggregated and analyzed statistically. Through such techniques, they claimed to be able to turn subjectivities into objectivities, rendering private taste sensations into reliable and stable facts about objects in the world.

Keywords

enology, Maynard Amerine, objectivity, quantification, sensory evaluation, statistics, subjectivity, taste panels

The man approached the wooden chair in front of the booth. The room had institutional-style linoleum floors; it was silent except for the air conditioners that kept the space at a constant 21°C. It was just after 11 a.m. and he was slightly hungry, having not eaten since his light, bland breakfast. He had not consumed alcohol in the previous 24 hours and was in good physical and mental condition. He had no dental fixtures or oral diseases and was a non-smoker. The room had skylights, as well as artificial lamps with broad-spectrum 'daylight' bulbs. It was quite bright, filled with natural northern light. A card and a pencil were the only objects on the booth's desk.

Correspondence:
Christopher J Phillips, Department of History, Carnegie Mellon University, 5000 Forbes Ave., Baker Hall 240, Pittsburgh, PA 15213, USA.
Email: cjp1@cmu.edu

Like many modern office cubicles, the booth was closed on three sides, painted gray, and featured a small sliding door just above the desk's surface. The booth was large enough to fit one person comfortably; the desk was about 3 ft^2, and its side walls were tall enough to block out anything else when its occupant looked forward. The man knew that others just like him were also at their own booths in the room, but he could neither see nor hear them. After he sat in the chair and filled out his name on the card, the door slid open, and a female hand pushed a gray tray onto the desk. The door closed.

On the tray was a dentist-style spittoon, as well as three glasses. The glasses were arranged as if they were the vertices of an equilateral triangle. All were identical: 235 mL, tulip-shaped and thin-walled. They had been washed the day before with hot, soft water and detergent, rinsed with distilled water, and allowed to dry fully by being hung upside down overnight; they had not had any contact with wood in the process. Each glass had been painted black, rendering the contents opaque, and the glasses were labeled on the tray with the numbers 90, 27, and 52.

The glasses contained wines at a temperature of 20°C. The man carefully noted the numbers 90, 27, and 52 on the left side of the card, and he then brought the glass labeled '90' to his lips, allowing about 10 mL to spread over the surface of his tongue but not yet swallowing. He began to consider how sweet the wine was, and he took his time doing so. He had been told that one of the three glasses held wine with a different level of sweetness than the other two, and he was now being asked to identify that wine, as well as to determine whether it was more or less sweet. After tasting this first glass, he spit the wine out into the spittoon and turned to the glass labeled '27' and then to the one labeled '52'. Then, he marked his card to indicate that the wine in glass 52 was sweeter than the identical products in glasses 27 and 90. Sliding open the door, he pushed the tray of glasses back through, along with his card, and patiently awaited the next test (Amerine et al., 1959; Amerine and Roessler [hereafter A&R], 1976; Pusais and Chabanon, 1974 [1969]).

This tasting bears little resemblance to typical settings in which wine is consumed. That's by design. Those most concerned with wine quality know that drinkers are always subject to possible 'errors' in their judgment: You can think that a poor wine is good because of the delicious dinner or delightful company; you can also think that a great wine is poor if you were told it was inexpensive or if it came in a styrofoam cup. Though little known outside the industry, there is a complex infrastructure of what's called the 'sensory evaluation of wine'. For those concerned, such factors as the brand or cost of wine can only introduce 'external' elements that might distract from the business of evaluating wine quality: Wine evaluation requires a 'context-free' setting.

The tasting booth was part of the experimental apparatus that mid-20th-century enologists created to turn the ephemeral, private, and undisciplined act of tasting into reliable and objective judgment about what wine is *really* like. These practices are at the juncture of lay and professional domains, yoking the act of tasting wine to institutions of applied and industrial science, advertising, and marketing. Making and marketing wine is different from much else in the modern economy, and the tasting booth is unlike other mid-century industrial laboratories in at least one key respect: Human sensory experience was inescapably at the heart of the exercise. Consequently, judges' diets, physical comfort, and mental states were part and parcel of the experimental apparatus itself (Amerine et al., 1959: 517). Arbitrary and undisciplined judgments of sweetness, of color, or of quality do little if anything in the quest for objectivity. That is why it is so common to say of such judgments that there's no arguing about taste. The sciences involved in wine evaluation provide a way into questions about how lines are drawn between unreliable, colloquial, 'subjective' knowledge and robust, scientific, 'objective' knowledge. Enologists wanted and needed reliable measures of sensory experience, independent of aesthetic preferences or of everyday contexts.

The scientific evaluation of wine was a key concern for Maynard A. Amerine (1911–1998), one of the most influential enologists of the 20th century and the central figure of the University of California, Davis' Department of Viticulture and Enology at mid-century. Amerine was consistently concerned – particularly from the 1940s to the 1970s – with the accurate description and communication of the sensory experience of wine (A&R, 1976: xii). Although he was based in California and focused much of his attention on the improvement of Californian wines, Amerine's techniques and publications gained an international following that eventually made him the sought-after 'eminent elder statesman of American wine' (Pinney, 2012: 192).

Amerine rejected the language of 'subjective evaluation' or 'subjective judgment' of wine in favor of 'sensory evaluation' in order to emphasize that the goal of his program was to move from the purely subjective – 'I like it or dislike it' – to a 'more lasting judgment' based on 'certain objective criteria' (A&R, 1976: 5). There will always be some variation in personal preference with something as complex and variable as wine, but, for Amerine, variation was simply a problem to be solved on the way to a scientific understanding (A&R, 1976: ix, 2). 'The individual's highly subjective opinion as to the desirability (quality) of a certain wine', Amerine wrote, 'should not be confused with the more or less objective assessment of panels of specialized judges' (Amerine et al., 1959: 496). A formal panel setting of specialized judges made 'more or less objective assessment' of wine possible, presuming that 'specialized judges' were available and reliably identifiable. Amerine suggested that wine might be taken as a 'splendid field for the study of aesthetic quality' and that his program could serve as a model for the objective processing of all sorts of sensory judgments (A&R, 1976: 4).

Amerine's process for objective assessment required carefully controlled settings and equipment, as well as statistical techniques for analyzing numerical data.[1] Only through aggregation and collection of taste data did he account for and thereby manage subjective experiences. Individual idiosyncrasies, and variations rendered by experts on different occasions, could be corrected for by the constitution of panels – balancing out residual idiosyncrasies – and by appropriate statistical processing of judgments.[2]

Amerine's 'panels of specialized judges' were not meant to be constituted by either connoisseurs of wine or typical consumers of wine. Amerine wrote very little about where to find judges of wine quality, apparently because they came from a population readily available to him – the students, staff, and faculty at Davis. This meant they began their participation largely 'naïve' to relevant aspects of wine, and became 'specialized' by being specially trained.[3] Judges were screened not only for basic sensitivity but also for honesty, reliability, memory, and attentiveness. The panels would have been much easier to construct had they simply been used to determine consumer or connoisseur preference, and indeed, judging had long occurred at state, national, and international competitions.

Amerine's project, however, was not to measure the 'pure enjoyment' of wines. '[F]or close evaluation of the relative quality with maximum efficiency', Amerine claimed, 'the sensory examination should be conducted under specially prepared conditions' (A&R, 1976: 56). Both aspects – 'specially prepared conditions' and 'maximum efficiency' – were important to the project. The first corresponded roughly to the various laboratory settings and experimental constructions and the second to the use of specific statistical techniques efficiently to make evaluations reliable and robust, that is, to make objective the subjectivities of sensory judgments.

This article is one of a pair; Steven Shapin has written an accompanying paper about Amerine's suggested reforms of 'wine talk', particularly changes associated with the relationships between tasters' descriptive terms and the chemical constituents of wine. Techniques like gas chromatography eventually changed how chemical constituents might be identified, although they didn't much affect Amerine's project in the 1950s and 1960s (Shapin, 2016). In this period, human tasters were required to make relevant determinations about a wine's characteristics and overall quality. My concern here is with how Amerine and his Davis colleagues understood the precision and reliability of individual judgments used to 'objectively assess' wine. These judgments could take many forms: a person might be asked to determine which wine was sweeter, which wine was closer to the ideal type of its variety or region, or which wine was better. All of these judgments were understood to be vulnerable to the corrupting intrusion of psychological and physiological factors. Amerine managed these possible intrusions by creating a setting in which sensory experience could be transcribed into numbers, aggregated, and statistically manipulated. Amerine and his colleagues believed that statistical analysis could reliably determine which evaluations were meaningful and which distinctions were objectively real.

The general outlines of this story – the narrative of how statistics came to be increasingly deployed in the 20th century to manage bias and chance – may seem familiar. Statistical analysis of aggregate data provides one way among many to create a 'view from nowhere/everywhere' or to facilitate 'intersubjectivity'; many of those pushing mathematical analysis into new domains did so with explicit reference to this virtue.[4] Amerine's project was distinct from the way statistics came to be used in agricultural experiments or clinical trials, however, in that the subjective experience – the private sensation of wine tasting – was not only unavoidable but also the object of inquiry. Subjectivity here was not just the 'grit in the knowledge machine', indicating the need for a good mechanical or rational cleansing: it was an element in the machine itself (Shapin, 2012a).[5]

Sensory evaluation as scientific practice

Amerine's best-known publication was his 1976 *Wines: Their Sensory Evaluation*, co-written with a Davis colleague in mathematics and frequent collaborator, Edward B. Roessler (1902–1993). The book was valedictory for both – Roessler having retired from Davis in 1970 and Amerine 4 years later. The book's disciplinary structure mirrored that of its authors' affiliations, divided evenly between a description of sensory evaluation practices and the 'statistical procedures' by which such evaluations could be interpreted and stabilized.

Amerine's fixation on the role of sensory evaluation grew out of his own experience judging wines in California state fairs in the years after Prohibition ended in 1933. He found the assessment and representation of wine characteristics and quality to be a pervasive problem, even as such evaluation was crucial to the development of the wine industry (Amerine and Marsh, 1965; Amerine and Singleton, 1965: 308–309; Pinney, 2012: 181, 186–187). Amerine's work as a judge ran alongside Davis' overall program to evaluate California's vineyards and improve the state's wines. He was a prolific author, writing for both scholarly and popular audiences, in venues ranging from the trade journal *Wines and Vines* to the more consumer-oriented *Wine and Food* (Pinney, 2012: 178). Roessler was also a wine judge for over two decades at California state fairs, and he shared Amerine's interest in making wine evaluation more objective.[6] (Amerine gave up judging at fairs in 1953, since he reckoned that industry people – once a good example had been set – should eventually learn to do it for themselves. He worried, however, that commercial interests were increasingly corrupting judging and that too many top prizes were being awarded.)

In Amerine's vision, judging was at once a report to winemakers on how well they were doing and a signal to consumers about what experts liked and what they, as consumers, might – perhaps *should* – like. How could you establish which wines were good, for example, both free from flaws and true to type? What sorts of sensory characteristics might you look for to establish goodness? Who was competent to make such judgments and on what grounds might they make that competence manifest?

These questions were closely tied to Amerine's larger project to raise the quality of Californian wine. Knowledge of the standards and processes of judgment might help winemakers predict which wines would be made well – wines whose properties would secure Amerine's approval and that might also succeed in the marketplace. As early as 1952, Amerine and Roessler noted that their focus was on 'relatively small panels' for sensory examination, rather than the large-scale 'consumer-acceptance' tests (A&R, 1952: 97). Their concern was not to affect the market directly; it was to affect winemaking by establishing reliable criteria of those characteristics that made for quality. It was impossible to raise standards without an agreed-upon way to evaluate wine.

To design practices of sensory evaluation, Amerine and his Davis colleagues drew on a variety of previous developments in the sciences of taste. Taste panels had been developed in industrial laboratories over the 1930s and 1940s to permit experimental control over the conditions for making flavors and taste measurable (Berenstein, 2016). Ranking scales were designed to measure soldiers' food preferences after the war and were later adopted more widely to provide a numerical basis for turning such preferences into measured judgments (Jones et al., 1955). Likewise, statistical methods for inference testing and measurement of the significance of preference rankings and distributions had been developed over the first half of the century. In 1953, a special issue of *Biometrics* was devoted to the experimental and statistical design of taste tests. By this point, Amerine and Roessler were aware of studies of how to present samples in scoring procedures (Hopkins, 1953), how to select panels for sensory evaluation (Bradley, 1953), how to properly control testing conditions (Mason and Koch, 1953), and how to select the most appropriate statistical tools for given situations (Harrison and Elder, 1950). While few experimental designs fit the standard requirements of statistical tests precisely, considerations of normal distribution, randomness, and independence were agreed to be close enough in practice. In fact, Amerine continued to cite this same research throughout the 1960s and 1970s, suggesting that the main decisions about designing taste panels had been made by the mid-1950s, even if crucial execution and interpretive problems remained.[7]

Experimental control of taste

Amerine believed that most undesired influences upon sensory evaluation could be physically manipulated and controlled through experimental design. This was the logic behind his extensive protocols governing the uniformity and regularity of the tasting booth. His views on experimental design drew on his work as a judge for California state fairs, during which he determined certain sensory stimuli to be 'irrelevant', and thereafter advocated for their elimination from tastings. Samples had to be served at set temperatures and in carefully specified and recorded orders (Amerine and Marsh, 1939). Those with an interest in the outcomes – winemakers or connoisseurs – were not allowed to encroach on the test site, so no special 'leniency' would be applied to appease (or not embarrass) individuals. Over time, Amerine and his colleagues also suggested requirements for the construction of dedicated tasting spaces, including specifications for ambient light, temperature, humidity, odor, and noise. Similarly, equipment had to be standardized so that identical glasses and preparation procedures were used in all settings (Amerine et al., 1959: 515–522; A&R, 1976: 62–67).

Other possible undesirable influences were managed by disciplining bodies. Given that the main apparatus was the human taster, he – it was presumed by those at Davis that the taster would be a 'he' – had to be regulated too. Prospective judges were required to control food and drink consumption before tastings, ideally arriving slightly hungry (hence the recommended late morning period for trials). They must not be rushed, must be in adequate physical and mental shape, and, in particular, have the ability to focus and recall (Amerine et al., 1959: 515–522; A&R, 1976: 49). Such ability and willingness to focus was a crucial human discriminator; sensitivities could be acquired, within limits, but just on the condition of agreeing to, and being able to, pay attention.[8]

The most concerning errors, however, arose from internal faults. Over time, judges would learn not to allow positive judgment on one aspect to affect their next judgment of a particular wine (making each judgment 'independent', an important criterion for statistical inference), although Amerine and Roessler admitted it was 'difficult to avoid this error' (A&R, 1976: 53). Likewise, novice judges could err by systematically avoiding the top and bottom marks in favor of rating everything as close to 'standard'. Such mistakes might help avoid a situation in which a judge would be criticized for marking a wine as 'excellent' when it should have been marked 'poor', but the overuse of middling grades tended to distort the distribution of scores and muddle results. These sorts of faults were real, but awareness of them could enable experienced judges to avoid them.

Amerine and Roessler noted that other psychological tendencies were more insidious and harder to manage. For example, the order in which wines were presented mattered, as did whether they were presented in a particular geometric configuration. Labeling could be suggestive: Marking one sample with the number 9 versus another with the number 4 would apparently affect their evaluation. Eventually, specific, standardized procedures could be deployed, from using randomly chosen two-digit numbers between

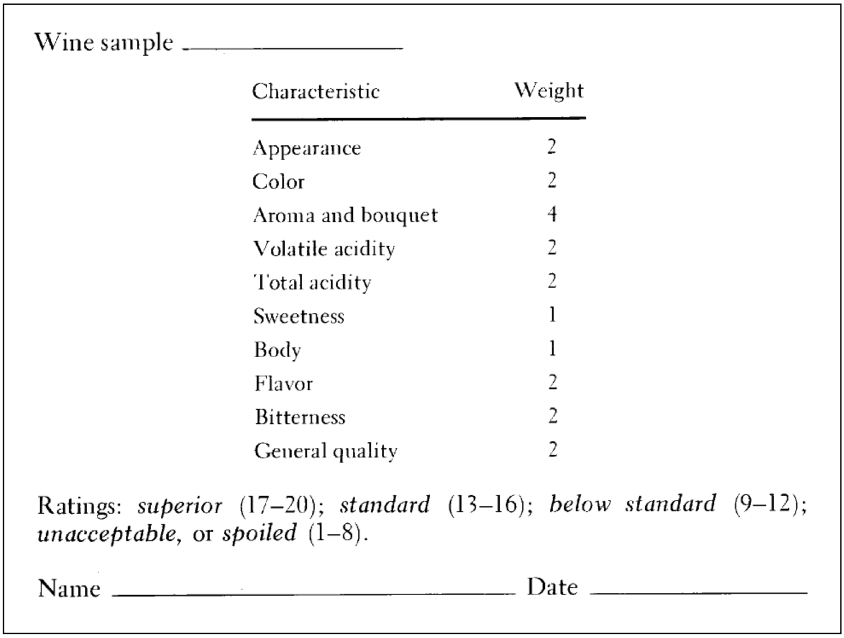
14 and 99 for labels, to carefully recording which wines were presented in which order (A&R, 1976: 62–67).

Judges were set different tasks depending on precisely what was being evaluated. Drawing on research in the broader food industries, Amerine and Roessler described three different tests central to sensory evaluation – the paired-sample test, the duo-trio test, and the triangle sample test. The paired test presented a judge with two samples and simply asked him to identify the one with ‘the greater intensity of a specific constituent or well-defined characteristic’ (A&R, 1976: 111; Pusais and Chabanon, 1974 [1969]). A judge might, for example, be asked to taste two samples and indicate on the scorecard which one was sweeter.

The duo-trio test began with a reference sample and then asked a judge to determine which of two subsequent samples matched the reference. Like the previous test, this drew attention to the judge’s ability to pick out relevant differences, but this time using an anchor point or standard. It was, in effect, a test of the measurement of similarity. (It also allowed similarity to be measured without needing to identify what caused the third sample to be odd: the differences between the odd sample and the reference might be multifaceted.) Likewise, the triangle test presented a judge with three samples, two of which were identical, and required him to pick out the odd sample. This was also a test of similarity, but a more difficult one since the standard was not specified in advance. Amerine and Roessler had been among the first to publish on the use of the triangle test for wine evaluation, promoting it in 1948 as an efficient way to compare wines (Amerine, 1948; Roessler et al., 1948).

A panel of judges might be called upon to compare one wine’s sweetness or color with another, but more typically these tests were used to determine competence itself. It was hardly convenient to spend time and money setting up an elaborate experiment simply to determine whether one wine was sweeter. But it was certainly useful to deploy the tests to evaluate whether a person could reliably make an important distinction. Once an effect was known to exist – that adding tannin changed the taste, for example – the test could be used to train a judge to recognize a difference (a sensed difference Y is labeled as a difference in the level of X) or evaluate whether he has really acquired the ability (if sample 1 is more acidic than sample 2, can he pick it out reliably, without knowing this in advance?). The determination of competence, like the experimental manipulation of tasting booth protocols, was ultimately an attempt to reduce subjectivity – what Amerine and Roessler called the management of ‘physiological factors’ and ‘psychological errors’ (A&R, 1976: 48–54).

Once deemed competent, judges were called upon to quantify a given wine’s overall quality, either by relative ranking or by providing grades on a standardized scorecard. Rankings were straightforward – putting wines in order from best to worst, say, on the basis of some set of criteria. More commonly, though, Amerine and Roessler wanted judges to evaluate wines by scoring specific characteristics and then adding up the values. Amerine, Roessler, and other colleagues at Davis had constructed their own scorecard, which – they claimed – when mastered by ‘highly skilled judges’ could produce results with ‘remarkable precision’ (Amerine and Roessler, 1976: 123). The greatest weight (4 points) in an early scorecard was given to aroma and bouquet, the least (1 point) to each of sweetness and body, and the rest were weighted equally (2 points):



Wine sample ________________

Characteristic	Weight
Appearance	2
Color	2
Aroma and bouquet	4
Volatile acidity	2
Total acidity	2
Sweetness	1
Body	1
Flavor	2
Bitterness	2
General quality	2

Ratings: *superior* (17–20); *standard* (13–16); *below standard* (9–12); *unacceptable*, or *spoiled* (1–8).

Name ______________________________ Date ________________

Figure 1. The Davis scorecard.

appearance, color, volatile acidity, total acidity, flavor, bitterness, and ‘general quality’. (Later, Amerine also promoted a ‘modified’ Davis scorecard, which added the category of astringency, deleted volatile acidity, and increased the weight given to aroma and bouquet (Figure 1).) Each of these scores for individual characteristics would be added to yield an overall score between 1 (‘spoiled’) and 20 (‘superior’).

There is some degree of incoherence in the Davis scorecard. Amerine and Roessler concede that ‘flavor’ – reckoned by many to involve a complex interplay between olfaction, gustation, and other sensory (and even cognitive) modes – was not clearly defined. Moreover, a category of ‘general quality’ within a scorecard meant to provide an overall measure of general quality might be thought strange (A&R, 1976: 125). Acknowledging that the scorecard was one of many then available, Amerine and Roessler surveyed a series of other contemporary scoring systems differing in crucial respects. A Seagram corporation scorecard gave a basis value to certain categories and then allowed upward modifications for positive attributes or downward modifications for defects; the scorecard of the Parisian Office International de la Vigne et du Vin started with 0 and only added points for defects (A&R, 1976: 126–127).

Once completed, a scorecard could serve as a ‘permanent record of the taster’s impression of the wine on a certain date’ (Amerine et al., 1959: 534). The scores were said to ‘stand for’ evaluative judgment and were then fungible, enabling integration or comparison with similar judgments from other scorecards, either from different people in the same setting or the same judge in a different setting. This is the usual benefit of flattening complex entities into numbers – they can then be easily isolated, moved, recombined, and preserved. Two scorecards might be compared to measure differences in a judge's consistency over time, differences among a panel of judges' views of one wine, or differences among judges' views of the wine over time. The scorecard is an inscription protocol that transforms sensory judgment into a measurement.

Scoring judgment

The scoring systems varied, but all were predicated on the belief that specific characteristics could be isolated, judged, and assigned numerical grades. Judgment of a wine's acidity was rendered reliable precisely because a judge's evaluation of acidity had been calibrated using paired, duo-trio, and triangle tests. Indeed, Amerine and Roessler reminded readers that even experienced judges will need to have their scores compared and calibrated, if standards were to be consistently applied (A&R, 1976: 126–127).

Judges on Davis panels were not in the business of deciding what constituted good wine; they were evaluating a given wine using predetermined criteria. That is, Amerine and Roessler believed there were objective differences between wines, and their goal was to establish a reliable procedure for making those distinctions on the basis of individual sensory experiences. If an evaluative test required knowing what counted as a score of 3 for aroma, then judges could be trained with the tests to distinguish a Pinot Noir with a 3 for aroma from one that should receive a 2 or 4. Judges might, of course, disagree about whether a wine should receive a 2 or 4, but ideally all judges would have been determined competent to make reliable judgments about that category.

Each experiment could be interpreted as a test of whether a judge can make determinations that were distinguishable from *chance*. That is, in order to show competence at determining which of two wines was sweeter, a judge had to select one wine at a rate exceeding that of an unskilled judge, who was presumed to pick randomly. A paired-sample test, for example, would have a 'null hypothesis of no difference', with researchers expecting that each wine would be selected about 50 percent of the time just by chance. Any determination about competence would require the calculation of a 'one-tailed' inference test of the alternative hypothesis – that, if one wine is indeed sweeter, then it will be selected significantly more than half of the time, with 'significantly' predetermined by reference to standard statistical tables. If a judge were instead asked to state a preference between the wines, a 'two-tailed' test would be deployed, since his preference might be expressed in either direction. Similar statistical methods applied to the duo-trio and triangle tests.

Almost 100 pages of the 1976 book – a section called 'Statistical Procedures' – are given over to defining and explaining the statistics involved in interpreting results of these tests. This part – about half of the book – resembles, in form and content, an introductory textbook, complete with basic descriptions of probability and statistics, as well as significance tests, suggested exercises and appendices at the end providing standard tables of distributions. Roessler doubtlessly bore primary responsibility for constructing this half of the text, having long encouraged the study of 'agricultural mathematics' – essentially statistics – at Davis (Roessler, 1936; cf. Alder et al., 1993).

Traditional inference testing had the serious drawback that, lacking a large sample size (without a judge being tested on many different occasions or without many judges evaluating a given wine), it was rare that differences could be objectively determined. Even guessing randomly, a judge would pick correctly 50 percent of the time in paired or duo-trio tests and 33 percent of the time in triangle tests. A competent judge needed to exceed that level significantly, and yet, it was not expected that many judges would be perfect or even close to perfect. In most judging situations, it was therefore impractical to take a large enough sample to determine significant differences.

The solution Amerine and Roessler proposed – sequential analysis – was drawn from wartime attempts to make statistical hypothesis testing more efficient for industry. Sequential analysis had been developed in large part by Abraham Wald at Columbia University in the 1940s under a classified Defense Department contract (Hunter, 2004; Statistical Research Group, 1945; Wald, 1947). Along with game theory – about which Wald was familiar, as it was developed by some of his close associates – sequential analysis was a method of making decisions under conditions of limited knowledge. Wald and his colleagues had developed sequential analysis for explicitly economic purposes: it would reduce by half, on average, the sample size required to achieve statistical significance (Giocoli, 2013; Wald, 1947: 1). The *economy* of sequential analysis offered two significant benefits for sensory evaluation. The method reduced the number of expert judges and of trials required, both expensive investments. Second, by halving the average number of trials, sequential analysis helped eliminate the complications arising from repeating trials. Reductions in the number of trials meant not only time and material saved but also intellectual efficiencies, enabling a researcher to hone in on relevant distinctions with fewer possibly confounding variables to worry about.

The default mode of sequential analysis was the ratio test, and its paradigmatic application was in the determination of whether a collection of widgets had a low enough percentage of faulty units to be acceptable. The original example in the 1945 publication introducing sequential analysis was the quality control of ordnance deliveries from military contractors (Statistical Research Group, 1945). Instead of sampling *n* units from the lot and finding the percentage of units in that sample that were faulty, one would test units, randomly drawn, one at a time: {Faulty, Good, Faulty, Good, Good, Good, … }. After every single test, a ratio of 'faulty' to 'good' units thus far drawn would be constructed. There were three options: (1) the ratio could be so large that the entire lot could be rejected (too many faulty items for each good one, so no matter what the rest of the lot looked like, it should be rejected); (2) the ratio could be close enough to 0 that the entire lot could be accepted (enough good items for each faulty one that it could be safely presumed the rest of the lot was fine); or (3) the ratio of faulty to good was neither high enough to reject nor low enough to accept, in which case the testing would continue.

By the 1960s, sequential analysis had moved far beyond its initial application to ordnance testing.[9] The 'stopping rule' aspect of sequential analysis – specifying in advance the conditions under which enough cases had been examined to make an inference – had proven useful, particularly in areas like clinical trials where the expense of treatment or non-treatment was very high (Simon et al. 1975). When deployed in medicine, there was a strong ethical component to avoiding the use of inferior treatments by bringing a trial to a close as soon as possible (Armitage, 1954, 1975: ix). For most applications, statistical techniques like sequential analysis were desirable as methods of minimizing or controlling bias and subjective influence. Doctors may be inclined to think a treatment worked if the first few patients responded well. Sequential analysis provided a way to know precisely how many cases needed to be examined before any such inferences might be deemed reliable.[10]

Amerine and Roessler used applications like sequential analysis to label a prospective judge's sensory judgment as good or faulty. It wasn't a way around subjective judgment but a way of efficiently determining its reliability. After every trial, the ratio of faulty to good judgments could be calculated. The resulting ratio could be so small as to deem the judge acceptable, so large as to reject him, or somewhere in the middle – an indication that the testing should continue. Sensory judgment, like ordnance delivery, was something that could only be observed under imperfect conditions of inherent variability.

How statistics might make the subjective objective

Amerine and Roessler were not optimistic about the overall prospects for training and selecting judges, especially in settings outside of Davis' highly controlled tasting booths. After all, they wrote, 'to qualify a judge unequivocally for the sensory evaluation of wines is probably impossible'. This was partially because the tests of ability were 'not infallible' and also because 'we admit that qualification tests for the selection of judges are usually impracticable and often impolitic'. Without a large and ready supply of judges willing to be tested (and possibly rejected) prior to their participation in wine evaluation, it was easier just to seat a panel of judges without prior formal testing. The promise of objectivity in this scenario came not from turning a single person into an objective sensory machine but from treating a group of people as a source of collectively manufactured objectivity: 'The solution is to use appropriate analytical procedures to determine statistical significance based on the combined assessments of the judges'. This may result in a judgment of no significant difference, but 'it is far better to publish results that are dull but meaningful than results that are interesting but meaningless'. In Amerine and Roessler's formulation, 'statistical significance' indicates 'a significant, objective quality difference', assuming that the judges' tastes are taken as 'conventional'. Moreover, statistical analysis might uncover aberrant judges through the 'anomalous data' they produce, enabling them to be excluded from later trials (A&R, 1976: 60–62). It was not practical to control for all possible idiosyncrasies and subjective biases of an individual judge, but a sufficiently large panel might produce a judgment that could be certified as objective – subject to statistical processing.

Amerine and Roessler were effectively differentiating subjectivity, distinguishing factors that could be eliminated through training or experimental protocols from the necessarily subjective judgments of taste. Statistical analysis was the solution for what remained inescapably subjective. Amerine was well aware, from his own work and that of a growing number of researchers in the taste sciences, that there would always be natural variations in motivation, memory, and, especially, sensitivity. Even a highly trained and experienced judge might rank the same wine differently on different occasions or change his preference from one wine to another simply by having been served them in a different order. Amerine and Roessler said that

> a statistical approach must always be employed in the analysis of sensory data because of the inherent variability of aesthetic judgments … The *only* way to evaluate the results and reach reliable conclusions is by the use of appropriate statistical procedures. (A&R, 1976: 49, 58–59; emphasis in original)

Inherent variability as a mark of subjectivity could be managed only by the statistical analysis of aggregate data.

The claim relied on an analogy between the removal of subjectivity and the control of bias or chance using random sampling. Amerine and Roessler implied that the inherent variability of aesthetic judgment was like the inherent variability of any particular parameter – heights in a given population, survival rates after a particular surgery, and so on. Regardless of how well trained and competent the judges were, there would inevitably be some variation in the scores they gave to two different wines. Nevertheless, the question was whether the panel of judges gave a different enough aggregate score to one wine to distinguish it from another on the basis of some characteristic or well-defined criteria. Using 'naïve' subjects, suitably trained for 'conventionality', rendered the analogy with the 'random' sampling of statistical inference precise. Random sampling requires every item in the population to have an equal chance of being in the sample. Amerine and Roessler's 'naïve' judges – without a long history of wine tasting or a stake in the industry – stood proxy for a randomly selected individual from the population of human wine tasters. In effect, the panel's scores for a particular wine are treated as if they were taken randomly from the set of all possible scores for that wine, meaning that they can be taken as objectively referring to the wine, rather than taken as an artifact of the vagaries of that particular panel of judges.

Statistics were also used to interpret the results of tests involving numerical evaluations on scorecards. When more than two wines are scored or ranked by more than two judges, can one distinguish the wines from each other in ways that point to real differences? Amerine and Roessler again turned to the statistical analysis of aggregate sensory data, equating 'meaningful' and 'real' with measures of statistical significance. They introduce the 'F-statistic', for example, which might compare the actual variance of scores given to three different samples by a panel of judges with the expected variance of scores had all three samples been taken from the same bottle of wine. Even if samples were taken from the same source, researchers would expect some variation in scoring due to the vagaries of chance, experimental design, and the inherent variability of individual sensitivity. A 'significant F-value implies that the evidence is sufficiently strong' to indicate real differences among the scored wines, however. Going into the rather tedious calculations of the various 'sum of squares' required – the one for the error given, for example, as $(\sum_{ij} X_{ij}^2 - \sum_i W_i^2 \; / \; n)$ – enabled them to at least provide the trappings of technical knowledge about the tests, although it was doubtful many readers rushed out to calculate such expressions themselves (A&R, 1976: 136–138). The inclusion of these formulations makes plain that they wanted to underscore the importance of statistical tests for making distinctions.

The statistical tests enabled an experimenter to be less concerned about whether two different judges rank the same wine slightly differently or that one judge varies in his evaluation over time. By looking at differences in the *distributions* of scores, one can calculate whether the same judge is giving distinguishing scores to different wines, as well as whether a group of judges is giving significantly different scores to different wines. The mathematics of statistical distributions is the same, whether a distribution is considered a set of errors from a true value or as irreducible variability in a population. There's always going to be some variability – for taste panel scorecards no less than for astronomical observations. For Amerine and Roessler, the problem of variance in sensory evaluation was rendered a problem of variation in measurement: subjective differences turned into objective ones.

There were similar tests for ranking procedures. When scorecards asked judges to list samples in order rather than provide a score for them, Amerine and Roessler introduced ‘Spearman’s Rank Correlation Coefficient’ as a tool for determining whether rankings pointed to real differences. As with scoring, the presumption was that there will always be variation in the rank order of wines, but some variations are meaningful while others are not. (That is, is the preference ranking C,E,A,D,B substantially different from C,D,A,B,E? Furthermore, even if nothing can be concluded about the group as a whole, can it still be judged that sample C is better than D?) As the authors explain, the ‘ranking of *k* wines by *n* judges is a very common procedure’, so tools like Spearman’s coefficient were required in order to figure out which results were significant and which might have been obtained by chance alone and should be ignored (A&R, 1976: 164).

The ‘Statistical Procedures’ section fits squarely in the mid-century genre of ‘off-the-shelf’ statistics, a genre that cut across a number of fields, from agriculture to psychology.[11] Examples of this genre typically focused on the powerful applications of formulae for making inferences rather than on the various mathematical niceties and technical requirements for the tests, such as randomness in sampling, careful interpretation of confidence intervals, and assumptions about population distributions. In the case of Henry Beecher’s experiments with pain, there was even another contemporary attempt to gain an objective understanding of subjective experience through statistical analysis. Beecher claimed to have found a reliable way to measure pain by the 1960s – and the effects of analgesics on it – through a placebo-controlled double-blind clinical trial, and he enlisted statistician Frederick Mosteller to help interpret the aggregate results statistically (Beecher, 1959, 1966). The Harvard anesthesiologist understood pain as a real entity – just as Amerine understood wine quality as real – but one whose origins and nature were poorly understood and whose effects were accessed primarily through subjective reports. Given these difficulties, Beecher, like Amerine, turned to statistical inferences from aggregate data to produce reliable, objective, results.[12]

Such parallels can be misleading. In the 1950s, practitioners across many fields saw great promise in new statistical techniques, but few had established protocols to draw upon. Beecher rejected the use of trained subjects, for instance, whereas Amerine and Roessler advocated training subjects whenever possible. Although pain investigations were often driven by the desire to achieve objective knowledge useful for the regulation of treatment regimes or approval of pharmaceutical agents, mid-century pain studies remained about people’s own experiences and therefore within the psychological tradition of making internal states measurable through experimental contrivance.[13] Beecher didn’t believe that pain existed outside of a holistic, subjective experience of it. Amerine and Roessler, however, were interested in the ‘object itself’, that is, wine (A&R, 1976: 5–6). If they could have eliminated the human tasters entirely, they would have (Amerine et al., 1959: 559). Whatever their differences in practices and assumptions, both Beecher and Amerine were part of a new wave of statistics enthusiasts, believing that the methods might help decipher complex phenomena – on the condition that it was possible to turn the phenomena into numerical data.

Amerine and Roessler were aware that many readers might just skip the statistical part of their book: Amerine recalled that it 'frightened many people' (Pinney, 2012: 186). And while admitting that one need not read the statistical sections to 'profit' from the book, the authors claimed one would gain 'much more' by reading the whole thing, 'especially if you are (or want to be) a wine professional' (A&R, 1976: x–xi). Aspiring professionals would want to understand the statistical procedures for making inferences from aggregate data, since such inferences were the only way for objective, publicly legible judgments to emerge from subjective and private taste sensations.

In an environment where there were legitimate and complex questions about which characteristics or sensations were real and which were imagined, artificial, or deceitful, the statistical tests were essential. By connecting real differences to 'statistically significant' differences, Amerine and Roessler could conclude that a tasting panel might yield judgments independent of individual subjectivities.

The distinction between judges and judgments was elided: The tests that could be deployed to measure whether judges were reliable were the same as those analyzing whether reliable differences existed in judgments about wines.[14] In practice, both judges and judgments were treated as interchangeable data points. If the same wine was poured into two glasses and then a bit of glycol added to one of them, the paired test could be used to determine whether a prospective judge could pick out the wine with the glycol. By the same logic, consider judges who had already been deemed trustworthy in their individual ability to distinguish astringency. Having a panel of them determine that one wine was more astringent to a given degree of statistical significance meant that the difference between these two wines was considered real. And this would be the case even if it were not known what, if any, chemical or physical distinction between them caused one to be more astringent than the other. If, on the other hand, qualified judges could not reliably distinguish between two wines in a test comparing levels of astringency, then that difference was effectively undetectable – it just didn't exist, as far as sensory evaluation was concerned, and there could be no judgment of preference of one wine over the other on that basis (Amerine et al., 1959: 523). The subjective differences among judges' tastes gained stability and reliability from having been transformed into statistical differences among observational measurements.

This statistical objectivity did not require turning people into machines or establishing consensus through the sharing and coordination of subjective experiences. Here, tasters were presumed independent, with their reliability established ideally through training or by gradual elimination of those whose judgments appeared 'anomalous'. Amerine and Roessler didn't know – or at least gave no published indication of knowing – about novel measures such as Jacob Cohen's κ, which attempted to establish a measure of inter-rater agreement (Cohen, 1960). Moreover, although Amerine used the term 'panel' for the purpose of aggregating judgments from a number of individuals, the judges themselves did *not* typically interact, even if they were evaluating simultaneously. Isolation of judgment was held to be important to the overall program. (This also had a statistical purpose, in that resulting data points could be considered independent.) In this way, although Amerine certainly wanted a standardized terminology to facilitate communication about wine (Shapin, 2016), his program was very different from other examples of taste panels, which depended on discussion and ultimately consensus for the establishment of objective knowledge (e.g. Liberman, 2013). In short, Amerine and Roessler never presumed that it would be possible entirely to calibrate judgments about taste or even to guarantee agreement about a given set of samples. There was simply too much inherent variation. But, it was possible to aggregate values, analyze those values statistically, and – if the circumstances were right – to make judgments about quality and difference that were reliable, stable, and objective.

Numbered knowledge

I have taken the story of scoring wines through the 1970s, with the Amerine and Roessler text representing the culmination of more than 20 years of research on sensory evaluation at Davis. This story, as Amerine understood, had a past, as scoring wines according to a range of their sensory aspects went back at least to the late 19th century. Some scoring systems were oriented, wholly or mainly, to faults and others (and especially so as time went on) to the summation of desired factors. The Davis 20-point system in standard use from the late 1950s possibly developed from early 20th-century French practices (reflecting French scoring systems in schools), but, by Amerine's time, you could pick from systems scoring out of 10, 50, or 100, as well as a 'star-system' building on the Michelin one-to-three-star evaluation of restaurants as well as an Italian system of one-to-three glasses – *bicchieri* (see Shapin, 2012b).

But the most consequential use of numerical wine scoring lay in the future. Amerine and his colleagues did not conceive of their scoring system as something that would or should circulate among ordinary wine consumers. Amerine had a low opinion of most consumers' ability to discriminate, and, while he thought that the objectifying practices of sensory evaluation and its scoring systems might, eventually, trickle down to the marketplace, he did not envisage a world in which consumers themselves would describe wine quality in a numerical idiom.

Yet, the scoring systems that were around in 1976 persist, and numbers have become a *lingua franca* of the wine world. The American Wine Society continues to follow Amerine and the Davis standard in using a 20-point 'wine evaluation chart' whose categories comprise appearance, aroma and bouquet, taste and texture, aftertaste, and an additional two points for 'overall impression'.[15] Some of the systems in current use represent the aggregate opinion of experts; few – outside of academia or sectors of the wine industry – seem to be produced through the statistical manipulations of aggregated evaluations described by Amerine and Roessler. Aspects of the disciplined tasting protocols they describe still remain important, for example, 'blind tasting', the control of the sensory environment, the numbers and orders of wines tasted, and so on (Jackson, 2009: 177–302). Just a few influential wine critics – mostly 'old school' and Old World – resist the whole idea of expressing quality on a numerical scale, but they have clearly lost the battle. Numbers are the norm.

Just a few years after the appearance of the Amerine–Roessler book, the Baltimore lawyer Robert M. Parker, Jr launched his subscription-only *Wine Advocate*, where he began vigorously to campaign against what he saw as arbitrary and probably corrupt systems of wine assessment. Parker fell in with Amerine in trying to find ways, so far as possible, to objectify taste and to discipline the extrinsic factors that tended to bias judgment. Parker opted for a 100-point scale that, he said, offered more flexibility and discriminatory capacity than the 20-point standard (McCoy, 2005). And, while much has been made – both by critics and enthusiasts – of the difference between 100 and 20, in fact the evaluative criteria were much the same between Parker and Davis, with two notable exceptions: that the judgments were rendered here by a single individual and that, understandably, the sorts of super-ripe, high-alcohol, highly oaked wines that Parker tended to anoint with high scores (especially 90 and above) did not equally impress Old World writers.

By the 1990s, Parker was celebrated as the most influential wine critic in the world, as a taste maker and as someone whose judgments not only evaluated wine but influenced the sorts of wines that were made and made available to be tasted and evaluated (Shapin, 2005). Winemakers now put their high Parker scores on their websites and back labels; wine shops now have shelves for ‘Parker 90s’ and ‘shelf-talkers’ advertising high Parker scores. Sommeliers at top restaurants have become used to diners asking for ‘nothing under a 90’, and customers can download a cell phone app that enables them to look up scores before making their selections.

One needn’t take sides in the contests over whether or not numbers should be regarded as a clear measure of quality, nor whether – even if one conceded that quality can be arrayed of a quantitative scale – Parker’s numbers are the right ones. But it is clear that the modern wine world’s fascination with both the appearance of objectivity in aesthetic judgment and the special form of objectivity that is numerical belongs to a lineage that includes the work of Amerine and his colleagues at Davis. The wine world is just one of many domains – courtrooms, emergency rooms, baseball clubs’ front offices – where statistical tools are now routinely deployed, even domains once defined by their irremediable subjectivity and their insulation from mathematics. In these fields, as in the case of Amerine’s work, neither subjectivity nor objectivity takes on an entirely pure, stable, or consistent form. Amerine’s and similar methods of objectifying subjectivities have nevertheless only increased in visibility and importance.

It is also clear that the wine world that Amerine and his colleagues did so much to create has wound up in irony. Once, the objective quantification of wine quality was commended as a way of disciplining those ‘extrinsic’ psychological and cultural factors that would corrupt judgment of wine quality ‘as it really was’. Now, widely distributed marketplace knowledge of the numbers assigned to wines has itself been recognized as one of those extrinsic distorting factors. If you know that the wine in front of you is a 95, it is not easy to taste it and find it mediocre. The taste machine has generated its own grit.

Acknowledgements

The author thanks Steven Shapin for access to his notes from the Maynard Amerine Papers and from his discussions with Professors Hildegard Heymann and Ann C. Noble of the University of California, Davis. Shapin also provided many helpful comments on earlier versions of this article, as did Ardeta Gjikola, audience members at the 2015 History of Science Society annual meeting, Sergio Sismondo, and the anonymous referees for *Social Studies of Science*.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. There has been almost no scholarly attention paid to the history of wine's sensory evaluation outside of Shapin (2012b) and Lahne (2016). A cursory mention is given in Stone (2014), and some chapters of Smith (2007) touch on historical developments. There is a far more extensive popular literature focused on present practices, for example, McQuaid (2015) and Spence and Piqueras-Fiszman (2014).
2. See Amerine and Marsh (1939), Amerine and Roessler (1952: 97, 101), Amerine and Roessler (1976: 7, 54, 58–59, 168), and Amerine et al. (1959: 486). The possible acquisition of reliability through the management of individuals has a long history in scientific practice – and in the relevant historiography – perhaps best represented by now-old debates concerning 17th-century Europe: for example, Daston (1991), Dear (1992), Shapin (1994), and Solomon (1998).
3. Personal communication from Ann C. Noble (Amerine's successor at Davis) to Steven Shapin, 31 July 2015.
4. Among many historical approaches to the question of objectivity, see especially the 1992 'Symposium on the Social History of Objectivity' in *Social Studies of Science*, especially Daston (1992) and Porter (1992); on historical forms of objectivity, see Daston and Galison (2007); on the role of numbers, see Porter (1995) as well as Gigerenzer et al. (1989); and for more recent analyses of the aggregation of statistics as a mode of managing subjectivity, see especially Bouk (2015), Igo (2008) and Porter (2011).
5. On subjectivity, taste and knowledge making more broadly, see McCormick (2009), Lamont (2009), Benzecry (2011), Roosth (2013) and Lemov (2015); and for the philosophical concerns involved in wine tasting, see Smith (2007).
6. *Maynard A. Amerine: The University of California and the State's Wine Industry*, an interview conducted by Ruth Teiser, 1972. Regional Oral History Project, pp. 39, 77, 81. Filed at University of California, Davis (UCD) Special Collections as rLD/781/D519/A45; 'Wine Day, California State Fair, Sacramento, 4 September 1951', UCD Special Collections, Amerine Papers D-060, Box MSS (1), 4-page typescript.
7. Even as late as 1976, Amerine and Roessler were still citing and following exactly Bradley (1953) on the statistics of sensory testing (A&R, 1976: 117). Bradley's work in turn relied on that of his student, Lombardi (1951).
8. Antoine Hennion has focused on the conditions by which bodies are disciplined for attention to tasting. See Hennion and Teil (2004) and Hennion (2001, 2005, 2007).
9. Jackson (1960) lists over 350 articles published on sequential analysis in the 15 years after its creation.
10. Some statisticians quickly realized, however, that sequential analysis techniques needed to be interpreted using Bayesian analysis to avoid bizarre situations when the result of a trial depended more on the 'stopping rule' than the underlying data (Cornfield, 1966).
11. Much of this genre was derived from Fisher's (1935) publications, especially *The Design of Experiments*, which presented the general outlines of inferential statistics for researchers; as late as 1976, Amerine and Roessler were still relying on a revised edition of Fisher's statistical tables (Fisher and Yates, 1974).
12. Beecher was writing against the use of apparatuses to induce 'experimental' pain on trained subjects, in favor of an understanding of pain as a clinical entity inseparable from

the individual experience of it. On Beecher's intervention and philosophy, see the work of Tousignant (2006, 2011); more generally on therapeutic reformers in this period, see Marks (1997).

13. For Beecher and the history of experimental apparatuses for pain analysis, see Tousignant (2006, 2011). On pharmaceutical testing as an example of regulatory objectivity, see Cambrosio et al. (2006, 2009).
14. The same mathematical test could be used to analyze the data derived from three different judges ranking 10 wines on 1 day as for one judge ranking 10 wines on three different days. The determination of a judge's consistency over time was procedurally identical to the determination of sufficient agreement among different judges.
15. American Wine Society scoring documents are available at http://c.ymcdn.com/sites/www.americanwinesociety.org/resource/resmgr/imported/AWS%20Wine%20Evaluation%20chart%202012.pdf (accessed 13 August 2015) and http://americanwinesociety.site-ym.com/?page=E3 (accessed 13 August 2015).

References

Alder HL, Krener AJ and Shepard L (1993) Edward B. Roessler, Mathematics: Davis. In: Krogh D (ed.) *University of California: In Memoriam, 1993*. Berkeley, CA: University of California Press, 153–154. Available at: http://texts.cdlib.org/view?docId=hb0h4n99rb (accessed 29 July 2015).

Amerine M (1948) An application of 'triangular' taste testing to wines. *The Wine Review* 16(5): 10–12.

Amerine MA and Marsh GL (1939) Wine judging methods at the 1939 fairs. *The Wine Review* 6: 20.

Amerine MA and Roessler EB (1952) Techniques and problems in the organoleptic examination of wines. *American Journal of Enology and Viticulture* 3(1): 97–115.

Amerine MA and Roessler EB (1976) *Wines: Their Sensory Evaluation*. San Francisco, CA: W.H. Freeman.

Amerine MA and Singleton VL (1965) *Wine: An Introduction for Americans*. Berkeley, CA: University of California Press.

Amerine MA, Roessler EB and Filipello F (1959) Modern sensory methods of evaluating wine. *Hilgardia* 28(18): 477–567.

Armitage P (1954) Sequential tests in prophylactic and therapeutic trials. *Quarterly Journal of Medicine* 23: 255–274.

Armitage P (1975) *Sequential Medical Trials*, 2nd edn. New York: John Wiley & Sons.

Beecher HK (1959) *Measurement of Subjective Responses: Quantitative Effects of Drugs*. New York: Oxford.

Beecher HK (1966) Pain: One mystery solved. *Science* 151(3712): 840–841.

Benzecry CE (2011) *The Opera Fanatic: Ethnography of an Obsession*. Chicago, IL: University of Chicago Press.

Berenstein N (2016) *Flavor Added: Synthetic Flavors and Flavor Science in the United States, 1870–1970*. PhD Thesis, University of Pennsylvania, Philadelphia, PA.

Bouk D (2015) *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual*. Chicago, IL: University of Chicago Press.

Bradley RA (1953) Some statistical methods in taste testing and quality evaluation. *Biometrics* 9(1): 22–38.

Cambrosio A, Keating P, Schlich T, et al. (2006) Regulatory objectivity and the generation and management of evidence in medicine. *Social Science & Medicine* 63: 189–199.

Cambrosio A, Keating P, Schlich T, et al. (2009) Biomedical conventions and regulatory objectivity: A few introductory remarks. *Social Studies of Science* 39(5): 651–664.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46.

Cornfield J (1966) Sequential trials, sequential analysis and the likelihood principle. *The American Statistician* 20(2): 18–23.

Daston L (1991) Baconian facts, academic civility and the prehistory of objectivity. *Annals of Scholarship* 8: 337–363.

Daston L (1992) Objectivity and the escape from perspective. *Social Studies of Science* 22(4): 597–618.

Daston L and Galison P (2007) *Objectivity*. New York: Zone.

Dear P (1992) From truth to disinterestedness in the seventeenth century. *Social Studies of Science* 22(4): 619–631.

Fisher RA (1935) *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fisher RA and Yates F (1974) *Statistical Tables for Biological, Agricultural and Medical Research*, 6th edn. London: Longman.

Gigerenzer G, Swijtink Z, Porter TM, et al. (1989) *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge: Cambridge University Press.

Giocoli N (2013) From Wald to Savage: *Homo Economicus* becomes a Bayesian statistician. *Journal of the History of the Behavioral Sciences* 49(1): 63–95.

Harrison S and Elder LW (1950) Some applications of statistics to laboratory taste testing. *Food Technology* 4: 434–439.

Hennion A (2001) Music lovers: Taste as performance. *Theory, Culture and Society* 18(5): 1–22.

Hennion A (2005) Pragmatics of taste. In: Jacobs MD and Hanrahan NW (eds) *The Blackwell Companion to the Sociology of Culture*. Oxford: Blackwell, pp. 131–144.

Hennion A (2007) Those things that hold us together: Taste and sociology. *Cultural Sociology* 1: 97–114.

Hennion A and Teil G (2004) Le goût du vin: Pour une sociologie de l'attention. In: Nahoum-Grappe V and Vincent O (eds) *Le Goût des Belles Choses*. Paris: Éditions de la Maison des Sciences de l'Homme, pp. 111–126.

Hopkins JW (1953) Laboratory flavor scoring: Two experiments in incomplete blocks. *Biometrics* 9(1): 1–21.

Hunter PW (2004) Connections, context, and community: Abraham Wald and the sequential probability ratio test. *The Mathematical Intelligencer* 26(1): 25–33.

Igo SE (2008) *The Averaged American: Surveys, Citizens, and the Making of a Mass Public*. Cambridge, MA: Harvard University Press.

Jackson JE (1960) Bibliography on sequential analysis. *Journal of the American Statistical Association* 55(291): 561–580.

Jackson RS (2009) *Wine Tasting: A Professional Handbook*, 2nd edn. Amsterdam: Academic Press.

Jones LV, Peryam DR and Thurstone LL (1955) Development of a scale for measuring soldiers' food preferences. *Journal of Food Science* 20(5): 512–520.

Lahne J (2016) Sensory science, the food industry and the objectification of taste. *Anthropology of Food* 10. Available at: http://aof.revues.org/7956

Lamont M (2009) *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

Lemov R (2015) *Database of Dreams: The Long Quest to Catalog Humanity*. New Haven, CT: Yale University Press.

Liberman K (2013) The phenomenology of coffee tasting: Lessons in practical objectivity. In: *More Studies in Ethnomethodology*. Albany, NY: State University of New York Press, pp. 215–266.

Lombardi GJ (1951) *The Sequential Selection of Judges for Organoleptic Testing*. MS Thesis, Virginia Polytechnic Institute, Blacksburg, VA.

McCormick L (2009) Higher, faster, louder: Representations of the International Music Competition. *Cultural Sociology* 3(1): 5–30.

McCoy E (2005) *The Emperor of Wine: The Rise of Robert M. Parker, Jr. and the Reign of American Taste*. New York: Ecco.

McQuaid J (2015) *Tasty: The Art and Science of What We Eat*. New York: Scribner.

Marks HM (1997) *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990*. Cambridge: Cambridge University Press.

Mason DD and Koch EJ (1953) Problems in the design and statistical analysis of taste tests. *Biometrics* 9(1): 39–46.

Pinney T (2012) *The Makers of American Wine: A Record of Two Hundred Years*. Berkeley, CA: University of California Press.

Porter TM (1992) Quantification and the accounting ideal in science. *Social Studies of Science* 22(4): 633–651.

Porter TM (1995) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Porter TM (2011) Statistics and the career of public reason. In: Crook T and O'Hara G (eds) *Statistics and the Public Sphere: Numbers and the People in Modern Britain, c. 1800–2000*. New York: Routledge, pp. 32–47.

Pusais J and Chabanon RL (1974 [1969]) *Initiation into the Art of Wine Tasting* (trans. JL Vaccaro). Madison, WI: Interpublish.

Roessler EB (1936) *Problems in Agricultural Mathematics*. Ann Arbor, MI: Edward Bros.

Roessler EB, Warren J and Guymon JF (1948) Significance in triangular taste tests. *Food Research* 13: 503–505.

Roosth S (2013) Of foams and formalisms: Scientific expertise and craft practice in molecular gastronomy. *American Anthropologist* 115(1): 4–16.

Shapin S (1994) *A Social History of Truth: Civility and Science in Seventeenth-Century England*. Chicago, IL: University of Chicago Press.

Shapin S (2005) Hedonistic fruit bombs. *London Review of Books* 27(3): 30–32.

Shapin S (2012a) The sciences of subjectivity. *Social Studies of Science* 42(2): 170–184.

Shapin S (2012b) The tastes of wine: Towards a cultural history. *Rivista di Estetica n.s.* 51: 49–94.

Shapin S (2016) A taste of science: Making the subjective objective in the California wine world. *Social Studies of Science* 46: 436–460.

Simon R, Weiss GH and Hoel DG (1975) Sequential analysis of binomial clinical trials. *Biometrika* 62(1): 195–200.

Smith BC (ed.) (2007) *Questions of Taste: The Philosophy of Wine*. Oxford: Oxford University Press.

Solomon JR (1998) *Objectivity in the Making: Francis Bacon and the Politics of Inquiry*. Baltimore, MD: Johns Hopkins University Press.

Spence C and Piqueras-Fiszman B (2014) *The Perfect Meal: The Multisensory Science of Food and Dining*. Chichester: John Wiley & Sons.

Statistical Research Group (1945) *Sequential Analysis in Inspection and Experimentation*. Report No. 255, AMP Report 30.2R. New York: Columbia University Press.

Stone H (2014) Sensory evaluation in the *Journal of Food Science* – 1936 to the Present. *Journal of Food Science* 79(1): iii.

Tousignant NR (2006) *Pain and the Pursuit of Objectivity: Pain-Measurement Technologies in the United States, c. 1890–1975*. PhD Thesis, McGill University, Montréal, QC, Canada.

Tousignant NR (2011) The rise and fall of the dolorimeter: Pain, analgesics and the management of subjectivity in mid-twentieth-century United States. *Journal of the History of Medicine and Allied Sciences* 66(2): 145–179.

Wald A (1947) *Sequential Analysis*. New York: John Wiley & Sons.

Author biography

Christopher J Phillips is an Assistant Professor of History at Carnegie Mellon University, where he teaches American history and history of science. He is the author of *The New Math: A Political History* (University of Chicago Press, 2015) and is working on a history of the spread of statistical methods at mid-century.