"The Making of Measurement": Abstracts

KEYNOTES

Nancy Cartwright: *The Theory of Measurement*

From work with Norman Bradburn and Rosa Runhardt.

Measurement, I shall argue, requires: 1) A characterization of the quantity or category: we have to be able to identify its boundaries and know what belongs to it and what does not (characterization); 2) a metrical system that represents the quantity or category (representation); and 3) rules for intervening in the world to produce measurement results (procedures). It is essential that we can defend these three mesh together properly. Representation theorems are crucial for the link between 1 and 2, and a great deal of empirical knowledge for linking 3 with the other two.

I shall next distinguish between concepts that refer to a single quantity or category that can be precisely defined and those that refer to things that are loosely related and for which the boundaries of the concept are not clear (Ballung concepts). When we make these precise for purposes of exact science, we generally leave behind a good deal of the original meaning. This suggest representing these with tables of indicators but at the cost of comparability.

Finally I shall consider advantages of purpose-built versus common metrics.

Terry Quinn: *From Artefacts to Atoms: The Basis of Reliable Measurements*

At the 26th General Conference on Weights and Measures, due to take place in 2018, it is planned to adopt a new definition of the International System of Units SI to be based on a set of fixed numerical values of constants of nature. This will be the culmination of more than two hundred years of metrology finally putting into practice the original ideas of those who created the metric system. The key is the replacement of the kilogram artefact of Pt-Ir by definition based on a fixed numerical value for the Planck constant. In this lecture I shall outline how this has come about and link it to the need for a system of measurement that is uniform and accessible world-wide for international trade, high-technology manufacturing, human health and safety, protection of the environment, global climate studies and the basic science that underpins all these.

Terry J. Quinn, CBE FRS was the Director of the International Bureau of Weights and Measures (BIPM) between 1988 and 2003. BIPM is an international standards organisation, one of three such organisations established to maintain the International System of Units (SI) under the terms of the Metre Convention (Convention du Mètre).

Graeme Gooday: ***A Measured Hearing***

What does it mean to 'measure' hearing? For what purposes would it matter that hearing could or should be measured - by whatever understanding of measurement might be involved? A conventional answer is that the clinical sub-discipline of audiology developed the relevant form of expertise as the distinctive intervention to quantify human hearing capacities. It did so by combining into a hybrid technical field insights from the physics of acoustics, the technology of telecommunications testing and the physiological expertise in hearing pathologies hitherto the province of the otologist. Put in historical context it is uncontroversial that the profession of audiology emerged in the mid-twentieth century, with audiologists using the audiometer as its principal instrument to quantify the extent and nature of hearing loss, doing so with a view to its amelioration by the prescription of appropriately configured hearing aids.

But what exactly did the audiometer measure? Certainly not the capacity to hear and understand human speech. So then what was signified by the audiometer's readings and how did that come to serve any useful clinical purpose? This paper sets out to answer that question, relating it chronologically to the massive increase of hearing loss in Second World War combatants, especially in the USA. To understand this story there is in turn a much longer term panorama to be explored of how the audiometer emerged in the late 19th century as a correlate of the telephone, and yet was not adopted for formal hearing testing until half a century later. This links in turn to the complex politics of the history of hearing loss that has so often been tangled up with the history of deafness in ways that have long obscured the huge variety of human capacities for hearing, and the multiple aetiologies for hearing loss.

Historical research in telecommunication history and disability history has recently started to uncover the diversity of human hearing capacities throughout the industrial era. These differential capacities were brought to focus by first encounters with the telephone in the late 1870s. This device only allowed for aural communication (stripping out all visual elements) and revealed a diversity of facility that had hitherto been masked by the starkly dichotomous language of deaf vs. hearing. So amongst all this diversity how was a notion of 'normal hearing' then created against which hearing loss could be diagnosed by relative comparisons? This initially contested notion of normalization was eventually canonized in the regular use of the audiometer in the 1920s not for clinical purposes, but is instead to deal with the civic problem of public 'noise', especially tin controversial campaigns to reduce alleged increases in civic street noise that have been studied by Karen Bijsterveld and Emily Thompson.

My paper will start with David Edward Hughes' proposal to the Royal Society in 1878 for devices to amplify human speech (the microphone) and an as yet unnamed device to measure human capacity to hear individual pure tones. This latter technique was soon taken up by the physician Benjamin Richardson's analysis in The Lancet, 1879, in which he christened the audiometer (initially just 'audimeter') proposing its adaptation to determine the hearing capacities

for certain key professions, and attempting to define relative deviation from 'perfect' hearing. However, I show instead how other methods of comparing hearing capacities were long preferred by physicians (tuning forks) and telephonic engineers (telephone-based devices). I then look at C.M.R. Balbi's attempt in The Lancet, 1925, to reinvent the audiometer to map in two-dimensional form the (drug-induced) variability of hearing loss independently of patients' own testimony. It was through such instrumentalisation of hearing tests that the audiometer could be used both in the clinic and laboratory to pathologize certain forms of hearing capacity as somehow less or greater than a new stipulated norm.

Nevertheless clinicians remained concerned that the audiometer only served to quantify human capacity to hear individual pitched sounds, not the much more significant human capacity to understand highly modulated multi-frequency speech. Hence my paper concludes by showing how the registrations of the audiometer came to serve as the surrogate for human hearing - and hearing loss - when the vast scale of combat-induced deafness in the Second World War obliged the American military services to adopt the audiometer in new specialist clinics for large-scale and high-speed evaluation and ameliorative prescription of reduced hearing capacities. Even so, the newly specialised body of audiologists that emerged to handle the long-term care of such deafened veterans did not lightly treat the audiometer as a simple measuring device - it was clear to them that a large amount of training was required to interpret the results of audiometric testing as well as special low-noise testing laboratories and a substantial division of labour in handling aftercare. Although this use of the audiometer was not then a full reductive 'industrialisation' of measurement (as discussed in Gooday, 2004) it is pertinent to note that 21st century audiology has moved way from intensively audiometric measurement regimes, returning to the more conversational bespoke evaluations of hearing loss that pre-dated the introduction of the audiometer.

PANELS

Charles Sanders Peirce on the Limits of Measurement in the Measurement of Limits

The Epistemology of Measurement looks beyond the axioms of measurement theory, the construction of scales, or the calculation of measurement error to the activity of measuring and the determination of measurements as fundamental to knowledge production. This widened scope of questions calls for a reconsideration also of those theorists who took as their starting point the juxtaposition of counting and measuring as fundamental epistemic practices - Hermann von Helmholtz, for example, or Charles Sanders Peirce.
The proposed session adopts an experimental format to engage and recognize Charles Sanders Peirce as measurement theorist. It brings together a physicist, a philosopher and historian of contemporary biomedical research, and a philosopher of technoscience to probe Peirce's views on the limits of precision and the heuristic significance of measurement error. Building on an exchange of manuscripts and ideas among the presenters in advance of the Cambridge conference, we will begin by providing a brief sketch of Peirce's relevance. This will be followed by two presentations on the determination of fundamental lower bounds for information processing in nanoelectronics and on the determination of baseline values, e.g., in clinical trial design. In the second half of the presentation on Peirce as philosopher of measurement, these two presentations will be used to explore the applicability and relevance of Peircean concepts. The three presentations will then be discussed together.

Improving Person-Centered Health Measuring Instruments: What is Needed?

The emergence of societal challenges such as the fair distribution of scare health care resources and the identification of effective treatments for the diseases of old age have created a need for outcome measures that go beyond the rudimentary quantification of survival and post-treatment morbidity. These measures, which usually take the form of self-reported questionnaires, attempt to quantify constructs such as well-being and quality of life that have formed a part of public discourse for millennia but have only recently become of interest to measurement scientists.

There is a deep dissatisfaction with the scientific properties of the most popular self-reported measures of health and well-being among many in the measurement community and increasingly, among philosophers of social science. The measures usually lack the basic standards used by the physical sciences such as unidimensionality, invariance and calibration. And yet the measures continue to be used in the face of such criticisms.

In this symposium we bring together three social scientists and a philosopher to diagnose some of the problems with these measures and comment on how we might improve them. Of specific interest is how and to what extent measurement

traditions in the natural sciences should inform our understanding of measurement in the social sciences.

In this collection of papers Leah McClimans and Stefan Cano diagnose the most serious problem with these instruments similarly: they lack and need a theory. McClimans situates this issue within the context of some of the recent literature on accuracy in the epistemology of measurement. She argues that without a theory we cannot even begin to concern ourselves with accuracy since it is unclear what variables confound quality of life measures, Cano approaches the lack of theory as one of three problems that beset the everyday use of these measures. First, he argues that scientifically weak measures continue to be used for extra-scientific reasons. How can we overcome this problem when it is entangled with powerful special interests? Secondly, social scientists disagree about what methods and criteria are best when developing these instruments. Such disagreement is not uncommon when scientists share a theoretical framework, but it is exacerbated when they do not which is the situation that these instruments face. John Browne and William Fisher's papers both ask what kind of relationship these measures ought to have to measures in the natural sciences. Browne argues that perhaps it is undesirable from a pratical perspective to insist on highly sensitive measures in the health sciences. Fisher, on the other hand, asks what kind of mutual informative dialogue might be had on the topic of measurement between the natural and social sciences? Drawing on work that began in 2008 he describes such a relationship in which Rasch methodology has been used to construct invariant linear measures and standard units with known uncertainties from ordinal observations.

We propose a roundtable format with 15 min presentations by each member of the symposium. We anticipate an hour for discussion following the presentations. We will precirculate 1) a 2,000-3,000 word paper in advance of the session which combines our respective interests (above) without asking the audience to read four lengthy papers and 2) our respective slides.

Points of Conversion: Quantification and Measurement in Germany and Mexico (1800-1940)

How is the metrication of a country achieved? How does it proceed at ground level? What meanings do the new measures carry for the actors concerned? The three papers in this panel explore these questions through case studies of rural India, Germany and Mexico. They investigate different historical moments at which conversion to new units and standards occurred, capturing the immediacy of process and the distinct meanings that precision and accuracy acquired in three agrarian settings.

Minakshi Menon follows the strategies of two Scottish *savants* as they tried to build a foundation for the East India Company state in south India. The first, Francis Buchanan had enormous difficulty converting local volumetric measures into linear measures in order to assess the revenue and population of Malabar. His compatriot, Alexander Walker, meanwhile, retained the poetic measures

common in the region while constructing a tax and property regime for the colonial state. Did Walker actually use indigenous forms of measurement? Or was his acceptance of the trans-empirical an indication of his willingness to negotiate with natives?

Anna Echterhölter's paper studies examples of rural measurements collected by the German linguist, Jacob Grimm. For example, a hen offered as tax payment would only be accepted if it could fly up on a barrel – a procedure Grimm classified as a measure of strength. He turned such old regime poetic and procedural measures into a polemic against the numerical forms of quantification imposed on German principalities by the French Code Civil in the wake of Napoleon's conquests. Echterhölter emphasizes that Grimm did not oppose quantitative to qualitative measures. Instead he argued that the absence of precision in Ancien Régime measurements guaranteed spaces of negotiation to the people. The fixed measures associated with the new law closed the space for face-to-face negotiation in settling disputes. They created a rigid, unrelentingly centralized, form of rule.

Héctor Vera analyzes the tactics of appropriation employed by lay people and non-experts as they confronted the decimal system of weights and measures, which was made mandatory in Mexico in 1895. People did not oppose metrication openly, but a 'surreptitious resistance' developed. Thus customary measures continued to be used, but they were 'adjusted' in a variety of ways, for example through *word substitution* (using the names of metric units to designate traditional indigenous measures); or *commensurability* (redefining customary units of measurements as exact fractions of metric units). Vera explores the implementation of a mandatory metric at ground level, showing how scientific ideas circulate across national boundaries and social classes. He argues that scientific and commercial globalization and nation-state formation were intertwined processes.

All three papers discuss encounters between divergent 'cultures of quantification' in widely different settings. Each context was, nevertheless, undergoing rapid economic change, making useful comparisons possible between ground-level practices in colonial India, Napoleonic Germany and early twentieth century Mexico.

PAPERS

Ann-Sophie Barwich: *The problem of coordination ≠ the problem of standardization: what sensory measurement is (not)*

Does sensory measurement deserve the label of measurement? I argue that it does. First, it complies with an epistemological view of measurement held in current philosophy of science. Sensory measurement faces the same epistemological challenges as 'nomic measurement' (e.g., temperature): the problem of coordination and the problem of circularity. These problems are resolved by similar procedures of 'epistemic iteration', i.e., a continually progressing coordination between theoretical concepts and their empirical basis. Second, I address the apparently insufficient reliability of sensory measurement. I argue for separating the problem of standardization from the problem of coordination. The problem of standardization characterizes the ambiguity of many sensory performance tests, while the problem of coordination relates to the correlation of theoretical concepts with an observational grounding in the measurement of sensory perception.

To exemplify my claims, I draw on olfactory performance tests, especially studies linking olfactory decline to neurodegenerative disorders. Changes in smell perception are a first symptom and a potential diagnostic tool for the pre-clinical detection of major neurodegenerative disorders. Differences in the course of hyposmias (reduced ability to smell) may also aid in differentiating disorders with similar clinical symptoms such as Parkinson's, Lewy bodies and Alzheimer's. A basic requirement for comparing abilities and differences of perception between healthy and ill test subjects is the design of standardized identification sets of test odors, and a reliable way to assess human sensory performances in identifying and discriminating odor qualities. Nonetheless, the measurement of odor perceptions in humans is distinctly difficult. Little agreement exists in how to best measure odor quality and how to deal with the apparent impossibility of eliminating bias in human sensory performances.

Distinguishing between the problem of standardization and the problem of coordination will prove a starting point for further analysis of measurement techniques in psychophysical studies. The practitioners' concerns are primarily methodological; however, their methodological concerns comprise three different epistemological issues that can be subsumed under the following three questions:

1. First, how can we link changes in qualitative sensory experiences to indicate and distinguish neurological disorders (coordination)?
2. Second, given the inherent variability of sensory experiences, how can we compare different tests and results (standardization)?
3. And, third, how reliable are our measurements, i.e., how objective are our strategies addressing the first two questions (reliability)?

   To address the epistemological issues underlying these questions, I take a closer look at the particular design and development of different olfactory performance tests: the *University of Pennsylvania Smell Identification Kit* (UPSIT), largely used

in North America, and the *Sniffin' Sticks*, predominantly used in central Europe. In fact, an extensive study of olfactory loss found that the choices of olfactory test kits produce different research outcomes. But what do such differences in measurement actually mean? Looking at their methodological design, studies of sensory performances make clear that when we analyze human test subjects as measurement instruments and the laboratory setting as measurement apparatus, the separation of standardization and coordination issues helps pinpointing sources of defect measurement.

Alessandra Basso: ***Two uses of robustness analysis in measurement practice***

The main idea of this paper is to distinguish between two functions of robustness analysis within measurement practice: i) to corroborate individual results and ii) to assess measurement procedures.

In philosophy of science, the use of robustness analysis is often described as being based on an inference to the best explanation. As the argument goes, if (i) multiple procedures converge on compatible results and (ii) the procedures are independent in some relevant sense, then the confidence in the correctness of the robust results (and of the converging procedures) increases. This conclusion is justified by saying that it would be an implausible coincidence if multiple independent procedures converged on the same result when the result is wrong or the procedures highly inaccurate. The argument includes a sometimes implicit "symmetry thesis", which says that the convergence of independence procedures simultaneously confirms the results and the procedures. This argument for robustness analysis is exposed to various objections that are well rehearsed in the literature, particularly that the presumption of independence is too strong and that the convergence of reliable procedures is uninformative, because it is the reliability of the single procedures rather than their convergence that makes us confident in the results.

The use of robustness analysis to corroborate individual measurement results is based on this argument. However, I argue against the symmetry thesis, by showing that the implications of this practice for the quality of the employed procedures remain ambiguous: the convergence of measurement procedure on some individual results does not automatically confirm the employed procedures.

I then introduce another way of using robustness analysis within *measurement assessment* practice, which has the function of evaluating the quality of procedures. This function of robustness received far less attention from the literature and deserves further investigation.

The practice can be briefly described as follows. Based on available information on the ways in which multiple procedures might fail to measure a certain property of interest, scientists formulate hypothesis about how their results converge. The expected convergence is then compared with the actual convergence of the procedures' results and a procedure is considered

(sufficiently) reliable if its outcomes converge with those of the others within the margins of expectation

I reconstruct the argument behind this function of robustness as follows: if (i) there is fit between the actual and expected convergence and (ii) the procedures have complimentary sources of error, then the procedures are (sufficiently) reliable.

The justification of this epistemic gain rests on the idea that a fit between actual and expected convergence shows that the sources of error do not affect the results more than what can be expected. This, in turn, is based on the complementarity of the sources of errors, which I describe as follows: two procedures have complementary errors if neither is "dominant" to the others with respect to all sources of error.

I argue that this use of robustness analysis is not exposed to the same criticisms as the other one, because it does not rely on the strong assumption of independence, and because the convergence of outcomes further constraints the errors attributed to each procedure, thereby providing new information about its reliability, information which would not be available without the comparison. Finally, the choice of examples used to illustrate the ideas contained in the paper allows me to explore some of the analogies between measurement practices in the physical and the social sciences.

Francesca Biagioli: ***Reconsidering The Philosophical Roots of Helmholtz's Theory of Measurement***

Hermann von Helmholtz has been acknowledged as one of the forerunners of measurement theory. However, his conception of measurement differs from later, representational conceptions, for several reasons. First, Helmholtz did not explicitly distinguish between numbers and numerals. Second, he used considerations regarding the origin of the decimal system in support of his empiricist philosophy of arithmetic. Third, his theory of measurement entails something more than a study of the conditions for using numbers in modelling measurement situations, because it implies that mathematical structures are common to both subjective experiences and objective ones. My suggestion is that these differences depend on Helmholtz's philosophical arguments. Not only do these arguments lend plausibility to some of the controversial aspects of Helmholtz's theory, but they provide us with a philosophical perspective on the main issues concerning measurement of his time. Helmholtz's paper on "Counting and Measuring" (1887) contains one of the clearest expositions of his naturalized interpretation of the Kantian theory of knowledge. He considered the concepts of number and of sum as taken only from the inner intuition of time in order to address the question: Under what conditions are mathematical symbols justifiably used to express metrical relations in the world? Helmholtz's answer was that such conditions can be formulated independently of the entities to be measured as laws of addition; nevertheless, the same laws are necessary for judgments about quantities to have a meaning. In other words, additive

principles play the role of constitutive principles of the objects of experience in Kant's sense, in that their domain of application defines what is measurable. At the same time, Helmholtz's argument differed from Kant's because Helmholtz excluded that a possible experience in general can be delimited a priori. Therefore, he argued for an extension of the laws of addition to all known physical processes as heuristic principle for empirical research. Although Helmholtz's universalistic account of measurability might seem to us to be contradicted by a variety of approaches to measurement, I point out that his view offered a new perspective on the discussion about the measurability of sensations which followed Gustav Fechner's and Wilhelm Wundt's attempts to measure psychological processes. Instead of arguing for measurability or incommensurability of sensations as such, Helmholtz replaced the metaphysical distinction between extensive and intensive quantities with a relative distinction between additive and nonadditive magnitudes in the sciences. The qualities of sensation offered the example of an attribute for which composition according to the method of addition was unknown, although neither the application of the same method nor a different form of quantification could be excluded. I suggest that the advantage of Helmholtz's empirical approach over both Fechner and the opponents of psychophysics – who argued for the incommensurability of sense qualities – lies in the fact that Helmholtz argued for a dynamical perspective on the problems concerning measurement.

Fabrizio Bigotti : *The Laboratory of Santorio: Measuring the body and its functions in seventeenth-century medicine*

Along with mechanics and astronomy, medicine played an important role at the beginning of the sixteenth century in the making of measurement. The pivotal figure in this sense can be considered the Italian physician Santorio Santorio (1561-1636) who, thanks to his work *Ars de statica medicina* (Venice 1614), converted the process of metabolism in mathematical terms. Santorio worked within the context of academic medicine, yet he explored new ways to study the so called 'insensible perspiration of the body' by evaluating the difference between the weight of ingested food and the one of the excretions. Santorio was quite aware of the modern idea of experimentation as he experimented daily for over thirty years. For the sake of scientific certainty, he felt also the need to invent and realize tools, namely the 'steelyard chair' (or 'statera medica'), the first graded thermometer, the hygrometer and the 'pulsilogium' (an early pulsimeter) to assess each of the many parameters involved in the complex calculation of the weight of the *perspiratio insensibilis*. All together these instruments form one of the most interesting and yet underestimated laboratory of the early modern science: *the laboratory of Santorio*. Thanks to his results Santorio's *statica medicina* extended the life of the concept of the balance converting it into the one of the proportion between two given quantities. As such he no longer needed to investigate the cause of a disease, being able to deduce the inclination toward health or sickness from the alteration of normal weight of *perspiratio*. According to Santorio's idea, the body is now capable of being studied as a clock, using *numero*, *pondere et mensura*. Shall we speak here about a 'paradigm shift' in the Kuhnian sense or we are rather dealing with a

plan of 'shared knowledge'? The answer to question depends on more general questions that intend to illustrate in my report this: What meant at the beginning of the sixteenth century measuring the body and its functions? What are the possible sources of such an approach and what kind of certainty Santorio's instruments were able to guarantee?

John Browne: ***Health measurement tools: when is good, good enough?***

Since its establishment in 1999 the National Institute for Health and Care Excellence in the UK has used the EuroQol health measurement tool to guide decisions about the cost-effectiveness of clinical treatments. In 2009 the English NHS broadened the use of the EuroQol to compare the performance of its hospitals when performing four elective surgical procedures (hip replacement, knee replacement, groin hernia repair and varicose vein surgery). The measurement properties of the EuroQol have received sustained criticism from psychometricians and clinicians but it contains to be used by a range of policy making bodies within and outside the UK. This is because scientific properties have become secondary to practicality when assessing the value of health outcome measures to end users. 'Practicality' here refers to the form of information required by policy makers, managers and economists when using the measures. The tools have been invented to help solve problems such as how to adjudicate between multiple expensive new technologies, and therefore it is a prerequisite that the output from these tools take a certain form, even if that requires significant scientific compromise. Meanwhile, a separate tradition within the social sciences, dominated by psychometricians, has largely ignored the practicalities demanded by end users and have produced scales that are increasingly impractical (longer and focused on narrowly defined constructs) but of a clearly superior scientific quality.

It is not clear how to reconcile the differing approaches to health measurement. In this talk a specific problem, the sensitivity of health measurement scales to clinically important differences between therapies, will be used to suggest a framework to guide end users when choosing one measurement paradigm over another. This framework requires answers to three questions. First, to what extent is one approach more likely than another to detect the signal of clinically important differences in outcome? Second, which perspective should we take when defining what constitutes a clinically important difference: societal or the perspective of the manufacturer of a therapy who will inevitably want ever more sensitive measurement tools to detect marginal benefits? Third, what counts as a good solution to a practical problem? Is it an absolute standard such as 'have we quantified the true difference between therapies as accurately as possible, regardless of the form that data takes''? Or is it a relative standard such as 'have we done better than the alternative', which might be anecdote, intuition, or simply allowing powerful vested interests to compete for the sympathy of policy makers for healthcare resources.

Eyja M. Brynjarsdóttir: ***Money as a Tool of Measurement***

The purpose of money is commonly considered to be to facilitate the exchange of goods and services. Its value, then, consists exclusively in exchange value, that which can be acquired in exchange for it, as opposed to use value. However, while exchange value seems to be solely quantitative, use value is largely qualitative. This brings us to a conundrum: How can money, as an exclusively quantitative instrument, represent goods and services that have qualitative value and be used to measure their value?

In his paper "The Moral Idea of Money" (1935),[1] Paul Goodman wondered how we could ever exchange one thing for another (Georg Simmel had brought up related concerns in his *Philosophie des Geldes* 35 years earlier[2]). While we could obviously never use a pair of shoes instead of a sack of flour, how come that we would ever consider the two things on a par so that we would be willing to exchange one for the other? Goodman argued that what really happens is that we take away the value and assign what he called indifference to the objects. Money serves the role of enabling such exchanges and in short, it might be said that when we use money to pay for goods or compensate someone for their work, we are assigning quantifiable value (amounts of money) to something that has qualitative value (human work, things that add to the quality of human life, etc.). What money is then measuring is not the actual qualitative value of the goods or services that are exchange, but this exchangeable quality; their *indifference*.

A related issue, dating back to Aristotle, is whether the ethical and the economical can ever be compared or whether they are by their nature concerned with values in different realms. If, as many have thought, quantifiable economical values are by their nature incommensurable with ethical values, or perhaps more broadly speaking with the normative realm or the qualitative realm, some further explanation may seem required for how money can function as a tool of measurement for goods and services that mainly have qualitative value.

My aim is to throw some light on how money as a quantifiable economical tool can be used to measure qualitative value, especially in terms of questions concerning the commensurability of value. I consider whether there is something unique in this respect when it comes to money or whether there are comparisons to be made to other standards of measurements, and whether the answer to the question about which things cannot be bought with money[3] has anything to do with difficulties regarding value measurement.

[1] Paul Goodman (1935), "The Moral Idea of Money", *The Journal of Philosophy* 32, 126–131.
[2] Georg Simmel (1978/[1900]), *The Philosophy of Money*, T. Bottomore & D. Frisby (transl.), Oxon: Routledge.
[3] Cf. Michael Sandel (2012), *What Money Can't Buy. The Moral Limits of Markets*, New York: Farrar, Straus and Giroux.

Stefan J Cano: ***Social Measurement: Three Problems, Five Challenges, One Solution***

The numbers generated by rating scales, questionnaires, and tests in the social sciences are increasingly used as measurements of the central dependent variables upon which high stakes decisions are made. This rise in profile has significant implications for instrument construction, evaluation, and selection, as well as for interpreting studies. The science underpinning instrument development (i.e., psychometrics) is well established. However, social measurement is hampered by three key problems: the continued use of instruments that are scientifically weak; a divided field of psychometrics with opposing methods, views, and criteria; and a lack of substantive theories underpinning latent variables. The first two of these problems can be illustrated by five current challenges:

- Which measurement paradigm to guide instrument development and evaluation? There are three main psychometric paradigms (i.e., CTT, IRT, RMT) that each encapsulate different approaches, different criteria for success and failure, and different considerations as to what counts as measurement.

- How to recognise the limitations of psychometric "statistics"? Irrespective of the psychometric paradigm chosen to develop or evaluate instruments, these statistical analyses do not inform us as to what is being measured. This is because statistical adequacy does not guarantee valid measurement. As such, psychometric statistics can be misleading when considered in isolation, and cannot be expected to produce consistently meaningful results when considered apart from qualitative scale content evaluations.

- How best to measure change over time? One of the most important properties of measurement instruments in the social sciences is their ability to detect change (commonly known as "responsiveness"). However, standard psychometric analyses of responsiveness generate illogical results and are flawed. Nevertheless, the relative responsiveness of competing instruments using these methods frequently drives instrument selection in social measurement.

- What sample sizes for scale development and evaluation studies? The appropriate sample sizes for ensuring adequacy in psychometric studies is unclear, the literature is vague and incomplete, but this has a major impact on the outcomes and inferences of instrument development studies.

- How to overcome the legacy of legacy instruments? There exists many thousands of measurement instruments in the social sciences, and although even in the instances where they are shown to be scientifically weak, they continue to be used for other reasons apart from the extent they are quality metrics.

The most pressing issue is the third problem; i.e. need for theory. To that end, there are two key requirements to advance our understanding of what instruments measure in the social sciences: explicit theories of the constructs; and explicit methods to test those theories. As such, instrument development

should be from the ground up (i.e., clear definitions), rather than the traditional top down approach (i.e., arbitrary statistically-driven methods of grouping items). This necessitates the complementary roles of qualitative and quantitative methods to ensure that construct definitions and subsequent theory determines instrument content, and psychometric analyses are used to assess the validity of the resultant construct theories. This "solution" provides the foundation to begin to address the current problems with, and challenges of, social measurement.

Laura M. Cupples: ***Applying Tal's Model-Based Account of Measurement to Nomothetic Quality of Life Measure***

Applying Tal's Model-Based Account of Measurement to Nomothetic Quality of Life Measures

Eran Tal has developed a model-based account of the epistemology of measurement. While Tal's work focuses on the measurement of time, he has suggested that this account might also apply to other measures as well, providing a unifying account of measurement across the physical and social sciences. I argue in this paper that his account does not extend unproblematically to measures in the social or medical sciences. As a case study, I examine nomothetic quality of life measures. Does the epistemic support models offer for these measures mirror that of Tal's physical measures, or are models playing a different role in these measures ?

According to Tal's model-based account, "a necessary precondition for the possibility of measuring is the specification of an *abstract and idealized model of the measurement process*" (Tal 2012). That is, our claims about measure validity and our judgments about measurement accuracy only become meaningful in reference to some model of the measurement process under consideration. Similarly, we can only meaningfully compare measurement outcomes when we have a model to contextualize those outcomes (Tal 2012).

Tal explains that he takes models to be abstract representations of local phenomena that are constructed based on theoretical, statistical, and pragmatic assumptions about those phenomena. He argues that models can function as mediators between abstract theory and concrete phenomena. They can also serve as instruments that help predict and explain the behavior of target systems (Tal 2012). This account suggests that, for Tal, models provide epistemic support for measurement primarily through theory articulation, i.e. taking scientific theory in concert with statistical and pragmatic assumptions and applying that theory locally.

However, many philosophers as well as thoughtful researchers and clinicians have complained that quality of life measures lack solid theoretical grounding. There is no generally agreed upon account of what well-being entails or how quality of life varies with life circumstances or our adaptation to those circumstances. There is also little theoretical grounding for our assumptions about how respondents understand and interact with these survey instruments,

i.e., how these measures tap into the phenomenon in question. Leah McClimans has argued that because respondents have varying conceptions of quality of life, they often interpret the questions posed by these survey instruments in unexpected and inconsistent ways (McClimans 2010).

It is clear that if models provide epistemic support for quality of life measures, it is not because they are mediating between abstract theory and concrete phenomena as Tal argues they do in physical measures. There is no well-developed or widely agreed theory of quality of life to serve as a target for mediation. Given this state of affairs, we should ask what it might mean to give a model-based account of quality of life measures and if it is still possible for models to provide epistemic support for claims about the validity, accuracy, and comparability of these measures.

Nadine de Courtenay and Fabien Grégis : ***Grappling with measurement uncertainties: philosophy of practice and practice of philosophy***

Calculations of measurement uncertainties heavily rely on probability theory. Their use of compound statistical models has recently nurtured extensive debates among metrologists on the proper way to calculate uncertainties. The International Organization for Standardization (ISO) published in 1993 a synthetic document which aimed to regulate the practices of uncertainty calculus in a large spectrum of scientific and applied fields: the *Guide to the expression of uncertainty in measurement* (GUM). However, far from closing the debates, the publication of the *GUM* has refuelled them after numerous commentators pointed out the lack of statistical consistency of the document. One of the most substantial points of contention revolves around the growing use of Bayesian statistics in metrology and their gradual overthrow of the traditional frequentist methods.

The draft we propose to discuss presents an analysis of the debates raised by uncertainty calculations in the metrological literature. After a brief survey of the core differences between the frequentist and Bayesian methods of measurement uncertainty and their philosophical ramifications, we would like to reach beyond the scientific and even the philosophical particulars of the debate to explore how the scientific, practical and philosophical agendas get articulated to one another within the metrological arena, and reflect on what this debate can teach us concerning the relation of philosophy of science with scientific practice. Indeed, our analysis shows that, while they often appear quite technical in nature, the metrological discussions on uncertainty explicitly deal with many philosophical topics and sometimes contain genuine epistemological developments which reveal that the practitioners themselves feel the need for firm conceptual and philosophical foundations.

More precisely, we would like to benefit from the insights of the participants to the conference on several topics among which we can mention the following ones:

– To what extent have the theoretical and technological means available to deal with the calculation of uncertainties contributed to shape the panel of conceptual and philosophical options discussed in the literature? To what extent other emerging possibilities make it questionable that a solution to the problem consists in choosing between the traditional frequentist-Bayesian philosophical alternatives? The ever-growing role of "Monte-Carlo" simulations in uncertainty calculations gives an example of how practical possibilities can drive the theoretical agenda much more that conceptual or philosophical (pre)conceptions do.

– The Bayesian approach did not impose itself on conceptual grounds. It emerged at the intersection of conceptual and philosophical concerns, technical constraints and practical goals driven by users' needs (researchers, engineers, industrials…). To what extent are the constraints we see at work in the conceptual and philosophical debates on uncertainty calculations in metrology of a same or different nature as the constraints that emerge in other areas of science? The kernel of this question lies in the fact that metrology addresses questions about the *methodology* of science and fixes the criteria of acceptability of our theoretical and practical statements, but is also oriented by practical and engineering goals. The publication of the *GUM* and its aftermath illustrate this role both on a theoretical and an institutional level.

Isobel Falconer ***: "No actual measurement … was required. No better method … has ever been devised". Cavendish, Maxwell, and the inverse square law of electrostatic attraction***

Traditionally, the foundation of the theory of electrostatics has been taken to be Coulomb's 1785 torsion balance experiments, reified as "Coulomb's Law". However, Coulomb's results, and interpretation, were frequently challenged, notably by Volta and Simon and, as late as 1836, by William Snow Harris.

In the first edition of the *Treatise on Electricity and Magnetism* (1873), Maxwell acknowledges Coulomb's experiments as establishing the inverse square law, merely to dismiss them again as demonstrating it only to a rough approximation. Instead he cites the observation that a charged body, touched to the inside of a conducting vessel, transfers *all* its charge to the outside surface of the vessel, as 'far more conclusive than any measurements of electrical forces can be' (#74). This assertion was based on mathematical proof that an exact inverse square law was a necessary condition for electricity to rest in equilibrium on the surface of a conductor.

The following year Maxwell acquired the hitherto unpublished electrical researches of Henry Cavendish, and found that around 1771 Cavendish had conducted a (fairly) rigorous test of the mathematically predicted null result concluding that the negative exponent in the force law could not differ from 2 by more than about 1/50. Maxwell and a research student, Donald McAlister, created their own version of Cavendish's experiment, achieving a claimed sensitivity of 1/21600. By the second edition of the *Treatise* (1881) his previous,

'…far more conclusive than any measurements…' has become '… a far more accurate verification of the law of force [than Coulomb's]' (#74). In his draft for the Cambridge Philosophical Society, Maxwell wrote, 'Cavendish thus established the law of electrical repulsion by an experiment in which the thing to be observed was the absence of charge on an insulated conductor. *No actual measurement of force was required. No better method of testing the accuracy of the received law of force has ever been devised*' (my emphasis).

More recently, Dorling (1974) has explored the sense in which it was rational for Cavendish and Maxwell to generate an entire law from a single (null) data point, while Laymon (1994) has pointed out the circularity of Maxwell's argument and located the actual measurement in the testing of the sensitivity of the electrometer.

Taking Laymon's and Dorling's critiques on board, and drawing on works and papers by Maxwell, Kelvin, Tait and Harris, this paper will examine how Maxwell and his contemporaries understood what Cavendish had done, what they thought the null method achieved, and the value to them of recreating the experiment.

William P. Fisher, Jr.: ***Toward A New Dialogue between the Natural and Social Sciences***

Given the demise of the universality formerly assumed characteristic of singular science, what potential might there be for a new, mutually informative and productive dialogue between the natural and social sciences? Four developments in this regard suggest new possibilities for the future.

First, instead of modeling themselves on the natural sciences, the social sciences should focus on the playful model-based reasoning practices used in everyday thinking and collective cognition that have been extended and refined in the natural sciences. The question becomes one of how to tap and extend everyday model-based reasoning in useful ways in psychology and the social sciences, instead of continuing fruitless efforts aimed at defining these fields in terms of, or in opposition to, the natural sciences.

Second, there are at least eight areas of advances in measurement research and practice that arguably should have had revolutionary, paradigm-shifting consequences in psychology and the social sciences. The eight points concern (1) the mathematical form of the models/laws, (2) Rasch's parameter separation theorem and proofs connecting it with axioms of additive conjoint measurement, and with the necessity and sufficiency of scores for the estimation of those parameters, (3) widely repeated experimental results producing linear units of measurement that retain their identities across samples, over time, and across instruments, (4) the integration of qualitative and quantitative data, methods, and results, (5) the reproduction of SI length, mass, and density units from ordinal observations, (6) predictive theories for multiple constructs explaining variation in item difficulty or agreeability, (7) improved measurement efficiency,

practicality, and meaningfulness, and (8) the widespread use of these models and methods in over 40 years of unpublished proprietary applications.
Third, the futile positivist expectation that data alone, and the equally fruitless anti-positivist expectation that theory and data together, are sufficient to the task of shifting paradigms are countered by the post-positivist recognition of the essential role played by portable instruments embodying theory-data relations expressed in a standard unit and common language distributed throughout a network of end users. A systematic post-positivist approach to devising consensus unit standards, ensuring the quality of measures traceable to them, and putting them in play in larger social, scientific and economic contexts builds on the eight points in part two and extends them in terms of the cognitive processes described in part one. Processes of noise-induced order that characterize the stability of a wide range of types of standards networks are of special interest.

Fourth, a start to a new dialogue between the natural and social sciences began in 2008 with presentations and publications co-authored by psychometricians and the physicists and engineers involved in the metrological work of international weights and measures institutes. Lessons in the fundamental importance of metrological traceability need to be learned in psychology and the social sciences, and the construction of invariant linear measures and standard units with known uncertainties from ordinal observations needs to be learned in metrology. Parity in this process of mutual teaching and learning across the sciences may create the context for exciting new innovations on a grand scale.

Jim Grozier: ***Are Angles Dimensionless?***

Angles are often said to be dimensionless – in other words, pure numbers. Yet they certainly have units – radians, degrees, minutes, revolutions etc – and hence, using Bridgman's construction for assigning dimensions, in which units and dimensions are inextricably linked, one is forced to conclude that they also have dimension.

In science, however, angles are commonly measured in terms of the *radian*, defined as the angle which subtends an arc of 1 unit of length at a radius of 1 unit of length, so that the numerical magnitude of an angle in radians is obtained by simply dividing arc length by radius. This means that angles in radians are not affected by changes in the unit of length, nor, naturally, are they affected by changes in the units of mass and time either, and hence in a 3-dimensional "mechanical" (MLT) dimension space they do appear to be dimensionless.

If we approach the measurement of angles from first principles, using Norman Campbell's method of specifying an ordering relation, a combination rule, and a unit, we see that our unit can in fact be anything we like, and there is nothing special about the radian. The radian was actually a fairly late arrival on the scene, having been first used by Euler in 1765 and named by James Thomson in the late 19th century.

I argue that the apparent contradiction can be avoided if we recognise that the above definition is that of a *unit*, and not of the concept of angle in general, and that angle should indeed be regarded as a base quantity with dimension. There are parallels here with the 19th century treatment of electrical quantities, in which Maxwell's definition of a unit of charge was also regarded as a definition of electricity itself.

Treating angular measure in this way has both advantages and drawbacks. In practical terms, a general format that is invariant to change of units – a property John Roche has termed "gauge invariance" – would introduce a dimensioned constant into the relationship between angle, arc and radius, as does Coulomb's Law today in the case of electric charge. When using radians, this constant would have the same dimensions but a numerical value of unity. However, if such a gauge-invariant system were adopted, the constant would appear in the Small Angle Approximation, and hence also in the derivatives of trigonometric functions, making any associated derivations rather unwieldy.

On the plus side, though, recognising angle as a fourth mechanical base quantity would remove an unfortunate degeneracy whereby some "rotational" quantities appear to have the same dimensions as other, quite distinct, quantities – such as torque and energy, or angular momentum and action. As Roche has pointed out, this degeneracy has been used as a counter-argument to the view of dimension as revealing something of the "essential nature" of a quantity.

Godfrey Guillaumin: ***From circles to ellipse: Epistemology in the historical development of measurement procedures in early modern astronomy***

An historical analysis of ancient practices of measurement in astronomy up to early modern era with Kepler shows an interesting regulative system of epistemic justification. From Hipparchus and Ptolemy to Kepler, measurement procedures for knowing planetary movements and cosmological distances and sizes were explicitly elaborated through a system of three intertwined components; namely, data, geometry, and ontological considerations. Even though these components have been well studied (Swerdlow & Neugenbauer 2012; Neugebauer 1972; Gingerich 1993), their underlying and multidimensional epistemology has not. It can be shown that through its historical development, this system was constituted and uphold by a complex dynamical regulative system of justification, which can be conceived in two directions: an horizontal and a vertical axes.

In the vertical axis, these three components exhibited their own and separate criteria of justification; *i.e.*, elements for reliable data are different and epistemologically independent from elements for justification of a specific geometrical configuration, etc. In the horizontal axis, there were feedback relations among these three components that conforms a kind of dynamic (or quasi iterative, see Chang, 2007) justification; *i.e.*, calculated data and detected data should match each other; or ontological considerations should be satisfied by geometrical configurations, etc. Because of historical and philosophical

relevant reasons it is reasonable and fruitful to consider this complex system like a dynamic regulative system of justification.

One relevant feature of this regulative system was that at the beginning of its historical conformation not all of its components and the relations among them were strongly and well established; in fact, just a few of them were explicit and well understood. It was through its historical development that almost all of its components and their relations were elaborated, modified, and mutually adjusted to finally accomplish an almost perfect integration (*i.e.*, a *cognitive integration*) in Kepler's procedures. Thus, globally the epistemic justification of Kepler's procedures represented an epistemological "phase change" in comparison with its predecessors. This case study clearly shows that the epistemic justification underlying the procedures of measurement is adequately conceived in terms of neither foundationalism nor coherentism alone. Therefore traditional epistemology is inadequate to study the epistemology underlying of the measurement procedures and its developments. Traditionally, the epistemology of measurement procedures has been insufficiently studied, with some exceptions (Boumas, 1999; Tal, 2012). The aim of this talk is to offer a rough account of the dynamic regulative epistemology that historically conformed the modern astronomical procedures of measurement.

Conrad Heilmann: ***The Making of Economic Measurement***

Numbers are ubiquitous in economics: numbers such as those that express GDP, unemployment, inflation, discount rates, and many more are frequently in the focus of interest. Yet, it is unclear to what extent such numbers really express quantities and whether they are the result of a procedure that one might call measurement. This is a pertinent problem not only for theoretical reasons, but also practical ones: it is impossible to imagine policy-making without numbers that are produced by economists. Yet, in policy-making such numbers can take on a life of their own: often, they are used without paying attention to what kinds of assumptions and limitations they entail.

In this paper, I defend a proposal for necessary and sufficient conditions that ensure that numbers used in economics are based on measurement. More specifically, I defend the following normative conception of good measurement in economics.

Economic Measurement. For any number used in economics, it is based on measurement if at least one of the following conditions holds:

(i) The number is based on the conceptually sound combination of quantities.
(ii) The concept expressed by the number is numerically representable. (iii) The number is based on a reliable and objective procedure.

This paper shows how these conditions embody important accounts of measurement from philosophy of science, investigates their relations, and demonstrates their applicability to important measurement problems in

economics.

In a first step, I show how each of the conditions captures important aspects of measurement, as discussed in the philosophy of science literature.

(i) Concerning the conceptually combination of existing quantities. There are many measurements of particular facts that can be combined in order to form a comprehensive picture of a given situation or system. Thus, economic measurement can be understood as the formation of numbers out of other quantified concepts. I demonstrate how the literature on concept formation in the philosophy of science and the philosophy of language can provide the necessary tools to make this criterion precise.(ii) Concerning the numerical representability of concepts. Minimally, in or- der to motivate a number, the concepts that the number expresses have to be representable. The Representational Theory of Measurement (RTM) provides tools to make this criterion precise. According to RTM, measurement consists in the construction of mappings between empirical structures and numerical structures. RTM expresses conditions of measurement as axioms on set-theoretic structures, and allows us to express precisely what kinds of numbers can in principle represent them.

(iii) Concerning the reliable and objective measurement procedures. In order to be able to replicate measurements and to compare numbers over a vari- ety of contexts and times, reliable and objective procedures are needed. In the literature on measurement instruments and measurement science, measurement is primarily a matter of having reliable and objective procedures. I demonstrate how the literature on model-based account of measurement can provide the necessary tools to make this criterion precise.

In a second step, I clarify the logical relationship between the three conditions, by applying them to competing accounts of the measurement of wellbeing, utility, welfare, and happiness in economics. I show that the vast literature of competing approaches in this area can be both captured and taxonomized by the account of economic measurement.

Gil Hersch: ***Filling the void: the need for accounts in defense of well-being measures***

If policy-makers want to determine the effects their policies have on well-being, they need some way to measure well-being. Social scientists have developed a variety of methods for measuring well-being (e.g. subjective well-being measures, economic measures, and objective indices). However, these methods sometimes significantly diverge in their assessments of the effects of a given policy on well-being. Because these measures play a central role in guiding public policy, it is important to figure out how to decide between them. The ongoing debate regarding well-being measures is an interesting case study in the philosophy of measurement because it allows us to both examine how the debate evolves as well as to influence how it does so.

Philosophers have debated how to characterize well-being for a long time and no resolution is soon to come. But even if we accepted some theoretical account of well-being, we face the further problem of representation—determining which measure best represents the given characterization of well-being. As both Eric Angner (2011) and Matthew Adler (2013) demonstrate, it is *possible* to view well-being as constituted by, for example, preference- satisfaction and still consider both subjective well-being measures and traditional economic measures as valid representations of well-being characterized as preference. That multiple measures might be considered appropriate representations of a single way of characterizing a concept is not new in the philosophy of measurement literature (see Chang (2004), Cartwright & Bradburn (2010), Tal (2013)), though this realization is often lost in the well-being measurement debate.

Because social scientists often conflate the characterization of a concept with how that characterization is represented, they often do not provide a clear account of why their chosen measure is a good operationalization of the well-being concept they have in mind. However, having such accounts are important because how a characterization of well-being is operationalized by a measure determines how well the measure corresponds to the characterization of the well-being concept. This correspondence is what we need to examine to adjudicate between the different measures. If we want to adjudicate between the measures to decide which measure to use to guide well-being policy, a defense of how the operationalization fits the concept as characterized is required.

Daniel Kahneman's work on objective happiness (2000, 2004) comes close to providing a defense of operationalization that does this, albeit for what Kahneman himself takes to be only a component of well-being—experienced utility. Kahneman characterizes experienced utility as a summation of moment-utilities and uses his Day Reconstruction Method to measures moment utility. In doing this Kahneman provides a concrete image of the abstract concept—experienced utility, and matches this image with an actual system of entities and operations— the Day Reconstruction Method. Kahneman's discussion of his operationalization of experienced utility can serve as a model for how to argue for a match between a characterization of a concept and its representation. Without such arguments being made explicit, it will be difficult to make progress in the well-being measurement debate.

Alistair Isaac: ***The Rise and Fall of "Volume": A Case Study in Psychophysical Measurement***

This talk examines the changing norms of psychophysical measurement through the example of auditory "volume." I argue that the trajectory of changing interest in volume is best explained by a change in attitudes toward our epistemic access to phenomenal attributes intimately connected to the introduction of multidimensional scaling techniques.

In the early 20th century, it was discovered that subjects could consistently assign volume as an attribute of auditory stimuli distinct from loudness or intensity. This prompted speculation that volume might be a distinct auditory quality, or even a truly amodal perceptual quality assignable to stimuli for any sensory modality (might their be "smell volume," for instance?). The lynchpin of research on volume was Stevens' (1934) discovery of a systematic relationship between volume, loudness, and auditory "density" (Boring, 1942). Stevens' result motivated further systematic investigation of volume in the following decades, e.g. Thomas' (1949) equal volume contours, Terrance and Stevens' (1962) volume scale, and Gulick's (1971) determination of subjective volume scales (Gulick, Gescheider, and Frisna, 1989).

Nevertheless, by the late 20th century, research on volume as an auditory quality had largely disappeared. If volume is mentioned at all in contemporary psychophysical texts, it is as a mere historical footnote. While the legitimacy of the claim that volume is an attribute of sound experience is typically not questioned, neither is it taken to be of interest for a modern study of sound perception. What explains this change in attitude toward volume?

Two possible explanations can be found in the classic criticisms of late 19th and early 20th century psychophysics: (1) measurement for measurement's sake (i.e. without a guiding theoretical framework) is inadequate for approaching truth—mere measurement of volume is no indication it is meaningful as a theoretical concept for understanding auditory perception; (2) overtraining of subjects contaminates data, producing "robust" but essentially meaningless effects—consistency in volume judgments was a mere artifact of overtrained subjects. I demonstrate that neither of these criticisms applies convincingly to volume research.

I argue that the decline of interest in volume may be better understood by placing it in the context of the change in data analysis techniques for psychophysics, in particular the introduction of multidimensional scaling. The program of Titchener, Stevens, and other early 20th century psychophysicists required the experimenter to posit fundamental perceptual attributes before they could be measured through experiment. In contrast, multidimensional scaling allowed researchers to instruct subjects to merely make judgments of similarity—the perceptual attributes which guided those judgments could then be derived mathematically through data analysis. The upshot of this new technique was that access to perceptual quantities was (perceived as) no longer hypotheticodeductive, but rather purely empirical. Nevertheless, I conclude, the new empiricism of psychophysics fails to answer key questions about the nature of perceptual experience—for instance, the reduction of volume and loudness judgments to a single "dimension" via multidimensional scaling leaves unexplained the initial data point, that the two qualities are systematically distinguishable by subjects.

Shaul Katzir: ***The second in the long reform of the SI base physical units***

The definition of the second played a unique role in the reform process of the international system (SI) base physical units. The reform, whose last stage has recently been approved, redefines these units by universal constants of nature, independent of specific material exemplars (like the kg prototype) and particularities of our planet. In principal, the suggested system allows any well-equipped laboratory determine the values of the units form nature, without a need to refer to a certificated standard.

One the one hand, time was the first unit that suggested the need for reforming the SI original definitions, and to offer more exact methods for its definition. Already in the 1920s astronomers and physicists suspected that the average daily rotation of the earth, which defined the second, is slowing down, and thus does not fit the role of a base stable unit. New highly accurate tuning fork and quartz clocks not only allowed unprecedented accuracy, but also independence from the earth's particularities. A property used in their early application in the field. These new techniques suggested replacing the earth's rotation in defining time. Still metrological reforms take long time. While it was clear that the earth's rotation varies, the first redefinition of time (from 1960) did not rely on another phenomenon but on its rotation in the year 1900. This definition contradicts the spirit of the future SI reform, since, unlike the original definition, it did not allow 'realization' of the second. Eight years later the unit was redefined by the frequency of transition in caesium atom, relying on the atomic clock technology in use for more than thirty years, which allowed 'realizing' the second according to its definition. As it was based on a universal constant of nature, the new definition fits the spirit of the reformed system.

On the other hand, the invariants by which the other base units are defined are not only universal, but are also 'fundamental constants'. By this, metrologist mean that the constants are not linked to any particular material system, or phenomenon, but reflect general relations in physics. These constants, like Plank's and Boltzmann's, are invariants of physics as such; they are linked to its fundamental theories. Time is exceptional in the new system as it relies on a particular atom, whose mere existence is not regarded as basic and necessary, unlike the constants of the fundamental theories of physics. Yet, the second is the cornerstone of the system since the definition of all other units are relied on it. The two properties might seem as contradictory, but actually the pivotal role of the second origins in its arbitrariness. The fundamental physic constants refer only to relations between magnitudes. Yet, to determine a unit one has to stipulate a magnitude above the relations, as done in defining the second. The second, thus, points at the limits of the attempt of basing metrology on 'natural units' alone.

Luca Mari: ***Measurement as a complex process: a structural view***

Measurement is overloaded of opposite stereotypes. On the one hand, the acquisition of information about physical quantities by means of sensors historically led to interpret it as a merely experimental activity, such that, once a suitable physical transduction effect is identified and embedded into a device, only technical operations are required for making the device interact in a controlled way with the object under measurement and then reading the outputs of the device itself. On the other hand, measurement has been formalized as representation in such an abstract way that it has become definitionally equivalent to morphic mapping, thus neglecting the assumption that its epistemic features of objectivity and intersubjectivity are guaranteed by a whole metrological system, which includes measuring systems traceable to measurement standards via calibration chains.

The practice of developing measurement processes for more and more complex systems has fostered an intermediate view, in which empirical and informational components are intertwined and driven by pragmatic targets: designing and then performing a measurement process involves goal-setting, theoretical assumptions, modeling, experiments, calculation, information interpretation and decision.

We propose here a conceptual framework (as sketched in the diagram below) aimed at identifying the main tasks expected in a measurement and their procedural inter-relations. The framework complexity, due not only to the multiplicity of the components but also to the structural feedback among them, explains the complexity of obtaining and validating measurement results. And while the naive black box, atomic models of measurement remain pragmatically acceptable whenever the required quality of results is much lower than the capability of the adopted methods and instruments, the framework provides a systemic illustration of the theory-ladenness of measurement and a justification of measurement uncertainty as epistemic tool for conveying reliable information on the reliability of the quantity values produced by the process. It shows that the public trust socially attributed to measurement has structural reasons, for which being a morphic representation or producing quantitative data are at most necessary, but definitely not sufficient, conditions.

As measurement is an acknowledged link between philosophy and science and technology, this framework might be intended as a tool to ground a shareable conceptualization able to highlight the epistemic and empirical conditions required to acquire knowledge from metrological information from experimental data.

Leah McClimans: ***Measurement in Medicine And Beyond: Quality of Life, Blood Pressure and Time***

Quality of life measures are popular with health policy makers in large part because of their ability to function as quantitative measuring instruments while also providing the patients' point of view. From a development perspective this attraction requires that these measures are epistemically and ethically sound. This double burden has proven difficult to achieve and these instruments have received significant criticism, mostly from those who develop and work with them. For instance, in 1995 the *Lancet* ran an editorial cautioning the use of these measures as end points in clinical trials, in 1997 Sonia Hunt's editorial in *Quality of Life Research* argued that they are misleading and probably unethical; more recently in 2007 Jeremy Hobart and colleagues argued in *Lancet Neurology* that almost all current measures are invalid. In my own work I have argued that they are invalid and difficult to interpret at least in part because they do not accurately represent the patients' point of view.

In this paper I ask why quality of life measures face these challenges. One explanation that researchers commonly invoke is that quality of life measures lack a 'gold standard' and are thus more difficult to measure than physical properties such as blood pressure. In what follows I examine and reject this explanation and offer a different one: the problems that quality of life measures encounter arise because quality of life lacks a theory that provides a representation of the measurement interaction, i.e. the relationship between the quality of life construct and its instruments.

To make this argument I examine Carolyn Gotay's argument that the lack of a gold standard makes quality of life more difficult to measure than "concrete" phenomena such as blood pressure. The kinds of difficulties that most commonly confound the accuracy of blood pressure readings stem from short-term biological fluctuations such as those that occur from white coat effect, e.g. anxiety. Researchers thus attempt to improve the accuracy of these readings by developing instruments and protocols that will reduce the effects of anxiety. But much depends on the belief that anxiety *obscures* (i.e. is not part of) one's true blood pressure reading. To be sure we have multiple overlapping theoretical frames that support anxiety's role as a confounding variable vis a vis blood pressure. But the same cannot be said when we examine a popular candidate for a confounding variable of quality of life: response shift.

To better understand the relationship between accuracy, measuring instruments and the variables that can confound them I turn to Eran Tal's account of measurement accuracy in the context of time. Tal shows how metrologists construct astonishingly accurate clocks without a gold standard. In lieu of a gold standard Tal argues that metrologists rely on what he calls the Robustness Condition (RC). I argue here that quality of life instruments are unable to even approximate the conditions of Robustness because it lacks the theoretical assumptions necessary to model the quality of life construct. Without these theoretical resources we cannot identify confounding variables or estimate their associated corrections. While the lack of theory is not unacknowledged in quality

of life research it is not given the attention it deserves. The lack of a theory is not persuasively linked to the need for one.

Teru Miyake: ***On Theory and Operationalization: The Seismic Mechanism Controversy and its Resolution***

This paper will examine a well-known controversy in the history of seismology over the question of exactly what happens physically at the source (i.e., the point of origin) of an earthquake. Although evidence bearing on this issue was available, and one group of seismologists had essentially the correct answer since the 1930's, the controversy was resolved only in the 1960's. Some seismologists (e.g., Frohlich 2006), looking back on this history, have been puzzled as to why the controversy took so long to resolve. As with most historical episodes, the reasons are complicated, but I think the resolution of this controversy is of interest to philosophers of measurement. I will argue that it was only resolved when the mathematical apparatus that was needed for operationalizing the seismic mechanism was developed, allowing for the measurement of parameters that represent the source mechanism.

The seismic source is usually located deep within the earth, so what happened at the source to cause the earthquake must be inferred from seismic wave observations made at the surface of the earth. Seismologists now usually represent the seismic source mechanism as an abstract mathematical object called a "double couple". From the 1930's through the 1960's, several different groups of seismologists attempted to infer the seismic source mechanism from observations of first motions of seismic waves at various seismographic stations on the earth's surface. Some concluded that the seismic source should properly be represented as a double couple, while others concluded that it should be a single couple. This controversy was well known among seismologists at the time, and it continued into the 1960's.

What is interesting to philosophers is that the controversy was not resolved straightforwardly through, for example, empirical observations. The paper that is widely perceived to have resolved the controversy once and for all is Burridge and Knopoff (1964). This paper is completely theoretical, the main result of which is showing that the elastic waves that radiate from a sudden dislocation along a fault are, under certain conditions, exactly equivalent to those that radiate from a double couple. Perhaps the most important aspect of this paper, however, is that it first developed the mathematical apparatus that would enable the measurement of stable parameters that represent the seismic mechanism of any particular earthquake, in the form of the "moment tensor", *if one makes the assumption that* the seismic source is a double couple. I will attempt to make the case that the controversy was resolved by the fact that making the assumption that seismic sources are double couples allowed for the operationalization (in Hasok Chang's sense) of the seismic mechanism, and ultimately it was the success of succeeding research programs in measuring moment tensors for large numbers of earthquakes, and the usefulness of these measurements in further seismic applications, that justified this assumption.

Alfred Nordmann: ***Charles Sanders Peirce and the Epistemology of Measurement***

When C.S. Peirce juxtaposes counting and measuring, counting is the capacity to discretize items or events, whereas measuring reflects the continuity of qualities. Especially interval and ratio measurement scales correspond to Peirce's conception ofcontinuity: Between any two points on the scale, there are infinitely many other points (Peirce refers to this property of a continuum as Aristolecity), the upper and lower bounds of a scale are constitutive limits that can only be approximated (Peirce refers to this property as Kanticity), and the elusive infinitesimals correspond to real magnitudes. Accordingly, Peirce does not consider measurement as the determination of a value which is then assessed in terms of measurement error. Instead, the distribution of measurement-error is part of the instrument which extracts a signal from a continuum and thereby always measures something that is subject to chance fluctuations. By the same token, any one measurement is only one of an in-principle infinite series of measurements. As measurements become more precise, there is a gain in sensitivity to detect the elements of chance that produce small deviations in the signal. Thereby, the act of measurement brings these elements of chance into view which, by definition, are as yet unknown but knowable. The increase of precision thus indicates a direction of research.

This idealized picture corresponds to Peirce's epistemology, theory of science, and his evolutionary conception of mind and reality. Since the physical meaning of the error-curve cannot be attributed in an *a priori* manner to external reality but only to a signal that is produced through the interaction of researcher, instrument, and world, a closer look is required at Peirce's views of measurement practice. These are informed by his scientific experience with measurement – beginning with his early experimental work in sense physiology and a series of measurements (with Jastrow) to critique the common view that there are thresholds of human sensitivity below which differences of sensation cannot be detected, continuing with his gravimetric measurements for the United States Coast Survey, including his role as the head of the Office of Weights and Measures. The other two presentations in this panel will be considered before the backdrop of Peirce's measurement practice as well as his epistemological views on measurement.

Guilherme Sanches de Oliveira: ***Data, Simulation, and Representation: Against Model-Based Measurement***

An important question arising from the recent philosophical literature on measurement concerns how measurement relates to other scientific activities. In this paper I examine the link between measurement and modeling, and I argue that scientific models are not means for measurement.

I begin by revisiting the distinction between data models and experiment models (or "models of data" and "models of phenomena," Frigg & Hartmann 2012), which I differentiate functionally for fulfilling *communicative* and *inquisitive*

purposes, respectively. I then draw a provisional connection with measurement by suggesting that, although some data models can be seen as organizing *measurement outputs*, a more interesting question is whether models (particularly experiment models) can be seen as a form of *measurement procedure* (following Tal 2013; also Morrison 2009 for simulation-based measurement).

Answers to the question "do models measure?" assume prior commitments about the nature of models and their relationship with the world. Current orthodoxy (the "representational view of models") holds that models are used for indirect investigations of real-world target phenomena because models *represent* their target (e.g. Morrison & Morgan 1999). Within this view, our question may be interpreted as asking "do models in themselves measure?," and accordingly be answered positively or negatively according to particular views of *representation*. Positively, "models in themselves measure" in virtue of the representational model-target relationship being dyadic (e.g. traditional *isomorphism* and *similarity* accounts). Negatively, "models do not in themselves measure" in virtue of the representational model-target relationship being triadic: scientists are a necessary component because they establish the representational mapping (e.g. updated versions of isomorphism (van Fraassen 2008) and similarity (Giere 2004), and also *inferential* (Suarez 2003), *interpretational* (Contessa 2007), and *semiotic* (Knuuttila 2010) accounts). This type of negative answer thus amounts to a positive answer to a different question, namely "can we measure with models?," such that measurement is possible given the proper three-place representational relationship holds.

But the representational view of models is riddled with problems, both general (e.g. accommodating idealizations and abstractions) and particular (e.g. the anything-is-similar-to-anything-else-in-some-way objection to pure dyadic representation, and the related anything-can-represent-anything-else-for-someone objection to triadic representation). A powerful but still widely neglected alternative to the representational view is the *artefactual* view of models, according to which models are not truth-bearing representations, but rather autonomous tools or devices. As such, models can be more or less useful for certain purposes, but they are useful in virtue of what they *present* rather than what they *re*present. Rather than being simply reduced to objective properties of the model, the "presentational force" of models is best understood in relational terms as encompassing models' affordances and the scaffolding role models play in understanding a target. After briefly articulating how the (anti-representational) artefactual view is a superior approach to scientific modeling, I argue that it can only answer the question "do models measure?" negatively. Even if models are built upon measurement outputs, because models are not representations of their target but rather autonomous artifacts, model-based measurements are never measurements of the target, only of the model itself.

Ilke Ercan: ***Calculating Fundamental Bounds and Measuring Limits in Nanoelectronics***

Emerging nanotechnology proposals promise to overcome limitations of current technologies. However, the realizations of such proposals face a broad range of challenges. In nanoelectronics, energy dissipation is arguably the most critical of all. There are various components of energy dissipation in a computing circuit, and a part of this dissipation comes from the unavoidable cost of implementing logically irreversible operations. This stems from the fact that information is physical – it is encoded in the physical states of a system – and manipulating it irreversibly requires energy. The unavoidable dissipative cost of losing information irreversibly fixes the fundamental limit on the minimum energy cost for computational strategies that utilize ubiquitous irreversible information processing. A relation between the amount of irreversible information loss in a circuit and the associated energy dissipation was formulated by Landauer's Principle in a technology independent form. In a computing circuit, in addition to the information-theoretic dissipation, other physical processes that take place in association with irreversible information loss may also have an unavoidable thermodynamic cost that originates from the structure and operation of the circuit. In conventional CMOS (Complementary Metal Oxide Semiconductor) circuits such unavoidable costs constitute only a minute fraction of the total power budget, however, in nanocircuits, it may be of critical significance due to the high density and operation speeds required.

The lower bounds on energy, when obtained by considering the irreversible information cost as well as unavoidable costs associated with the operation of the underlying computing paradigm, may provide insight into the fundamental limitations of emerging technologies. In order to establish these limitations, measurements are important to support the fundamental bounds we obtained for models of nanocircuits. In view of Peirce's theory of measurement, the presentation therefore addresses the issues surrounding the measurement of the energetic equivalent of information. Why do various claims to verify the Landauer principle by experiment fail to satisfy? In light of the difficulties of measuring the energy equivalent of information processing, what is the physical meaning of the lower bound? How might simulation modeling provide indirect proof by measuring the "real" energy costs in a technical system and showing co-variance with the derived bounds?

Lara Huber: ***How constant are values of measurement? The case of baseline values***

Precision measurement presupposes stability: Units and other standards of measurement are said to be robust and of constant value. It was therefore, that Charles Sanders Peirce addressed the weakness of artefact standards, given that prototypes of measurement such as metallic bars might change significantly due to time and use. To identify a valid baseline value could be regarded as another major challenge of measurement, especially in the biomedical sciences. Baseline values are necessary preconditions to allow for measurement in general and to

identify significant changes of therapeutic intervention in particular. For instance, a set of data found at the beginning of a study is used for comparison with later data in order to establish a relative rather than absolute meaning to data. Pharmaceutical trials often try to achieve statistical significance (precision) by introducing a placebo baseline period at the beginning of a given clinical trial. Taking Peirce's general considerations about precision measurement and constancy of values as a starting point, this paper explores iterative processes of setting standards (here: baseline values) and validating practices of measurement in the biomedical sciences. The paper elaborates on practical considerations of ensuring sufficient reliability with regard to clinical trial design. The paper presents different strategies of identifying and defining baseline values, including such that refer to physical data of a single subject and, e.g., are mandatory for diagnostic purposes or therapeutic monitoring. Against this background, Peirce's general considerations are revisited with regard to the conceptual framework of evidence-based medicine (EBM) that regards the very design of randomised-controlled trials (RCT) as gold standard for demonstrating clinical efficacy.

J. Brian Pitts: Empirical Significance in Some (Nearly) Fundamental Physical Theories

Fundamental physical theories' empirical content is not always transparent (even neglecting quantum mechanics!). It isn't clear why rods and clocks can exist, and empirical significance is often problematized by a multiplicity of physically equivalent descriptions and/or algorithms for dealing with them.

First, it isn't obvious how objects that we treat as rods and clocks---complicated physical devices that they are---manage to conform so well to geometrical structures applicable in the first instance to primitive infinitesimal idealizations. As probably first suggested by Poincaré as a motive for conventionalism, it could have been (or even might be) the case that different matter types exhibit different geometries. While Einstein defaulted to treating rods and clocks as primitive in General Relativity on pragmatic grounds, he continued to worry about the justification for doing so. The idea of rival geometries in one world has reappeared in various ways in Brans-Dicke scalar-tensor theories (1960s) and in philosophers' debates on chrono-geometric significance (Grünbaum, Harvey Brown). Under what circumstances is there a simple answer to questions about chrono-geometric significance and the possibility of rods and clocks? Much of the answer lies in coupling to matter, which usually is assumed to see exactly one geometry. However, physicists' work on multi-metric theories of gravity since 2010 has discussed the possibility that matter couples to more than one metric. If such theories are viable, then new experimental possibilities or even Finslerian geometry (a quartic analogue of the Pythagorean theorem) might arise.

Second, modern physical theories often include different descriptions of the same state, with the descriptive conventions varying over space and time ('gauge freedom'). Measurable quantities shouldn't depend essentially on descriptive conventions. Must observables be gauge-invariant, or merely gauge-covariant

(translatable)? Does it matter whether the different conventional gauge choices are external (relating different place-times) or internal (each point isolated)?

Third, some formulations of important physical theories introduce auxiliary quantities not used in more economical formalisms. In Hamiltonian electromagnetism, some of these auxiliary quantities are often called "the electric field", a name suggesting direct empirical content. How, if it all, does one measure quantities which in some sense aren't even needed? What if auxiliary quantities disagree with more fundamental quantities about gauge choices, and the auxiliary quantities apparently behave better?

Fourth, some theory formulations, such as Hamiltonian General Relativity, have been claimed to lack real change, because such change as they have is diagnosed as an artefact of arbitrary descriptive choices. Yet clearly we observe change. Do we manage to observe something that is not really there?

These questions and their answers reveal an interplay between theory and experiment. Some of the mysteries above are resolved by diagnosing theorists' mistakes---mistakes of which the existence (though not the nature) is made certain by measurement practices involving voltmeters, street maps and British Summer Time. But attention to a broad range of possible theories of space-time and gravity calls attention to the metaphysical possibility, perhaps a live possibility in the real world, that measurable geometry might be ambiguous, material-dependent, or non-Pythagorean.

Anna Echterhölter: ***Measurement as Negotiation: Jacob Grimm on the Procedural Quantification of Possessions, c. 1810***

In his lesser known studies on the History of Law the German editor, collector, and linguist Jacob Grimm listed pages and pages of examples of "measurements" from a rural context. The curious quantification routines he encountered in old law decrees utilized sounds, wielded objects or the movements of animals to determine over possessions. The eldest of three brothers, for example, would inherit only a little more than a third of the estate. But this small share of land was determined in a rather surprising way: It comprised the surface a flying rooster crosses in the air before landing on the ground. Sometimes the animals themselves were measured. The strength of a hen proved crucial on certain festive dates, when a so-called 'tax-hen' had to be delivered to the local authorities. These would only accept the animal as valid payment, when it was still capable of flying onto a barrel in its own might. The procedures provide a script of action as well as a rough frame indicating amounts and values. But what is generally missing is the exact determination of the quantity in question. Grimm highlights this lack of precise information as the crucial point, since it opens a space of negotiation between interested parties in each individual setting.

His account of rural measurements lies at the core of Grimm's theory of ancient law and resonates with democratic issues of the period. On a first level this

critique is directed against the new metric system. Exact numbers replace a democratic procedure of settling disputes. On a second level Grimm expands his argument to the character of the French Code civil and the imposed waves of administrative modernization: German principalities, which had suffered recent defeat by Napoleon's troops, had to introduce the metric system and profoundly reformed jurisdiction at the same time. It is here that Grimm's most interesting argument sets in: the procedural, descriptive, or ‚poetic' measurements are portrayed as exceeding more rationality than the metric ones of the Enlightenment. In the cases of economic quantification necessary in rural contexts, the rational law brought by the French Revolution discourages negotiation. In opposition to this, the regime of the old *ius commune*, at least as Grimm and a considerable fraction of the German Historical School of Law had it, facilitated democratic solutions.

It is the main aim of the paper to outline this constellation that subverts the dichotomy of qualitative versus quantitative measurement. Quantification is now opposed to explicit negotiation, while the procedures that Grimm deems 'poetic' are described as rather functional in their agrarian setting. What is more, Grimm's juridical metrology conceptualizes both – the old and new forms of economic quantification – as regimes. If a 'tax-hen' is presented to a local authority this procedure symbolically sustains and confirms the existing structure of power. The modern administration may rely on abstract metric units, and does no longer require a personal encounter. Regimes are less visible, but Grimm still correlates numerical quantification with a centralization of political power.

Hector Vera: ***Making the Metre Their Own: Laypeople and the Appropriation of the Metric System in Mexico, 1895-1940***

The decimal metric system of weights measures (a scientific language invented by mathematicians and astronomers at the end of the eighteenth century) was made the exclusive and mandatory system of measurement in Mexico in 1895. This paper analyzes how laypeople received, learned, and appropriated the metric system in their everyday lives. It analyzes who accepted the new measures voluntarily, who had to be coerced into it, and what forms of opposition appeared. In Mexico the reception of the metric system was slow and difficult; for the most part people did not oppose metrication openly, but a very effective "surreptitious resistance" became pervasive. The public at large, small merchants, and some local authorities simply ignored the regulations that banned customary measures and kept on using them.
The paper also describes the tactics of appropriation that non-experts developed to cope with the metric system in their daily activities and how they reacted to the enforcement and policing strategies displayed by the government to secure the actual employment of the metric system.

These tactics are organized in these analytical categories:

1) commensurability: redefining customary units of measurement as exact fractions of metric units;
2) rounding: modifying the magnitude of a customary unit of measurement by slightly increasing or decreasing its value to make it identical to a metric unit;
3) word substitution: using the name of metric units to designate traditional indigenous measures;
4) metric surnaming: the creation of hybrids, combining the name of the old unit with a sort of metric "surname", designating a new value to it;
5) arithmetic continuity: maintaining the arithmetic of a technique of measurement, but changing the base unit employed;
6) monetary equivalence: estimating the amount of a commodity in relation to its price according to its pecuniary valuation;
7) functional equivalence: using a metric unit to measure a different physical dimension than the one it was designed for.

Using data from archival research, ethnographic reports, and census information, the paper shows concrete instances of how these tactics were actually employed in the countryside and among indigenous, non-Spanish speaking communities.
The process of implementing a mandatory metric adoption at ground level reveals many aspects of how scientific ideas circulate across national boundaries and social classes. The history of the metric system, in this regard, intertwines processes of scientific and commercial globalization with the birth of modern national states. It is a history of the relationship between capitalism, statecraft, and science. The diffusion and appropriation of the metric system helps to illustrate how scientific ideas are distributed not only among members of an enlightened elite, but also through wider groups.

This paper is thus focused on how this scientific language (the metric system) was appropriated, reinvented, and manipulated by the Mexican population at large: peasants, peddlers, clerks, lawyers, municipal employees, bakers, bricklayers, bazaar buyers, carpenters and so on. I want to show how the "man on the street" (persons who operate according to a vague "knowledge of recipes," following procedures that can be trusted even if they are not clearly understood) coped with the imposition of an exotic, technical lingo that began to appear in shops and government offices.

Isaac Record and Boaz Miller: *Measurement and Knowledge from Instruments*

In recent years, there has been a growing body of literature in the philosophy of measurement that explains the nature and possibility of measurement without ontologically committing to realism about the measured magnitudes. The central idea behind such accounts is cashing out measurement in terms of reaching agreement between measuring instruments that work on different principles. While reading measurement results off instruments is a major way in which we acquire knowledge from instruments, it is not the only way. There is still a need

for a general account of how we acquire knowledge from instruments, which would integrate the insights from the recent work in the philosophy of measurement into it. In our paper, we provide such an account, namely, a novel account of knowledge from instruments. We first critically examine the received view in analytic epistemology, according to which knowledge from instruments is reducible to true beliefs generated by justified inductive inference from beliefs generated by perception. While this view has some merits, we end up rejecting it to argue that instruments constitute a distinctive source of knowledge not reducible to other sources. As an alternative to the received view, we argue that obtaining knowledge from instruments depends on epistemic subjects' ability-knowledge. The relevant abilities include instrument users' successfully operating their instruments and reading off information from them, and instrument-makers' reliably manipulating available material capacities, including for reaching agreement between instruments that work on different principles. Because instrument makers and users rarely interact directly, the relevant abilities are distributed in an epistemic community. We suggest that this relatively thin abilities/capacities ontology explains how technology effectively affects standards of justified belief, and how epistemic responsibilities related to technology can be allocated on a principled basis.

Klaus Ruthenberg: ***Acidity, a multifarious chemical conception***

If we pour nitric acid over copper fillings, the former colourless liquid turns blue while the fillings vanish, and a brown gas emerges. The typical reaction of an *acid* – its ability to dissolve metals – has been proved another time. Nevertheless, there are puzzling aspects of acidity when it comes to chemical theory. The former copper atoms have been transformed into ions with the oxidation number +2. However, these ions are not naked, rather they carry with them six bound water molecules, $Cu(H_2O)_6^{2+}$.What matters here is that we are confronted with another actual acidity concept which as a matter of fact denotes the Cu^{2+}-ions (which we address here only simplifying) as *acidic* and the water molecules as *basic*. Moreover, to complicate the situation even more (poor students...), a third theoretical concept lists Cu^{2+} as *borderline acid* and H_2O as *hard base* [1]. What the *protonists* Brönsted and Lowry call acid (our example: nitric acid) has no such label in the frameworks of the *electronists* Lewis and Pearson, and what the latter call acid (here the copper ions) cannot be manufactured as a pure, proper substance, and is no acid at all in the classic sense [2][3].

The measurement of acidity with respect to the *protonic* theories of Arrhenius, Brönsted, and Lowry (the *pH*, which rather is a tag number than the result of a proper theoretical effort) is not sufficient to allow for the description of the full set of reactivity, strength, and other properties considered characteristic or pertinent of an acid, see, in our example, the unusual formation of the brown nitrogen dioxide [4]. On the other hand, these determinations are a widespread and sometimes very important everyday routine in chemical and biochemical industry, environmental protection, food production, clinics, etc., that is, they

have gained invaluable practical success.

There is a wide open gap between the application oriented, more empirical concepts (Arrhenius, Brönsted, Lowry) and the more theoretical concepts (Lewis), some of which are forgotten or restricted to theoretical circles (Usanovich, Pearson). In order to understand better how the mentioned conceptual coexistence developed, the operational (and instrumental) background of the early history of acidity up to the crucial year 1923 will be examined in the present contribution.

The *protonic* and the *electronic* theoreticians are talking about different entities. Hence, a uniform natural kind by the name of *acid* does not exist [5]. It seems, moreover, that the incommensurabilist has only very small chances to tell his or her story this time.

Raphael Scholl: ***Measurement and the method of difference***

Mill's method of difference offers a credible first-pass description of much scientific practice. We infer from a contrast of two situation where an effect occurs in one but not in the other, and where a single antecedent difference exists, that the antecedent difference is the cause of the difference in the effect. However, in actual practice inferences are complicated by the fact that both antecedent differences and effects are generally not easily observable: this has been labeled the "problem of inferred differences" by Peter Lipton. Thus, both antecedent differences and effects need to be intervened on and detected by complex causal interactions. How this type of causal access in experiments is established is, by its nature, a problem of measurement. For example, when introducing a restricted DNA fragment into a cell culture in order to test its causal role, we need to be able reliably to measure the frag- ment's presence in order to know that the relevant antecedent difference has been established. The present paper will look in detail at two case studies from the life sciences: one from 19th century bacteriology and one from 20th century molecular biology. The goal is to outline an account of some of the methods by which scientists establish the causal access required for actually applying Mill's method of difference.

Aashish Velkar: ***'Inching towards the Metre': Comparing metric conversion policies in industrialised and late-industrialising countries (c1950-80)***

The 1875 International Treaty of the Metre transformed the metric system from a French system of weights and measures into an international metrological system. Over 150 years, they diffused piecemeal across the globe such that by the end of the twentieth-century almost all countries had defined their legal metrology in terms of SI units. Even in the US and UK, SI units are the fundamental legal units although everyday measurements continue to be made using 'customary' or Imperial units.

This paper examines the discourse surrounding metrication between c1950 and c1980. During this period there was a 'wave' of metrication and currency decimalisation, mostly in newly independent nations in Africa and Asia. Political independence was a strong factor in such conversions, just as political unification encouraged metrication in nineteenth-century Europe. Many former colonies were using the metric system even before independence, inheriting it from French, Spanish and other European colonisers. British colonies had used either customary measures or developed hybrid metrological systems. Once independent, late-industrialising countries like India or Singapore had strong techno-economic reasons to establish national metrological systems based on the SI units. Against this backdrop, metric debates during 1950-80 in industrialised countries such as UK, US, Canada, etc. seems like an aberration. The paper probes why metrication became a policy issue in the industrialised nations and how discourses differed between industrialised and late-industrialising nations. The primary focus of the paper is on metric discourse in the UK and USA comparing it with similar discourses in India around the same time.

Discourse around reforming legal metrology is situated in an interdisciplinary context. Metrology can be studied as a technology of governance (Latour, Mitchell) or constitutive element of state formation (Curtis), resulting in strong or weak metrological 'chains'. Equally, measurements form fundamental elements of economic transactions (Kula) such that producers often invest considerable efforts in securing standards that give them a strategic or competitive advantage (Velkar). In this way 'measurement' can be distinguished from 'metrology' (Gooday) Measurement units are socially embedded objects that have a shared cultural meaning that leads to inertia, resistance or protests against metrological change (Crease, Sheldon).

Using a constructivist approach this paper examines how the discourse around metrication was framed, especially but not exclusively in the mass media. Supplementing this with other sources of public information it draws comparison across discourses such as 'rule of experts' versus 'rule of reason', voluntary versus compulsory metric change, internal markets versus export trade, competitiveness versus calculability, industrial metrology versus everyday measurements. It probes how such competing 'frames' informed metric debates in this period. It probes why proponents of change considered it a feasible or necessary policy and why opponents (especially in the UK and US) strongly resisted change.

Methodologically, the paper argues that study of discourse, as distinct from study of practice, can provide a rich understanding of how measurements are 'made'. Further, as measurements are often sites of protest and conflict, the study of measurement is useful in revealing larger social, economic and political concerns of the period.

Veronica J. Vieland: ***Measurement of Evidence in Theory and in Practice***

In *Inventing Temperature*, Chang considers the history of thermometry in terms of "nomic" measurement, or measurement of a quantity based on application of a law. The challenge was to discover a function f that mapped an observable quantity (e.g., the volume V of liquid in a thermoscope) onto an unobservable quantity (the temperature T), without any way of knowing T except through application of the law $T=f(V)$. Loosely following Chang, we can distinguish various stages in the solution to this problem. During Stage 1, "empirical" measures t were developed, which could be more or less directly verified as tracking in the same direction as the temperature. E.g., V could be seen to expand as the environment was heated, so that $\Delta t = \Delta V$ could be used to assess temperature change at least within certain limits, but with numerical values of Δt dependent upon the type of thermoscope. During Stage 2, different thermoscopes were empirically calibrated against one another at specific reference points (e.g., the steam point of water). But this still left unresolved the possibility of differences in the rate of change of t across the temperature range between and beyond reference points, both within and across thermoscopes, as well as the problem of establishing empirical measures beyond the limits in which direct verification was feasible. What I'll call Stage 3 was accomplished with Kelvin's derivation of the absolute scale T.

My group works on the nomic problem of measuring the strength of statistical evidence, particularly for applications in the biomedical sciences. One of our aims is to derive an absolute measure for evidence E on the model of Kelvin's absolute T. We have thus far derived a particular mathematical equation in the form $E=f(\mathbf{X})$, where $\mathbf{X}$ comprises directly observable (computable) properties of the likelihood ratio graph for given data under a specified statistical model. We have spent considerable time articulating the evidential version of Stage 1, and demonstrated that E alone among alternative candidates (p-values, likelihood ratios, Bayes factors) satisfies our basic Stage 1 criteria. E also appears to satisfy Stage 3, that is, it appears intrinsically to be on an absolute scale. But Stage 2 has us stumped! We are unable to ascertain whether E is empirically calibrated across statistical models at reference points, or indeed, how we would know if it were not.

In this talk I will focus on the epistemic conundrum of attempting to conquer Stage 2 for a live, as-yet unresolved nomic measurement problem, highlighting both similarities to and differences from the corresponding problem in thermodynamics. One intriguing distinction between the two problems involves the mathematical nature of the intended object of measurement in our case. T is defined mathematically within the framework of thermodynamics, and the thermometric difficulty is relating T as it appears in its natural, purely theoretical, habitat to the corresponding physical phenomenon. But E directly captures features of systems that are already inherently mathematical. What is unclear is whether this simplifies or complicates the measurement situation.

Ted Vosk : ***Forensic Metrology: Scientific Measurement and Inference in the Courtroom***

Forensic measurements play a prominent role in today's criminal justice system. They are relied upon to investigate and prove an array of civil and criminal charges, from speeding, through driving whilst intoxicated with alcohol or drugs to murder. As powerful a tool as measurement is for the discovery of factual truth though, it is often misunderstood and misapplied in the courtroom. Forensic scientists commonly present measurement results in a simplistic fashion as accurate representations of a quantity's value without any indication of the limitations of the inferences such results actually support. Judges, lawyers and jurors often naïvely accept results presented in this manner as establishing a quantity's true value. No measurement, though, no matter how sophisticated or carefully performed, can ever reveal a quantity's true value. To facilitate the discovery of truth in the courtroom, forensic measurements must not only be properly validated and performed, but the relationship between measured and true quantity values determined and clearly communicated. Failure to do so undermines efforts to rationally and appropriately weigh measured results.

The results of measurements have traditionally been reported as if their referent is an actual physical state of nature. In this context, their limitations are conveyed in terms of systematic and random error associated with the measurement performed. Unfortunately, the error associated with a particular measurement is as unknowable as the measured quantity's true value. Accordingly, the traditional approach to measurement leaves doubt about how well the result of the measurement represents the value of the quantity being measured.

Metrology, the science of measurement and its application, overcomes this difficulty by providing a modernized framework for building epistemologically robust knowledge through measurement. This framework includes the elements of measurement validation, traceability and uncertainty. Adherence to its principles in the development and performance of measurements yields high-quality information which places limitations on the inferences a result can support. Once a result has been obtained, this framework provides rules for determining the inferences and conclusions that are supported by it. This permits an unambiguous, quantitative characterization of the relationship between the values attributable to a quantity and those obtained through measurement. Finally, metrology provides a precise language by which results and the inferences they support can be clearly communicated.

In the modern paradigm, however, the focus is no longer on the actual physical state of a measurand which is unknowable. Rather, the referent is what our state of knowledge about a measurand's physical state permits us to conclude about it. This makes explicit the fact that while one cannot say what a quantity's value actually is, they can determine what the information obtained through measurement permits them to infer and justifiably believe about that value.

Application of these principles to the investigation and prosecution of crime constitutes forensic metrology. Forensic metrology is essential if measured results are to facilitate the determination of truth in the courtroom. This presentation addresses foundational principles of metrology and their application within the courtroom.

Ioannis Votsis: *Why Immaterial Standards Matter*

In a well-known passage in the *Investigations*, Wittgenstein makes the following claim: "There is *one* thing of which one can state neither that it is 1 metre long, nor that it is not 1 metre long, and that is the standard metre in Paris." (2009, p. 29e) [original emphasis]. The standard meter, Wittgenstein reasons, is an 'instrument' of our language. Qua an instrument, it provides a means through which length can be represented, though it is not itself representable. It is thus illegitimate, he claims, to ask whether the standard meter is a meter long. I begin this talk by showing how Wittgenstein's concerns become immaterial in the face of modern measurement theory. That's because standards nowadays are set by definitions, not samples. I then proceed to explore several advantages of the definitional approach, focusing, among other things, on the stability it offers over the old sample-centric approach.

Let's travel back in time to the 1950s. How would one find out whether something was a meter long back then? By laying it against some sample meter like a ruler. And how were these sample meters constructed to ensure they were a meter long? By machines whose operation was calibrated against so called 'working standards', i.e. some meter-long sample used for industrial purposes. These were in turn checked against so called 'secondary standards', themselves presumably adjusted to match the standard meter in Paris. But what about the standard meter itself? Surely, the standard meter cannot be laid against itself. Thus, it seemed right back then to claim, like Wittgenstein did, that we cannot ask whether the standard meter is a meter long. Commenting on this point, Beaney (2006) concurs with this assessment. Indeed, he goes on to generalise: "it is illegitimate to say of any sample that it either possesses or lacks that property of which it is a sample" (ibid.).

Fast forward to today. Luckily for us, standards are not material anymore. Instead of a sample meter, we rely on a definition that makes reference to a fundamental physical constant: "The meter is the length of the path travelled by light in vacuum during a time interval of 1/299 792 458 of a second" (NIST). Such definitions are now commonplace and they include the standards for the units of time and of temperature.[1] Thus, the issue of comparing a sample to itself that so puzzled Wittgenstein and Beaney doesn't even arise. But beyond this rather minor point, notice that the main advantage of the definitional approach is that it overcomes a problem that plagues all physical samples, namely mutability. It does so by making reference to fundamental physical *constants* in the said definitions.

It might be objected that the problem of mutability does not vanish if, as some physicists have speculated, the values of the parameters we identify as fundamental constants do indeed vary over time. One redeeming feature of the aforementioned definitions is that they do not only make reference to such parameters but also to relations between them.[2] This means that even in the case where the parameters turn out not to be constant, unless the related parameters co-vary, any non-negligible change in their values would be reflected in our measurements. This would in turn force us to opt for definitions that cite parameters that are either truly worthy of the fundamental constant label or at the very least those that cite parameters that come as close to being worthy of it as possible.

Johanna Wolff: ***Realism about measurement and realism about magnitudes***

A realist about measurement, roughly speaking, holds that measurements give us information about, or epistemic access to, the way the world is. Measurement, on such an account is objective. A realist about magnitudes (understood either as properties or relations) holds that the way measurements provide such objective knowledge is by tracking features of the world, namely certain quantitative properties or relations. Does realism about measurement require realism about magnitudes, or can we be realists about measurement without any additional commitment to magnitudes?

Operationalists (Bridgman 1927, Stevens 1935) and nominalists (Field 1980) have traditionally held that it is possible to be a realist about measurement without being a realist about magnitudes. Realism about measurement without realism about magnitudes also seems to be promoted by Representationalism about measurement (Krantz et. al. 1971-90).

Against these attempts to divorce realism about measurement from realism about magnitudes, Mundy (1987), Swoyer (1989), and more recently Peacocke (forthcoming) have argued that realism about measurement, while conceptually distinct from realism about magnitudes, nonetheless requires a commitment to magnitudes.

My primary question in this paper is how exactly we should understand realism about measurement and realism about magnitudes respectively. A secondary aim will be to see how traditional arguments in favor of realism about magnitudes fare, depending on how we understand these two realisms.

At first glance realism about magnitudes seems to be an ontological claim, whereas realism about measurement makes an epistemological claim about the objectivity or epistemic reliability of certain practices. But if the objectivity and reliability of such practices ultimately derives from getting things right with respect to how the world is, then the step from realism about measurement to realism about magnitudes appears to be a small one. Magnitudes might just be those features of the world about which measurement gives us information, and

measurement is successful when it gets things right with respect to such magnitudes.

Perhaps a better way to understand the contrast is to take the disagreement to be over which entities are required to guarantee objectivity of measurement. A nominalist version of representationalism holds that objects are all that is required, whereas realists about magnitudes believe that properties, understood as universals, are required for the objectivity of measurement. This seems to be the main way in which proponents of realism about magnitudes have understood the debate, and the point of disagreement to which their arguments respond.

I will propose a third way of contrasting these two realisms, suggesting that difference between realism about measurement and realism about magnitudes is best understood in terms of priority: which is prior, measurement or magnitudes? After clarifying different ways in which such priority might be understood, I re-evaluate the arguments from realism about measurement to realism about magnitudes.