Sensory Measurements: Coordination and Standardization

Ann-Sophie Barwich [1,*]

Email ab4221@columbia.edu

Hasok Chang [2]

Email hc372@cam.ac.uk

[1] Center for Science and Society, Columbia University, New York, NY, USA

[2] Department of History and Philosophy of Science, University of Cambridge, Cambridge, UK

Abstract

Do sensory measurements deserve the label of "measurement"? We argue that they do. They fit with an epistemological view of measurement held in current philosophy of science, and they face the same kinds of epistemological challenges as physical measurements do: the problem of coordination and the problem of standardization. These problems are addressed through the process of "epistemic iteration", for all measurements. We also argue for distinguishing the problem of standardization from the problem of coordination. To exemplify our claims, we draw on olfactory performance tests, especially studies linking olfactory decline to neurodegenerative disorders.

AQ1

Keywords

Circularity

~~Coordination~~

Epistemic iteration

~~Measurement~~
Olfaction
Psychophysics
Reliability
Sensory perception
~~Standardization~~

Introduction: The Challenge of Sensory Measurements

Inhale deeply. Do you smell anything? Would you have smelled something if we had not directed your attention to it? And if you detected something, can you name it? Have you smelled that odor before, perhaps, and does it help you identify it if we show you the source of the smell, or give you a list of words to describe it? All these questions point to different aspects of your perceptive experience, the objects of your perception, and the stages of the physiological perception process. Can such qualitative features of experience, and of reality, be quantified in a meaningful way? Can sensations be objects of measurement? The apparent divide between qualitative phenomena and quantitative research practices, as discussed from various perspectives throughout this thematic issue, has been a central motif in the development of sensory studies. Smell, for example, may seem like the ultimate qualitative property, but there have been credible attempts to quantify it. By looking at how different olfactory performance tests relate to sensory qualities and their physiological basis, we propose to analyze how sensory measurement deserves the label of measurement in terms of its epistemic nature.

Sensory measurements are notoriously difficult. They encompass many complex and changeable parameters, and controlled measurements under laboratory conditions may not reflect ordinary conditions of perception. Because of their causal multidimensionality and limited degree of control, sensory measurements are rarely considered a decisive factor in settling theoretical disputes about perception and perceptual mechanisms.[1] Data from performance studies are not considered strong evidence and are often approached with suspicion, especially when significant claims are made on their basis. Too much seems to depend on

methodological decisions and differences in measurement setups. Sensory measurements may seem like the scientific counterpart to the mythical creature of the Lernaean Hydra: once one problem is solved, three more require consideration.

Nonetheless, there are good reasons for trying to make and use sensory measurements, even in relation to experiences such as smell that are apparently so qualitative and unmeasurable. For example, olfactory performance studies are increasingly useful for medical and neuroscientific research. The deterioration of olfactory performance has been systematically linked to a range of neurodegenerative disorders such as Alzheimer's and Parkinson's, and also to epilepsy, multiple sclerosis, and depression. Using such indicators requires good measurement: how can you tell whether you are losing your sense of smell, and to what degree? Does it matter what kinds of odor you become unable to perceive, and the order or timeframe in which the decline occurs?

These are important problems, and they are problems of measurement. Yet, when philosophers discuss measurement they almost exclusively focus on the measurement of physical quantities, or at most the impersonal quantities that occur in the social sciences. When philosophers discuss sensory experience they primarily think of matters related to consciousness and cognition. They rarely speak of epistemological concerns such as the question of what sensory measurements are (and what they are not), or whether sensory measurements deserve the label of measurement. Even though questions about the nature of sensory measurement have occupied practitioners of psychology and other scientific disciplines throughout the past two centuries, this discourse has become forgotten in philosophical and historical studies of science.[2] We wish to call for renewed philosophical attention to sensory measurements, and to the epistemological problems they face, which are in fact notably similar to those encountered in the measurement of physical quantities.

Before we go on, let us be clearer about what "sensory measurement" means. As part of psychophysics, sensory measurements aim at understanding the relationship between physical stimuli and perceptual experiences. However, this description is ambiguous as to what exactly is measured and how. Psychophysical measurements

are based on behavioral responses of perceiving test subjects. These responses are interpreted as giving some information either about the perceiver's experience of a stimulus or about the workings of the perception system. Sensory measurements, thus, appear to measure two different things (cognitive and physiological processes) by a third (behavioral, including verbal, responses). Behavioral responses tracked in sensory measurements are quantified; they are evaluated statistically by grouping human test subjects into sensory panels based on relevant performance-affecting factors such as gender, age, training, ethnicity, and habits (e.g., smoking). The reliability and justification of inferences to cognitive processes through quantified behavioral responses has been subject to great debate ever since the birth of psychophysical measurement.

Worries about the measurability of cognitive processes persist in recent literature. Measurement of cognitive categories appears troublesome because it is not obvious how to assess subjective or ambiguous categories through empirical operations. Eran Tal says:

> The philosophical study of *standardization* is still in its nascent stages, and much work lies ahead. To provide a single example, the standardization of subjective measures of psychometric constructs currently poses new challenges. Attempts to validate questionnaires for measuring subjective well-being and quality of life raise something similar to the problem of *coordination*: are the questionnaires measuring what they should? Should the construct be defined in terms of the best-correlated questionnaires? *It is doubtful whether these questions can be answered through a process of iterative stabilization similar to the one encountered in the standardization of physical quantities*. (Tal 2013, p. 1163; emphasis added)

Tal's concern resonates with ~~John~~Joel Michell's (1999) criticism that the mere design of tests does not answer the principal question of whether our operations provide meaningful categories of measurement, i.e., whether they in fact measure something and whether they measure what they claim to measure (or produce more

or less arbitrary data instead). For this reason, psychophysical and psychometric measurements[3] appear to be riddled with an inextricable entanglement of theoretical constructs (of cognitive processes), arbitrarily assigned operations (of measurement tests) and empirical data (of statistical evaluations of tests). There appears to be no reliable grounds to ensure objectivity and validity.

Sensory measurements present a great challenge to scientific practice. You have to start from somewhere, fixing a first point of reference, and then you gradually test, compare, and adjust from there. Again, and again, and again. But how can you fix this initial point of reference? How do you compare your measurements and test your criteria of measurement without entering into circular reasoning? Sensory measurements seem mired in methodological difficulties.

Though we would not deny the difficulties of sensory measurements, we want to argue that these difficulties are not unique to sensory measurements. There is the admitted complication that human subjects can serve as both the object and the instrument in sensory measurements, but the epistemological issues that plague sensory measurements are the same as those in which philosophers of science have become increasingly interested when analyzing any kind of measurement as scientific practice. In any process of quantifying a quality, even in the physical sciences, there are difficult issues to be dealt with. In the remainder of the article we will identify two different problems that any successful sensory measurement must solve: the problem of coordination, and the problem of standardization. These problems are mentioned by Tal almost interchangeably, but we think there is some benefit in keeping them more clearly distinct. We will argue that in each case both the problem and the solution are similar to what we see in physical measurements, drawing especially from our previous work on the case of temperature measurement.

The Case of Olfactory Measurement

Before tackling philosophical concerns about sensory measurement systematically, let us introduce some issues concretely through the consideration of examples in the field of olfaction. Olfaction turned from an obscure footnote in the history of sensory studies into a modern model system in the past couple of decades

(Firestein 2001). Earlier, difficulties of measurement such as the control and evaluation of stimuli had greatly hampered the study of smell perception (Wise et al. 2000; Sell 2005). The widely held opinion of smell as a sense in decline and unimportant in human evolution had further contributed to its unfortunate long-standing neglect (Darwin 1871, p. 24; Pinker 1997, p. 183).[4] Recent research in neuroscience, anthropology, and cognitive science, however, has started to extend our knowledge about olfaction and set out to correct misconceptions about its role in human perception and cognition (Shepherd 2004, 2012; Majid and Burenhult 2014; Wnuk and Majid 2014; Majid 2015).

Olfaction is significant in practical contexts such as fragrance chemistry, the food industry, and clinical tests, as well as in academic research in neuroscience and psychology. There are very interesting issues about how all the different smells that humans experience should be identified and classified. For instance, the so-called "trillion odors paper" in *Science* last year (Bushdid et al. 2014) received wide attention in scientific and popular media (Morrison 2014). In this paper it was suggested that humans detect as many as one trillion odors. Within that same year, the claim was rebutted as a gross overestimation based on an unjustified calculation for the discrimination of sensory space(s) (Meister 2014, 2015). ~~Indeed, an even more recent response~~A further response by Magnasco et al. (2015) to both Meister and to another critical paper by Gerkin and Castro (2015) was posted on bioRxiv on July 6, 2015, opening up further methodological discussion of how to model perceptual dimensions in olfaction.

What we want to focus on in this article are the measurements of *olfactory performance*: how good is someone at smelling? An important application of olfactory performance tests is in medical diagnostics. Olfactory decline is a first symptom and a potential diagnostic tool for the preclinical[5] detection of major neurodegenerative disorders, as mentioned above. Differences in the course of hyposmias (reduced ability to smell) may also aid in differentiating disorders with similar clinical symptoms such as Parkinson's and Alzheimer's (Hawkes et al. 1999; Doty 2013). A range of measurement techniques and tests are available, depending on what exactly one wants to measure and how. Among the various olfactory test kits, the most prominent are: the University of Pennsylvania Smell

Identification Kit (UPSIT) largely used in North America (Doty et al. 1984), and the Sniffin' Sticks predominantly used in central Europe (Hummel et al. 1997). Further tests and modifications of tests are in development; some try to factor in physiological and cultural differences between human test subjects, leading to the development of test kits for specific geographic regions, for example South Korea (Cho et al. 2009).[6]

But the measurement of odor perceptions in humans is distinctly difficult (Wise et al. ~~2001~~2000). A basic requirement for comparing perceptive abilities between healthy and ill test subjects is the design of standardized identification sets of test odors, and a reliable way to assess sensory performance in identifying and discriminating selected odor qualities. Little agreement exists in how to best characterize odor quality and how to deal with the apparent impossibility of eliminating bias in performance tests (Sell 2005).

AQ2

The fundamental issue for the practitioners is to figure out whether theoretical concepts of sensory perception are successfully correlated with observational methods. Their methodological concerns reflect deeper epistemological issues. First, how can we link changes in qualitative sensory experiences with stages of the perceptual mechanism and developments of neurological disorders? In general terms, this is a problem of *coordination*. Second, given the inherent variability of sensory experiences, how can we compare different tests and results? This is a problem of *standardization*. The lack of established measurement methods in both the profiling of odor quality and the assessment of human sensory performances poses a severe handicap to neurobiological research for exploring how changes in sensory experiences may indicate and distinguish neurological disorders. In fact, an extensive study of olfactory loss found that the choice of olfactory test kits (UPSIT or Stiffin' Sticks) affects research outcomes (Lötsch et al. 2008).

It is instructive to take a closer look at the design of the test kits, which reveals a major difference. UPSIT is designed to be quickly self-administered and consists of four paper booklets, each containing ten test substances that are compared by scratching and sniffing. In comparison, the Sniffin' Sticks test kit, using felt pens,

consists of three subtests, each measuring a different aspect of olfactory function. These three aspects are odor *threshold* (i.e., the perception of odors at low concentrations), odor *discrimination* (i.e., the nonverbal distinction of different smells), and odor *identification* (i.e., the ability to name and/or associate an odor). While the Sniffin' Sticks test methodologically divides olfactory performance into three measurable components, UPSIT merges them into one.

This poses an interesting question. With these two tests, are we measuring the same thing differently (Sniffin' Sticks providing a more detailed and elaborate way than UPSIT), or measuring different things? There seems to be more than a pragmatic choice involved here. The two tests are based on different views of olfactory processing and its physiological basis. UPSIT is based on the assumption that the three components of olfactory performance distinguished by the Sniffin' Sticks have a common source of variance and can be merged into one olfactory-performance measure. But there are sufficient grounds to doubt the assumption of a common source of variance. For instance, differences have been observed in the decline of these three olfactory faculties in studies on test subjects with focal brain excision. These subjects showed impaired identification abilities but unchanged threshold performance (Jones-Gotman and Zatorre 1988).

What we perceive when we smell is linked to various stages of olfactory processing in the nervous system. Each stage of the olfactory pathway constitutes a specific circuit in which certain aspects of odor perception are generated. Identifying what part of the pathway is responsible for what part of the olfactory impression is crucial to understanding how olfactory decline can be correlated with specific neurophysiologic disorders. What mechanisms are responsible for specific olfactory impressions and functions being realized (or hindered)? For instance, which odoriferous molecules are detected depends on physiological factors such as receptor expression in the nasal epithelium. How is the olfactory information of the receptors collected and distributed in the bulb (a multilayered neural structure situated at the frontal lobe of the brain) or, as Richard Axel put it: "How does the brain know what the nose is smelling?" (Axel 2005, p. 236).[7]

When we think of smell as having multiple sensory dimensions (i.e., threshold,

quality, intensity, hedonic tone, etc.) that correlate with stages in the perception process, each of those stages gives insight into differences in olfactory function that can be traced and measured (Barwich 2014). The identification between aspects of olfactory experience with stages of the physiological process of olfaction then allows us to pose the question of olfactory decline in operational terms: when is something (not) detected, identified, and further discriminated, and to what extent? Recall, for instance, the question of whether odor threshold, identification, or discrimination is affected. Here these distinctions relate to differences (normal and pathological) in how odors are detected (receptors), odor impressions are identified (glomeruli, microcircuits),[8] discriminated (glomeruli, mitral cells, and microcircuits), and further associated with cognitive processes such as memory (cortex) (Shepherd 2012, p. 67).[9]

Analyzing what happens when, for how long, and in what stage of olfactory processing helps us understand the difference between normal and abnormal variance in olfactory performances between test subjects. Is the measured non-perception of particular odors, for instance, grounded in normal individual variations of receptor expression patterns, or due to pathological degenerations in the glomerular layer?

From this perspective, having multiple ~~operations, i.e.,~~ measurement tests, is not a hindrance to sensory measurement and is further endorsed by practitioners (Tourbier and Doty 2007). Pluralist measures of olfactory performances are a useful starting point for tackling the coordination of sensory responses with possible physiological processes. Understanding what each test is modeled to measure aids in a comparison and mutual adjustment of different tests. This is conducive to an iterative process of knowledge production and correction—as well as more reliable diagnosis of health issues underlying human test subject performances.

For instance, it can be unclear whether an inability to smell specific odors is because of partial anosmia[10] due to a lack of appropriate receptors expressed, or because of a possible neurodegenerative disorder of the frontal lobe. In tackling this problem, one suggestion is to compare results of the UPSIT (measuring

whether certain odors are or are not detected) with results from the so-called “sniff magnitude test”. The sniff magnitude test does not rely on reports that require cognitive processing of verbal tasks like the UPSIT.[11] Instead it measures whether the sniff rate remains stable or increases/decreases with exposure to certain odors such as malodors. Positive sniff magnitude responses to odors that are not recorded through UPSIT help to exclude anosmia as a cause of non-perception (Frank et al. 2003).

The Problem of Coordination

Having introduced some difficulties and the potential usefulness of sensory measurements through the example of olfactory performance tests, we now want to return to general philosophical considerations of sensory measurements. As noted in previous sections, there are two main epistemological problems to be solved in sensory (or other) measurements, namely the problem of coordination and the problem of standardization. In addressing the problem of coordination, it will be instructive to see how basic physical measurements are subject to it, too. It is usually considered that the virtues of physical measurements derive from the use of reliable measuring instruments. The design of such instruments, however, is based on an inevitable form of circularity. In our previous work, we have encountered such a problem most strikingly in the case of temperature measurement.

Temperature, as a quantity, cannot be measured directly. As is generally the case, our evaluation of why an instrument (thermometer) measures a phenomenon (temperature) correctly is fundamentally based on our theoretical understanding of what this phenomenon is (and does). When we have made a thermometer, how do we know that its readings really indicate temperature, as opposed to something else? How do you test whether thermometric liquids expand uniformly with the increase of real temperature, without already possessing an instrument (i.e., a thermometer) capable of giving an ~~indication of whether liquids expand in such a way when heated~~indication of real temperature? We have called this the “problem of nomic measurement”, nomic measurement meaning a method of measurement that relies on an empirical law, in this case linking the real temperature and the

volume of the thermometric liquid (Chang 2004, Chap. 2). Such measurement can only be correct if the law it relies on is correct; however, the correctness of the law cannot be tested and confirmed without already having a means of measuring the variable in question—where do we get the values of real temperature, without a trusted thermometer? So we are caught in a tight logical circle. Solving the problem of coordination would require the justification of the measurement law, and it seems that the only way to have such justification would be to turn the measurement law into a tautology, by defining the quantity to be measured in terms of the measurement law itself.

Pragmatic responses by scientists to this fundamental problem of measurement have been analyzed as a continually progressing development of coordination between theoretical concepts and their measurement (Johansson 2014). In the case of temperature, such progressive coordination took place over a long period of time. For example, by the late 18th century it was a very common practice, yet not entirely justified, to trust the expansion of mercury in a thermometer to indicate "real temperature". Thermal physics advanced on this provisional basis, using the mercury thermometer to collect a great deal of empirical data. With the accumulation of empirical knowledge, as well as some pertinent theoretical developments, the idea took hold that the behavior of gases exhibited true correlation with real temperature; this resulted in the adoption of gas thermometers as the primary standards, which were then used to correct mercury thermometers. Further investigations on the thermal physics of gases led to the development of classical thermodynamic theory, by means of which William Thomson (later Lord Kelvin) defined his concept of absolute temperature around 1850; by the time Kelvin's concept was theoretically and operationally established, gas-thermometer readings were deemed to be only approximately coordinated with absolute temperature, which was considered to be the real temperature. (Chang 2004, Chap. 2, Chap. 4)

Characterized as "epistemic iteration", such progressive development comprises processes of knowledge production "in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals… [T]he whole chain exhibits innovative

progress within a continuous tradition" (Chang 2004, p. 226). Although successive stages do not follow logically from their preceding ones, their manifestation is nevertheless informed by and grounded in previously established techniques and concepts. Consisting of more than mere repetition of central concepts, the process of epistemic iteration is progressive insofar as it reflects the efforts by which researchers revisit their knowledge claims and which, ~~within~~in successive stages of scientific development, lead to their improvement. This improvement is not exclusively empirical but also comprises other epistemic advances, such as ~~accuracy~~simplicity, unifying and explanatory power, consistency, scope, etc. (Elliott 2012). On this account, a successful research strategy need not be one that produces robust empirical data; alternatively, it can be one that fosters question-driven investigation, data collection, and technology-oriented research.

Now we wish to show that sensory measurement, too, deals with the same epistemological problem (circularity), and solves it through the same kind of process (epistemic iteration). Sensory measurement might deal with phenomena of a completely different nature than physical quantities, but it is still a form of measurement. Psychologists have been aware of this issue for a long time. Modeled on the scientific ideal of physics, the question of whether psychophysical methods deserve the status of measurement occupied a large proportion of the reports of the 1932 and 1940 meetings of the British Association for the Advancement of Science (~~BAAS~~BAAS 1932; Ferguson et al. 1940), where a joint panel between the psychology and the physics sections was held in order to investigate "the quantitative relation of physical stimuli and sensory events". The problem of coordination was noted clearly: how to determine the "correlation of the sense experience of the individual with the processes and patterns investigated by the physicist" (Drever in BAAS 1932, Section A. Mathematical and Physical Sciences, pp. 300–301).

S. S. Stevens later summed up the situation as follows:

> An orderly and systematic account of sensory communication must include a delineation of *what* is perceived as well as an explanation of *how* perception is accomplished. In this sense, psychophysics

> defines the challenge: it tells what the organism can do and asks those who are inspired by such mysteries to try, with scalpel, electrode, and test tube, to advance our understanding of how such wonders are performed. (Stevens 1961, p. 1)

Adopting No indentation here. such a naturalist position, research on sensory measurement is leaving certain "long-standing prejudices" about the nature of sensory perception, arising from a "chronic dualistic metaphysics", that "have triggered a variety of stubborn objections whenever it has been proposed that sensation may be amenable to orderly and quantitative investigation" (1961, p. 1). We solve the problem of coordination, and the rest will follow, he seems to be saying.

In fact, the circularity issue in solving the coordination problem was recognized right from the start in psychophysics. Gustav Theodor Fechner, who has been called the founding father of psychophysics, aimed at establishing a nomic relation between cognitive process and sensory responses. He observed regularities between the (increase of) intensity of a physical stimulus and its perception; the resulting Weber–Fechner law states that, "we can understand the relative change of stimulus and sensation best by expressing it as a general function relating to the successive constant increments of the sensation the successive variable increments of the stimulus" (Fechner [1860] 1966, p. 52). Basically, the intensity of a sensation must be doubled from an initial value for the change to be perceived as a difference in quality.

Fechner's work received thorough criticism from psychologists and physicists alike. Is the Weber–Fechner law reliable, and how can its validity be tested? Here the problem of coordination was clearly present. How ~~we can~~can we correlate our theoretical concept of what is measured (cognitive processes) with what is observed (behavioral responses)? It was not possible to assess the measured characteristic (i.e., intensity) independently of the Weber–Fechner law due to the absence of alternative measurement operations (Michell 1999). The application of the law was far form obvious, either. In Eleanor Acheson McCulloch Gamble's (1898) study on *The Applicability of Weber's Law to Smell* ~~(1898)~~, for instance, a

major question was how to define objective differentiae for subjectively perceived distinctions such as weak and strong odors. Recording responses to particular mixtures, regularity of exposure, and other factors, Gamble—an unfortunately forgotten female pioneer in olfactory research—started from the following observations:

> (1) Weak smells have vague differences of intensity. For example, vanilla and coumarine soon reach a maximum of intensity which cannot be increased. Greater concentrations simply become unpleasant. (2) Individual differences are more evident for weak smells. (3) The daily variations of sensitivity are more evident for weak smells. (4) Exhaustion has more effect on weak smells. (5) Strong smells hide the weak. (Gamble ~~and McCulloch~~ 1898, p. 93) Insert footnote after brackets (Gamble 1898, p. 93):
> Gamble was adequately vague about what she meant by weak and strong stimuli. She mainly refers to stimulus concentration.

These Remove indentation here. criteria do not lead to precise categories.

Despite such methodological difficulties, the Weber–Fechner law was considered the most promising attempt at quantifying sensations among psychologists at that time. The law bestowed them with a basic testable input–output relation. However, we are left to wonder in a counterfactual fashion how different the methodological struggle of early psychophysics might have developed if practitioners had adopted a different account of sensory quality measurement. The retrospect by Stevens was quite scathing. Stevens attacked Fechner for his dominant role in the early stages of psychophysics, silencing other proposals, more fruitful and adequate in Stevens' view:

> Another difficulty is that psychophysics had an unfortunate childhood. Although Plateau in the 1850s made a half-hearted attempt to suggest the proper form of the function relating apparent sensory intensity and stimulus intensity, he was shouted down by Fechner, who saddled the infant discipline with his erroneous "law"

> that bears his name. (Stevens 1961, p. 2)

Given that Fechner becomes almost singularly associated with psychophysics in the occasional "historical origins paragraph" in many philosophy and science papers, Stevens' criticism is of serious concern and marks a valid point.

A more careful look at how the practices have developed and are executed, however, reveals a more pragmatic solution, grounded in the strategy of epistemic iteration. Many aspects of olfactory perception that initially posed difficulties for psychophysical measurement have turned into a target of enquiry in parallel with growing knowledge about the constituents of the perceptual mechanism. Physical stimulus complexities (e.g., mixtures often change odor quality with slight changes in concentration or purity) and inconsistency in behavioral responses (indicating intra- and inter-subjective factors) hindered an obvious association of perceptual phenomena with their physiological basis. An evaluation of the emergence of perceptual phenomena was further diminished by the lack of insight into the perceptual mechanism. Only after the discovery of the olfactory receptors (Buck and Axel 1991) and subsequent insight into signal transduction and higher brain processing (Firestein 2001) was it possible to start correlating what is perceived with how it is perceived and processed. Indeed, this coordination process is still ongoing!

Manipulating stimuli and observing a range of different odor responses raise a range of interesting questions: What happens when you mix two odor stimuli? Do they blend qualitatively into a synthesized impression, or can you still discriminate the individual components? And does their blending effect depend on a qualitative similarity between them? Is the intensity of the mixture increased in an additive way (based on the magnitude of the individual stimuli)? In parallel with studies of stimulus responses, questions emerge as to how we know whether a perceptual effect is, for instance, based on the particular workings of the receptors rather than another processing stage. Different disciplinary approaches have gradually formed informative collaborations on such questions of olfactory processing.

Psychophysical findings, sometimes in contrast with comparative measures of other sensory modalities such as different "modes of counteracting" in olfactory

responses (addition, compromise, compensation),[12] have been correlated with in vitro experiments on olfactory receptors (Oka et al. 2004).[13] Combinatorial studies of psychophysical, genetic, and neuroscientific techniques have been used either to link ~~stimulus induced~~stimulus-induced behaviors and preferences to a genetic basis (Keller and Vosshall 2004) or to indicate potential lines of new enquiry in little-understood mechanisms such as olfactory cortex processing (Bowman et al. 2012). Proposals through genetic and neuroscientific approaches likewise beget questions that can be further addressed psychophysically, presenting a process of mutual grounding in the coordination of theoretical concepts and observational practices. Such mutually informing disciplinary entrenchments are also visible in the integration of psychophysics into biomedical institutes (e.g., Leslie Vosshall's ~~laboratory~~Laboratory of Neurogenetics and Behavior at Rockefeller University).[14]

Perception can neither be measured directly nor is there a straightforward way to link perceptions to their physiological basis. It is an ongoing enquiry in which successive insights into the workings of the ~~olfactory~~sensory system are gradually correlated with its apparent effects. These insights and correlations are continuously challenged through subsequent findings, resulting in a progressive iteration of psychophysical and neuroscientific studies.

The Problem of Standardization

Let us now turn to the problem of standardization. Here it may seem that we have more of a unique problem for sensory measurement, due to the subjectivity and contextuality of many sensory experiences. Due to the inherent variability of sensory experiences, ~~especially~~such as olfactory ones (Barwich 2014), generalizations drawn from sensory performance studies are often accompanied by severe caution, even from the practitioners themselves (Sell 2005). The lack of a uniform basis of measurement and its evaluation constitutes one of the main reasons for contesting the status of sensory measurement as proper measurement. A lot of attention has been directed at methodological concerns on how to eliminate individual differences across observers. On the BAAS panel mentioned above, there was concern about how to conduct a "study of the differences between

individuals in sensory experience" (Drever in BAAS 1932, Section A, p. 301). Dominated by physicists, the panel judged proposals of psychophysical measurement patronizingly as "not very accurate", but "much better than nothing" (Houstoun in BAAS 1932, Section A, p. 302).

But once again, it can be seen that physical measurements also suffer from the same kind of problems, though perhaps not to the same degree. Standardization, more specifically the selection and maintenance of measuring instruments and procedures, is something that physical scientists have been very concerned about. It was in astronomy, as much as in psychology, in which differences in individual human sensory experience ~~was~~were first problematized in the context of measurement (Canales 2009). Astronomers took great care in guarding against these differences, by means of various ingenious procedures including the estimation of the "personal equation" for each observer. In the case of thermometry, it is easy to see how different versions of the same instrument (~~"the thermometer", "the liquid-in-glass thermometer", "the mercury thermometer"~~the liquid-in-glass thermometer or the gas thermometer, the mercury thermometer or the alcohol thermometer, or even ~~"~~the mercury thermometer graduated with 0° and 100° set by the melting point of ice and the boiling point of water~~"~~, etc.) gave divergent results from each other. Epistemologically this is not so different from different human observers giving different reports.

In such cases, the absence of a convincing solution to the problem of coordination exacerbates the problem of standardization. As mentioned above, if there were a perfect solution to the coordination problem, it would be possible to say which of the competing standards are true and which ones false. Disagreements among standards would be resolved into different amounts of error attributable to each standard. However, without an agreed-upon idea of what exactly each standard is meant to be measuring, it is impossible to say whether each one is measuring correctly or not. Consider the case of kinetic-energy measurement in early quantum physics (Chang 1995), in which different measurement methods (magnetic deflection, electrostatic retardation, and material retardation) delivered different results from each other. In the late nineteenth century, the well-established theory of classical physics would have given a clear verdict on whether

each alleged method of energy measurement was really measuring energy, and if not, how much the systematic error would have been. Such a clear solution to the coordination problem did not exist in the early twentieth century, when the old classical theory had been knocked out but the new quantum-mechanical theory had not been firmly established yet. Such situations are bound to be common in periods of major scientific change. Interestingly, these days physical scientists tend to be oblivious to such problems of measurement, consigning that business to "metrologists" and instrument manufacturers.

How is the problem of standardization solved? Again, we want to argue that it has been dealt with in similar ways in both physical and sensory measurements, namely through epistemic iteration. The standardization problem could be solved in a reductionist way through an exceptionally good solution to the coordination problem, by corrections and calibrations rooted in firm theoretical knowledge in each measurement setup. That is not likely, even in the case of the physical sciences. More usually, a key element of the iterative solution to the standardization problem is the mutual grounding of measurement standards. According to this scheme, one must start by creating various standards possessing some initial plausibility, and then locate a group of them that accord well enough with each other. Standards in such a group provide a sort of provisional justification for each other, and their collective verdicts provide a sufficient basis on which further investigation can be carried out.

Such "mutual grounding" process can be seen in the early development of pyrometry, namely the measurement of high temperatures (Chang 2004, Chap. 3). Beyond temperatures at which mercury boiled and glass softened, traditional thermometers could not be employed any more, which created real difficulties for the control of various industrial processes. Josiah Wedgwood, renowned porcelain maker of the late 18th century, made headway here with his clay pyrometer, based on the observation that pieces of clay exposed to high temperatures shrank, and seemed to shrink more when exposed to higher degrees of heat. Wedgwood's innovation was highly praised (resulting in his election as a fellow of the Royal Society, among other things), but there were a number of other proposed pyrometric methods, too, getting at the temperature of a very hot object through

the time it takes to cool down, the amount of ice melted by it, the degree of heating of cold water in which it is immersed, and so on. In the end Wedgwood's pioneering method was rejected, in favor of a set of other pyrometric standards that agreed reasonably well among themselves but not with Wedgwood's. In this process there was no absolute right or wrong, and the main concern was to devise a sufficiently stable set of mutually grounded standards, which allowed further investigations to be carried out. The results of such investigations can lead to adjustments in the initial set of standards, and this process can be repeated in a way that is typical of epistemic iteration. If there is suitable theoretical progress arising from the further investigations, then the iterative adjustment may take the form of better theoretical justification of the standards that remain in the set.

There is no reason why this sort of iterative attack on the problem of standardization cannot be made credibly in the psychological sciences. As with the case of physical measurements, different methods of measurement are applied in mutual adjustment of each other (i.e., using different evaluative statistical methods as well as different experimental setups exposing human test subjects to stimuli). For example, the "g" factor measured in intelligence tests, introduced by Charles Spearman in 1904 under the heading of "general intelligence", was a statistical composite made from various tests assessing specific tasks, the results of which seem to go together well enough (see Deary et al. 2008). There is nothing wrong with this way of proceeding, although the identification of g as "the" measure of intelligence has been highly problematic. Epistemic iteration on the basis of mutual grounding has been a pattern in olfactory performance tests, too.

Adjusting for factors such as age and performance In the proofs the line-break was "in-fluencing", making it look like "performance-in-fluencing". In print, please make the line-break before "in" so that it doesn't look like performance-in-fluencing. -influencing behavior (e.g., smoking), first smell identification tests were based on initially rather rough statistically representative samples that were further refined through increasing data from successive applications of these tests. The reference standard (normal perception) initially fixed thereby was continuously calibrated. This process of standardization at first was relatively independent of deeper theoretical insight into olfactory processing (coordination through observational grounding).

At the time of the establishment of the UPSIT in 1984, for instance, insight into the second messenger pathway in olfaction was only beginning to emerge (Pace et al. 1985; Sklar et al. 1986; Jones and Reed 1989, Firestein et al. 1991) and neither were the olfactory receptors discovered (Buck and Axel 1991) nor was much known about ~~glomerulus and cortex processing~~higher brain processing of odors (Firestein 2001). Indeed, the latter poses a matter of ongoing inquiry (Stettler and Axel 2009; Chen et al. 2014).

Gradually corrective procedures underlying sensory performance studies are obviously at work when it comes to tackling sensory measurement problems such as human bias. Nonetheless, in order to impose first standards, to ensure the comparability of results and assess sources of variation, it is important to tailor the "instrument" (i.e., the way human test subjects are used) to its purpose. To ensure a level of control over human bias most psychophysical studies use *trained* test subjects. It is not uncommon that subject panels consist of the researchers conducting the study or their colleagues. This strategy seems counterintuitive at first, but it allows the practitioners to exclude factors that are potentially distorting results (i.e., the subject misunderstanding the nature of the task or not being observant).[15] It thus sets a standard of comparison, for instance of the observational abilities and level of training of subjects, that allows for a more accurate analysis and comparability of data. Using experts as test subjects further facilitates the adjustment of the original test design, thus having a self-corrective effect.[16] This strategy is conventionally found in psychophysical research on detailed aspects of normal perception (e.g., testing antagonist perception), setting an observational standard through which perceptual and measurement categories are developed and their compatibility envisaged. As a result, "[m]ost olfactory psychophysical tests are positively correlated with one another and measure common attributes …." (Tourbier and Doty 2007).

In comparison, test panel selection differs in some clinical or cross-cultural studies as their purpose differs. Their corrective is not so much focused on counteracting human bias in a small controlled setting, but ~~to test~~on testing for the presence or absence of a general effect in a wider population (clinical tests looking at average rates, not specific differences in perception; Keller and Vosshall 2004, p. R876).

Clinical and biomedical studies that investigate the applicability and range of performance tests use large patient groups. Here the aim is to get as much normative data as possible for the evaluation of test kits such as the UPSIT (Deems et al. 1991 evaluated the responses of 750 patients) and the Sniffin' Sticks (Hummel et al. 2007 evaluated the performances of over 3000 subjects). A similar case holds for cross-cultural comparisons, e.g., where general differences of cultural background are used for the modification of test kits. Such application-based studies are often less fine-tuned or rigorous than groundwork psychophysical studies.

What is considered an adequate standard, defining the quality by which tests allow for judgments about the underlying phenomenon, thus depends on the purpose of its design. The establishment and maintenance of such a standard, however, is grounded in the same corrective and gradual process of epistemic iteration. In all of the above cases, practitioners start from stipulating a representative standard through a statistically representative sample (of human test subjects that are either experts or different groups of laypeople) to determine normal, i.e., constant, perceptual responses to specific stimuli. Based on this standard, they measure whether something meets this standard (i.e., the extent to which phenomena of sensory perception appear constant or what anomalies occur) by observing the scope and degrees of variation in human responses to stimuli. Successive psychophysical studies and the growth of statistical data then allow for successive modifications of interpretations based on initial samples as well as a revision of the test kit or study design.

Taking the Challenge Forward

Having seen how problems of coordination and standardization have been dealt with, what can we say about the crucial question that persists, namely that of the reliability of measurement? Let's look back on the sources of unreliability in sensory measurement. Even though the problems of standardization and coordination are in principle separable, in practice we often do not know whether possibly defective measurement is badly standardized, or is instead badly coordinated with the phenomenon. When we have doubts about the reliability of

measurement, we cannot always say in a decisive way whether the difficulty we have is with standardization or coordination.

Still, it is helpful to separate out the problem of standardization and the problem of coordination. These two issues are not always clearly distinguished but often analyzed alongside each other (BAAS 1932; Michell 1999; Tal 2013). The problem of coordination~~, which we discussed in "The Problem of Coordination" section,~~ concerns the grounds on which recorded effects correspond to the phenomenon in question that is to be measured. The problem of standardization, in contrast, concerns the assignment of operations that allow us to compare, repeat, and correlate studies through the specification of ~~mutual or~~mutually compatible standards. Whereas the problem of standardization resides at the level of operations, the problem of coordination surrounds the way in which knowledge of how the instrument works links the operational level with the theory of the phenomenon in a meaningful way.

As a measurement practice, sensory performance studies are obviously less stable and unambiguous than the measurement of physical quantities. However, the lack of standardization in sensory measurement is distinct from the problem of coordination. This distinction, as we have shown in this paper, can help us understand and tackle the sources of unreliability in sensory measurements. For instance, is it because of defective coordination between theoretical concepts and observational grounding (i.e., lack of insight into the nature of the phenomenon)? Or is it because of differences in the standardization of sensory performance studies? And so on. When we look at the problem of standardization and the problem of coordination separately, we will find that these two problems resonate with two separate epistemological difficulties: circularity (in the coordination of theoretical concepts and observational grounding) and ambiguity (in the standardization of coordinating operations). It is the overcoming of these difficulties in a process of epistemic iteration that earns good sensory measurements the label of "measurement".

Sources of unreliability in sensory measurement do not come down to bias in human perception per se. Any measurement is *ultimately* based on human

perception, and difficulties concerning human perception have to be dealt with somewhere, somehow. It might be useful to remember, again, the humble origins of physical measurements. The parallel that we have drawn between physical and sensory measurements is in fact more than just a parallel. There is real continuity between the two, as is shown, again, in the case of temperature measurement (Chang 2004, Chap. 1). In the Aristotelian ontology temperature was seen in purely qualitative terms, with hot and cold as opposite qualities not reducible to each other. Later the notion of "degrees of heat" was established, assuming a one-dimensional spectrum of hotness on which phenomena could be ordered as more or less hot than each other. This defined an "ordinal scale" of temperature, coupled with instruments such as the early thermometers with rather arbitrary numbers attached on their scales, which are more properly called "thermoscopes" rather than "thermometers". Thermoscopes enabled observations that established the "fixed points", which then allowed the creation of a meaningful numerical scale. This case of the stepwise quantification of temperature recalls the influential approach to the problem of coordination in psychophysical measurement by Stevens, who defined measurement as "the assignment of numerals to events or objects according to rule" (1961, p. 4). Proposing four different types of measurement scales (nominal, ordinal, interval and ratio), Stevens (1946) focused on possible empirical operations through which one can assign various kinds of scales to phenomena.[17]

Stevens' "operational measurement theory" has been criticized as a form of conventionalism lacking any commitment to "truth"—that is, without any claim as to the existence of what is allegedly measured. Especially Joel Michell (1999) finds fault with established psychophysical measurement in such a manner, as it only meets what he calls the "instrumental challenge". Without demonstrating first whether sensory perceptions really are quantifiable, psychophysicists, in his opinion, fail to prove whether what they do is in fact measurement. However, Michell greatly underestimates the potential in Stevens' kind of approach; it is through the iterative process that quantification is made and measurability is established.

Limitations in experimental control and decisiveness persist in sensory

measurements. However, therein also lie occasions for further research. ~~Debate~~Debates on the adequacy of elements in the measurement apparatus or conceptual interpretations present an opportunity for revisions of central concepts in the measurement of sensory perception. Results of revisions might lead to suggestions of new or refined correlations in the coordination of theoretical concepts and observational practices. Less-than-perfect agreements provide "a useful starting point for concept building, rather than hindrance to reliable measurement" (Chang 1995, p. 165).

As an epistemic practice, sensory measurement is an invaluable tool for investigating and establishing correlations between different sets of phenomena (i.e., physiological processes and cognitive effects through behavioral responses). Methodological issues related to standardization or coordination, and incoherence or inconsistencies in the setup and results of sensory performance studies, provide sources for further inquiry about the nature of the measured phenomena. Recall the example of the different test kits for measuring olfactory performance. In such cases, researchers are challenged to revisit their observational basis: are we measuring the same thing just differently, or are we measuring different things?

Therefore, even if sensory measurements exhibit less precise concepts and less coherent test designs than physical measurements, they comply with a conception of measurement as an epistemic activity. Sensory qualities may not be the same as physical quantities, but they can be measured. Sensory measurements should be of great interest for current philosophical analysis of scientific practices in general. Many epistemological questions are waiting to be asked about the conceptualization of sensory perception, sensory qualities, stimulus choice, and stimulus control, as well as the multitude of possibilities and limits in the measurement thereof.

AQ3

Acknowledgments

This paper has benefitted greatly from comments on previous versions by Olivier Morin, Ingvar Johansson, John Dupré, Stuart Firestein, and two reviewers for the journal. Andreas Keller kindly answered our questions about recent developments

in olfactory psychophysics. The work was made possible by funding from the KLI Institute. Special gratitude belongs to Werner Callebaut (†), a dearly missed cartographer of knowledge.

References

Axel R (2005) Scents and sensibility: a molecular logic of olfactory perception (Nobel lecture). Angew Chem Int Ed 44(38):6110–6127

BAAS (1932) Report of the British Association for the advancement of science. John Murray, London

Barkai E, Wilson D (2014) Odor memory and perception. Elsevier, Oxford

Barwich A-S (2014) A sense so rare: measuring olfactory experiences and making a case for a process perspective on sensory perception. Biol Theory 9:258–268

Baylor DA, Lamb TD, Yau KW (1979) Responses of retinal rods to single photons. J Physiol 288:613–634

Bechtel W (1986) The nature of scientific integration. In: Bechtel W (ed) Integrating scientific disciplines. Kluwer, Dordrecht, pp 3–52

Bowman NE, Kording KP, Gottfried JA (2012) Temporal integration of olfactory perceptual evidence in human orbitofrontal cortex. Neuron 75:916–927

Buck LB, ~~Richard A~~Axel R (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65:175–187

Bushdid C, Magnasco MO, Vosshall LB, Keller A (2014) Humans can discriminate more than 1 trillion olfactory stimuli. Science 343:1370–1372

Canales J (2009) A tenth of a second: a history. University of Chicago Press, Chicago

Chang H (1995) Circularity and reliability in measurement. Perspect Sci 3:153–172

Chang H (2004) Inventing temperature: measurement and scientific progress. Oxford University Press, Oxford

~~Chen F-F, Zou D-J,~~Chen FF, Zou DJ, Altomare CG et al (2014) Nonsensory target-dependent organization of piriform cortex. Proc Natl Acad Sci USA 111:16931–169316

Cho JH, Jeong YS, Lee YJ et al (2009) The Korean version of the Sniffin' stick (KVSS) test and its validity in comparison with the cross-cultural smell identification test (CC-SIT). Auris Nasus Larynx 36(3):280–286

Darwin C (1871) The descent of man, and selection in relation to sex. John Murray, London

Deary IJ, Lawn M, Bartholomew DJ (2008) A conversation between Charles Spearman, Godfrey Thomson and Edward Thorndike. Hist Psychol 11:122–142

Deems DA, Doty RL, Settle RG et al (1991) Smell and taste disorders, a study of 750 patients from the University of Pennsylvania Smell and Taste Center. Arch Otolaryngol Head Neck Surg 117(5):519–528

Doty RL (2013) Smell and the degenerating brain. The Scientist, October 1. http://www.the-scientist.com/?articles.view/articleNo/37603/title/Smell-and-the-Degenerating-Brain/. Accessed 15 July 2015

Doty RL, Shaman P, Dann M (1984) Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. Physiol Behav 32:489–502

Doty RL, Applebaum SL, Zusho H et al (1985) Sex differences in odor identification ability: a cross-cultural analysis. Neuropsychologia 23:667–672

Elliott KC (2012) Epistemic and methodological iteration in scientific research. Stud Hist Philos Sci Part A 43(2):376–382

Fechner GT ([1860]1966) Elements of psychophysics, vol. 1. Adler HE (trans), Howes DH, Boring EG (eds). Reprint. Holt, Rinehart and Winston, New York

Ferguson A, Myers CS, Bartlett RJ et al (1940) Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. Rep Br Assoc Adv Sci 2:331–349

Firestein S (2001) How the olfactory system makes sense of scents. Nature 41:211–218

Firestein S, Darrow B, Shepherd GM (1991) Activation of the sensory current in salamander olfactory receptor neurons depends on a G protein-mediated cAMP second messenger system. Neuron 6:825–835

Frank RA, Dulay MF, Gesteland RC (2003) Assessment of the Sniff Magnitude Test as a clinical test of olfactory function. Physiol Behav 78(2):195–204

Gerkin RC, Castro JB (2015) The number of olfactory stimuli that humans can discriminate is still unknown. eLife 4:e08127

Hawkes CH, Shephard BC, Daniel SE (1999) Is Parkinson's disease a primary olfactory disorder? QJM: Int J Med 92(8):473–480

Hecht S, Shlaer S, Pirenne MH (1942) Energy, quanta, and vision. J Gen Physiol 25(6):819–840

Hummel T, Sekinger B, Wolf SR et al (1997) ~~'Sniffin' sticks'~~'Sniffin' sticks': olfactory performance assessed by the combined testing of odor identification,

odor discrimination and olfactory threshold. Chem Senses 22(1):39–52

Hummel T, Kobal G, Gudziol H, Mackay-Sim A (2007) Normative data for the "Sniffin'Sticks" including tests of odor identification, odor discrimination, and olfactory thresholds: an upgrade based on a group of more than 3,000 subjects. Eur Arch Otorhinolaryngol 264(3):237–243

Johansson I (2014) Constancy and circularity in the SI. Metrologybytes. www.metrologybytes.net/PapersUnpub/OpEds/Johansson_2014.pdf. Accessed 15 July 2015

Jones DT, Reed RR (1989) Golf: an olfactory neuron specific-G protein involved in odorant signal transduction. Science 244:790–795

Jones-Gotman M, Zatorre RJ (1988) Olfactory identification deficits in patients with focal cerebral excision. Neuropsychologia 26(3):387–400

Keller A, Vosshall LB (2004) Human olfactory psychophysics. Curr Biol 14(20):R875–R878

Klein SA (2001) Measuring, estimating, and understanding the psychometric function: a commentary. Percept Psychophys 63(8):1421–1455

Lettvin JY, Maturana HR, McCulloch WS, Pitts WH (1959) What the frog's eye tells the frog's brain. Proc Inst Radio Engr 47:1940–1951

Lötsch J, Reichmann H, Hummel T (2008) Different odor tests contribute differently to the evaluation of olfactory loss. Chem Senses 33(1):17–21

Magnasco MO, Keller A, Vosshall LB (2015) On the dimensionality of olfactory space. bioRxiv July 6. doi: http://dx.doi.org/10.1101/022103. Accessed July 15 2015

Majid A (2015) Cultural factors shape olfactory language. Trends Cogn Sci (in

press)

Majid A, Burenhult N (2014) Odors are expressible in language, as long as you speak the right language. Cognition 130(2):266–270

McCulloch Gamble EA (1898) The applicability of Weber's Law to smell. Am J Psychol 10(1):82–142 Correct is: Gamble EAM (1898) The applicability of Weber's Law to smell. Am J Psychol 10(1):82–142
AND: Ref needs to be between (Frank et al. 2003) and (Gerkin and Castro 2015) in reference list

Meister M (2014) Can humans really discriminate 1 trillion odors? arXiv, 1411.0165 and 1411.0165v2. Accessed 15 July 2015

Meister M (2015) On the dimensionality of odor space. eLife 4:e07865

Michell J (1999) Measurement in psychology. A critical history of a methodological concept. Cambridge University Press, Cambridge

Morrison J (2014) Human nose can detect 1 trillion odours. Nat News 20 March. doi:10.1038/nature.2014.14904

~~Nevid J (2012) Essentials of psychology: concepts and applications, 4th edn. Cengage Learning, Wadsworth~~

Oka Y, Omura M, Kataoka H, Touhara K (2004) Olfactory receptor antagonism between odorants. EMBO J 23(1):120–126

Pace U, Hanski E, Salomon Y et al (1985) Odorant-sensitive adenylate cyclase may mediate olfactory reception. Nature 316:255–258

Pinker S (1997) How the mind works. Norton, New York

Sell C (2005) Scent through the looking glass. In: Kraft P, Swift KAD (eds)

Perspectives in flavour and fragrance research. Wiley-VCH, Zurich, pp 67–88

Shepherd GM (2004) The human sense of smell: are we better than we think? PLoS Biol 2(5):e146

Shepherd GM (2012) Neurogastronomy: how the brain creates flavor and why it matters. Columbia University Press, New York

Sklar PB, Anholt RR, Snyders SH (1986) The odorant-sensitive adenylate cyclase of olfactory receptor cells. Differential stimulation by distinct classes of odorants. J Biol Chem 261(33):15538–15543

Sorowska A, Sorokowski P, Hummel T (2014) Cross-cultural administration of an odor discrimination test. Chemosens Percept 7(2):85–90

Spehr M, Schwane K, Heilmann S et al (2004) Dual capacity of a human olfactory receptor. Curr Biol 14:R832–R833

Stettler DD, Axel A (2009) Representations of odor in the piriform cortex. Neuron 63:854–864

Stevens SS (1946) On the theory of scales of measurement. Science 103:677–680

Stevens SS (1961) The psychophysics of sensory function. In: Rosenblith WA (ed) Sensory communication. Contributions to the symposium on principles of sensory communication. July 19-August 1, 1959, Endicott House, M.I.T. MIT Press, Cambridge, pp 1–34

Tal E (2013) Old and new problems in philosophy of measurement. Philos Compass 8(2):1159–1173

Tourbier I, Doty RL (2007) Sniff magnitude test: relationship to odor identification, detection, and memory tests in a clinic population. Chem Senses

32:515–523

Wise PM, Olsson MJ, Cain WS (2000) Quantification of odour quality. Chem Senses 25:429–443

Wnuk E, Majid A (2014) Revisiting the limits of language: the odor lexicon of Maniq. Cognition 131:125–138

[1] Exceptions prove the rule. In 1942 a landmark study showed that retinal receptors detect single photons but require a higher threshold to be recognized as having a conscious effect (Hecht et al. 1942). The results of this early psychophysical test were only confirmed physiologically almost 40 years later (Baylor et al. 1979).

[2] Michell (1999) is a salient exception. A recent debate between Hatfield, Feest, and Chirimuuta concerns psychophysical studies of visual perception. Nonetheless, their focus remains on conceptual categories that relate to the problem of introspection and consciousness, rather than epistemological issues of sensory measurement. We will therefore not refer to their discussion in this paper.

[3] Psychometrics studies psychological categories such as personality traits, skills, abilities, etc. It also concerns the methods employed in psychological measurement and questions regarding their (limited) objectivity. Psychophysics is about the relation between physical stimuli and their perceptions. ~~Both fields~~The two fields are related: “The psychometric function, relating the subject’s response to the physical stimulus, is fundamental to psychophysics” (Klein 2001, p. 1421).

[4] ~~Alas, it seems that consumer behavior~~Consumer behavior does not seem to care about such academic dismissal, as the fragrance industry thrives “with sales of scented products constituting an annual market of over $25 billion dollars in the United States alone” (Keller and Vosshall 2004, p. R875).

[5] “Preclinical” means the stage prior to the manifestation of specified symptoms that facilitate the diagnosis of a disease or disorder.

[6] The design of test kits for olfactory function, especially in medical applications, has to factor in the cultural background when measuring smell perception, for instance regarding the subjects’ familiarity with an odor. Cross-cultural comparisons showed similar odor responses (e.g., familiar odors are considered more pleasant). Such comparisons also showed that subject groups from different cultures performed differently in threshold and discrimination tasks based on their

familiarity with an odor (Doty et al. 1985; Sorowska et al. 2014).

[7] Axel's quote refers to the classic paper, "What the Frog's Eye Tells the Frog's Brain" (Lettvin et al. 1959).

[8] Glomerulus (plural: glomeruli) is a term for a spherical neural structure formed by the concurrence of olfactory sensory nerves that are expressing the same receptor gene. Mitral cells are neurons in the olfactory bulb. Microcircuits are ~~characterised~~characterized as functional modules of cell collections in the brain that carry out specific actions.

[9] The assignment of functional stages to specific constituents of the olfactory system is not clear-cut and remains in dispute. Yet, the concentrated distribution of different processing stages in specific structural components of the olfactory system allows for a rough functional compartmentalization (Barkai and Wilson 2014).

[10] Anosmia is the inability to smell. Partial anosmia characterizes the inability to detect specific odors. Hyposmia is the reduced ability to smell. (Erratum to Barwich 2014, where anosmia and hyposmia have been accidentally placed in reverse order on p. 266.).

[11] This can also be problematic for testing children and people with cognitive decline.

[12] "The estimated intensity of the smell of the mixture of two odorants is frequently perceived as being non-additive. This phenomenon is called counteracting. There are three types of counteracting: partial addition, in which the mixture smells more intense than the stronger component; compromise, in which the smell intensity of the mixture is in between the intensities of the components; and compensation, in which the mixture smells less intense than the weaker component" (Keller and Vosshall 2004, p. R877).

[13] Besides a few attempts (Spehr et al. 2004), however, work on the relation between antagonism behavior at the receptor level and counteracting responses in odor perception has come to a halt in the past 10 years. (We thank Andreas Keller for making us aware of this.)

[14] See Bechtel (1986) for a thorough discussion of the epistemological and sociological aspects in similar cases of interdisciplinarity and scientific integration.

[15] Sociological factors such as the physical location of the research group (e.g., well-connected versus isolated institutes) and the availability of human test subjects (such as grad students) also play a role, of course.

[16] We thank Stuart Firestein for pointing that out to us.

[17] On an interval scale, the placement of the zero point is conventional but the same size of interval has the same meaning everywhere (e.g., the difference between 10° and 15°, and the difference between 85° and 90°). On a ratio scale, the value zero means the real absence of the

quantity, and it is meaningful to apply the arithmetic operation of multiplication to a magnitude. The familiar temperature scales (centigrade, Fahrenheit, etc.) are interval scales, and Kelvin's absolute temperature scale is a ratio scale. It may be said that the temperature concept has evolved through all four of Stevens' scale types.