The quantification of intelligence in nineteenth-century craniology: An epistemology of measurement perspective

Abstract

Craniology – the practice of inferring intelligence differences from the measurement of human skulls – survived the dismissal of phrenology and remained a widely popular research program until the end of the nineteenth century. From the 1970s, historians and sociologists of science extensively focused on the explicit and implicit socio-cultural biases invalidating the evidence and claims that craniology produced. Building on this literature, I reassess the history of craniological practice from a different but complementary perspective that relies on recent developments in the epistemology of measurement. More precisely, I identify two aspects of the measurement culture of nineteenth-century craniologists that are crucial to understand the lack of validity of craniological inference: their neglect of the problem of coordination for their presupposed quantification of intelligence and their narrow view of calibration. Based on my analysis, I claim that these methodological shortcomings amplified the impact of the socio-cultural biases of craniologists, which had a pervasive role in their evidential use of measurement. Finally, my argument shows how the epistemology of measurement perspective can offer useful tools in debates concerning the use of biological evidence to foster social discourse and for analyzing the relationship between theory, evidence, and measurement.

Keywords

Craniology; quantification; measurement; intelligence; coordination; calibration; validity

1. Introduction

The practice of measuring skulls originated in the late 1700s as a tool for comparative anatomy to develop a systematic classification of human races (Banton, 2007; Richards, 2018; Vermeulen, 2015). In the early nineteenth century, the materialist view of the mind put forward by phrenologists introduced the core assumption of a relationship between skull size and form, brain, mental faculties, and behavioral traits, which had a pervasive influence in science and society (Erickson, 1977; Kornmeier, 2017; Van Wyhe, 2017).[1] Although phrenology was eventually dismissed, skull measurement became the source of large quantities of data that were gathered to answer questions concerning mental differences among human groups.[2] This epistemic practice, generally known as craniology or craniometry, established itself as a part of physical anthropology, which emerged in the first half of the nineteenth century with the goal of quantifying all human traits, physical ones as well as behavioral and mental ones. The very possibility of

[1] Phrenology was first developed by the Viennese doctor Franz Joseph Gall [1758-1828] and his collaborator Johann Caspar Spurzheim [1776-1832]. It is often characterized as the direct ancestor of nineteenth-century craniology, but with a much wider popular resonance and less preoccupation with accurate measurement than the latter (cf. Bittel, 2019; Parssinen, 1974; Shapin, 1979; Shortland, 1987).

[2] By the mid-nineteenth century, physiologists offered experimental proof that the brain is a homogeneous organ and it is not composed of various organelles, each regulating a certain mental faculty, as argued by phrenologists (Young, 1990). Although the functional unity of the brain meant the end of phrenology, its central tenet concerning the overall proportionality between skull or brain size and mental worth persisted as an entrenched assumption of craniological practice.

quantifying these traits – such as intelligence – by means of physical parameters was, thus, a corollary of craniology as a branch of physical anthropology, situated at the confluence of comparative anatomy, physiology, and psychology. In this sense, nineteenth-century craniologists took skulls as their primary source of evidence to quantify differential intellectual abilities among human individuals and groups.

Craniology flourished between the 1830s and the 1870s, but towards the last quarter of the century several internal and external factors started to weigh against its claims of intelligence differences among human groups. First, an increasing amount of recalcitrant evidence, gathered by craniologists themselves, was threatening the coherence of the assumption that there was even an approximate correlation between brains or skulls and intelligence. Second, the anthropologist Franz Boas [1858-1942] found evidence that environmental factors, such as health and nutrition, impact cranial shape and size and consequently mental faculties, which directly contradicted the hereditarian view held by most physical anthropologists. Finally, the assumption of a correlation between brain size and intelligence was directly attacked by a group of scientists guided by the English mathematician Karl Pearson [1857-1932]. These factors, in parallel with the birth of mental testing and of more refined statistical techniques, led craniology to lose its evidential grip compared to the performance-based measures developed by the emerging science of intelligence at the beginning of the twentieth century (Gould, 1981; Carson, 2007).[3]

Although historical and methodological overviews of the techniques of nineteenth-century craniology had already appeared by the 1950s (e.g., Hoyme, 1953; Shapiro, 1959), from the 1970s the history of craniological measurement became increasingly central to socio-historical analyses (e.g. Fee, 1979; Gould, 1978, 1980, 1981).[4] These seminal contributions successfully uncovered the entanglement of craniologists' epistemic practices with contemporary social pressures and pervasive cultural values. According to these authors, craniologists strove both consciously and unconsciously to cover up for the effects of their biases by adopting unsound epistemic strategies, often coupled with an overemphasized positivistic rhetoric that stressed the centrality of quantification as the golden standard of physical anthropology. More precisely, many craniologists were driven by the pressing aim of finding new justification for the existing social hierarchies on biological grounds, under the supposition that the prestige of science would put those hierarchies on a safer and less questionable footing. The impact of this research has been far-reaching, stimulating further historical and critical scholarly work on how socio-cultural and epistemic factors interacted in situated craniological practices (Anderson & Perrin, 2009; Challis, 2016; Douglas, 2008; Fabian, 2010; Geller & Stojanowksi, 2017) and, more generally, in nineteenth-century racial science and sexual science (Daston, 2008; Geller, 2020; Perrin & Anderson, 2013; Russett, 1991; Schiebinger, 1989; Tuana & Peterson, 1993). The long echo of this research also reached the public sphere, as in the case of Stephen J. Gould's reassessment of Samuel G. Morton's craniological research (Kaplan et al., 2015; Lewis et al., 2011; Mitchell, 2018; Weisberg, 2014; Weisberg & Paul, 2016).

[3] Even though the relevance of craniology as a research program aimed at establishing intelligence differences declined, craniological practices and the interest in cephalic indexes survived well into the twentieth century. Notably, cephalic indexes continued to be used to classify humans according to sex (e.g., Parsons & Keene, 1919), race (e.g., Coon, 1939; Parsons, 1922), and even nationality (e.g., Parsons, 1919).
[4] See also Blanckaert (1987, 1989), Carson (1999), and Kremer-Marietti (1984).

Indeed, classic socio-historical analyses of nineteenth-century craniology have successfully uncovered several forms of negligence, malpractice and misconduct perpetrated by craniologists in the attempt to save their claims against recalcitrant evidence. In addition, recent contributions have greatly clarified some of the epistemic limitations rooted in the lack of adequate justification for the evidential use of measurement by craniologists (Kaplan et al. 2015), as well as the relationship between craniologists' practices and their underlying views of intelligence (Carson, 1999, 2007: ch.3). However, a comprehensive epistemological analysis of the issues related to the inferential and justificatory structure of nineteenth-century craniology *qua* measurement practice, as well as of craniologists' approach to these issues, is still lacking. More precisely, certain structural features at the root of measurement issues that craniologists were unable – and often unwilling – to face, are yet to be properly spelled out. For this reason, analyzing the history of craniology from a measurement perspective, informed by the recent developments in epistemology of measurement (cf. Tal, 2013), would be greatly fruitful.[5] This broadly coherentist and practice-oriented literature has offered a set of conceptual tools that are, in my view, helpful in assessing how nineteenth-century craniologists approached some core measurement issues that were affecting the validity of their inferences. This, in turn, will shed light on the specific contribution of craniologists' measurement culture to the dynamics of kind-building fostered by their research program.

In this paper, I will analyze two interconnected epistemological aspects of nineteenth-century craniological measurement that have not received sufficient scholarly attention and that might be of interest to philosophers of science: coordination, viz. the process by which quantitative concepts acquire meaning through measurement (cf. van Fraassen, 2008), and calibration, the process that, in the terminology used by contemporary epistemologists of measurement, encompasses all the activities aimed at modeling a measurement procedure (cf. Boumans, 2007; Mari, 2000; Tal, 2017a). I will argue that craniologists neglected the importance of the problem of coordination for their presupposed quantitative notion of intelligence, and that their narrow view of calibration led them to place an unjustified epistemic burden on their instrument readings. I will show how understanding these two points is crucial to appreciate how and why craniologists embraced methodologically unsound escape routes in the attempt to preserve their preferred hierarchies of intelligence. Finally, I will claim that these two methodological shortcomings strengthened the influence of the socio-cultural values of craniologists, which had a pervasive role in their evidential use of measurement.

The impact of my analysis will extend beyond the domain of epistemology of measurement, in that it will contribute to understanding how measurement problems interact with the dynamics of kind building in the social domain, particularly with respect to the notion of race. To this day, the biological concept of race has been subject to decades of critique, starting with seminal works in the genetics of variation, most notably with the arguments by Lewontin (1972, 1974). In addition, substantial philosophical literature has uncovered several ways in which biological evidence has been used to foster racial social discourse and, more generally, has discussed how the biological and social level of discourse dynamically interact in generating social and racial kinds (e.g., Hacking, 2007; Kaplan 2010, 2011; Kaplan & Winther, 2013, 2014;

[5] In a similar vein, Carson (2014) uses some contemporary metrological insights to revisit the history of the development of IQ testing. From the historical point of view, this paper may be viewed as complementary to Carson's, in that it applies this perspective to the prehistory of IQ measurement, that is, craniology.

Pigliucci & Kaplan, 2003; Winther & Kaplan, 2013). Certainly, mainstream academic debates recognizes the validity of the arguments against racial naturalism, and its focus has shifted to discussing race as a purely social category. However, every now and then, the appeal to biological evidence – usually in the form of novel or reappraised measured data – still makes its appearance in the public arena, as in the case of the recent flare surrounding the Morton-Gould controversy. By means of a historical case study, I will show how the epistemology of measurement approach can provide additional tools to uncover the specific role of methodological measurement assumptions in contributing to enhance and normalize the illegitimate use of biological evidence to foster social discourse. Indeed, creditable research has long ago discarded the view of intelligence as a single, biologically inherited quantity, and racial naturalism is widely contested. Nevertheless, unjustified attributions of meaning to relationships among quantities can still lead to highly problematic uses of evidence, and particularly so in contexts where the kind-building assumptions are contested.

In Section 2, I will first introduce how issues relative to the use of evidential measurement in nineteenth-century craniology have been discussed with reference to the so-called Morton-Gould controversy. Then, I will rely on the metrological distinction between instrument readings and measurement outcomes to disentangle craniologists' general inferential structure from skull measurements to claims of intelligence differences. This analysis will be crucial to identify the two issues of craniological measurement that I will discuss in later sections. In Section 3, I will draw on recent literature discussing the notion of coordination with respect to issues of circularity and reliability in measurement to show how craniologists neglected the problem of coordination for their implicitly quantitative notion of intelligence. In Section 4, I will introduce a twofold distinction relative to the metrological notion of calibration, viz., into broad and narrow calibration, to discuss how craniologists' narrow view of calibration resulted in their attribution of an excessive evidential burden on instrument readings. In Section 5, I will summarize my results and tease out some general implications of this case study for the broader topic of the relationship between theory, evidence, and measurement.

2. Theory, evidence, and the scaffolding of craniological inference

In this section, I will provide a reconstruction of the general scaffolding underlying the inferences that nineteenth-century craniologists drew from their measurement practice to their claims of intelligence differences among human groups. This is required to identify the two specific aspects of craniological measurement that I will discuss in later sections, as well as to situate their significance with respect to more general debates in philosophy of science and race. Since several key aspects of craniological inference have been discussed in the context of the so-called Morton-Gould controversy, I will start my reconstruction from there.

2.1 Some lessons from the Morton-Gould controversy

Between the 1830s and the 1850s, the American physical anthropologist Samuel G. Morton [1799-1851] measured the skulls of his collection at various times. His aim was to rank different racial groups based on their average cranial capacity, which he took as a proxy of brain size and, thus, of intelligence. Morton's methods of cranial measurement became internationally recognized (Poskett, 2015), while his racial hierarchies of intelligence were widely used as scientific support against anti-slavery movements (Brown, 2015).[6] In *The Mismeasure of Man* (1981), Stephen J. Gould famously argues that these rankings are scientifically unsound because Morton's averages reflect his unconscious racial biases concerning mental worth. Gould identifies three main sources of methodological bias:

i. **Measurement bias**: to obtain the measurements of cranial capacity used as evidence for his first ranking in *Crania Americana* (1839) Morton's procedure consisted in filling the skulls with white pepper seeds. However, for his *Catalogue of Skulls* (1849), Morton measured a slightly different and larger sample of skulls which he filled with lead shot, a procedure that he deemed more reliable. Gould (1981) notices that this change of measurement procedure resulted in the increase of the average cranial capacity of all racial groups, but in a larger increase for Africans. According to Gould, the earlier seed-based procedure left more room for Morton's own bias to produce unsystematic measurement errors (for instance, by compressing seeds in African skulls more than in others), therefore leading to a larger increase in the 1849 African averages, where the measurements were taken by using the less malleable lead shot.

ii. **Sampling bias**: Morton measured skull samples from different races, but the size and composition of the samples was highly variable. According to Gould, comparing averages from larger samples to averages from smaller ones or from samples with disproportionate representation of the sexes inevitably skewed the comparison. For instance, since females have smaller average cranial capacity than males, samples with more females had lower averages.

iii. **Omissions and miscalculations**: Gould points out several mistakes in the calculation of average means. In particular, he argues that Morton included or excluded certain racial subgroups from their larger families to match his expectations concerning the ranking of averages.

After his critique, Gould recalculates Morton's averages and shows that there are no significant differences among mean cranial capacities across races in Morton's skull collection, thus leaving Morton's racial rankings of intelligence without any substantial evidential base.

In more recent times, Gould's own recalculations became the subject of an acrimonious controversy. Following up on Michael's (1988) early critique of Gould, Lewis et al. (2011) remeasured the skulls of Morton's 1849 sample and argued that Morton's measurements were objective. On these grounds, they claimed that Morton's work was free from racial bias, while Gould's reanalysis was skewed by egalitarian bias. Weisberg (2014) defended Gould's critique of Morton on several grounds and argued against Lewis and colleagues by pointing out that their argument is not sufficient to rehabilitate Morton's work as unbiased. In fact, showing the reliability of the 1849 measurements does not falsify Gould's claim that the earlier 1839 measurements were affected by Morton's unconscious racial biases concerning mental worth, due to the more unreliable procedure used by Morton at the time (Weisberg & Paul, 2016). Two further

[6] For historical overviews of the American school of physical anthropology and the so-called science of race, see, among others, Bay (2000), Dain (2002), Gossett (1963) and Stanton (1960). For further perspectives on nineteenth-century racial science, see, for instance, Stepan (1982) and Tucker (1994).

contributions, while rejecting the claim of Lewis and colleagues, also show the limitations of Gould's own conclusions. Since my argument builds directly on these views, I will present them in more detail.

According to Kaplan et al. (2015), Gould rightly claimed that Morton's evidence was inadequate to answer his questions on race, cranial capacity and intelligence. They show that Gould's analysis of the shortcomings in Morton's data gathering is, for the most part, correct and that it was largely misrepresented by Morton's recent defenders.[7] However, while arguing that the main source of this inadequacy were Morton's implicit biases, Gould overlooked how the lack of justification for the theoretical and statistical background assumptions underlying Morton's inferences invalidated his rankings. More precisely, Kaplan and colleagues emphasize that Gould himself failed to offer a better answer to Morton's question, because "Given how the skulls were actually collected, there are no interesting ways to summarize the dataset in order to draw broader conclusions about the world" (2015: 23). In other words, Morton and Gould shared the same mistake of believing that, given the craniological data available from Morton's sample, a valid inference concerning the relationship between race, cranial capacity, and intelligence in real populations could be drawn. In fact, for Morton's limited data to count as evidence for intelligence differences among races, he would have also required: 1) sound independent evidence to identify biologically meaningful populations (i.e., evidence for kinds), whereas Morton's own distinctions among races were based on anecdotal and unscientific ethnographic grounds; 2) a justifiable method of gathering a representative sample of skulls from the relevant populations in order to take the required measurements (i.e., evidence for representativity of samples), while Morton's samples had been collected without knowledge of the characteristics of the real population, thus making it impossible even in principle to factor in the relevant confounding factors, such as the statistical effect of sexual dimorphism; 3) a justifiable method to generate a population average for cranial capacity (i.e., evidence for representativity of averages), which Morton lacked since he had no justifiable grounds to assign a certain average cranial capacity to a well-defined population.

Taking a different angle, Mitchell (2018) provides new historical data relative to several, previously unidentified, specimen of skulls belonging to Morton's 1839 measurement sample, data that were not available to Gould when he developed his analysis. According to Mitchell, the new data support the claim that the errors in the 1839 measurements were significant, but likely random, thus putting pressure on Gould's claim that Morton's 1839 sample was affected by systematic measurement bias.[8] That being said, Mitchell views Gould's core claim that an *a priori* conviction of a race hierarchy guided Morton's work as indeed well-founded. This is evident once we establish a comparison between Morton's results and the work of his contemporary, the German craniologist Friedrich Tiedemann [1781-1861]. In fact, while both of them worked on very similar samples, obtained very similar measurement results and carefully explained

[7] For example, Kaplan and colleagues (2015: 25) rightly emphasize, contrary to what Lewis et al. (2011) seem to assume, that Gould never accused "Morton of wanting biased results, or of consciously trying to manipulate data" in his published works and, rather, that he respected Morton's intellectual honesty and continuous strive to improve his measurement procedures so as to be less easily manipulated by unconscious bias.

[8] Kaplan et al.'s (2015) statistical reanalysis of the data does, instead, confirm that the discrepancy between the averages of the 1839 and 1849 samples was very unlikely to be random and, thus, that the 1839 measurements were very likely affected by some systematic bias. In their view, however, this does not necessarily entail that the source of the discrepancy was the racial bias with which Gould charged Morton.

their methods of measurement, neither of them "justified their respective choices of statistics upon which to base their differing interpretations, whether ranges or averages" (Mitchell, 2018: 9), but they only implicitly held assumptions about the explanatory validity of the different statistics of variation justifying their inferences. As a result of this lack of justification for their background theoretical and statistical assumptions, Morton and Tiedemann could eventually draw opposite inferences from very similar data, as Tiedemann concluded that there were no inter-racial differences in intellectual faculties.[9]

In my view, the main take-home message of this debate is that, while Gould's major argument against Morton based on unconscious measurement bias may be less convincing than expected, this should not leave any room for doubt as to the presence of value-laden inferential choices in Morton's work, as well as in that of all nineteenth-century craniologists. In fact, on the one hand, conscious and unconscious biases may affect the production of data not only while performing the concrete procedure of measuring, but at any step of an inferential process involving measurement. Hence, the importance of focusing on the overall inferential scaffolding of craniology, and especially on its background assumptions and on their justification, as strongly emphasized by Kaplan et al.'s and Mitchell's contributions. On the other hand, even when we focus on concrete measurement procedures, we cannot forget that the representational character of measurement, i.e., the possibility to measure a quantity in terms of another quantity, presupposes a choice of theoretical and statistical assumptions that often requires value judgments and, thus, attaches some meaning to the data even before their interpretation. I will clarify this point further in the next section, after my reconstruction of the general scaffolding of nineteenth-century craniological inference.

2.2 The scaffolding of craniological inference and its structural limitations

In what follows, I will outline a model of craniological inference that generalizes from Morton's case and can be applied to all nineteenth-century craniological practices. In my view, the scaffolding of nineteenth-century craniological inference can be subdivided into the following four major inferential "steps":[10]

1. Inference from individual instrument readings of volume to individual values of cranial capacity
2. Inference from individual values of cranial capacity to average values of cranial capacity of the group samples
3. Inference from average values of cranial capacity of group samples to average values of cranial capacity of populations
4. Inference from average values of cranial capacity of populations to relative positions of a population on an ordinal scale of intelligence

[9] Cf. Tiedemann (1836). In this sense, Tiedemann was sharing with Morton a similar racial classification, as well as the same core assumption of a correlation between skull size and intelligence (Schmutz, 1990). Plausibly, the fact that he drew opposite conclusions is partly rooted in Tiedemann's adoption of some inferential assumptions carrying the cultural influence of the Enlightenment ideals still surviving in the German context, rather than that of the urges of racial differentiation of American society (Richards, 2018).
[10] These should be viewed as logical steps and not necessarily as chronological ones. For the sake of simplicity, I use "cranial capacity" as a placeholder for the several different cranial measures that were proposed as alternatives to cranial capacity.

Before going into details, I would like to emphasize that craniologists' evidence was inadequate to address their target questions because the justification for crucial assumptions involved in these inferential steps was insufficient. This claim does not imply that other research programs lacking sufficient justification for some of these inferential steps were or are making an inadequate use of measured evidence; nor does it rule out that several contextual factors, such as the inductive risk connected to an evidential claim, must be taken into account to assess what counts as sufficient or insufficient justification for certain inferential assumptions. Evidently, it is not my intention here to embark in an attempt to provide general demarcation criteria. Rather, my goal is to analyze nineteenth-century craniological inference as a paradigmatic case of problematic scientific inference and emphasize its structural epistemological shortcomings from within its context of inquiry.

In the second and third step of this model we can identify the shortcomings highlighted by Kaplan et al. (2015) with respect to Morton's use of evidence. As Kaplan and colleagues emphasize, inferring group means of cranial capacity from individual measurements of cranial capacity presupposes a classification of the relevant groups or kinds to which the individuals of the measured samples belong. The assumption that a certain biological kind classification is meaningful is usually guided by some epistemic purpose and is made against the backdrop of theory (Kaplan & Winther, 2014). In Morton's case, the purpose was clear – to provide a ranking of races based on average cranial capacity. However, scientifically adequate theoretical justification for the assumption of his classification of races was evidently not available to him in the first place, at least by our own standards (Kaplan et al., 2015). Indeed, all nineteenth-century craniologists would have required biological knowledge that was out of their reach to justify the anthropological kinds among which they wanted to establish intelligence differences. In the case of racial kinds, the current consensus is to reject the very possibility of justifying *any* meaningful racial classification on biological grounds given our own contemporary standards of knowledge, and the use of racial kinds ultimately draws its legitimacy not from biological data or theory, but from social discourse (Kaplan, 2011; Kaplan and Winther, 2013; Winther & Kaplan, 2013). This undermines recent reappraisals of nineteenth-century craniology as a valid source of evidence and, retrospectively, the work of nineteenth-century craniologists, as the very meaningfulness of their kind classifications was presupposed by the question concerning the relationship between human groups, intelligence, and skull size that they asked, rather than being validated by criteria of kind-building based on independent and reliable empirical evidence.

That the issue of background justification for kind building was hardly a concern of craniologists at all is also demonstrated by their approach to a core issue involved in this inferential step, i.e., the identification and classification of skulls. The absence of independent scientific evidence for kind assignment was not deemed as an issue by those craniologists who emphasized that skull classification by sex and race could easily be made based on pure observation, as if they belonged to different species (e.g., Vogt, 1864).[11] Even

[11] A similar line of reasoning was followed, for instance, by the German anthropologist Hermann Schaaffhausen [1816-1893] who advanced a criterion of kind identification based on the distinction between primitive and advanced skulls by *postulating* that male skulls are more advanced than female skulls and European skulls are more advanced than the skulls of other races (Schaaffhausen, 1868). Clearly, this criterion of classification excludes, by definition, the chance of assigning possible outliers to the appropriate kind – as in the case of exceptionally large female skulls – when independent evidence, such as the rest of the skeleton, is not available.

though some cautious craniologists voiced skepticism in this respect early on (cf. Fee, 1979), only in more recent times has the systematic misattribution of sex and race to skulls been clearly addressed as a pervasive problem in anthropology (Birkby, 1966), while the issue of classifying skulls with unknown background, especially in the absence of the rest of the skeleton, is still problematic for today's forensic anthropologists (Spradley & Jantz, 2011). These considerations also illustrate that the proper sampling method can hardly be identified without enough background knowledge of proper kind building.

The availability of sound justification for kind classification is at the root of the third inferential step, too. More precisely, the inference from ranking sample averages of skull capacity to ranking population averages of skull capacity rests on the assumption that the sample average is representative of the population average. In the absence of knowledge of features such as the general composition and boundaries of what meaningfully counts as one population, it is difficult to know how to build a representative sample of it, that is, what individual skulls to include in the sample, or exclude from it, to make the sample representative of the general population. This was rightly discussed as a fatal shortcoming of Morton's work by several commentators (Gould, 1981; Kaplan et al., 2015; Mitchell, 2018; Weisberg, 2014). However, the lack of statistical knowledge itself, on top of the lack of background empirical knowledge of the populations from which samples were taken, largely affected the craniological research program in general, and the issue of sampling error inevitably tainted the calculation of sample means as representative of real populations.[12]

We see now how the second and third steps of nineteenth-century craniological inference were irremediably affected by the shortcomings identified by Kaplan and colleagues: the lack of independent evidence for meaningful kind building, the lack of representativity of the samples, and the lack of representativity of the averages. These issues obviously affect also the fourth step, the one from population averages of cranial capacity to rankings of intelligence, as it presupposes the validity of the previous ones. However, extant analyses of nineteenth-century craniology have not sufficiently clarified the inferential issues specific to the fourth step, apart from general references to craniologists embracing forms of circular reasoning (Gould, 1981; Russett, 1991; Tuana & Peterson, 1993). In addition, the connections between this fourth step and the apparently unproblematic first step, i.e., the one from instrument readings of volume to values of cranial capacity, have not yet been properly spelled out. To do that, I will now introduce a distinction between instrument readings and measurement outcomes that will be helpful to identify two underspecified aspects of nineteenth-century craniological measurement related to the first and fourth inferential steps.

Measurement procedures are often characterized as physical interactions between one or more epistemic subjects, a material apparatus, and a phenomenon occurring in an environment. At the same time, the epistemic subjects purport to *represent* a certain relationship between quantities by means of the physical process taking place during the measurement interaction, as when we represent temperature in terms of the length of a column filled with mercury. Recent overarching accounts of measurement have focused on the process by which justification for the representational relationship between the outcomes of a measurement process and the quantity of interest is obtained (e.g. Chang, 2004; van Fraassen, 2008). This aspect will be

[12] For example, even a careful experimenter like Tiedemann tended to exclude from his samples "unusually" large skulls, which were more frequent in certain racial groups (Caucasians and Malay). Tiedemann was plausibly guided by genuine sampling concerns, but operated on a purely subjective basis, thus leaving his selection vulnerable to his own biases (Richards, 2018).

relevant to the fourth step of craniological inference, which bears on how outcomes of cranial volume were made to represent intelligence as a quantity.

Furthermore, recent works in the epistemology of measurement have focused on the inferential relationship between instrument indications and measurement outcomes (e.g., Frigerio et al., 2010; Mari, 2003; Tal, 2016, 2017b, 2019). These authors characterize *instrument readings* as observations of the states of the material instrument used to provide a quantitative representation of a certain phenomenon, once the physical process enacted during the measurement procedure has arrived at its end-state. *Measurement outcomes* are inferred from certain instrument readings by means of abstract and idealized models of the measurement procedure, which constructed and tested by modelling uncertainties and systematic errors of the procedure (or across procedures measuring the same quantity). The modeling of a measurement procedure, viz., calibration, often impinges on theoretical and statistical assumptions that are required to build models of measurement.

In the light of these clarifications, let us consider the first inferential step of the model, from instrument readings of volume to measurement outcomes of cranial capacity. Cranial capacity is nothing but the internal volume of a skull. Therefore, this inferential step does not involve the representation of a quantity in terms of another quantity. However, as I have explained in Morton's case, the reliability of measurement procedures was highly variable even in the case of the direct measurement of skull volume. Craniologists generally used some small-sized material (sand, seed, shot) to fill the entirety of the cranial cavity, and then emptied it into graduated containers to finally note readings of volume. Yet, the fact that the readings of volume from the graduated containers were taken directly as values of cranial capacity does not mean that no inferential step was required. Evidently, at this stage, the only quantity involved was volume. In this sense, this procedure presupposed measurement only in the commonsense meaning of number assignment according to a pre-established scale, viz., that of volume. However, an inference was made in that the volume readings from the graduated container were taken as reliable measurement outcomes of cranial capacity, where this inference must be justified, among other things, by sufficient knowledge of the possible measurement errors that might affect the reliability of the physical measurement procedure. Therefore, even the step from readings of volume to values of cranial capacity presupposes some form of modeling of the measurement procedure. While craniologists did not develop full blown models of their measurement procedures, they certainly resorted to calibration activities in order to improve their accuracy. These calibration activities involved certain implicit and explicit background assumptions, whose analysis can be informative of the general approach to measurement of craniologists. I will focus on these aspects in Section 4.

Finally, let us get back to the fourth and most problematic inferential step, the one concerning the relationship between the quantity of cranial capacity and the real quantity of interest, that is, intelligence. In the case of Morton's rankings, this relationship is not explicitly discussed, and the ranking of the average values of cranial capacity is intended to directly mirror the ranking of average intelligence among races. To believe this, Morton, as well as most craniologists, had to presuppose the existence of a direct correlation between skull or brain size and mental abilities. However, craniologists never systematically investigated the relationship between values of cranial capacity and intelligence as a quantitative notion. Thus, given their lack of theoretically justified definitions or independent measures of intelligence, they incurred into a specific form of circularity, one by which they were presupposing, rather than establishing, that their measurement procedures were capturing their quantity of interest. My goal in the next section will be that of analyzing in detail the epistemic dynamics at the root of this sort of circularity, which is not an unusual feature of the early developmental stages of novel measurement techniques.

Before turning to these two aspects, it is important to stress their relevance for the following point. Even if nineteenth-century craniologists had had biologically and statistically appropriate evidence for their presupposed kinds, as well as for the representativity of their sampling and of their averages, the question they wanted to ask concerning the relationship between skull size and intelligence differences across human groups could not have been answered. In the rest of the paper, I will show how analyzing craniologists' approach to coordination and calibration in measurement, themselves interconnected, is essential to understand the failure of nineteenth-century craniology from a methodological point of view. In addition, this analysis will show that these methodological shortcomings were critical in reinforcing the value-laden background assumptions that their inferential model carried.

3. The problem of quantity coordination and its relevance to nineteenth-century craniology

3.1 Nineteenth-century craniology and views of intelligence

Devising a reliable quantitative method to capture intelligence differences was central to the goal of craniologists. Throughout the nineteenth century, the increasing interest of naturalists and physical anthropologists for differences among human groups – especially among races, but also among sexes, nationalities, and social classes – brought them to focus on intellectual faculties as a key trait for classification and on cranial features as the parameters that would enable their quantification. Even beforehand, skull features, together with other skeletal traits, had been viewed as a more appealing source of evidence than other superficial traits, like skin color, to justify the drawing of lines across distinct kinds, in virtue of the fact that they were "more than skin deep" (Schiebinger, 1989). However, their use for quantifying intellectual abilities finds its roots in the process of naturalization of reason from a metaphysical absolute into an ability manifested in degrees, viz. intelligence, a notion imported from zoology and then progressively used to arrange humans and animals on a unitary, hierarchical, and gradual scale of mental ability (Blanckaert, 1987; Carson, 1999, 2007: ch. 3; Richards, 1987: ch. 1). As we have seen, the assumption of a physiological causal link between brain size or shape and intelligence came to craniology through the medium of phrenology, which identified skulls as material markers of intelligence. Yet, it is with craniology that skull volume and other cranial features became veritable measurement parameters and, as such, extremely powerful tools to classify human kinds via a single, measurable, naturalistic criterion of mental ability. The very possibility of quantifying intelligence, thus, emerged as a corollary of this biological and hierarchical view of intelligence: "Its connotations of global mental power, varying by degrees and related to the brain's physical nature, allowed measurable external characteristics, such as cranial capacity, to be related to an internal mental feature that could plausibly account for a people's place in the racial hierarchy" (Carson, 2007: 89). This view was already well-established before the advent of Darwin's theory of evolution by natural selection, which did not challenge its resulting hierarchies of intelligence differences across human groups, while hereditarianism, a byproduct of Darwinian thinking, rather contributed to consolidate them.[13]

Given this context, most craniologists seemed well-aware that the existence of a precise relationship between values of certain skull measures and values of intelligence was central to the validity of their measurement practice. The importance of the correlation between cranial size and mental faculties introduced by the materialist paradigm of phrenology was evident to them, and efforts towards a more precise characterization of it generated internal debates even in the early days of craniology (Fee, 1979; Gould, 1981). However, craniologists seemed much less aware of the fact that independent evidence would have been required to establish whether cranial capacity or any other measure of the skull was indeed a reliable measure of intelligence. Indeed, their views of the cognitive correlate to their naturalistic conception of intelligence were generally vague. Although with important differences depending on social context, craniologists often borrowed their language of intelligence, heavily loaded with morally evaluative notions, from ethnographic accounts assessing the degree of civilization of populations, and they usually referred to intelligence in the singular, as a unitary faculty (Carson, 2007).[14] Depending on the circumstances, craniologists equated intelligence with whatever more specific intellectual ability that made white male Europeans more civilized and advanced, while the real focus of scientific interest, as well as the justification for social hierarchies, remained on the natural, physiological differences in brains and skulls.[15] What is more, none of them recognized that they were lacking an appropriate form of coordination between intelligence as a quantifiable cognitive ability and their skull-based measurement procedures.

3.2 Measurement and coordination: the example of thermometry

To measure a physical quantity, we often infer its value from the values of other quantities, as when we infer measurement outcomes of temperature from indications of length of a thermometer column. This inference is based, among other background assumptions, on knowledge of the physical law that describes the relationship between the quantities of temperature and length in a specific physical interaction, which is often called a *measurement law*. The more precisely scientists can identify a measurement law relating

[13] Several theoretical principles that were developed against the background of Darwin's theory of evolution by natural selection were invoked by craniologists to justify the view that intelligence is an innate or even a purely hereditary character and that its different distribution across human groups reflects some natural evolutionary pattern, for the most part adequately mirrored by social hierarchies (Russett, 1991; Shields, 1982). One example of how these differences came to be viewed as natural, is that Darwin himself, in his *Descent of Man* (1871), identified natural selection, and not sexual selection, as the origin in the mental differences between the sexes. He postulated a hierarchy of mental faculties resulting from evolution by natural selection, which were taken as a source of independent evidence for male superiority (Fee, 1979; Tuana & Peterson, 1993).

[14] Carson (2007: 97-98) stresses that the socio-historical pressures behind the work of American and French craniologists differed and that this is mirrored by the languages of intelligence they used. For further historical discussion of intelligence views in French craniology and physical anthropology, see Carson (2007: 97-108) and references therein.

[15] This does not mean that no attempt to clarify the notion of intelligence was made from the side of craniologists. One example was put forward by the French craniologist Gustave LeBon (1879): "[…] a formula for measuring intelligence […] can be appreciated by the degree of aptitude for associating […] the greatest number of ideas, and perceiving as clearly and rapidly as possible their analogies and differences." Clearly, this definition would have been extremely hard to operationalize, given the knowledge available to craniologists, as well as to use it to test against their cranial measurements.

the two quantities, the better and more accurate measurement scales they can develop based on this correlation. However, these crucial empirical regularities need not be fully theoretically understood before measurement can take place, since progress in their precise characterization and advancements in measuring techniques usually go hand in hand through an iterative process of mutual refinement (Chang, 2004; van Fraassen, 2008).

In the case of thermometry – by now a classic example in history of measurement – progress towards the identification of the relevant empirical relationship among quantities was attained through various steps (e.g., Chang, 2004; Sherry, 2011). From a basic and rough distinction warranted only by bare sense-perceptions of heat and cold, a successful upgrade was achieved by means of thermoscopes. The use of thermoscopes enabled the correction of the highly fallible perceptual judgments, although within the limited scope of an ordinal measurement scale, which only permits to rank order among quantity values.[16] The development of thermometers marked the setting of new measurement standards that enabled the collection of a great deal of empirical data. The creation of interval scales of temperature (Celsius, Fahrenheit, etc.) allowed for the representation of the degree of difference between quantity values and it went hand in hand with the systematic study of the expansion of different materials, mercury and air being the prominent ones. At a later stage, with the development of classical thermodynamic theory, an overarching theory provided justification for the law of expansion of gases that could eventually be taken as a measurement law to infer (absolute) values of temperature on a ratio scale from the indications read out of gas thermometers.

The case of thermometry shows that, in the early stages of development of quantitative measurement, multiple measurement procedures can coexist in the absence of a precise and independently established empirical regularity that univocally justifies inferences from values of the representing quantity to values of a represented quantity. However, identifying that there is some empirical relationship between the representing and the represented quantity seems crucial to get measurement started in the first place. In this respect, craniology may be fruitfully compared to the phase of thermoscopy, as both were aimed at ranking different values of a quantity.[17]

3.3 Circularity, alternative indexes, and craniologists' escape routes

Most craniologists were mainly interested in ranking the (average) intelligence of different human groups, a purpose for which an ordinal scale of intelligence would suffice. To do that, they were relying on values of absolute cranial capacity or of other skull features. The point is the following: On what basis could craniologists reliably identify the relative position of certain individuals or human groups on a scale of intelligence from values of a physical skull measure? Thermoscopes provided a ranking of temperature values by relying on a certain empirical relationship between temperature and changes in pressure of a fluid, a regularity that, albeit only roughly identified, seemed to confirm our perceptual experience. In an

[16] Cf. Stevens' (1946) standard fourfold classification of measurement scales: nominal, ordinal, interval, ratio.

[17] Here I am abstracting away from the different problem of the reification of the intelligence construct, which was evident within the materialist paradigm of craniology, but also subtly affected subsequent episodes of the history of intelligence science (cf. Gould, 1981). For a comparison between contemporary psychometrics and the development of thermometry in the 1840s, complementary to my own analysis, see Bringmann & Eronen (2016).

analogous way, craniologists could rank values of intelligence of different human groups only by assuming that there exists a certain empirical relationship between intelligence and brain size. Craniologists considered the existence of such a relationship as a matter of fact, also due to the influence of phrenology, and it may even be argued that it had a statute of certainty on a par with the perceptual judgement of heat and cold differences against which thermoscopes could be tested. Therefore, cranial measurement, in their view, would not only serve the purpose of ranking human groups according to their intelligence, but also that of refining what they viewed as only a rough characterization of an empirical relationship between brain size and intelligence, by providing more accurate measures.

Generally, craniologists assumed that the relationship between their favoured skull or brain measure (be it absolute size/capacity, or any of its alternatives) and intelligence was a linear correlation. However, the only evidence that all craniologists could offer in support of the validity of this relationship was the same evidence that they were using also to establish (or reject) intelligence differences among human groups. In other words, they incurred in circularity because they took for granted that certain measures of the skull or brain constituted evidence that there were (or were not) intelligence differences among human groups while, at the same time, these same measurements were taken as evidence for the linear correlation between values of skull or brain features and values of intelligence. Evidently, craniologists did not recognize that their evidence was fulfilling, at the same time, two different and incompatible epistemic functions (cf. Gould, 1981).

However, as I emphasised above, the risk of incurring in this sort of circularity is not infrequent at the early stages of development of quantitative measurement. This is certainly due to the lack of precise definitions of the quantity of interest that, ideally, would require reference to independently established empirical regularities. Yet, it can also be viewed as the result of difficulties in identifying what exactly a certain procedure is measuring. For this reason, a strategy often used at the early stages of development in measurement is what Chang (1995) has called the "mutual corroboration" of measurement procedures, whereby different procedures that supposedly measure the same attribute are compared in search for convergence on robust fixed points and as a basis to study relevant empirical regularities underlying the procedures themselves.[18] As it is evident, craniologists did not recognize that their core assumption of an empirical relationship between brain or skull measures and intelligence was involved in a form of circularity that threatened the very possibility of establishing intelligence differences. In addition, they failed to see how, in the absence of any agreed-upon definition of intelligence, resorting to other measures of intelligence as a cognitive ability, independently of skull features, could provide a crucial tool to assess both their own skull-based measures and their measurement assumption. Most importantly, they were unable to take the lack of convergence of their different skull-based scales as a sign that their core measurement assumption was problematic. Finally, even craniologists' opponents struggled to realize that independent evidential support for the relationship between skull measures and intelligence, at least in the form of alternative measures of intelligence, was crucial. For instance, Tiedemann, did not deem it altogether necessary to

[18] For more recent accounts discussing the appeal to robustness in the identification of accurate measures see, for instance, Basso (2017), Bokulich (2020) and Tal (2017a).

provide independent evidence for the correlation between intelligence and cranial capacity presupposed by his own measurements, based on which he claimed that there are no intelligence differences among races.

Obliviousness to this issue is evident in how craniologists dealt with the so-called elephant problem. This problem arose from the recognition that, if intelligence is proportional to brain size, animals with brains of a larger absolute size than humans should also be more intelligent. Craniologists first tried to evade this undesirable logical consequence by restricting the criterion only to the human species (Russett, 1991). However, this did not help them to face the issue of recalcitrant data-points within the human domain. Craniologists often found themselves with unusually small brains or skulls coming from renowned scientists or men of intellect, or of very large brains belonging to criminals, or unusually large female skulls, etc., sometimes impacting the group averages to the point of altering their expected position on the scale of intelligence (Gould, 1981; Tuana & Peterson, 1993). In other words, those data could not be coherently accommodated on an ordinal scale of intelligence constructed by taking the core assumption of proportionality as the basis for their intelligence scale, let alone do that in a way that preserved the expected ordering of the human groups on the scale. When facing this issue, craniologists generally did not reflect on whether their criterion of proportionality between cranial measures and intelligence could be flawed, nor did they express the necessity to test it by means of alternative measures of intelligence. Rather, they adopted two alternative and equally unsound strategies. The most important French craniologist and neurologist, Paul Broca [1824-1880], fervently supported the strategy of reaffirming the linear correlation between absolute cranial capacity and intelligence by stressing that it only held in rough terms, thus underplaying the epistemic role of the linear correlation as a measurement law (Broca, 1861, 1868).[19] As pointed out by many commentators, this strategy led Broca to explicitly fall in the trap of circular reasoning without realizing how this jeopardized his attempt at being a good positivist (Gould, 1981; Russett, 1991). However, the circularity result could not be avoided even by those craniologists who embraced a different strategy, since they tried to preserve the preferred ordering relations of intelligence in the face of unwelcome evidence by shifting the physical parameter taken as a measure. The naturalist and anatomist Georges Cuvier [1769-1832] introduced his facial angle scale based on the relative proportion of the cranial bones to the facial bones exactly to get away with the elephant problem (Cuvier, 1837).[20] However, the rate of appearance of alternative measures spiked starting from the early 1870s, when craniology entered its "Baroque" phase (cf. Fee, 1979), or the beginning of the paradigm crisis, in Kuhnian terms. Faced with mounting recalcitrant data, craniologists responded with more measurements, both in terms of amount of measured data and of alternative measurement parameters.[21] Yet, the shift to an alternative physical

[19] For overviews of the French school of anthropology, see, for instance, Kremer-Marietti (1984), Stocking (1968), and Williams (1985).

[20] The facial angle as a skull measure to classify human beings was first introduced by the Dutch artist, naturalist, and anatomist Peter Camper in 1768. The connection between facial angle and mental capacity was first put forward by Cuvier and Saint-Hilaire in 1795 (Blanckaert, 1987). For more on Cuvier and his facial angle scale see Coleman (1964).

[21] In addition to the facial angle scale, another alternative that had gained popularity by the 1860s was the scale based on the ratio between brain size and body weight (cf. Fee, 1979, Gould, 1981). During the Baroque phase, many more – and sometimes quite eccentric – scales were put forward as often as they were thrown away, sometimes even by the same craniologist, as it happened with Topinard and the cephalic index scale that he had himself adopted (Topinard, 1885).

parameter does not, as such, alter the assumption of a linear correlation between that parameter and intelligence understood as a single, quantifiable cognitive capacity.

All of these alternative scales shared the common purpose of preserving the traditional rankings of intelligence among human groups by shifting to a measure that would accommodate recalcitrant data, as it has been adamantly shown by historians and sociologists (Carson, 2007; Fee, 1979; Gould, 1981; Russett, 1991; Tuana & Peterson, 1993). In this sense, establishing a coordination between their measurement procedures and a particular cognitive ability that these procedures were supposed to measure was not a central concern of craniologists, as they were not particularly interested in precisely identifying the trait in the first place. These physical measures were indeed considered as the empirical basis to infer intelligence values to be placed on an ordinal scale. However, their choice was made primarily in the light of their capacity to accommodate the data to fixed pre-ordered positions of anthropological kinds on the intelligence scale, rather than for their capacity to pick out more precisely the quantitative structure of the trait of interest. Previous commentators have insufficiently stressed the connection between the strategy of shifting the measurement scale and the failure or disinterest of craniologists in the identification of the potential threats of a lack of coordination and independent validation of the relationship between the quantity of intelligence and any of the measures of the skull or brain used to build the alternative scales.

3.4 The collapse of craniology and the significance of craniologists' neglect of coordination

A final confirmation of the centrality of coordination to assess the craniological research program comes from the very scientists who managed to expose the internal contradictions of craniological practice. In 1901, Alice Lee [1858-1939] a student and collaborator of the English mathematician Karl Pearson [1857-1932], published the first paper in which she provided evidence against the correlation between skull capacity and intelligence. In this paper, she showed that several skulls belonging to a group of female undergraduates had larger cranial capacity than some male faculty members of the University College. This paper had a great impact, because it proved the inevitability of the choice between rejecting traditional rankings of intelligence and rejecting cranial capacity as a measure of intelligence. However, she could not, through this strategy, directly undermine the validity of the linear correlation between absolute skull size and intelligence. In fact, this could only have been achieved by fully acknowledging the lack of coordination undermining craniological practice, that is, by providing alternative and reliable measures of intelligence independent of skull measures to test for the reliability and accuracy of the latter. This crucial step was taken one year later by Pearson (1902) who, for the first time, introduced an independent performance-based measure of intelligence to assess the fit of skull-based parameters as a measure of intelligence. Pearson compared cranial measurements of a group of undergraduates with their examination test scores and found no significant correlation, thus directly ruling out the core assumption of linear correlation between cranial capacity and intelligence as a spurious regularity. In a further series of papers, Lee, Pearson, and Marie Lewenz [1876-1955] provided evidence of the unfoundedness of other craniological intelligence indexes, including the ratio of body weight to brain size (e.g., Lee et al. 1903; Lewenz & Pearson, 1904).

In sum, the neglect of the potential threats coming from the lack of coordination is the root of several epistemic discrepancies that craniologists tried to circumvent by implementing unsound strategies based on evading recalcitrant evidence, embracing confirmation bias, or introducing ad hoc hypotheses. The notion of coordination clarifies in what sense the assumption of the correlation between skull or brain measures and intelligence can be understood as an unreliable measurement law, that it was tangled up in a specific sort of circularity. In fact, craniologists were using the same evidence, i.e., their cranial measurements, to fulfil two incompatible epistemic functions at the same time, that is, finding empirical support for their claims of intelligence differences (or lack thereof) among human groups, and finding support for the very empirical regularity that was justifying the representational character of their measurement practice. In addition, it sheds light on the epistemic dynamics by which the measurement practice of craniologists, notwithstanding the level of technical precision achieved, could not make progress towards an improved quantification of intelligence as a cognitive ability (more on this in the next section). Since the naturalistic view of intelligence on which craniologists were founding their measurement practice lacked any meaningful connection with any operational definition or cognitive-based measure of intelligence, there was no viable ground for meaningful inferences from values of cranial measures to values of intelligence. Not even craniology's opposers managed to effectively address the relevance of coordination until very late, and this shows how pervasive this issue was. Although coordination is an epistemological problem, its neglect was not a methodological fault with mere epistemic consequences. On the contrary, it contributed to shaping the intelligence concept as "a singular, real, measurable, physical entity, one open to appropriation by a range of scientific practitioners with a variety of agendas" (Carson, 2007: 78), as well as to implicitly justifying and consolidating the socially-driven classifications of human kinds that were lying in the background.

4. Narrow calibration and its influence on craniologists' view of measured evidence

In this section, I will make a final point on nineteenth-century craniological measurement. This point is again related to the representational character of measurement that was discussed above with respect to the problem of coordination. However, I will now focus on its relation with the practices of calibration implemented by craniologists.

In contemporary epistemology of measurement, calibration indicates the process through which models of the measurement procedure are constructed and tested, by modeling confounding factors, as well as systematic and unsystematic errors of a procedure under idealized statistical and theoretical assumptions (Boumans, 2007; Frigerio et al., 2010; Giordani & Mari, 2012, 2019; Mari, 2003; Tal, 2017a).[22] The aim of calibration is (ideally) to account for all possible sources of measurement error given the best standards of precision available and, therefore, to improve the accuracy of a measurement procedure.[23] Based on the

[22] Tal (2017a) points out that, in this sense, calibration amounts to more than the theoretical practice of instrument making.

[23] Accuracy and precision are two key aspects of the reliability of measurement outcomes. Although different meanings of measurement accuracy have been identified among practicing scientists (Tal, 2011), the model-based approach to epistemology of measurement characterizes it as the closeness of agreement among values reasonably attributed to a quantity given available empirical data and background knowledge (Giordani & Mari, 2012). In this view, precision is one component of accuracy, referring to the minimization of the measurement error due to the uncontrolled variations in the indications produced by the physical measurement procedure over repeated trials.

results of calibration, measurement outcomes are inferred from certain instrument readings. Evidently, one central aspect of calibration concerns the improvement of the reliability of the measurement instruments in producing precise indications (i.e., readings), that is, it concerns the modeling of the measurement interaction as a physical process. However, an equally crucial aspect of calibration involves the representational character of measurement. As we have seen, in a measurement process where we infer measurement outcomes of one quantity (e.g., temperature) from instrument readings of another quantity (e.g., length), the identification and modeling of possible measurement errors partly depends on how accurately the empirical relationship between the two quantities has been captured (Tal, 2017a). Although these two aspects of calibration are not separate in practice, for the purposes of my analysis I will refer to the former as *calibration in the narrow sense*, and to the latter as *calibration in the broader sense*.

In previous sections, I have emphasized that craniologists were far from being thorough when it came to provide empirical justification for several assumptions involved in their evidential use of measurement and that this attitude was pervasive during all phases of craniology until its very collapse.[24] The other side of the coin of this attitude has been defined by previous commentators as an "obsession" with quantification on the part of craniologists, a somewhat compensatory reaction to the inconsistencies of their results based on an obstinate strive for even more precise measurement (Fee, 1979; Russett, 1991). In my view, this reaction can be more adequately characterized as directed towards the material aspects of the measurement process, to the detriment of its non-material components, that is, the host of inferential presuppositions and modeling activities involved by its representational use. To discuss this point, I will briefly reconsider the example of absolute cranial capacity and spell out craniologists' implicit approach to calibration to better understand the origins of the field's obsession with precise measurement.

As I mentioned while discussing the Morton-Gould controversy, Morton had himself realised that the seed-based procedure that he and his assistant used in 1839 could lead to inaccuracies, since the characteristics of the seeds used to fill the skulls, such as their compressibility, influenced the reliability of the indications read out of the graduated containers (Gould, 1981; Mitchell, 2018). For this reason, Morton turned to lead shot to measure his skulls in 1849 and found they produced much more reliable values of cranial capacity. However, when reading Morton's account of his techniques for measuring cranial capacity in *Crania Americana* (1839), it is impossible not to appreciate the subtlety of the calibration activities that he implemented to obtain precise measurements of cranial capacity. The first step of craniological inference, from instrument readings of volume to values of cranial capacity, discussed in Section 2.2, can here be seen in all its complexity. First, Morton carefully describes the graduated container used to take measurements of volume, including the calibration procedure adopted to build the instrument and determine precise units of volume. Then, he describes how skulls were prepared for measurement by putting cotton in the foramen magnum and how seeds were poured up to the surface "and then pressed down with the finger until the skull would receive no more". The seeds were then transferred to the graduated cylinder, "which was well shaken in order to pack the seed" (Morton, 1839: 253). Finally, Morton goes on to describe all the precautions to set the skulls in fixed and stable positions in order to be properly manipulated, as well as the

[24] This is not to say that they sought no justification at all. As Russett (1991) discusses, several pieces of then available theory were used as justification for the evidential use of cranial measurements. However, these pieces of theoretical background were hardly subjected to empirical testing themselves.

specific manipulations and additional instrumentation devised to measure the capacity of the different cranial cavities, such as the coronal region and the anterior chamber.

All these activities, enacted by Morton with the objective of producing as precise measurements of cranial capacity as possible, belong to the category of narrow calibration introduced above. Morton provides all the details concerning the calibration of his measurement instrument as well as the modeling of the physical measurement procedure, including the preparation and manipulation of the skulls. Given the great variety of individual differences in shape, structure, and size of the skulls and of their internal parts, this activity entailed a process of standardization, so that the skulls could be compared based on certain features. Even granting that the standardization of these procedures was successful, [25] a problematic aspect concerns how Morton operated the selection and abstraction of those features that he deemed as relevant for his purpose, and the consequent discard of all the others. As Carson (1999) has pointed out, Morton's approach in this sense was a markedly "reductionistic" one, since a very small number of cranial features – most notably, those used to identify the race and measure cranial capacity – were chosen, standardized, and made to signify what Morton required so as to produce measurements that could function as evidence for his racial hierarchies of intelligence. Yet, when describing his measurement practice, Morton does not indulge in explanations as to how exactly these features, and not any of the other several measurable (and non-measurable) traits that his skulls retained, could become bearers of the meaning Morton gave them (Carson, 1999). In other terms, Morton showed little awareness of the fact that, for measured data to mean something, it is not sufficient to operate a selection and abstraction of certain parameters, but that justification is required for narrowing down their range of possible meanings. This point is certainly connected with the discussion above concerning the problem of coordination, as Morton was working under the assumption that the linear correlation between intelligence and cranial capacity would itself give meaning to his hierarchies of cranial capacities. Yet, the point here is slightly different. What seemed to escape the attention of Morton and of nineteenth-century craniologists in general is that meaning does not automatically arise by increasing the precision of the measurement procedure, nor does it become clearer. In this sense, the reductionistic approach and the narrow view of calibration are two sides of the same issue. Instead of dedicating some of their efforts towards a greater precision to the theoretical and statistical presuppositions that were making their chosen measurable features meaningful already while performing concrete measurement operations, craniologists remained stuck in their view of measured data as somehow pure bearers of meaning, corroborated by the assumption that the only relevant modeling of the measurement process concerned the physical procedures and the material aspects of measurement.

The late stages of the history of craniology are a manifest example of how a narrow view of calibration can lead to a dead end. The increasing disunity of craniometry, reaching its peak during the Baroque phase, shifted the focus of attention from absolute cranial capacity towards several alternative physical parameters.

[25] As several commentators have pointed out, the highly irregular internal structure of the skull means more material can always be packed in it, by filling a hidden cavity through some shakes or readjusting the distribution of the material (Gould, 1981; Tuana & Peterson, 1993). This may be viewed as an instance of a metrological issue that often characterizes the early stages of development of a measurement technique and concerns the correct identification of the end-state of a measurement procedure (Tal, 2017a). The lack of clarity concerning *how* and *when* a measurement procedure terminates may lead to both systematic and unsystematic measurement errors, thus causing the epistemic subject to read misleading indications out of the measurement apparatus.

This further exacerbated craniologists' attention to the material aspects of measurement, mostly in the attempt to produce the best instruments to precisely measure the different cranial angles and indexes rapidly crowding the craniological canon. The mounting difficulties in standardization led to an ever increased attention to the procedural errors of their physical measurement procedures. However, it was their disregard of calibration in the broader sense, involving the representational use of measurement – and, ultimately, the coordination between represented quantity (intelligence) and a representing quantity (any skull-based measure) – that prevented them from identifying the reason why their evidence did not, and could not, coherently fit their measurement scales. Instead, their increasing obsession with physical measurement led them to force meaning on their measured data by placing an unwarranted epistemic burden on their instrument readings, to the detriment of the inferential assumptions underlying their measurement procedures.

5. Conclusion

The collapse of the craniological research program and its failure at quantifying intelligence by means of physical measures did not prevent the assumption of quantitativity of psychological attributes from making its way into the development of early psychological testing, most importantly intelligence testing (Boring, 1961; Carson, 2007, 2014; Gould, 1981). Even when measures of intelligence by means of standardized testing started to appear, independent evidence of its quantitative structure proved far from easy to obtain (Michell, 1997).

By analyzing the structure of craniological inference through the lenses of contemporary epistemology of measurement, I was able to clarify the attitude of nineteenth-century craniologists towards two important aspects of measurement. My first point was that craniologists neglected the threats coming from the lack of coordination between what they were treating as a quantitative attribute, i.e., intelligence, and the procedures through which they were measuring it. When confronted with the lack of convergence of their different skull-based scales, rather than investigating its roots in depth, they protected their core assumption of a linear correlation between skull features and intelligence at the cost of falling into circularity. In this sense, craniology can be characterized as both an incoherent and an unsuccessful measurement practice, as it was unable to maintain its internal consistency and to accumulate reliable evidence for its purported aims. A by-product of this attitude was craniologists' obsession with precise measurement, partly in the genuine hope that this would, by itself, lead to a better understanding of the nature of intelligence differences; partly because it helped deflecting the attention from the internal inconsistencies of their research program, while conveying a superficial image of rigor and objectivity, a strategy that has been frequently adopted in other contexts of inquiry (e.g., Porter, 1996). My second point clarified the nature of this obsession as limited to a restricted class of activities that can be implemented to model a measurement procedure, what I called narrow calibration. Craniologists' preoccupation with improving the precision and reliability of their physical procedures and material instruments was not counterbalanced by an equal attention to the assumptions embedded in their measurement process, which surreptitiously transformed their very selected set of physical features into bearers of meaning.

These two points add to the debate surrounding the socio-cultural biases of nineteenth-century craniologists, albeit from a different angle compared to classic critiques. Whether or not Gould's claim of unconscious racial bias against Morton was overstated, the existence of pervasive explicit racial, sexual, and class biases in the work of craniologists has been amply demonstrated and, in my view, does not require further support. Instead, my contribution has focused on some epistemic preconditions that enabled these biases to have such an extensive role. The neglect of the lack of coordination as a potential threat to the validity of craniological inference and craniologists' narrow view of calibration were two characteristic attitudes of their measurement culture. In fact, craniologists attributed a central place to measurement and quantification – very much in line with the positivistic spirit of the time – as the source of incontrovertible evidence for their claims. At the same time, they believed that their measurement practice amounted to little more than data-gathering, thus mistaking it for a value-free epistemic activity and overlooking the pervasive role that background theory plays in all the stages of measurement. Indeed, the measurements they produced were not neutral, or meaning-free, but they were carriers of conceptual pre-categorizations that reflected their biases rather than creditable theoretical views. Notably, the neglect of coordination and the narrow view of calibration permeating craniologists' measurement culture enabled them to ground their claims on an enormous quantity of measured data while understating the depth of the methodological flaws affecting their evidential use of those measurements.

Considering these two aspects as enabling conditions is not to say that the measurement culture of craniologists had a marginal role in reinforcing the socio-cultural values shared by craniologists. This can be clearly seen with respect to the case of racial categorizations. Through the measurement practices of craniologists, the very existence of racial kinds as biological entities was further legitimized by appealing to evidence that could be regarded – although, as we have seen, only superficially so – as external and independent from the assumption of a hierarchy of races, i.e., the evidence of intelligence differences among human groups. In this respect, my analysis introduces a new angle to the debate about circularity and kinds in the social realm, by drawing conceptual tools from the epistemology of measurement. In fact, the case of nineteenth-century craniology exemplifies how the attitude towards measurement embraced by a scientific community can function as a conduit for value-laden epistemic goals. This is not to say that more attention to the lack of coordination and a less restricted view of calibration would have sufficed to open the eyes of craniologists on the inherent flaws of their scientific enterprise, considering how pervasive the interests guiding their research program were. Yet, this case study provides an insight on how analyzing the measurement culture of an epistemic community, most importantly their approach to the representational character of measurement, can help us understand how methodological issues can become platforms for social agendas.

A final, more general point can be made about what the history of nineteenth-century craniological measurement can teach us concerning the relationship between theory, evidence, and measurement. Classic philosophical works on the theory-ladenness of measurement and more recent contributions on data-intensive science (e.g., Leonelli, 2012, 2015; Pietsch, 2015) have emphasized how theory plays multiple roles in the production, dissemination, and curation of data. Craniologists' view of the relationship between theory and evidence was relatively unsophisticated. This was reflected, as we have seen, in their measurement culture, since they largely underestimated the justificatory function of background assumptions for their cranial measurements not simply to count as evidence for their claims of intelligence differences, but to function as data in the first place. The force of their research program, particularly when inconsistencies started to pile up, lay in the quantity of measurements produced. However, concerns for their quality were raised mostly in relation to the material aspects of the measurement process as if, once the right physical procedure were identified, the measured data could almost automatically be accumulated and would be self-explanatory. Although progress in theorizing without progress in measurement may be considered as empty, the history of nineteenth-century craniology should be taken as a cautionary tale, warning us that progress in measurement without progress in theorizing can be blind and, in some cases, have dangerous consequences.

References

Anderson, K., & Perrin, C. (2009). Thinking with the head: race, craniometry, humanism. *Journal of Cultural Economy*, 2(1-2): 83-98.

Banton, M. (2007). The classification of races in Europe and North America: 1700-1850. In: Gupta, T. D., (Ed.), *Race and Racialization: Essential Readings*. Toronto: Canadian Scholars' Press, 15-23.

Basso, A. (2017). The Appeal to Robustness in Measurement Practice. *Studies in History and Philosophy of Science Part A*, 65: 57–66.

Bay, M. (2000). *The White Image in the Black Mind: African-American Ideas about White People, 1830–1925*. New York: Oxford University Press.

Birkby, W. H. (1966). An evaluation of race and sex identification from cranial measurements. *American Journal of Physical Anthropology*, 24(1): 21-27.

Bittel, C. (2019). Testing the Truth of Phrenology: Knowledge Experiments in Antebellum American Cultures of Science and Health. *Medical History*, 63(3): 352-374.

Blanckaert, C. (1987). «Les vicissitudes de l'angle facial» et les débuts de la craniométrie (1765–1875). *Revue de Synthèse*, 108(3), 417-453.

—— (1989). L'Indice céphalique et l'ethnogénie européenne: A. Retzius, P. Broca, F. Pruner-Bey (1840–1870). *Société d'Anthropologie de Paris: Bulletins et Mémoires, n.s.* 1**,** 165–202

Bokulich, A. (2020). Calibration, Coherence, and Consilience in Radiometric Measures of Geologic Time. *Philosophy of Science*, 87(3): 425–56.

Boring, E. G. (1961). The beginning and growth of measurement in psychology. *Isis*, 52(2): 238-257.

Boumans, M. (2007). Invariance and calibration. In: Boumans, M., (Ed.), *Measurement in Economics: A Handbook*. Amsterdam: Elsevier, 231-248.

Bringmann, L. F., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology*, 26(1): 27-43.

Broca, P. (1861). *Sur le Volume et la Forme du Cerveau Suivant les Individus et Suivant les Races*. Paris: Hennuyer.

—— (1868). On Anthropology. *Anthropological Review*, *6*: 35-52.

Brown, B. R. (2015). *Until Darwin, Science, Human Variety and the Origins of Race*. London: Routledge.

Carson, J. (1999). Minding Matter/Mattering Mind: Knowledge and the Subject in Nineteenth–Century Psychology. *Studies in the History and Philosophy of the Biological and Biomedical Sciences* 30, 345–76.

—— (2007). *The Measure of Merit: Talent, Intelligence, and Inequality in the French and American Republics, 1750–1940*. Princeton: Princeton University Press.

—— (2014). Mental testing in the early twentieth century: Internationalizing the mental testing story. *History of psychology*, 17(3), 249-255.

Challis, D. (2016). Skull Triangles: Flinders Petrie, race theory and biometrics. *Bulletin of the History of Archaeology*, 26(1).

Chang, H. (1995). Circularity and reliability in measurement. *Perspectives on Science*, 3: 153-172.

—— (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford & New York: Oxford University Press.

Coleman, W. (1964). *Georges Cuvier, Zoologist: A Study in the History of Evolution Theory.* Cambridge: Harvard University Press.

Coon, C. S. (1939). *The Races of Europe*. New York: The Macmillan Company.

Cuvier, G. (1837). *Leçons d'Anatomie Comparée*. *Tome 2*. Paris: Crochard, 2nd ed.

Dain, B. (2002). *Hideous Monster of the Mind: American Race Theory in the Early Republic.* Cambridge: Harvard University Press.

Daston, L. (2008). Die Quantifizierung der weiblichen Intelligenz. In: Tobies, R., (Ed.), *Aller Männerkultur zum Trotz: Frauen in Mathematik und Naturwissenschaften*. Frankfurt am Main: Campus-Verlag , 81-96.

Douglas, B. (2008). Climate to Crania: science and the racialization of human difference. In: Douglas, B., & Ballard, C., (Eds.), *Foreign Bodies: Oceania and the Science of Race* 1750-1940. Canberra: ANU Press, 33–96.

Erickson, P. A. (1977). Phrenology and physical anthropology: the George Combe connection. *Current Anthropology*, 18(1): 92-93.

Fabian, A. (2010). *The Skull Collectors*. Chicago: University of Chicago Press.

Fee, E. (1979). Nineteenth-century craniology: The study of the female skull. *Bulletin of the History of Medicine*, 53(3): 415-433.

Frigerio, A., Giordani, A., & Mari, L. (2010). Outline of a general model of measurement. *Synthese*, 175(2): 123-149.

Geller, P. L. (2020). Building nation, becoming object: The bio-politics of the Samuel G. Morton Crania Collection. *Historical Archaeology*, 54(1), 52-70.

Geller, P. L., & Stojanowksi, C. M. (2017). The vanishing Black Indian: Revisiting craniometry and historic collections. *American Journal of Physical Anthropology*, 162(2): 267-284.

Giordani, A., & Mari, L. (2012). Measurement, models, and uncertainty. *IEEE Transactions on Instrumentation and Measurement*, *61*(8): 2144-2152.

—— (2019). A structural model of direct measurement. *Measurement*, 145: 535-550.

Gossett, T. F. (1963). *Race: The History of an Idea in America*. New York: Oxford University Press.

Gould, S. J. (1978). Morton's ranking of races by cranial capacity. Unconscious manipulation of data may be a scientific norm. *Science*, 200(4341): 503-509.

—— (1980). Women's brains. In: *The Panda's Thumb: More Reflections in Natural History*. New York: Norton.

—— (1981). *The Mismeasure of Man.* New York: W.W. Norton and Company.

Hacking, I. (2007). Kinds of people: Moving targets. *Proceedings of the British Academy*, 151: 285-318.

Hoyme, L. E. (1953). Physical anthropology and its instruments: an historical study. *Southwestern Journal of Anthropology*, 9(4): 408-430.

Kaplan J. M. (2010). When Socially Determined Categories Make Biological Realities: Understanding Black/White Health Disparities in the U.S. *Monist*, 93 (2): 283–99.

—— (2011) "Race": what biology can tell us about a social construct. In: *Encyclopaedia of Life Sciences*. Wiley, Chichester. doi:10.1002/9780470015902.a0005857

Kaplan, J. M., Pigliucci, M., & Banta, J. A. (2015). Gould on Morton, Redux: What can the debate reveal about the limits of data?. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 52: 22-31.

Kaplan, J. M., & Winther, R. G. (2013). Prisoners of abstraction? The theory and measure of genetic variation, and the very concept of "race". *Biological theory*, 7(4): 401-412.

—— (2014). Realism, antirealism, and conventionalism about race. *Philosophy of Science*, 81(5): 1039-1052.

Kornmeier, U., (Ed.), (2017). *Schädel Basis Wissen II: Texte zur Wissensgeschichte eines Knochens*. Berlin: Kulturverlag Kadmos.

Kremer-Marietti, A. (1984). L'Anthropologie physique et morale en France et ses implications idéologiques. In Rupp-Eisenreich, B. (ed.), *Histoires de l'Anthropologie: XVIXIX siècles.* Paris: Klincksieck, 319–352.

Le Bon, G. (1879). Recherches anatomiques et mathématiques sur les lois des variations du volume du cerveau et sur leurs relations avec l'intelligence. *Revue d'anthropologie,* 2nd ser., 2: 27–104.

Lee, A., Lewenz, M. A., & Pearson, K. (1903). On the correlation of the mental and physical characters in man. Part II. *Proceedings of the Royal Society of London*, 71(467-476): 106-114.

Leonelli, S. (2012). Classificatory theory in data-intensive science: The case of open biomedical ontologies. *International Studies in the Philosophy of Science*, 26(1): 47-65.

—— (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, 82(5): 810-821.

Lewenz, M. A., & Pearson, K. (1904). On the measurement of internal capacity from cranial circumferences. *Biometrika*, 3(4): 366-397.

Lewis, J. E., DeGusta, D., Meyer, M. R., Monge, J. M., Mann, A. E., & Holloway, R. L. (2011). The mismeasure of science: Stephen Jay Gould versus Samuel George Morton on skulls and bias. *PLoS Biology*, 9(6), e1001071.

Lewontin, R. C. (1972). Apportionment of Human Diversity. *Evolutionary Biology*, 6: 381–98.

—— 1974. *The Genetic Basis of Evolutionary Change*. New York: Columbia University Press.

Mari, L. (2000). Beyond the representational viewpoint: a new formalization of measurement. *Measurement*, 27(2): 71-84.

—— (2003). Epistemology of measurement. *Measurement*, 34(1): 17-30.

Michael, J. S. (1988). A new look at Morton's craniological research. *Current Anthropology*, 29(2): 349-354.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88(3): 355-383.

Mitchell, P. W. (2018). The fault in his seeds: Lost notes to the case of bias in Samuel George Morton's cranial race science. *PLoS Biology*, 16(10), e2007008.

Morton, S. G., & Combe, G. (1839). *Crania Americana; or, a comparative view of the skulls of various aboriginal nations of North and South America: to which is prefixed an essay on the varieties of the human species*. Philadelphia: J. Dobson; London: Simpkin, Marshall.

Parsons, F. G., & Keene, M. L. (1919). Sexual differences in the skull. *Journal of Anatomy*, 54(1): 58-65.

Parsons, F. G. (1919). Anthropological Observations on German Prisoners of War. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 49: 20-35.

—— (1922). 14. The Cephalic Index of the British Isles. *Man*, 22: 19-23.

Parssinen, T. M. (1974). Popular science and society: the phrenology movement in early Victorian Britain. *Journal of Social History*, 8(1): 1-20.

Pearson, K. (1902). On the correlation of intellectual ability with the size and shape of the head. *Proceedings of the Royal Society of London*, 69(451-458): 333-342.

Perrin, C., & Anderson, K. (2013). Reframing craniometry: human exceptionalism and the production of racial knowledge. *Social Identities*, 19(1) : 90-103.

Pietsch, W. (2015). Aspects of theory-ladenness in data-intensive science. *Philosophy of Science*, 82(5): 905-916.

Pigliucci, M., & Kaplan, J. M. (2003). On the concept of biological race and its applicability to humans. *Philosophy of Science*, 70: 1161–72.

Porter, T. M. (1996). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.

Poskett, J. (2015). National types: The transatlantic publication and reception of Crania Americana (1839). *History of Science*, 53(3): 264-295.

Richards, R. J. (1987). *Darwin and the Emergence of Evolutionary Theories of Mind and Behavior*. Chicago: University of Chicago Press.

—— (2018). The beautiful skulls of Schiller and the Georgian girl: Quantitative and aesthetic scaling of the races, 1770-1850. In: Rupke, N., & Lauer, G., (Eds.), *Johann Friedrich Blumenbach: Race and Natural History 1750±1850*. London: Routledge, 142-176.

Russett, C. E. (1991). *Sexual Science*. *The Victorian Construction of Womanhood.* Cambridge, MA: Harvard University Press.

Schaaffhausen, H. (1868). On the primitive form of the human skull. *The Anthropological Review*, 6(23): 412-431.

Schiebinger, L. (1989). *The Mind Has No Sex? Women in the Origins of Modern Science*. Cambridge, MA: Harvard University Press.

Schmutz, H. K. (1990). Friedrich Tiedemann (1781–1861) und Johann Friedrich Blumenbach (1752–1840): Anthropologie und Sklavenfrage. In: Mann, G., & Dumont, F., (Eds.), *Die Natur des Menschen: Probleme der Physischen Anthropologie und Rassenkunde (1750–1850)*. Stuttgart & New York: Gustav Fischer, 353-365.

Shapin, S. (1979). Phrenological knowledge and the social structure of early nineteenth-century Edinburgh. *Annals of Science*, 32(3): 219-243.

Shapiro, H. L. (1959). The history and development of physical anthropology. *American Anthropologist*, 61(3): 371-379.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science Part A*, 42(4): 509-524.

Shields, S. A. (1982). The variability hypothesis: The history of a biological model of sex differences in intelligence. *Signs: Journal of Women in Culture and Society*, 7(4): 769-797.

Shortland, M. (1987). Courting the cerebellum: Early organological and phrenological views of sexuality. *The British Journal for the History of Science*, 20(2): 173-199.

Spradley, M. K., & Jantz, R. L. (2011). Sex estimation in forensic anthropology: skull versus postcranial elements. *Journal of Forensic Sciences*, 56(2): 289-296.

Stanton, W. (1960). *The Leopard's Spots: Scientific Attitudes Toward Race in America, 1815-1859*. Chicago: University of Chicago Press.

Stepan, N. L. (1982). *The Idea of Race in Science: Great Britain, 1800–1960.* Hamden: Archon Books.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684): 677-680.

Stocking, G. W. (1968). French Anthropology in 1800, in *Race, Culture, and Evolution: Essays in the History of Anthropology.* New York: The Free Press.

Tal, E. (2011). How Accurate Is the Standard Second?. *Philosophy of Science*, 78(5): 1082–96.

—— (2013). Old and new problems in philosophy of measurement. *Philosophy Compass*, 8(12): 1159-1173.

—— (2016). How Does Measuring Generate Evidence? The Problem of Observational Grounding. In *Journal of Physics: Conference Series* (Vol. 772, No. 1, p. 012001). IOP Publishing.

—— (2017a). Calibration: Modelling the measurement process. *Studies in History and Philosophy of Science Part A*, 65: 33-45.

—— (2017b). A model-based epistemology of measurement. In: Mößner, N., and Nordmann, N., (Eds.), *Reasoning in Measurement*. London & New York: Routledge, 233-253.

—— (2019). Individuating quantities. *Philosophical Studies*, 176(4): 853-878.

Tiedemann, F. (1836). On the Brain of the Negro, Compared with that of the European and the Ourang Outang. *Philosophical Transactions*, 126: 497–527.

Topinard, P. (1885). *Éléments d'Anthropologie Générale*. Paris: A. Delahaye et É. Lecrosnier.

Tuana, N., & Peterson, M. J. (1993). *The Less Noble Sex: Scientific, Religious, and Philosophical Conceptions of Woman's Nature.* Bloomington, IN: Indiana University Press.

Tucker, W. H. (1994). *The Science and Politics of Racial Research.* Urbana: University of Illinois Press.

Van Fraassen, B. C. (2008). *Scientific Representation: Paradoxes of Perspective*. Oxford & New York: Oxford University Press.

Van Wyhe, J. (2017). *Phrenology and the Origins of Victorian Scientific Naturalism*. London: Routledge.

Vermeulen, H. F. (2015). *Before Boas: the Genesis of Ethnography and Ethnology in the German Enlightenment*. Lincoln: University of Nebraska Press.

Vogt, K. C. (1864). *Lectures on Man: His Place in Creation, and in the History of the Earth*. London: Longman, Green, Longman, and Roberts.

Weisberg, M. (2014). Remeasuring man. *Evolution & Development*, 16(3): 166-178.

Weisberg, M., & Paul, D. B. (2016). Morton, Gould, and Bias: A Comment on "The Mismeasure of Science". *PLoS Biology*, 14(4), e1002444.

Williams, E. A. (1985). Anthropological Institutions in Nineteenth-Century France. *Isis*, 76: 331–348.

Winther, R. G., & Kaplan, J. M. (2013). Ontologies and politics of biogenomic 'Race'. *Theoria: A Journal of Social and Political Theory*, 60(3): 54–80.

Young, R. M. (1990). *Mind, Brain, and Adaptation in the Nineteenth Century: Cerebral Localization and its Biological Context from Gall to Ferrier*. Oxford & New York: Oxford University Press.