Studies in History and Philosophy of Biological and Biomedical Sciences 52 (2015) 22–31

Contents lists available at ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

journal homepage: www.elsevier.com/locate/shpsc

Gould on Morton, Redux: What can the debate reveal about the limits of data?

Jonathan Michael Kaplan [a,*], Massimo Pigliucci [b], Joshua Alexander Banta [c]

[a] School of History, Philosophy, and Religion, Oregon State University, Corvallis, OR 97331, United States
[b] City University of New York, Graduate Center, Philosophy Program, United States
[c] Department of Biology, University of Texas at Tyler, United States

ARTICLE INFO

Article history:
Available online 7 February 2015

Keywords:
Research bias
Scientific controversy
Skull sizes
Race
Cranial capacity
Nested analysis of variance

ABSTRACT

Lewis et al. (2011) attempted to restore the reputation of Samuel George Morton, a 19th century physician who reported on the skull sizes of different folk-races. Whereas Gould (1978) claimed that Morton's conclusions were invalid because they reflected unconscious bias, Lewis et al. alleged that Morton's findings were, in fact, supported, and Gould's analysis biased. We take strong exception to Lewis et al.'s thesis that Morton was "right." We maintain that Gould was right to reject Morton's analysis as inappropriate and misleading, but wrong to believe that a more appropriate analysis was available. Lewis et al. fail to recognize that there is, given the dataset available, no appropriate way to answer any of the plausibly interesting questions about the "populations" in question (which in many cases are not populations in any biologically meaningful sense). We challenge the premise shared by both Gould and Lewis et al. that Morton's confused data can be used to draw *any* meaningful conclusions. This, we argue, reveals the importance of properly focusing on the questions asked, rather than more narrowly on the data gathered.

1. Introduction

Samuel Morton was a 19th century physician, sometimes credited with founding physical anthropology in the U.S., who cataloged and measured skulls. Gould (1978) famously argued that Morton's work (Morton, 1849) reflected unconscious manipulation to fit racist stereotypes. Lewis et al. (2011) posthumously rehabilitated Morton, arguing that it was Gould who fudged the results. Lewis et al.'s article received significant press, both popular and scientific. Much of the press took their work to have "debunked" Gould's claims regarding the influence of Morton's unconscious biases on his analyses. As the New York Times put it: "Study Debunks Stephen Jay Gould's Claim of Racism on Morton Skulls" (Wade, 2011); Nature claimed that Lewis et al.'s research showed that "Gould's staunch opposition to racism, and desire to make an example of Morton, may have biased his interpretation of Morton's data" (Mismeasure for Mismeasure, 2011).

Lewis et al. note that "were Gould still alive, we expect he would have mounted a defense of his analysis of Morton"; this is not that defense.[1] Rather, while we agree with Lewis et al. that Gould's statistical analysis of Morton's data is in many ways no better than Morton's own, we believe that Lewis et al.'s work is at least equally problematic. Gould was, in our view, right to recognize that there was something very wrong with Morton's analysis; but he went wrong himself in trying to find a "better" analysis. Lewis et al. are right that Gould's analysis isn't better, but wrong to think that Morton's is appropriate. Further, both Lewis et al.'s analysis of the role that Gould's work on Morton plays in the literature, and of the

* Corresponding author. Tel.: +1 541 224 2994.
E-mail address: jonathan.kaplan@oregonstate.edu (J.M. Kaplan).

[1] Michael Weisberg's "Remeasuring Man" (2014) comes rather closer to providing such a defense.

http://dx.doi.org/10.1016/j.shpsc.2015.01.001
1369-8486/

role played by the measurements of the skulls themselves, are, at best, misleading. Finally, the uncritical "exoneration" of Morton by Lewis et al. incorrectly implies that there was nothing very wrong with either Morton's methods, or with his overall project. We reject both implications.

We have two main goals in this paper. The first is to note the ways in which the adequacy of the evidence gathered depends critically on the questions one is trying to answer. Gould suggests that Morton had a very specific question in mind regarding race, average cranial capacity, and intelligence. But Morton's evidence was not adequate to address this question, and Gould's attempts to find better analyses of the data are equally problematic. Lewis et al. fail to appreciate the problems with Morton's data, and give the impression that while Gould's analysis is mistaken, the project as a whole is reasonable. A more careful consideration of the relationship between the data gathered and the questions that can be answered can help make clear why Morton's data cannot be used to answer the questions attributed to him.

Our second goal is to counter some important problems with the Lewis et al. piece. As noted above, the Lewis et al. article received significant attention in the popular media. But many of the claims made by Lewis et al. in their article are misleading in important ways, and, as we make clear, much of the media attention focused on the most misleading aspects. It is impossible, reading Lewis et al., not to be led to the conclusion that Gould's work was badly flawed, and that Morton's was broadly correct. This is, for the reasons we suggest below, not the case. But the kind of sloppiness that Lewis et al. engaged in has real consequences—e.g., members of the White Supremacist website "StormFront" immediately trumpeted Lewis et al.'s results as proving that Gould was "a fraud," and took them to be broadly supportive of their explicitly racist agenda,[2] a view apparently shared by many in related communities.[3]

We begin this paper with Lewis et al.'s re-measurements of the skulls in Morton's collection. Their discussion of the re-measurement takes up a significant portion of their paper, and much, indeed most, of the media coverage focused on this aspect of their work. We argue that this re-measurement was completely irrelevant to an evaluation of Gould's published analysis of Morton; the exercise was pointless, and there was no legitimate reason to feature the results of that work. The space Lewis et al. devote to their re-measurement of the skulls, as well as the media attention it garnered, form part of a larger pattern of a reframing of Gould's criticisms of Morton that is, again, at best misleading. We next explore briefly some of the ways in which Lewis et al.'s article misrepresents Gould's basic claims, as well as misrepresenting the ways in which Gould's claims are generally interpreted and used.

Gould's actual disagreement with Morton, we maintain, was a disagreement about the correct methods to deploy in the analysis of Morton's data; Gould argued that Morton's choices (which skulls to include and which to exclude, how to compute averages, etc.) were the result of unconscious biases on Morton's part. Lewis et al. counter that Morton's choices, far from being the result of racist biases, were objectively sound, and that Gould's choices were influenced by his own biases, and were unsound. We argue that the methods deployed by Morton and Gould were both inappropriate. 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; questions about the average sizes of the skulls Morton happened to have in his collection are, we maintain, not worth asking, let alone trying to answer. We note here as well that while the application of modern statistical techniques to the skulls in Morton's collection can tell us something about that collection, at best such methods provide more reasons to think that the data from the collection cannot be used as it stands to answer the sorts of questions Gould believed Morton was asking.

2. Remeasuring skulls[4]

Lewis et al. (2011) remeasured 308 skulls from Morton's collection; their results, correcting for systematic changes due to methodological differences, align well with the results Morton reported in his 1849 "catalog" of skulls (Morton, 1849). While their report of this undertaking is only about a quarter of the substantive text of their article, it was the focus of most of the media commentary on their work. For example, Nicholas Wade, writing for the New York Times, declared:

> They identified and remeasured half of the skulls used in his reports, finding that in only 2 percent of cases did Morton's measurements differ significantly from their own. These errors either were random or gave a larger than accurate volume to African skulls, the reverse of the bias that Dr. Gould imputed to Morton. (Wade, 2011)

Writing for *Wired*, Brandon Keim claimed that:

> In a study published June 7 in *Public Library of Science Biology*, researchers led by anthropologists Jason Lewis of Stanford University and the Paleoanthropology Institute's David DeGusta re-measured 308 skulls on which Morton had published data. Their conclusion: Morton's numbers differed significantly from their own in just 7 cases, and those few mismeasurements didn't favor the narrative of Caucasian superiority that Gould ascribed to Morton's motivation. (Keim, 2011)

A New York Times editorial noted that:

> Now a team of six physical anthropologists has filled almost half the skulls with pellets and concluded that Morton's data were generally reliable and not manipulated. ("Bias and the Beholder," 2011)

And an editorial in *Nature* claimed:

> Now, in a paper published on 7 June, Jason Lewis, an anthropologist at Stanford University in California, and his colleagues test Gould's assertions in detail. They remeasured the volume of some 300 skulls in Morton's collection, which survives at the University of Pennsylvania's Museum of Archaeology and Anthropology in Philadelphia, while taking care to blind themselves to knowledge of the population that each skull came from. Comparing their measurements to Morton's, they find no evidence that his were distorted by bias. ("Mismeasure for Mismeasure," 2011)

Finally, note that in a press release for an exhibition on race from the Penn Museum, where one of the study's co-authors (Janet Monge) is a curator and was a consulting scholar on the exhibition being announced, it is claimed that:

> Gould's charges, the first to popularly discredit Morton's scientific methodology, were not challenged until researchers at Penn

[2] See e.g. https://www.stormfront.org/forum/t861796-8/#post9942864 accessed 11/16/2014.

[3] See e.g. http://www.occidentaldissent.com/2011/06/14/the-mismeasure-of-man-stephen-j-gould-refuted/ and http://www.theoccidentalobserver.net/2011/06/stephen-jay-gould-next-to-judas-iscariot-brutus-and-cassius-in-the-devil%E2%80%99s-mouth-at-the-center-of-hell/. Accessed 11/16/2014.

[4] While developed independently, our analysis here is very similar to Weisberg's (2014).

> Museum decided to perform Morton's measurements anew. That team found them largely accurate… ("Morton collection of human crania" 2013)

But from the standpoint of evaluating Gould's published claims, the re-measurement was completely pointless. Gould never claimed that Morton's shot-based measurements, which is what Lewis et al. compared their new measurements to, were unreliable. Rather, Gould explicitly stated that he assumes "as Morton contends, that measurements with shot were objective and invariably repeatable to within 1″3" (Gould, 1978 507). In his *The Mismeasure of Man*, Gould is even more straightforward, and states simply that after Morton switched to lead shot, Morton "achieved consistent results that never varied by more than a single inch for the same skull" (Gould, 1981 53). Gould did not "bother" to re-measure the skulls, because Gould explicitly stated that, once Morton developed a method that made the unconscious "fudging" of the results difficult, the results became reliable. For Gould what was of interest was the difference in the kinds of results obtained when a less reliable method (seed) was used, and those obtained when a more reliable method (lead shot) was used; Gould hypothesized that the less reliable method permitted more room for unconscious bias to influence the results (Gould, 1978 505).

This is also why no one should have considered Michael's previous re-measurements of the skulls in Morton's collection (Michael, 1988) a "refutation" of Gould. Insofar as Michael was testing "whether Morton accurately measured cranial capacity using shot" (p. 350), he was testing something about which Gould never expressed any doubts. It is unfortunate that this aspect of Michael's work received, as far as we can tell, essentially all the attention garnered by that article; as in Lewis et al., remeasuring the skulls was, from the standpoint of evaluating Gould's published claims, a complete waste of time, and detracted from Michael's other (in some cases much more incisive) considerations regarding Gould's analysis of Morton.[5]

Lewis et al. were certainly not ignorant of Gould's claim that the shot-based measurements were to be trusted; what reason, then, did they give for the re-measurements? They claim that if "Gould's hypothesis that Morton physically mismeasured some skulls due to racial bias were correct, we would expect the mismeasured crania to be non-randomly distributed by population" (Lewis et al. 2011 3). Since, for the most part, Morton's measurements using shot were accurate, and, for all but one population, "Morton's errors were random with respect to population," Lewis et al. conclude that "[t]hese results falsify the claim that Morton physically mismeasured crania based on his a priori biases" (Lewis et al. 2011 3).

Again, recall that Gould explicitly stated that the shot-based measurements, unlike the seed-based measurements, were trustworthy; Gould simply never claimed that the shot-based measurements were subject to manipulation via unconscious bias at all. Gould did not claim that, once Morton switched to shot, Morton "physically mismeasured some skulls"—he in fact states the exact opposite of this. Lewis et al. are here falsifying (their word) a claim that Gould never made.

Had their re-measurement of the skulls turned up systematic errors on Morton's part, that would have been an interesting result; it would have implied that Gould was in fact wrong to trust even Morton's shot-based results. But that is not what they found. While it is important to publish negative results, framing those results in ways that suggest that they refute other people's claims, when they do no such thing, is at least misleading, if not dishonest. The physical re-measurement of the skulls, providing evidence that Gould was right to trust Morton's shot-based measurements, cannot possibly have warranted more than a short footnote based on the real intellectual interest of the result. Unless Lewis et al. meant their findings to be misunderstood, it is hard to imagine why they included them so prominently, and why they wrote in ways that strongly imply that they have somehow shown Gould to be in error, when they must have known that they had done no such thing.[6]

3. Seed versus shot measurements: room for bias?

Gould notes that when Morton moved from using seeds to using shot, there were changes in the average skull volumes measured; that these changes were not uniform across Morton's "racial" groupings is, Gould thought, some evidence of bias. Gould claimed that "Indian" skulls originally measured with seed that were then remeasured with shot increased by an average of 2.2"3, and that a reasonable estimate of the change in "Caucasian" skull measurements between seed and shot is about 1.8"3. For "Africans," however, Gould's best estimate for the increase moving between seed and shot based measurement is closer to 5.4"3 (Gould, 1978 506–507). This, Gould argued, is some evidence that, when a method that was more easily subject to unconscious bias was used (seeds), the native "Africans" were systematically disadvantaged (Gould, 1978 507). Gould suggested that this systematic disadvantage was, of course, avoided once Morton switched to a reliable method of measurement, one not subject to unconscious manipulation.

Except for the Native American skulls, Morton did not list individual measurements in his 1839 book (Morton, 1839), and hence Gould did not have access to the individual skull measurements taken with seed that formed the corresponding samples. Indeed, since not all skulls from either the "Caucasian" or "African" samples from 1839 were remeasured with shot, Gould was forced to make some guesses about which skulls from the 1849 catalog (Morton, 1849) formed the 1839 samples. (Lewis et al. reconstruct Gould's reasoning, and find 18 skulls in the 1849 sample that were likely remeasured from the 1839 sample). Since not every skull included in the 1839 estimate was remeasured, Gould was unable to determine what, precisely, the seed-based average from that remeasured subset of the 1839 samples was. He argued that it was unlikely to be

[5] Some authors cited by Lewis et al. (2011) criticize Gould for failing to take Michael's work seriously, but their focus is entirely on Michael's re-measurements, ignoring Michael's much more important contribution—the recognition that the assumptions Gould made in his statistical (re)analysis of the skull volume data were no better justified than those of Morton, and that no statistical analysis would (or even could), in this case, provide the "right" answer. Michael has since published a four part blog post on issues related to Gould, Lewis et al., and Morton, and, while we do not agree with all of his claims, we again find his views to be much more subtle and well-considered than those expressed in Lewis et al. (2011), DeGusta & Lewis (2011), and the various authors of Lewis et al. (2011) in interviews.

[6] Other writers have explored possible reasons for the inclusion of, and the prominence given to, the skull re-measurements by Lewis et al. (2011), focusing on the problematic way in which the skulls were "acquired," and the continued controversy over arguments for the repatriation of, especially, native American remains. See, e.g. Jason Antrosio's comments on this at his "Living Anthropologically" Blog http://www.livinganthropologically.com/2011/06/14/mismeasuring-gould/ (accessed 7/17/2013). Similarly, an editorial in *Nature* notes that "[s]everal in the group have an association with the University of Pennsylvania, and have an interest in seeing the valuable but understudied skull collection freed from the stigma of bias (although, as for many nineteenth-century museum collections, its ethically dubious assembly will remain an issue). (Mismeasure for Mismeasure, 2011 419).

Along these lines, Monge, one of the study's authors and "Curator-In-Charge and Keeper of the Physical Anthropology Section, University of Pennsylvania Museum of Archaeology and Anthropology," stated in an interview that Morton's collection may be "more useful now than in Morton's time," because "the skulls give a snapshot of humanity in the 1800s" ("Report: Skull Study" 2011). We argue below that they provide no such snapshot.

much lower than the estimate that would be generated by the 18 skulls Morton did remeasure, since there is no reason to think that the subset that was remeasured was particularly large or otherwise special; in any event, Gould maintained, in order to account for most of the discrepancy between the difference in moving from seed to shot in Native Americans and Native Africans, one would have to make heroic assumptions about the smallness of the Native African skulls that were not remeasured (Gould, 1978 507).

Michael (1988) argued that a) since there is no sign of mismeasurement in the 1849 shot-based data, and b) the statistical analyses that Morton performed were not, *contra* Gould, clearly biased, there is no reason to suspect that the 1839 seed-based measurements were biased, minor (and statistically untested) anomalies notwithstanding (p. 353). This is the same line of reasoning adopted by Lewis et al., who note that the seed-based measurements were far more variable and suggest that the small sample size and large errors in seed measurements might account for the discrepancy. They argue that "[r]ather than bias, the source of changes between Morton's seed-based and shot-based cranial capacities is more likely that stated by Morton himself: mistakes in the seed measurements" (Lewis et al. 2011 4).

But as Gould noted in his 1978 article, the larger error in seed-based measurements "will increase the variance, but it need not alter the mean for a series of skulls" (Gould, 1978 505). Here, the mean did change, and did so substantially. Without knowing the individual seed-based skull measurements for Native Africans in the 1839 sample (nor even the standard deviation of the sample), it is impossible to directly test the likelihood of generating such different results by chance alone. But, since we do have access to a large collection of individual seed-based skull measurements that were remeasured (the Native American sample), what we can do is to test Lewis et al.'s hypothesis that the larger errors inherent in seed-based measurements, coupled with the smaller sample size, makes such differences in the changes found when moving between seed and shot relatively likely. We chose samples of 18 skulls at random from those Native American skulls measured with seed and remeasured with shot, and computed the change in the mean size between seed and shot. The results indicate that the chances of getting a change of over 5 cubic inches was tiny (far less than 1%; see Supplemental Materials, online). The higher standard deviation of the seed-based data and small sample size are therefore, on their own, very unlikely to account for the difference between the average increase in skull size in the Native American samples and the average increase in the Native African samples. While we cannot single out the sort of unconscious bias to which Gould appealed as the only possibility (there is no way to rigorously test for the possibility of some combination of unfortunate skull selection and chance, and other scenarios surely exist), the possibility of bias in the seed-based (not the shot-based) measurements remains.[7]

In short, the claim that the discrepancy between the seed-based and shot-based results can be attributed merely to the higher variance of seed-based results and small sample size is refuted by our analysis (Weisberg makes a qualitatively similar point; 2014 172). Again, we feel it is important to stress that insofar as (unconscious) bias was a factor in the seed-based measurements, Gould argued that Morton himself recognized that the seed-based measurements were unreliable, wanted to find a more reliable method, and successfully addressed this problem by switching to shot.

[7] Michael notes (personal communication) that so far, at least, no one has tested whether the kinds of unconscious manipulations Gould suggested Morton and his assistant might have engaged in are easier using seed than using shot.

4. Misrepresentations and oddities

To anyone familiar with the published work of Gould on Morton that Lewis et al. cite, the following claim from Lewis et al. strikes an oddly false note: "But it was Morton's work on human skulls that drew first Gould's interest, then his ire" (Lewis et al. 2011 1).[8] This is a strange line, because, if anything, the impression one gets from Gould's published writing on Morton—the work Lewis et al. cite—is not that Morton drew Gould's "ire" but rather than Gould had a real (if sometimes grudging) respect for Morton and his work. In his 1978 paper, after all, Gould is not interested in "fraud" or "scientific misconduct" (more on this below), but rather in the sorts of unconscious biases and short-sightedness that even the most scrupulous and honest researchers can suffer from. And his recommendation for dealing with these kinds of biases is, explicitly, that we act more like Morton: "I only raise what I regard as a pressing issue with two hopes for alleviation—first, that by acknowledging the existence of such a large middle ground, we may examine our own activity more closely; second, that we may cultivate, as Morton did, the habit of presenting candidly all our information and procedure, so that others can assess what we, in our blindness, cannot" (Gould, 1978 504–505).

Gould credits Morton with recognizing that the measurements performed with seed were unreliable, and switching to a more reliable method, one less easily manipulated by unconscious biases. Gould writes that: "Indeed, we know that Morton himself began to worry. He had hired assistants to measure the Indian crania, but, distressed by errors and inconsistencies, he later took to making all measurements himself with lead shot" (Gould, 1978 505). Gould never accuses Morton of wanting biased results, or of consciously trying to manipulate data; rather, he suggests, in his 1978 article and in *Mismeasure*, that Morton went out of his way to try to get accurate answers, and to avoid unreliable results when he could. Gould writes that he finds "no indication of fraud or conscious manipulation… [Morton] explained everything he did, and published all his raw data…" (Gould, 1978 509). Nowhere in the article, nor in Gould's *The Mismeasure of Man*, can we find any hint of Gould's suggesting that Morton was anything other than a careful and honest researcher, albeit one blind to his own unconscious biases and pre-conceptions about racial hierarchies, biases that were, after all, quite common in the cultural context in which Morton was born and raised.[9]

[8] Michael forcefully reminds us (personal communication) that after the publication of *Mismeasure*, Gould's public talks and interviews on Morton painted Morton in a much less flattering light than does his published work. But in the work cited by Lewis et al., Gould remained carefully respectful of Morton, and gives no hint that he thought of Morton as anything but honest and careful. That Gould in fact said things more similar to what Lewis et al. attribute to him in other, uncited sources, does not excuse their misrepresentation of his published works.

[9] As noted, in our view, Gould's published writing reveals a real respect for Morton's intellectual honesty; Gould clearly stresses, in his written work, that Morton both wanted to get things right, and tried to do so. While in the article itself, Lewis et al. pretend to a similar respect for Gould, there is good evidence that at least some of the authors are being disingenuous about their actual views here, and believe Gould to be guilty of *conscious* fraud and data-manipulation. Holloway, one of the co-authors, is quoted as saying that "I just didn't trust Gould… I had the feeling that his ideological stance was supreme. When the 1996 version of 'The Mismeasure of Man' came and he never even bothered to mention Michael's study, I just felt he was a charlatan" (Wade, 2011). And *Nature*, in their editorial on the study, notes that "Although the new paper does not accuse Gould of intentionally misrepresenting Morton, some of its authors have raised this possibility in interviews, noting that Gould's oversights would be less troubling were he known to be a less meticulous scholar…" (Mismeasure for Mismeasure, 2011 419). We note in passing that, while he was alive, Gould was called many things, but "meticulous scholar" was not often among them, even by his supporters.

Lewis et al. claim that, based on Gould's analysis, "Morton is now viewed as a canonical example of scientific misconduct" (Lewis et al. 2011 1). They reiterate later that "Morton has become a canonical example of scientific misconduct and an oft-told cautionary tale of how human variation is inevitably mismeasured" (Lewis et al. 2011 2), and finally that "Samuel George Morton, in the hands of Stephen Jay Gould, has served for 30 years as a textbook example of scientific misconduct" (Lewis et al. 2011 5). These are odd claims, not least because it was very important to Gould's stated aim in his 1978 piece that Morton *not* be guilty of "scientific misconduct." It is, rather, critical to Gould's project of revealing the ways in which unconscious biases and assumptions can result in honest, careful scientists producing mistaken results, that Morton be an honest, careful scientist, trying to generate correct results[10].

Might it be the case, for all of Gould's care in laying out his project, that his work was misinterpreted as showing that Morton was guilty of scientific misconduct? It is possible, of course, but the only authors we can find who make that mistake are authors who are attacking Gould's work. Lewis et al. themselves cite no textbooks that imply that Gould's analysis shows that Morton committed scientific misconduct or fraud, and we have been able to find none; Lewis et al. do cite three works in support of their *claim* that Gould's analysis is often used to attribute scientific misconduct and fraud to Morton, but what none of these authors do is provide any references or citations to articles or books that in fact make that mistake. Cook, whose work is cited to support the claim that "Samuel George Morton, in the hands of Stephen Jay Gould, has served for 30 years as a textbook example of scientific misconduct" in fact mentions neither scientific misconduct, nor fraud, nor any other obvious synonyms; she claims that her "experience in teaching Gould's paper to undergraduates has been that Gould unfairly brands Morton as racist" (Cook, 2006 40), not that Gould's analysis has made Morton out to be an exemplar of scientific misconduct. Lewis et al.'s reference to her is especially odd given that they state that "Morton indeed… assigned a plethora of different attributes to various groups, often in highly racist fashion" (Lewis et al. 2011 5); we hope it is obvious that it is hard to reconcile Morton being unfairly branded as a racist with this claim. One can argue about whether Morton's obvious racism is a moral failing, given how common such views were at the time he was working, but, given what Morton writes about the "character" of non-white peoples, we cannot imagine how calling him a racist is anything other than descriptively accurate.[11]

Buikstra (2009) is similarly interested in defending Morton, and in her introduction to the reproduction of Morton's *Crania Americana* does so with vigor. What she doesn't do, however, is provide any evidence in support of Lewis et al.'s claim that "Gould's analysis of Morton is widely read, frequently cited, and still commonly assigned in university courses" (Lewis et al. 2011 2), which is the claim Lewis et al. cite her work to support. Brace (2005) likewise presents a spirited defense of Morton's work, understood in context. Again, however, what this does not do is provide any evidence for the above mentioned claim by Lewis et al., except insofar as Brace in fact cites Gould in order to criticize the latter's interpretation of Morton. Still less does Brace suggest that Gould's analysis can be related to claims of "scientific misconduct" on Morton's part—indeed, Brace is quite explicit that Gould's interpretation is that Morton was not guilty of fraud or misconduct, but of unconscious bias; interestingly, Brace attributes Gould's poor treatment of Morton to this same kind of unconscious bias on Gould's part (Brace, 2005 88).

In short, there is a troubling disconnect between what Gould actually claimed, and what Lewis et al. attribute to him, as well between what other authors have in fact written about Gould on Morton, and how Lewis et al. interpret Gould's place in the literature on Morton. This disconnect is especially striking in light of the claim, noted in the editorial in *Nature*, that the Lewis et al.'s manuscript "spent eight months in the review process at PLoS Biology" (Mismeasure for Mismeasure, 2011 419).

5. Questions and the statistical techniques to address them: why there is no right answer

Gould wrote as if it was obvious that Morton's purpose in measuring the volumes of the skulls was to demonstrate that the "races" Morton recognized could be ranked by intellectual ability, using brain size as the proxy. With this question in mind, Gould attempted to show that Morton's statistical manipulations were poorly justified. If Morton actually had a question like Gould's in mind, Gould was surely right that Morton's statistical analysis falls far short of providing a defensible answer. But Gould's own statistical manipulations are no better able to answer the question Gould imagines that Morton is asking. Gould is simply wrong to argue that the question he supposes Morton is asking can in fact be answered by the data Morton had gathered. We would suggest, in line with Brace (2005), that it was at least in part because Gould was able to get an answer that he liked better than Morton's that Gould was unable or unwilling to see that his own statistical manipulations—his own decisions about which skulls to include, which to exclude, how to group them into populations, and how to perform the analyses in question—were (at best) no better justified than Morton's. In this section, we present our reasons for thinking that Morton's skulls cannot be used to answer (some of) the sorts of questions people have tried to answer using them.

If Morton's project was driven by a desire to find a way to objectively rank the races by intelligence (Gould, 1978 503), quibbles over the best statistical approaches, and which particular subpopulations or individual skulls to include or exclude, seem misguided. The more important problem is that reaching conclusions about the "average" cranial capacity of the various "races" requires that one has a defensible position regarding what it means to generate a population average for cranial capacity, a defensible way of recognizing biologically meaningful populations, and a defensible way of gathering a representative sample of skulls from the relevant populations in order to take the relevant measurements. But no such defenses are available in this case. Indeed, the lack of clarity about what constituted a "race," how the "races" and populations identified as sub-populations within races were related to each other, what, precisely, the question being asked about average capacities even was, and the implausibility of treating the skulls gathered as representative samples of the populations to which they were assigned, render the statistical quibbles moot, and make Gould's answers no more defensible than Morton's. If Morton's project was, as Gould believed, to rank the "races" by intellectual ability on the basis of skull volume, the project was hopelessly confused from the start.

[10] Much of our argument regarding the important distinction between "misconduct" and "unconscious bias," as well as Lewis et al.'s failure to cite work that actually defends their claims, was anticipated by Jonathan Marks' (2011) blog post "Plotz biology" on his "Anthropomics" blog (Marks, 2011), and we have made use of some of his analysis in developing our own; while any errors introduced here are our own, much of the credit should go to Marks.

[11] "Racism" is a contested term (see e.g. Doane, 2006), and it is beyond the score of this paper to engage in the literature on defining racism. But Morton attributes morally loaded characteristics to "racial" populations in quite wildly offensive ways, ways that surely count as racist on any reasonable definition (see e.g. various descriptions of populations in Morton, 1839). Again, whether, given the time period in which he was writing, we should think of his racism as a (major) moral failing is a different issue.

Consider: if Morton was trying to find the average cranial capacity of members of the "Indian" or "Native American" race, what, exactly, was he trying to find the average *of*? Perhaps he meant the average adult cranial capacity, and hence should exclude children (Morton in fact sets the cut-off at the age of 16, arguing that the brains of people under this age were still growing; see 1849 VII). Should we also exclude skulls from those individuals identified as "insane" (by whom?) or from so-called "idiots"[12] or from those with badly "deformed" skulls (what counts as "deformed" here? What standard should be used as the cut-off?)? Why or why not? Should we exclude African Americans, many of whom are a mixture of African and European heritage (ignoring for the moment the fact that "Africans" themselves do not constitute either a population or a race)? Why or why not? (Morton in fact treats them as a separate sub-population, at the same level as the "Native African Family," see 1849 VI and VIII.) Should we include first-generation offspring of "native" Americans and (nearly) contemporary Europeans? Those with "at least" three "native" grandparents? At least two? Why or why not? Should we worry about the "admixture" of different "native" populations *within* the Americas? Why or why not?

Turning from the difficulties with deciding what sorts of individuals should be part of our averages, what is it, precisely, that we are trying to find the average *of*? Should we, for example, be worried about the fact that the total population of "native" Americans in 1839 was vastly reduced from its historical high, due to genocide, disease, and other "influences" from European settlers? The demographics—which "sub-populations" were more numerous, which rarer, which entirely extinct—had changed radically from the historically common distributions, and were still, in the middle of the 19th century, undergoing rapid and fairly unpredictable changes. What, under these circumstances, does it mean to speak of the "average" cranial capacity of an "Indian"?

Even if one were to make some principled decisions regarding the questions of appropriate sampling for determining the average cranial capacity of the "populations" identified, a more fundamental problem would remain. The groups identified by Morton are not obviously "races" (nor even "populations") in any biologically respectable sense. Leaving aside some of the more egregious (but perhaps historically understandable) errors, such as lumping Australian aboriginals and Native Africans together in one "race" (Morton, 1849 vi) (see also Gould, 1978 508), it simply isn't at all clear what the (actual) biological status of the "groups" that Morton identified might be. So, for example, Morton treats "Mexicans" as one of two subgroups in the "Toltecan Family" (along with "Peruvians"), which itself is a member of the Native American "race" (or "Group"—more below on Morton's often confused and inconsistent taxonomic practices). Should we interpret this as a claim about the actual taxonomic status of these groups? If so, it is clearly wrong, but worse, it is not clear what would be right (who count as "Mexicans" in this context?). Even contemporary practices with respect to identifying and naming significant populations below the species level are controversial and often confused, and this kind of enterprise is still often contentious; are the groups identified supposed to be (perhaps temporary) clades (Andreasen, 2004)? Population-genetic clusters (Rosenberg et al. 2002)? Ecotypes (Pigliucci & Kaplan, 2003)? A population identified by interactions (Millstein, in press)? Whether a particular group is biologically meaningful depends critically on the kind of group one is looking for, and on one's beliefs about what constitutes a biologically meaningful group (Kaplan & Winther, 2013, 2014; Winther & Kaplan, 2013; Winther, 2014). Morton's answers to these questions are indefensible, and there remains no straightforward way to answer them.

Indeed, it is not always clear how Morton thought about or referred to the various "levels" of population organization he studied. For example, the nesting of "Families" within "races" and subgroups within "Families" is not particularly clear in Morton. Putative populations are identified by different names in different places, it is often unclear from the text into which group a particular individual should be placed, or how smaller populations should be put together into larger "Families," etc. Brace, who provides a vigorous defense of Morton against "Whiggish" interpretations of his work, especially Gould's, notes that there is "complete" confusion in Morton's distinctions between "species," "races," "groups," "primitive varieties" and the like (Brace, 2005 86).[13] Lewis et al. note in passing that "studies have demonstrated that modern human variation is generally continuous, rather than discrete or 'racial,' and that most variation in modern humans is within, rather than between, populations" and hence it is only "with substantial reluctance that we use various racial labels" (Lewis et al. 2011 2). But the labels that they use, and the groups that they identify, are not obviously the same as those used by Morton, nor the same as those used by Gould, nor did Gould always follow Morton's (inconsistent) classificatory practices. It is clear in neither Gould nor Lewis et al. whether the attempt is to recreate Morton's reasoning, as closely as possible, or if it is to apply more modern standards to Morton's work; both Gould and Lewis et al. seem to slide uncomfortably between these very different projects, and the impossibility of rigorously recreating Morton's less-than-fully consistent classificatory schemes make the entire project that much more difficult. This is especially true in light of modern, biologically informed concepts of populations and races to which Morton did not have access.

Again, though, even if one were to make an informed and principled decision regarding what kinds of groups one cared about, and what the biological criteria for those groups might be, Morton's dataset does not include the necessary information that could permit us to categorize the skulls into those groups. This isn't surprising; Morton himself simply could not have organized his samples in ways that would make sense to us, since the concepts necessary to do so were not fully developed until after the Modern Synthesis in evolutionary biology and the techniques to identify and sort individuals by sub-population are still being actively developed and debated about today (Kalinowski, 2011; Rosenberg et al. 2002).

The problem of sampling compounds the troubles noted above: even if we had a principled reason to specify some particular question from among the possible questions suggested, figuring out how to go about answering it, given the need to specify (sub) populations from which to sample, seems at best difficult. When the problem of "acquiring" skulls is taken into account, the hopelessness of the sampling problem is made obvious: if one's sampling technique is "I include whatever skulls people send to me"

[12] Morton also excluded "idiots" from the tables of averages, but gives little sense of who counts as an "idiot" in this context (see Morton, 1849 IX and Morton, 1850 246). More generally, what method of testing should be used to identify idiots? If one population has a larger proportion of "idiots" than another, should this count against them in some way? Either answer seems defensible, depending on what one wanted to do with the answer. More on this problem below.

[13] Confusions surrounding these levels may bare on some of the arguments surrounding Gould's treatment of Morton's work, especially Gould's claims regarding Morton's failure to calculate certain "Indian subsample means" (Gould, 1978 505), claims that Lewis et al. (2011) argue are clearly false.

there is no reason at all to think that the samples will be statistically appropriate for *any* purpose, let alone the purpose specified.[14]

The absurdity of Gould's "solution" to this problem—calculate the mean of each "subpopulation" identified by Morton separately, and then average those—should now be obvious. Despite what some authors have claimed, Gould's analysis was not wrong because he deployed statistical methods that were too sophisticated or anachronistic. Cook (2006), in her defense of Morton, may be right that the "concept of grouped means is exceptionally difficult for students who lack a quantitative bent," but that, even if true, is surely irrelevant. The problem of unequal sample sizes is a serious one, and needs to at least be addressed—even Morton recognized that. It was why, after all, he didn't calculate a "grand mean" for the "Caucasian" race: "No mean has been taken of the Caucasian race collectively, because of the very great preponderance of Hindu, Egyptian and Fellah skulls over those of the Germanic, Pelasgic and Celtic families" (Morton, 1849 9). Morton followed up on this reasoning later, noting that no fair "collective comparison" could "be instituted between the Caucasian and Negro groups in such a table, unless the small-brained people of the latter division (Hottentots, Bushmen and Australians) were proportionate in number to the Hindoos, Egyptians and Fellahs of the other group" (1850 248). Gould's (reasonable) complaint was that Morton did not apply this same logic to the Native American group. As Weisberg notes, Gould's complaint that "Morton included small-headed Inca Peruvians in the American mean, but excluded small-headed Hindus from the Caucasian mean," and that "[t]his allowed Morton's American sample to appear smaller and Caucasian sample to appear bigger than it might have been otherwise," is surely justified (Weisberg, 2014 173)[15].

The issue with Gould's "mean of means" approach is that the statistical techniques he deployed were applied to a problem that only superficially resembles the problems for which they are appropriate. The groups identified are *not* independent entities, they do *not* represent populations of equal size (neither in 1839/49, nor historically before the populations were decimated), most of the samples are too small for it to make sense to treat them as representative (even excluding those that consist of a very few, or even a single, individual), and the groupings of "races," "populations," etc., bear no obvious resemblance whatsoever to biological reality. Under these conditions, there is no way to generate a reasonable approximation to the values of the populations from which the samples were drawn, and the "mean of means" approach just generates an entirely new meaningless metric based upon the old meaningless metrics.

All of this, however, presupposes that Gould was right about the reason that Morton took to measuring skulls and reporting "racial" averages, that is, that Morton really was attempting to calculate the average cranial capacity of members of the various major "races" in order to support a hierarchical ranking of the intellectual capacity of members of those "races." But Lewis et al. join various others (Buikstra, 2009; Cook, 2006) in denying this. Buikstra (2009) in particular argues that claims about "intellectual capacity" formed at best a small fraction of Morton's writing, and that for all the attention Morton's cranial capacity measurements have received, these too were a fairly minor part of Morton's overall project (p. 27). She writes that:

> In Crania Americana, the small table reporting means and ranges for cranial capacity appeared almost an afterthought, at the end of the brief discussion of Morton's conclusions and did not affect them. One can almost hear the author thinking that he had collected all these data and since others may be interested, he should include them. They were not referenced earlier in the text and were certainly not fundamental to his arguments. (Buikstra, 2009 28)

DeGusta and Lewis argue that rather than hoping to establish that there was an objective racial hierarchy based on intelligence as governed by skull size, "Morton hoped to determine whether different human populations were one species or many, and thus whether the divine creation had been singular or a play in several acts" (DeGusta & Lewis, 2011). Similarly, Brace argues that Morton in fact was a strong proponent of polygenism, and that his advocacy for that position colored many of his views and interpretations of ethnographic data (Brace, 2005 83–86). Cook (2006 38) however, denies that Morton was even particularly interested in defending the theory of polygeny (though he clearly supported it), let alone in developing empirical arguments for the presumed intellectual rankings of the races. Morton's intentions are thus not at all clear.

If Morton's goal in collecting and measuring the skulls of different races was to support polygenist theories, it is obvious that the project was hopeless, and that arguments over which skulls to include or exclude, how to group the skulls, or even whether the skulls represent a reasonably representative sample of the populations from which they were drawn, are completely moot. There are no measurements that could, given what we know today, support polygenist theories, because those theories, requiring as they do that species be the result of separate creation events, are at odds with basic evolutionary biology.

If we accept that Morton wished to use his skulls to make inferences about the populations from which those skulls were drawn, do his measurements of cranial capacities give us reasons to think that these inferences will be particularly robust or meaningful? The answer, we submit, is clearly no. Here, to underscore the futility of the arguments surrounding the best ways to summarize the data, we performed an analysis of variance (ANOVA) on Morton's skull volume data. Again, this does not imply that we believe Morton's data can be used for any useful purpose; rather we wanted to see if there was any substantial systematic variation in skull sizes among Morton's poorly sourced, non-randomly sampled, and ill-conceived, "races" at all. We found that the vast majority of the variance in skull size is attributed to *individual* differences (differences between individuals) within Morton's "families." Only a small part is associated with the average differences between "families" within Morton's "races," and an even smaller part is associated with the "races" themselves (see Supplemental Materials, online). In fact, the variance in skull volumes among individuals is 44 times larger than the variation among so-called "races".[16] Morton's "races" therefore explain

[14] This is, we admit, an unfair summary of Morton's sampling technique; Morton used skulls found in archeological burial sites, "acquired" from near-contemporary cemeteries, as well as those that were sent to him, often accompanied by notes on the providence, by various people in the field. In addition, he actively sought skulls from populations that he felt were underrepresented in his samples (see Morton, 1849). But our point is that none of these is a "good" sampling method, and there is no reason to think that Morton's collection of skulls was representative of the populations from which the skulls were drawn, always keeping in mind that the concept of a "population" is contentious in this context.

[15] Some of Gould's suggestions and criticisms of Morton's work, it should be clear, *are* anachronistic, and perhaps deliberately so. His removal of the "Australoid family" from the "Negro mean," on the grounds that we now know those populations not to be closely related is one example (Gould, 1978 508), and his notes concerning the correlation between brain size and body size (see e.g., Gould, 1978) 506) are another place where Gould's methods rely on information of which Morton could not have been expected to be aware.

[16] Weisberg makes a similar point qualitatively, noting that "since there is much variation within races, and relatively small differences between the racial aggregates," Morton ought to have been "much more cautious about reporting means for each race" (Weisberg, 2014 173).

virtually none of the variation in skull sizes, which should not be surprising considering the problems with his dataset that we have already described. So in what sense, then, could Morton (or Gould) possibly be "right" about anything, if there was virtually no systematic variation in skull sizes among Morton's "races" to begin with? It is obvious why different approaches to analyzing this data yield such radically different answers; the small proportion of the variance associated with Morton's groups makes any analysis of the data overly sensitive to trivial differences in decisions regarding which skulls to include, how to treat the different groups, what groupings to use, what statistical methodologies to deploy, etc.

If Morton wasn't collecting and measuring skulls to support the intellectual ranking of the various races, nor to support polygenist theories, why was he doing it?[17] It is possible that Morton had no particular purpose in mind. A fascination with skulls and measurements could easily result in reams of data being gathered with no guiding hypothesis to be tested. And given Morton's strong interests in "races," summarizing the data along those lines might have been, as Buikstra (2009) suggests, something of an afterthought. In this context it is perhaps worth noting that at the same time Morton was working so assiduously on his human skull collections, he was also publishing detailed descriptions of e.g. fossil crocodile skulls (Morton, 1845).

But if the human skulls were gathered and measured without any clear purpose for which a summary of the data would be useful, the idea that there is a "correct" way to analyze the resulting data should seem even less plausible. If the data collection scheme was not guided by specific questions, then it is hard to see how any statistical approach(es) could rescue that data for any particular purposes later on. And the questions guiding the data collection have to be well thought out too. To take a trivial example, if a particular genus of plants, say, a kind of tree, has two extant species, and we ask how tall a fully grown average member of the genus is, it is unclear how one should answer—even if one knows everything there is to know about the trees in question. If one species is far more common, and you interpret the question to be about the average height of the individual trees in the world, a reasonable answer might be approximately the average height of members of the much larger population (the smaller one being statistically swamped). But we might be interested not in the average height of extant trees, but in the average height of the species within the genus; here, Gould's "mean of means" would be appropriate. If the two species live in different areas, we might be interested in how tall each kind grows, on average, where it is "native," or we might be interested in how tall they would grow under some more equal environmental regime (the latter would be especially pressing if I were considering the two trees as options for planting in front of my house, for example). Whether we care about the proportion of "juvenile" trees, and their heights, will again, depend critically on the force of our question. Without a better specified question—which often demands that we know what we want to do with the answer, that is, why we are asking the question—it isn't at all clear what methodology should be deployed.

In the end, the question "Whose statistical approach to summarizing the skull volume data was right? Morton or Gould?" is the wrong question to ask. It is hard to see how a set of skulls, collected unsystematically, often of uncertain provenance (how far should we trust the descriptions of where the skulls came from that were sent to Morton along with the skulls?), and identified in pre-modern fashion with no way to tie them back to meaningful biological groups, could be usefully deployed to answer any meaningful question about the larger populations from which they were drawn. Neither Gould's nor Lewis et al.'s analysis is appropriate for answering many questions that a reasonable person might want answered, in a modern context where biological definitions of "race" and "population" are used instead of arbitrary pre-modern delineations. Debates about how best to summarize measurements of those skulls should be seen as hopelessly misguided. That Gould permitted himself to get sucked into such a debate is telling—the availability of another way of analyzing the data clearly made doing so tempting to him. Lewis et al. were right to call Gould out on this failure, but in the end, they followed the same garden path, refusing to see that trying to make use of Morton's skulls for any interesting purposes is just pointless.

The lesson here can be generalized: without knowing what question(s) a particular dataset is meant to answer, it is hard to evaluate the quality of the data. Data that would be reliable for some purposes might be quite unreliable for others. Arguments over the "best" ways of treating data should be seen as secondary to evaluating whether the data is adequate to the purposes one hopes to use it for. Where the data is simply inadequate to the purposes to which is it pressed into service, *that* is the only conclusion that ought to be drawn.

6. Skulls and brain size in context: race and racism in the contemporary literature

What does it mean to write that the "data on cranial capacity gathered by Morton are generally reliable, and he reported them fully" (Lewis et al. 2011 6)? Is the claim merely that Morton accurately measured those skulls that happened to fall into his collection, and whose measurements he happened to think were worth including? Or is the claim rather that since "Morton's methods were sound," we expect that if we surveyed cranial capacities using larger and statistically sound samples from the same locations and time periods, etc., and attempted to match the populations to those that Morton thought he was sampling, we would find that the average cranial capacities were similar to those reported by Morton?[18] Lewis et al. are not entirely clear about this. And this seems something that one ought to be very clear about.

Whatever Morton believed about the relationship between cranial capacity, intelligence, and race, the belief that differences in average cranial capacity explain differences in intelligence both within and between the so-called "races" remains alive today. While this position is defended explicitly by relatively few serious academics, "relatively few" is not none, and those few academics garner a significant amount of attention from outside academia, much of it laudatory (any reader unconvinced of this need only perform a simple Google search on the names of leading "researchers" supporting these positions). And, in this literature, Morton's results are treated not as interesting historical curiosities,

[17] It is clear that Morton had many beliefs about the intellectual abilities of the different groups he identified and that his interpretations of skull features followed this (see Weisberg, 2014 168–16); similarly, Morton's support of polygenism is undeniable. But whether this is why he collected, measured, and reported on skulls is another matter. We take no stand on this question, and merely note that what one thinks Morton was up to will influence what kind of data one thinks Morton needed to gather given his interests, and what kinds of decisions would therefore be better and worse justified.

[18] We take Gould's argument to be that one would *not* expect the population averages to reflect the averages reported by Morton. We remain convinced that that is true, but have argued that Gould overstepped when he suggested that, given the skulls in Morton's collection, some *different* average should be expected (e.g., Gould's 1978 Table 6). Rather, given the poor dataset, we suggest that *no* legitimate inferences to "natural" populations can be drawn.

but as just another part of the evidence for these hypotheses. For example, Rushton & Rushton (2003) write that "Morton (1849) … found that Blacks averaged about 5″3 less cranial capacity than Whites. These results have stood the test of time" (p. 141) (see also Rushton & Jensen, 2005 255; Rushton & Ankney, 2009 693 & 710).

It is difficult to take seriously the claim that Lewis et al. in fact "find other things to admire in Gould's body of work, particularly his staunch opposition to racism" (Lewis et al. 2011 5) when the thrust of their article is an utterly uncritical acceptance of Morton's "results" as accurate, and Gould's as uniquely flawed.[19] Lewis et al. note that Morton "assigned a plethora of different attributes to various groups, often in highly racist fashion," (Lewis et al. 2011 5) but these assignments are never critically explored, nor are the relationships between these assignments and contemporary research programs noted, even in passing. And it seems to us to be irresponsible to claim, as Lewis et al. do, that "studies have demonstrated that modern human variation is generally continuous, rather than discrete or 'racial,' and that most variation in modern humans is within, rather than between, populations" (Lewis et al. 2011 2) without noting the ways in which this consensus has been recently challenged, and the role that those challenges have played in contemporary debates surrounding biological racial realism (Edge and Rosenberg, in press; Edwards, 2003; Kaplan & Winther, 2013, 2014; Ludwig, in press; Pigliucci, 2013; Spencer, in press; Weiss and Fullerton, 2005; Winther, 2014; Winther et al., in press).

Whatever Morton's goals in gathering and summarizing skull-volume data by race, his work was, verifiably, used by racists to defend e.g. the continued practice of slavery. As Brace notes, towards the end of his life, Morton corresponded with John Calhoun, an "outspoken slavery proponent," and, as "a result of his correspondence," Morton sent Calhoun his major works, which "confirmed Calhoun's racial bigotry" and provided intellectual grounding for his pro-slavery position (Brace, 2005 92). And, as noted above, Morton's work continues to be used to support positions associated with what are (in our view) appalling political positions on the basis of wildly inadequate evidence. This is the context in which a "defense" of Morton's assumptions, methods, and results must be understood.

For all that, continuing to ignore Gould's real mistakes and flaws would, indeed, be intellectually dishonest; no matter how much one might prefer the results of Gould's statistical summary, the assumptions used in generating those summaries are simply not supportable. But, there is also no reason to accept the assumptions used by Morton in generating his original estimates, nor those used by Lewis et al. in generating their (very similar to Morton's) estimates.

Acknowledgments

We thank Weisberg for his comments on an earlier draft of this paper. We thank Jake Michael for providing feedback on a draft of a paper, and helpful discussion. Jay Odenbaugh's questions and comments at a talk based on an early draft of this paper were very useful. Rasmus Grønfeldt Winther provided helpful comments on several drafts, and useful discussion of some key issues. Janet Stemwedel's suggestions were helpful. We thank four anonymous reviewers for their thoughtful comments and suggestions.

[19] And again, see above on the comments (some of) the authors of Lewis et al. have made in interviews regarding Gould being a "fraud" and a "charlatan," comments that, as we noted above, have been picked up on and used by members of explicitly White Supremacist websites and blogs.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.shpsc.2015.01.001

References

Andreasen, R. O. (2004). The cladistic race concept: A defense. *Biology & Philosophy, 19*, 425–442.
Brace, C. L. (2005). *"Race" is a four-letter word: The genesis of the concept*. New York: Oxford University Press.
Bias and the beholder. [Ediorial]. (2011, June 14). (p. 4). New York Times.
Buikstra, J. E. (2009). *Introduction to the 2009 reprint edition of crania Americana* (pp. i–xxxvi). Davenport, Iowa: Gustav's Library.
Cook, D. C. (2006). The old physical anthropology and the new world: A look at the accomplishments of an antiquated paradigm. In J. E. Buikstra, & L. A. Beck (Eds.), *Bioarchaeology: The contextual analysis of human remains* (pp. 27–71). Amsterdam: Elsevier.
DeGusta, D., & Lewis, J. E. (2011). *Gould's skulls: Is bias inevitable in science?*. New Scientist.
Doane, A. (2006). What is racism? Racial discourse and racial politics. *Critical Sociology, 32*(2–3), 255–274.
Edge, M. D., & Rosenberg, N. A. (2015). Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* (in press)
Edwards, A. W. F. (2003). Human genetic diversity: Lewontin's fallacy. *BioEssays, 25*, 798–801.
Gould, S. J. (1978). Morton's ranking of races by cranial capacity: Unconscious manipulation of data may be a scientific norm. *Science, 200*, 503–509.
Gould, S. J. (1981). *The mismeasure of man*. New York: W. W. Norton and Company.
Kalinowski, S. T. (2011). The computer program structure does not reliably identify the main genetic clusters within species: Simulations and implications for human population structure. *Heredity, 106*, 625–632.
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*, 401–412.
Kaplan, J. M., & Winther, R. G. (2014). "Realism, antirealism, and conventionalism about race." Jonathan M. Kaplan and Rasmus Grønfeldt Winther. *Philosophy of Science, 81*, 1039–1052 (December 2014)
Keim, B. (2011, June 14). The mismeasures of Stephen Jay Gould. Retrieved January 27, 2015, from http://www.wired.com/2011/06/gould-morton-revisited/.
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*.
Ludwig, D. (2015). Against the new metaphysics of race. *Philosophy of Science* (in press)
Marks, J. (2011, June 17). Plotz biology. Article posted at Anthropomics blog. http://anthropomics.blogspot.com/2011/06/plotz-biology.html. Accessed 28.08.13.
Michael, J. S. (1988). A new look at morton craniological research. *Current Anthropology, 29*, 349–354.
Millstein, R. (2015). Thinking about populations and races in time. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* (in press)
Mismeasure for mismeasure: A critique of the work of Stephen Jay Gould should serve as encouragement to scrutinize the celebrated while they are still alive, 2011. [Ediorial]. Nature, 474, p. 419.
Morton collection of human crania in the spotlight with year of proof: Making and unmaking race, 2013. [Press Release]. http://www.penn.museum/press-releases/894-making-and-unmaking-race.html. Accessed 28.08.13.
Morton, S. 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.
Morton, S. G. (1845). Description of the head of a fossil crocodile from the cretaceous strata of New Jersey. *American Journal of Science and Arts, 48*, 265.
Morton, S. G. (1849). *Catalogue of skulls of man and the inferior animals* (3rd ed.). Philadelphia: Merrihew and Thompson Printers.
Morton, S. G. (1850). Observations on the size of the brain in various races and families of man. *American Journal of Science and Arts, 9*(26), 246–249.
Pigliucci, M. (2013). What are we to make of the concept of race? Thoughts of a philosopher scientist. *Studies in History and Philosophy of Biological and Biomedical Sciences, 44*, 272–277.
Pigliucci, M., & Kaplan, J. (2003). On the concept of race and its applicability to humans. *Philosophy of Science, 70*, 1161–1172.
Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., et al. (2002). Genetic structure of human populations. *Science, 298*, 2381–2385.
Report: Skull study right, results misused. (2011, June 14). Retrieved January 27, 2015, from http://www.upi.com/Science_News/2011/06/14/Report-Skull-study-right-resultsmisused/UPI-25651308086338/.
Rushton, J. P., & Ankney, C. D. (2009). Whole brain size and general mental ability: A review. *The International Journal of Neuroscience, 119*, 692–732.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law, 11*, 235–294.
Rushton, J. P., & Rushton, E. W. (2003). Brain size, IQ, and racial-group differences: Evidence from musculoskeletal traits. *Intelligence, 31*, 139–155.
Spencer, Q. (2015). Philosophy of race meets population genetics. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* (in press)
Wade, N. (2011). Scientists measure the accuracy of a racist claim. *New York Times*. July 11, 2013.
Weisberg, M. (2014). Remeasuring man. *Development and Evolution, 16*(3), 166–178.
Weise, K. M., & Fullerton, S. M. (2005). Racing around, getting nowhere. *Evolutionary Anthropology, 14*, 165–169.
Winther, R. G. (2014). The genetic reification of 'Race'? A story of two mathematical methods. *Critical Philosophy of Race, 2*(2), 204–223.
Winther, R. G., Giordano, R., Edge, M. D., & Nielsen, R. (2015). The mind, the lab, and the field: Three kinds of populations in scientific practice. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* (in press)
Winther, R. G., & Kaplan, J. M. (2013). Ontologies and politics of bio-genomic 'race'. *Theoria, 60*, 54–80.