The under-recognized implications of heterogeneity: Opportunities for fresh views on scientific, philosophical, and social debates about heritability

Peter J. Taylor
Programs in Science, Technology & Values and Critical & Creative Thinking
University of Massachusetts, Boston, MA 02125, USA
617 287 7636; 617 287 7656 (fax); peter.taylor@umb.edu

Short-title: Heterogeneity and debates about heritability

To appear in *History and Philosophy of the Life Sciences*

Abstract

Despite a long history of debates about the heritability of human traits by researchers and other critical commentators, the possible heterogeneity of genetic and environmental factors that underlie patterns in observed traits has not been recognized as a significant conceptual and methodological issue. This article is structured so as to stimulate a wide range of readers to pursue diverse implications of underlying heterogeneity and of its absence from previous debates. Section 1, a condensed critique of previous conceptualizations and interpretations of heritability studies, consists of three core propositions centred on heterogeneity and six supplementary propositions. Reference is made to agricultural evaluation trials in which a number of different genetically replicable varieties are raised in multiple replicates in one or more locations. In such analyses, the best case for illuminating genetic and environmental factors can be achieved; analyses in human genetics, in contrast, fall far short of the ideal. Section 2 identifies a wide range of questions that invite philosophical, historical, sociological, and scientific inquiry. These are organized under four headings: debate over the conceptual implications of heterogeneity; history of translation of methods from agriculture and laboratory breeding into human genetic analysis; racialized imaginaries in the analysis of differences among groups; and areas of scientific inquiry that may allow more attention to underlying heterogeneity.

Introduction

Claims that some human trait, say, IQ test score at age 18, show high heritability derive from analysis of data from relatives. For example, the similarity of pairs of monozygotic twins (which share all their genes) can be compared with the similarity of pairs of dizygotic twins (which do not share all their genes). The more that the former quantity exceeds the latter, the higher the trait's "heritability." Researchers and commentators often describe such calculations as showing how much a trait is "heritable" or "genetic." However, no genes or measurable genetic factors (such as, alleles, tandem repeats, chromosomal inversions, etc.) are examined in deriving heritability estimates, nor does the method of analysis suggest where to look for them. Indeed, even if the similarity among twins or a set of close relatives is associated with similarity of yet-to-be-identified genetic factors, the factors may not be the same from one set of relatives to the next, or from one environment to the next. In other words, the underlying factors may be heterogeneous. It could be that pairs of alleles, say, AAbbcbDDee, subject to a sequence of environmental factors, say, FghiJ, are associated, all other things being equal, with the same outcomes as alleles aabbCCDDEE subject to a sequence of environmental factors FgHiJ (Figure 1).

Some prominent geneticists have noted that heritability estimates are not helpful in identifying the specific genetic factors involved (e.g., Rutter 2002, 4), but the possible heterogeneity of factors that underlie patterns in observed traits has not been identified as an important issue. This intrigued me—it seemed that despite the long and politically charged history of debates about the heritability of human traits something significant may have been overlooked. If the underlying factors are heterogeneous, what can researchers do with information about a trait's heritability? What can clinicians do, or policy makers? Further conceptual, methodological, historical, and sociological questions have followed—far more than one person could address; thus this article.

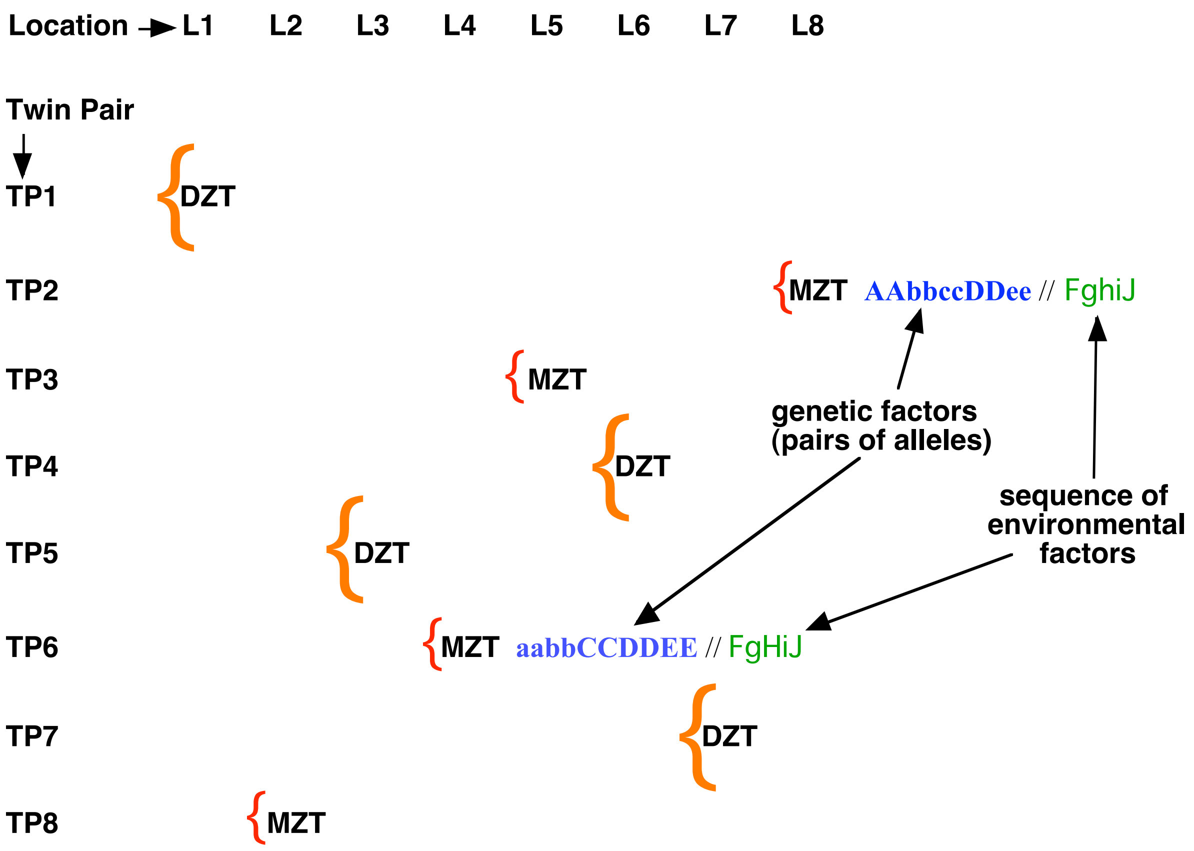

Figure 1. Factors underlying a trait may be heterogeneous even when identical (or monozygotic) twins raised together (MZT) are more similar than fraternal (dizygotic) twins raised together (DZT). The greater similarity is indicated by the smaller size of the curly brackets. The underlying factors for two MZ pairs are indicated by upper and lower case letters for pairs of alleles (A-E) and environmental factors to which they are subject (F-J).

The conventional approach for an author who thinks something significant has been overlooked is to advance an explicit argument through well-supported theses and rebuttals of previous work. A different, less agonistic approach is taken here. I write in the spirit of inviting readers from philosophy, history, and sociology of science and various fields of science to contribute to the development and sharing of ideas, arguments, narratives, and new lines of inquiry. To this end, the article combines two expository tacks. Naturally, I have to present enough of the conceptual critique to convey that underlying heterogeneity is worth more attention—this is goal of section 1. But not so much detail that only aficianados of heritability estimation read the paper and no space remains for the complementary tack of section 2. There I identify many angles of inquiry opened up by the possibility of underlying heterogeneity and the absence of this consideration from previous debates. Combining these tacks in one article means that the critique (sect. 1) has to be condensed—I do not try to pin down a line of argument in detail, and I can only lay out, not pursue, let alone wrap up, the angles of inquiry (sect. 2). To be accesible to a wide range of readers, I avoid technically detailed data analysis that would seem to be the province of specialists. At the same time, to disturb what many researchers and critical commentators have taken as settled, I dig deeper than simple themes such as the oft-cited "genetic does not imply unchangeable." Notwithstanding the expository choices and departures from conventional expectations, I hope to move even those philosophers and researchers who are skeptical that the possibility of underlying heterogeneity could make significant difference to established results and interpretations. Perhaps they will not agree with my heterogeneity-centered critique of heritability studies, but at least they may be moved to make explicit their own counter-arguments.

1. Heterogeneity-centered Conceptual Critique of Heritability Studies

1.1 Preliminaries

1.1.1. Exposition

In research and discussions on heritability one finds colloquial notions (e.g., "surely it makes sense that some traits are more influenced by genes than environment") intersecting with technical discussion of what data analyses would be needed to assess various claims about separate and interacting genetic and environmental causes. This section attempts to find a middle ground as it balances three expository considerations:

1. examine issues of data analysis so as to identify where problems arise in making science out of colloquial notions. (The emphasis on data analysis is necessary because heritability is defined and estimated through data analysis. By paying attention to what data analysis can and cannot show, we derive an antidote to discussions of heritability that move too quickly to what genes and environment do or what they can make humans do);

2. make the discussion accessible to readers who are not specialists in data analysis. (My thinking about accessibility is that under-appreciated technical limitations point to conceptual issues that are worthy of wider consideration); and

3. condense the exposition so as to preserve space for the complementary discussions of section 2. (More detailed exposition of the conceptual critique is provided by Taylor 2006a, b, c, d.)

With these considerations in mind, two issues that may be of strong interest to certain readers—competing senses of the term "heterogeneity" and path analysis as an alternative to Analysis of Variance—are addressed as notes inserted at appropriate points in the text.

1.1.2 Terminology

The various uses of the term “genetic” create potential for confusion, but this risk can be reduced if genetic is reserved as an adjective in reference to entities or “factors” that are transmitted from parents to offspring and that can, in principle, be measured. Measurable genetic factors include the presence or absence of variants (alleles) at a specific place (locus) on a chromosome, repeated DNA sequences, reversed sections of chromosomes, and so on. In a similar spirit, “environmental” can be taken to refer to measurable factors, which can range widely, say, from average daily intake of calories to maltreatment as a child. To complement these adjectival choices, I use the agricultural terms “variety” and “location” instead of the more common terms “genotype” and “environment." The agricultural terms do not suggest what needs, in fact, to be established, namely, that the quantities estimated through analysis of data about observed traits have a relationship with measurable genetic and environmental factors influencing the development of the trait. (Similar thinking leads me to refer to “trait” not “phenotype.”) Indeed, use of the agricultural terms also invites further thinking about what the common terms actually mean (see end of sect. 1.3). For now, it will suffice to think of a variety as a group of individuals whose relatedness by genealogy can be characterized, such as offspring of a given pair of parents, and a location as the situation or place in which the variety is raised, such as a family. In neither case is it assumed that researchers can specify the genetic or environmental factors that influence the trait in the various variety-location combinations.

Not only do I use agricultural terms, but at various points I refer to agricultural evaluation trials in which a number of different genetically replicable varieties are raised in multiple replicates in one or more locations. In such analyses, the best case for illuminating genetic and environmental factors can be achieved; this provides a contrast for analyses in human genetics, which fall far short of the ideal.

Finally, when I discuss the statistical Analysis of Variance (AOV or, synonymously, ANOVA) of traits, I stick with the conventional term "effect," despite its misleading causal connotations. A variety effect can be thought of as a single value of the trait for each variety that best conveys its average difference from other varieties. Typically, this is given by the average value of the trait for the variety over the range of locations in which varieties are observed minus the average value for all varieties over all locations. Similarly for location effects.

1.2 Core Propositions

The under-appreciated implications of underlying heterogeneity for analysis of human variation stem from three core propositions:

1. Statistical "effects" are distinct from measurable factors. Because the AOV for observed traits involves no reference to measurable genetic or environmental factors involved in the development of those traits, the quantities estimated by an AOV—variety ("genetic") and location ("environmental") "effects"—cannot be equated with such factors. Heritability is a quantity derivable from variety effects (or equivalently from path analysis)[1], so it too is distinct from genetic factors.

This is not a new point, but the distinction is not always preserved, even by critical commentators (e.g., Turkheimer 2000; see also the reinterpretation of Lewontin's much-cited agricultural thought experiments in Taylor 2006a, online appendix);

[1] Heritability is given by the ratio between the variance of variety effects and the variance of the trait over all varieties, locations, and replicates. Heritability can also be estimated through path analysis, a data analysis technique that quantifies the relative contributions—"path coefficients"—of variables to the variation in a focal variable once a certain network of interrelated variables has been accepted; Lynch & Walsh (1998, 823). When the same assumptions are used, AOV and path analysis estimate the same quantities; see Taylor (2006a, online appendix 1, part 5).

2. Statistical effects are an unreliable guide to hypothesizing about underlying measurable factors if those factors might be heterogeneous. It is possible that the measurable factors that underlie the development of the trait are heterogeneous in the sense conveyed by the example at the start of the Introduction.[2] If that is the case, it will be fruitless to try to translate from the variety effects (or from heritability) to a single hypothesis about the genetic factors involved. If it is not known whether the underlying factors are heterogeneous, translation from effects to hypotheses may or may not be fruitful.

Given this uncertainty, what can researchers do on the basis of knowing a trait's heritability? Propositions #4, 6, and 7 in the next section address this question from a number of angles, but the problem is best illuminated through an approach in agricultural research:

3. Statistical effects can guide hypothesis generation for agricultural evaluation trials when cluster analysis is used to group varieties by similarity in responses across all locations (Byth et al. 1976). Such clustering minimizes underlying heterogeneity and allows researchers to hypothesize about the group averages—about what factors in the locations elicited basically the same response from varieties in a particular variety group, responses that distinguish one group from another. (Of course, knowledge from sources other than the data analysis is always needed to help researchers generate any such hypotheses about genetic and environmental factors.)

To repeat an example from Taylor (2006a), suppose the varieties in a group yielded poorly in locations where rainfall occurred in concentrated periods on poorly drained soils and those varieties all have in their ancestry a particular parental stock more susceptible to plant rusts (a

[2] Underlying heterogeneity should be distinguished from three other uses of the term "heterogeneity" in the arenas of statistics and of genetics (Kaplan 2000, 18): Statistical methods often assume equality or "homogeneity" of variances from one sample to the next; mutations in a gene may be heterogeneous in the sense that they occur at a variety of points in the gene and the clinical expression of such mutations can vary significantly; and different genetic factors may be expressed as the same clinical entity. This last form of heterogeneity can be viewed a special subset of underlying heterogeneity, which also considers environmental factors acting in conjunction with genetic factors when allowing for the possibility that different underlying factors may be expressed as the same clinical entity.

form of parasitic fungi). The obvious hypothesis about genetic factors modulated by environmental factors is that the varieties share genes from the ancestral stock that are related to rust susceptibility and this susceptibility is evident in the yields where the rainfall pattern in a location enhances rust growth. It may be possible, through additional research comparing the variety and parental genomes, to identify specific sets of genes that are shared and investigate whether and how each one contributes to rust susceptibility. (See Byth et al. 1976, 224ff for actual hypotheses after analysis of an international wheat cultivar trial. See Taylor 2006a for further discussion of heterogeneity, grouping, and generation of hypotheses.)

In human quantitative genetic research, varieties can at most be replicated in two locations (i.e., identical twins separated at birth) and these locations differ from one variety (twin pair) to the next. This means that grouping of human varieties by similarity of responses across locations is impossible, leaving us back with proposition #2 and unreliable translation from effects to hypotheses.

Taken together the three core propositions mean that, unless we can rule out the possibility that measurable factors that underlie patterns in observed traits are heterogeneous, translation from statistical analyses to hypotheses about measurable factors is unreliable (not to mention difficult). The six supplementary propositions in the next two sub-sections amplify and extend this conclusion.

1.3 Supplementary propositions not centered on underlying heterogeneity

Of course, there are traits for which the underlying factors are not heterogeneous. Some traits are largely determined by the genes at a single locus, more or less independently of the individuals' upbringing (so called "high penetrance major genes")—presence of extra digits (or polydactyly) is an example. However, these traits can be detected through examination of family trees; heritability analysis is not involved. There may, in addition, be traits in which many underlying genetic factors each of small influence turn out to be similar for all individuals who show the same value for the trait—or, at least, are similar for all individuals within some defined population. Suppose that researchers decide to investigate the molecular genetic basis of a trait (not of the simple, high-penetrance kind) on the assumption that the underlying factors are homogeneous, even though they know that the underlying factors may actually be heterogeneous and, if that were the cae, the investigation would turn out to be frustrating. Could these researchers be guided by the idea that the higher the estimated heritability the better candidate the trait is for inquiry into its underlying molecular genetic basis (Nuffield Council on Bioethics 2002, chap. 11)? Alas no; choosing to inquire into the genetic basis of traits with high heritability relies on a contention that lacks support:

4. Even in the ideal case of agricultural evaluation trials, support is lacking for the contention that high heritability is an indication that measurable genetic factors have more influence on variation in the trait than measurable environmental factors. To gather such support would, in the absence of prior knowledge of how genetic and environmental factors influence the development of the trait, require: consideration of a range of models of that factors influencing development; calculation of heritability for a representative range of values of each model's parameters; and discovery of associations between the heritabilities and the corresponding genetic and environmental factors that are robust across models (Taylor 2006a, sect. 4.2).

To speak of considering a range of models is to imply that alternatives exist to standard models presented in quantitative genetic texts. The textbook models are constructed through a sequence of steps beginning with a trait governed by a pair of alleles of a single gene (i.e., at a single locus) and raised in a single location (Lynch and Walsh 1998). An example of an alternative is that a disease trait could be modeled as occurring when the combined “dosage” from many loci exceeds a threshold, where each pair of alleles contributes a full, zero, or half dose according to whether the alleles are, respectively, both the same for one variant, same for the other, or one of each (Taylor 2007; see Taylor 2006b, online appendix, for a more complex example).

Even if consideration of the alternative models could be put to the side, the contention builds on a questionable intuition that the effects for each variety estimated through an AOV are related to the level of some genetic factor or composite of genetic factors that have yet to be exposed. (Similarly for location effects.) However, a variety effect is not simply a property of the variety:

5. Analyses of observational data and interpretation of the results are conditional on the particular sets of varieties and locations. The calculation of effects in an AOV depends on averages of observations for the trait over a set of varieties and over a set of locations, so effects—and thus any hypotheses drawn from them about measurable factors—are

conditional on those particular sets. (Similarly for coefficients calculated through "path analyses" based on additive models related to those in AOV; Lynch and Walsh 1998, 827ff).

This proposition is widely acknowledged. Conditionality can, however, be viewed as even more constraining than this. Because the use of additive models in AOV and path analysis is analogous to a linearization of more general, but unknown dynamics, the resulting approximation is conditional not only on the particular set of varieties and locations, but also on the unknown dynamics remaining close to the original situation. Close here means that the only difference is "noise" equal to the residual or "error" effects in the AOV model (Taylor 2006a).

In light of the possibility of underlying heterogeneity and the other limitations in hypothesizing about measurable factors conveyed in propositions #4 and 5, the origins and durability of the heritability concept warrant explanation. Notice that in agricultural and laboratory settings there are two mitigating considerations:

6. Heritability can be a useful predictor of advances through selective breeding in agricultural and laboratory settings where researchers have the ability to replicate varieties and locations (give or take some variability of weather from season to season in field trials) and select among varieties for the next generation on the assumption that the environmental factors will remain unchanged. Comparison of the predicted advance under different breeding plans (e.g., mating of half-sibs versus mating of cousins) can inform breeders' decisions about which plan to implement.

Because heritability is a summary of observations made at one point of time for a specified set of varieties and locations, it can be expected to be an imperfect predictor of advances from one generation to the next under selection (which changes the mix of varieties) and under breeding (which produces new genetic combinations). However,

7. when selective breeding does not achieve the predicted advance, breeders can compensate for the discrepancy between predictions and outcomes—they can discard the poor offspring, breed the good offspring, and continue.

Neither of these mitigating considerations (i.e., #6 and 7) applies, however, to research on humans.

The utility of heritability estimates is even more limited than conveyed in propositions #4-7, because

8. analysis of observed traits is only the first step on a long path to interventions based on well-founded claims about the causal influence of genetic or environmental factors.

Suppose that hypotheses have been derived about measurable factors (even if, as must be the case for human traits, AOV has provided no reliable guidance [#2]). The next step for researchers would be to investigate associations with measurable factors through statistical regression analysis and experimental trials. In both cases, conditionality (#5) applies, now extended to conditionality on the set of factors measured as well. By choosing significant factors from the regression analysis to be manipulated in experimental trials, researchers are assuming that this manipulation does not modify the structure of the overall dynamics within which the factors were statistically associated with the observed traits. (Manipulations or interventions that preserve the same dynamics seem more plausible for agricultural and laboratory trials than for human social relations; Freedman 2005). Insights from these experimental studies can, in turn, contribute to research on the ways that pathways of growth and development are affected by the genetic makeup of varieties and the environmental factors in the locations. Such research might, in turn, provide a basis for interventions outside the typically well-controlled conditions in which research on causes in growth and development is undertaken. The sequence of steps in this paragraph is summarized in Table 1.

Table 1: Connections from one kind of data analysis to the next.

Kind of data to be analyzed	Agricultural evaluation trials (varieties each replicated over a number of locations)	Human studies of twins and other relatives
Observations of a trait that differs across different varieties and locations I v	AOV + Cluster analysis + knowledge from sources outside data I V hypotheses about measurable factors	AOV (& path analysis) not helpful in generating hypotheses about measurable factors[a] (hypotheses about factors drawn from other sources)
Observed associations with measurable factors	Significant factors from regression analysis I v	Significant factors from regression analysis I v
I v	factors for testing through experimental trials	(Same as on the left[b])

Experiments that vary measurable factors l v	Significant factors l v insights for investigation of dynamics of development	(Rare) ?
Factors observed over the course of development [rarely-realized ideal] l v	Significant factors in development in controlled research conditions l v candidates for interventions in less controlled situations	? ?

a. The solid line underneath denotes the disconnect between the data analysis and the generation of hypotheses about measurable factors.
b. It is more questionable for humans than for agricultural species whether factors can be manipulated without modifying structure of dynamics.

There is one way that heritability can be given causal significance without the quantity being translated into terms of measurable genetic and environmental factors. In the broad sense of a cause as a difference that makes a difference, a difference between two variety effects makes a difference among the observed traits. More strictly, the difference between two variety effects is associated with differences among the average value of the trait for the variety across locations and replicates. This could be viewed as tautologous and uninformative given that variety effects are estimated from observations of a trait by averaging across locations and replicates. Suppose, however, we put that objection to the side. Given the conditionality of the effects on the particular sets of varieties and locations (#5), this difference-between-effects form of causality corresponds to a situation in which the noise is the only thing that can vary from the original to any “rerun” (Taylor 2006a) (a situation equivalent to more complicated, but unknown dynamics staying close to the original situation, as mentioned under #5).

This circumscribed sense of causality highlights the role of replicability in the concept of variety and location. In agriculture, a variety refers not only (as defined in sect. 1.1.2) to a group of individuals whose relatedness by genealogy can be characterized, such as offspring from repeatable mating of a certain sire and dam, but also to a group of individuals whose mix of genetic factors can be replicated, as in an open pollinated plant variety. Locations are the situations or places in which the varieties are raised. Give or take variability in weather affecting field sites from season to season, locations can also be replicated. Now, for human research, replication is limited or impossible, so when methods are used to analyze variation across "genotypes" and "environments"—here: varieties and locations—, this entails a thought-experiment in which replication of varieties and locations is imagined to be possible. (Eventually the analysis of variation in a trait may help identify the measurable genetic or environmental factors that influence that trait and lead to data on those factors being brought into the analysis [see Table 1], but heritability studies can and usually do proceed from data about the trait alone. As noted earlier, in the definition of variety or location it is not assumed that researchers know or can specify the genetic or environmental factors that influence the trait for any variety-location combination.) The role of this thought-experiment in the origin of heritability studies and the subsequent history of being applied to human populations warrants interpretation (see sect. 2.1 and 2.2).

1.4 Differences between groups

Discussions about heritability in humans get most contentious around the issue of explaining differences among groups (e.g., Jensen versus Lewontin in Block and Dworkin 1976, Jencks and Phillips 1998). Given that effects from AOV and heritability estimates provide no reliable guidance in hypothesizing about measurable factors behind observations of human traits within one group of varieties (#2), they can provide no reliable guidance about measurable factors associated with differences between two groups. This alone might be enough to discount the relevance of heritability to discussions of group differences (Taylor 2006b). Nevertheless, by examining further what is involved in attempting to find genetic factors that explain differences between groups, some deeper issues can be exposed.

Consider first the case of agricultural evaluation trials in which the observations of the trait are used to cluster culitvated varieties (or "cultivars") by similar responses across locations (#3). By minimizing the possibility of heterogeneity, researchers can hypothesize about the group averages, that is, about what factors in the locations elicited basically the same response from varieties in a particular variety group (see discussion of #3). Figure 2 conveys the relationship between factors and patterns in data that underlies such hypothesizing. Within any location there is a range of responses for each variety group, but the spread is not so large as to eclipse the difference in the averages. The genetic and environmental factors that underlie the responses also have a range but from each variety and location group to the next they are distinct.

The researcher hypothesizing about what these underlying factors might be is able to relate insights about one group in one location through contrasts to insights about other groups in that location and about the same group in another location. (The plant example under proposition #3 illustrates the idea of contrasts. Varieties susceptible to rust yielded poorly in locations where rainfall occurred in concentrated periods on poorly drained soils; varieties not susceptible yielded better than them in those locations.)

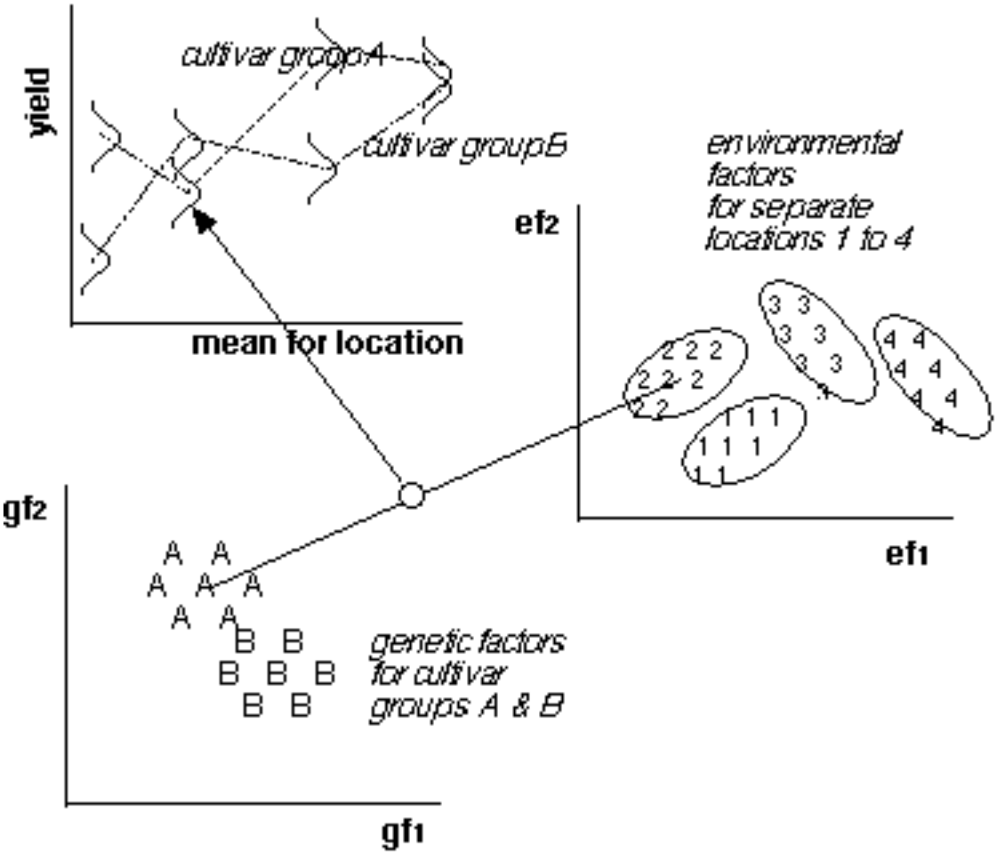
yield
cultivar groupA
cultivar groupB
mean for location
environmental factors for separate locations 1 to 4
ef2
ef1
gf2
gf1
genetic factors for cultivar groups A & B

Figure 2. Generation of hypotheses about genetic and environmental factors underlying patterns in data when those factors are homogeneous within groups. See text for discussion.

However, if varieties are not grouped by similarity of responses across locations, the possibility of heterogeneity of factors (#2) needs to be entertained. The relationship between factors and patterns in data that underlies any hypothesizing in this situation may be very difficult to disentangle (figure 3). To undertake such hypothesizing is akin to hypothesizing about the difference between group averages as if the wide and overlapping spread (variance) of values (in AOV: the within-variety-group effects) were noise (figure 4). Such a typological worldview, whether held deliberately or inadvertently, warrants interpretation (see sects. 2.1 & 2.3).

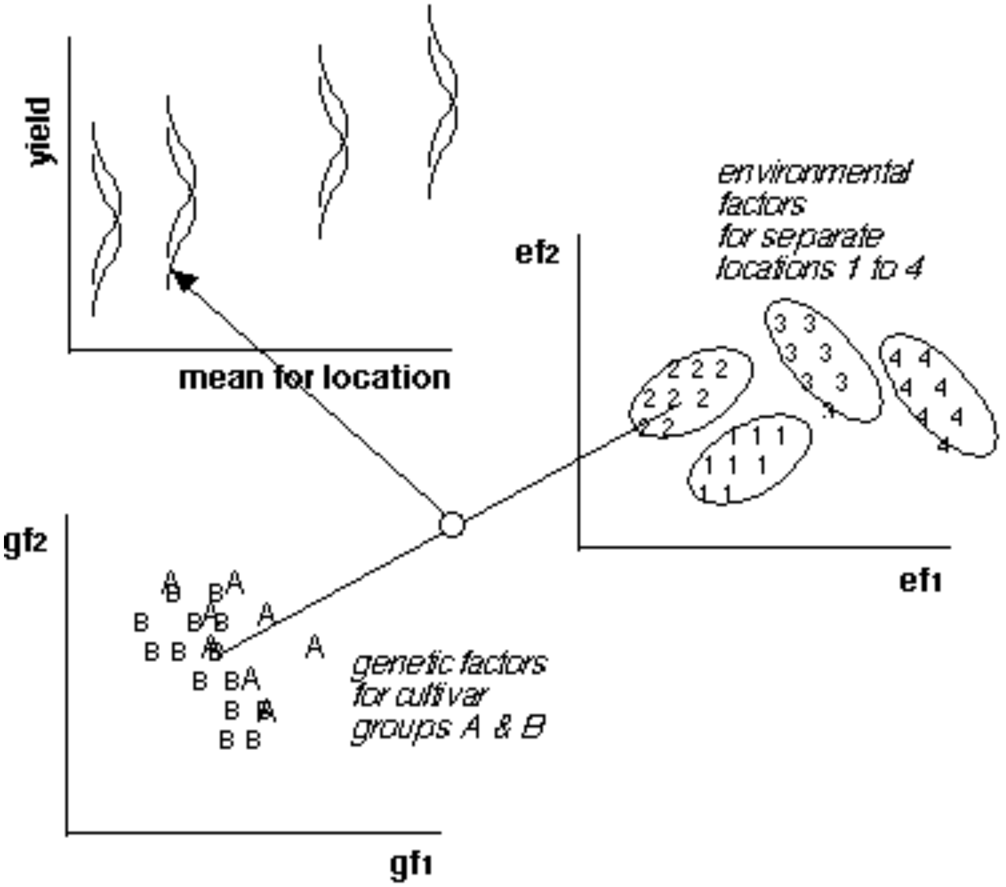
yield
mean for location
environmental factors for separate locations 1 to 4
ef2
ef1
gf2
gf1
genetic factors for cultivar groups A & B

Figure 3. Generation of hypotheses about genetic and environmental factors underlying patterns in data when the genetic factors are heterogeneous. Note: variety groups A and B have not been formed by cluster analysis and are different groups from those in Figure 2. See text for discussion.

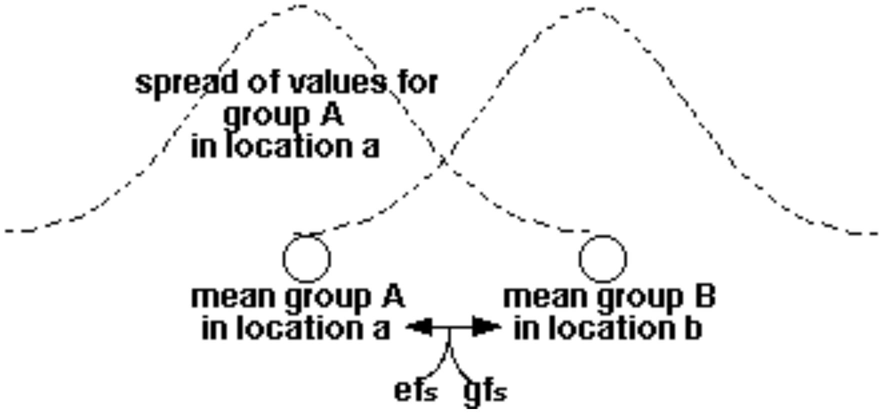

Figure 4. Generation of hypotheses about genetic and environmental factors underlying differences among groups under a typological worldview. (gfs and efs refer to measurable genetic and environmental factors.) The spread of values within groups is not taken into account.

If the possibility of heterogeneity has not been minimized, then, by extension, it must also be difficult to gain insights from one group that shed light on underlying factors in other groups or on factors for the first group in other locations. The prospects become even worse when replications of all varieties in a group are limited to the same subset of the locations (as must be the case for human racial groups given that a person's location includes their experience of membership in the racial groups). In that case, two bell curves from two different pairs in figure 3 would have to serve as the basis for any hypothesizing about the genetic and environmental factors. This means that, unlike the example of plants susceptible or resistant to rust in damp locations, contrasts among varieties within a location are not available to guide (or constrain) the researchers' hypothesis generation and subsequent inquiries; nor are contrasts available for a single group of varieties across locations.

Lindman's (1992) textbook illustrates a cautionary note about "nested" AOV (i.e., when each variety is replicated in one location only) with an example of high school students' test scores in algebra viewed in relation to their teacher and school. The students within a school are randomly assigned to a teacher in their usual school. Lindman notes that a significant location (school) effect "is likely to be interpreted as due to differences in physical facilities, administration, and other factors that are independent of the teaching abilities of the teachers themselves… [However, d]ifferences between teachers in different schools are part of the [location or school] effect, and the observed differences between schools could be due entirely to the fact that some schools have better teachers [or] some schools have smarter children attending them" (Lindman 1992, 194).

Lindman could have added that the observed differences between schools could be due entirely to combinations of factors, such as students responding worse to teachers whose attention is distracted because their school's administrators insist more on detailed documentation of student performance and so on. The point in common is that nested AOV cannot help researchers hypothesize about the difference in the average scores from one school to the next when the teachers are replicated (in the form of their students' test scores) only within schools. Translated into the concerns of this article:

9. Nested AOV cannot help researchers hypothesize about the difference in the average scores from one subset of locations to the next when the varieties are replicated only within one subset. Researchers might just as well conduct a separate statistical analysis for each subset of varieties and location—or, in the context of racial differences, for each combination of group of individuals and experience of membership in different racial groups. (To respect this methodological limitation of nested analysis is not to make the claim that disjunct kinds of causes must be operating in the different racial groups.)

Although Lindman's note and the preceding discussion and diagrams center on AOV, the points about the possible heterogeneity of factors (#2), about membership in different groups being analyzed as different locations; and about the limitations of nested analysis (#9) might also apply to drawing hypotheses and insights from regression analysis and experimental trials (see Table 1). Exploring this idea is one of the lines of inquiry raised in section 2 to follow.

2. Some Possible Angles for Investigation by Interpreters of Science and Scientists

The critique of heritability studies presented in section 1, although condensed, should convey that the possible heterogeneity of factors that underlie patterns in observed traits is worth more attention. Some scientists or philosophers might be moved to show that it makes no significant difference to established results and interpretations; others might want to pin down the arguments, address counter-arguments, and provide a careful, digestible exposition. Yet some historians, sociologists, and philosophers of science and some scientists might be ready, without waiting for such elaboration, to tease out questions raised by the possibility of underlying heterogeneity and the absence of this consideration from previous debates. For these readers, section 2 lays out a wide range of angles of inquiry that follow from the critique. This reveals my sense of what is interesting and potentially important; indeed, I have begun to pursue some of the angles. However, rather than elaborate on any preliminary findings or speculate on the impact of such interpretative and scientific inquiries, I leave the questions open in the spirit identified in the Introduction, namely, of inviting others to develop and share ideas, arguments, narratives, and further lines of inquiry.

2.1. Conceptual reconstruction and extensions

The possible heterogeneity of factors is not mentioned as an issue in the extensive entry on heredity and heritability in the Stanford Online Encyclopedia of Philosophy (Downes 2004), in the key sources cited therein (e.g., Sarkar 1998; Kaplan 2000), or in Sesardic (2005)'s rebuttal of many critiques of heritability studies. There is room, therefore, for philosophers of biology to be drawn into debate about the conceptual steps identified in section 1 (see also Taylor 2006a,b,c,d, 2007, 2008) and to rebut, refine, rethink, or extend the arguments and their conceptual basis. Questions for debate might include: Can parsimony justify the assumption that the same genetic factors underlie traits that appear similar? Even without knowledge of the underlying factors, can useful interpretations or actions sometimes be drawn from the size of heritability (Taylor 2007, 2008)? Or, from variation in heritability across economic classes (Turkheimer et al. 2003)? Or, from the smallness of the effects due to the members of a family growing up in the same location relative to the residual or “non-shared” effects (i.e., those not associated with the shared location or with differences among varieties) (Plomin and Asbury 2006; but see Taylor 2007). How do the propositions in section 1 affect claims that individual’s genetic makeup contributes to the environments they experience (Plomin and Asbury 2006)? Indeed, how are each of the nine propositions addressed or overlooked in previous studies and critiques?

There may well be skepticism about the relevance of agricultural methods to the analysis of human variation. Yet, human heritability estimation is based on data that are less ideal than agricultural evaluation trials, so how can human estimates support claims about more general notions of genetic or environmental causality? The agricultural case, with the control entailed in replicating varieties and locations, also seems to be well suited for clarifying discussion of the kinds of realizable intervention entailed by the project of making inferences about causality from observations of traits. Given the emphasis in recent philosophy and in social science to the causation-intervention relationship (Pearl 2000, Woodward 2003), there should be interest in the argument that heritability estimation and the AOV on which it is based presuppose a circumscribed sense of causality in which everything is kept close to the original situation (see the observation under proposition #5 and elaboration in Taylor 2006a).

Do the possibility of underlying heterogeneity and the limitations I identify in causal inference from quantitative data in heritability studies have wider relevance in social science and epidemiology? For example, as an extension of proposition #2, should we question the methodological assumption in epidemiology that, when similar responses of different individual types are observed, similar conjunctions of risk and protective factors have been involved in producing those responses? Does the "close to the original situation" condition (see under #5) apply to <u>any</u> attempt to start from statistical patterns based on observations and move through inferences about causal factors to policy interventions? If that is so, how can that condition be reconciled with the likelihood that most policy interventions, if implemented, would alter the structure of the relations that produced the phenomena observed, and thus the patterns and causal inferences derived from the observations? As mentioned at the end of section 2, in what ways do the points about the possible heterogeneity of factors (#2), about membership in different groups being analyzed as different locations (#9); and about the limitations of nested analysis (#9) apply—or not apply—to drawing hypotheses and insights from regression analysis and experimental trials?

Another conceptual issue to explore is the relationship between lack of attention to the possibility of heterogeneity and a typological or essentialist worldview that "conceptualizes diversity as 'deviation' from a natural state or path of change" (McLaughlin 1998, 25). Notice that Lindman, even as he performs the valuable role of cautioning readers about nested analyses (end of sect. 1.4 above), perpetuates the typological worldview in refering to "the observed differences between schools" when he means the observed differences between <u>averages for schools</u>. It is still commonplace to hear typological expressions of the kind "men are taller than women," "men tend to be taller than women," or "men are, on average, taller than women." Some might reject the label typological, saying that the implicit variation is well appreciated and that nothing would be gained by wordier statements making the variation explicit, such as, "the variation among men's heights centers at a point that is greater than center of the variation among women's heights," or "the variation among men's heights and the variation among women's heights overlap, but some of the men's varation lies to the right of the women's and some of the women's lies to the left of the men's variation." Yet is it simply linguistic convenience to use simple expressions that put group or class membership first and leave deviation as implicit or secondary? The wordier alternatives could help us keep in mind the possibility that the factors underlying the pattern in the data could vary among men and women and need not include factors solely possessed by one sex or the other (see also 2.3 below). Is that possibility something that statistical analysis has to ignore in order to derive results?

Finally, a more modest question—one of the sociology of knowledge—is to consider the ways that discussion among philosophers of science might have obscured the relevance of heterogeneity, say, through visual and verbal conventions that emphasize types over variation, or by over-reliance on scientists to set the terms of issues on which philosophers focus their efforts in conceptual reconstruction.

2.2. History of translation from agriculture and laboratory breeding to human genetic analysis.

Heritability estimation was first used in selective breeding in agricultural and laboratory settings, a context in which researchers have the ability to replicate varieties and locations (Fisher 1918, Wright 1920, Lush 1945). Indeed, when agricultural researchers compare varieties and make recommendations to farmers and when they select among varieties for the next round of evaluation trials, they do so on the assumption that the environmental factors will remain more or less unchanged. For observations of human traits, however, such replicability of varieties and environmental factors is not possible (requiring the thought experiment discussed under #8). This last observation opens up the historical question of how such restrictive conditions were discounted or forgotten in the translation of heritability estimation from selective breeding to human genetics.

For example, when Wright (1920) presented his original formulation of heritability estimation he used the notation E to refer to "environmental factors that are common to litter mates" of guinea pigs that he bred. To translate heritability estimates into predictions of future changes under selective breeding, these "factors" had to remain constant from one generation to the next. (In the terms of this article, Wright meant that the effects from the AOV or path coefficients had to remain constant—no measurable factors were involved in the analysis.) "E" is now used, however, to denote environmental factors without reference to Wright's restricted conditions. One part of an historical investigation would be to trace Wright's notation from its origin through its adoption in human genetics (see Burks 1928; Lush 1947), where it remains commonplace to discuss the relative influence of G (genes) and E (environment) in accounting for the variation among individuals and groups. Other historical investigations might consider the influence in the other direction (D. Paul, pers. comm.) or the separation of heritability from the context of selective breeding (S. Downes, pers. comm.). (A historical sidebranch would be to explore the grounding of Wright's shifting balance theory of evolution in the search for ideal approaches to animal breeding; see, e.g, Lush 1945, p. 433-435.)

The historical question about the forgetting or discounting of the restrictive conditions of selective breeding could extend into a critical revisiting of the long and politically charged history of scientific and policy debates about the heritability of IQ test scores (and other human traits) and about genetic explanations of the differences between the mean scores for racial groups. (For key points in the debate, see the exchange between Jensen and Lewontin reprinted in Block and Dworkin 1976; Jencks and Phillips' 1998 review of research on the black-white test score gap; Parens' 2004 even-handed overview of past and potential contributions of human behavioral genetics to discussions of social importance well beyond IQ tests; and philosopher Sesardic's 2005 critique of critiques of human behavioral genetics.)

2.3. Racialized imaginaries in the analysis of differences among groups.

Another historical issue arising from the critique of heritability studies concerns the persistence of interest in explaining differences among the averages for human groups defined on racial grounds. Questions related to this issue are more speculative than the concept-centered philosophy and history in sections 2.1 and 2.2 and invite a more interpretive, cultural approach.

Consider first that, because the ranges within human racial groups are large and overlapping (as in Figure 4), it is hard to make policy out of any finding about the differences between averages for groups unless individuals are treated (by teachers, social workers, medical practitioners, etc.) on the basis of their group membership. What else can people do with the patterns that researchers find in observations of human relatives assigned to groups? (Recall that the possible heterogeneity of factors makes heritability estimates within groups unreliable, if not irrelevant, for developing or supporting hypotheses about differences between averages for groups [#9]; see also Taylor 2006b.) When researchers do not address heterogeneity, are they making typological or essentialist assumptions? Does a racially essentialist imagination facilitate the transfer of conventional statistical tools from agricultural to human research? Might it be possible to pinpoint paths not taken or objections not taken up in scientific and public debates about group average differences? Such blind spots might, in turn, be interpreted in terms of the persistence of racial types as an organizing category in American social and scientific thought. Similarly, might the transfer of tools from selective breeding in agriculture and the laboratory to analysis of human variation (see sect. 2.2) speak to persistence of eugenic hopes and fears? How are responsibility and causation conceived by people when they talk of individuals in terms of their group membership? In light of the connection between causation and intervention (sect. 2.1), just who is empowered to do something as a result of analysis of average group differences—and who is given license not to have to do anything?

2.4. Areas of research that may have potential to allow for the possible heterogeneity of factors.

As any interpreter of science in its social context knows, new ideas or arguments do not realize their transformative potential without the social structure of the field changing in ways that overcome the inevitable resistance from the mainstream. Perhaps the philosophical, historical, and sociological interpretations emerging from inquiries like those in sections 2.1-2.3 can contribute to such changes. Moreover, alternative research programs usually have to be opened up before many researchers begin to shift—critique is rarely sufficient for a dominant paradigm to be abandoned. In that light, let me identify three areas of inquiry that may have potential to allow for the possible heterogeneity of factors. I am not claiming that these areas overcome the limitations of heritability studies, let alone that the brief sketches are sufficient to show that this is the case. By including them in an article for a history and philosophy journal, my intention is primarily that some readers join in nudging natural and social scientists to be more explicit about the ways their methods and models address—or suppress—the possible heterogeneity of underlying factors. As mentioned earlier, perhaps this possibility is one that statistical analysis has to ignore in order to derive results, but that choice and its implications could be made transparent. Conceptual analysis by philosophers might, in this spirit, help articulate the possibilities and limits of generating empirically validated models of developmental pathways whose components are heterogeneous and differ among individuals at any one time and over generations.

2.4.1. Multivariate developmental models in education and mental illness.

Woodhead (1988) summarizes studies explaining how the IQ test score increases produced by Head Start preschool programs tend to be transient, but in the long term, through social support systems initiated or enhanced during the Head Start years, the children end up with significantly higher high school graduation rates, employment, and many other socially valued measures. Ou (2005) has put that conclusion on a quantitative basis in finding associations among preschool participation and other measures taken through the course of schooling and development to adulthood. (These measures include: basic skills scores at kindergarten, classroom adjustment age 9-10; parent involvement for children age 8-12; abuse/neglect reports age 4-12; school quality for children age 10-14; number of school moves age 10-14; commitment to school at age 10 or 15; grade retention through age 15; achievement age 15; highest grade completed by age 22.)

Ou (2005, 604) remarked on her model’s limitations in the areas of "the correlational nature of the data, possible alternative models, and generalizability." It might also be noted that the factors in Ou’s analysis would traditionally be labeled environmental. In a different context, Kendler et al.'s (2002) comprehensive developmental model incorporated factors that could be labeled genetic and was able to account for 52% of the variance in liability to episodes of major depression. The models of Kendler et al., like those of Ou, provide a picture of development that is rich and plausible, but clarification is warranted of the class of changes or interventions in which it makes sense to construe the factors in the models as causes (see sect. 2.1). Indeed, Kendler et al. (2002, 1133) show admirable reserve in concluding that their "results, while plausible, should be treated with caution because of problems with causal inference, retrospective recall bias, and the limitations of a purely additive statistical model." Interestingly, they did not remark on the absence of variables that correspond to therapeutic interventions (as if to suggest that these had no effect on the etiology of depression and risk factors implicated in that etiology) or to social changes that have led to the rising incidence of depression. Inclusion of interventions and social changes would seem important in any analogous comprehensive developmental model of IQ test scores and other outcomes subject to educational influence.

Kendler et al. (2002) take an additional step in characterizing different paths to the outcome to be explained, namely, depression, e.g., "Paths Reflecting a Broad Adversity/Interpersonal Difficulty Pathway to Major Depression." Although they identified the paths by eye, the exercise opens up the possibilities of identifying paths that operate heterogeneously across social groups, across individuals within any social grouping, and, in relation to the Flynn effect (large increases in average IQ scores from one generation to the next; Flynn 1994), heterogeneously across generations.

2.4.2. Life course analyses in epidemiology

In a field initiated by the epidemiologist Barker at the University of Southampton, a large number of researchers are now studying associations between nutritional deficits during critical periods in utero and diseases of late life, including heart disease and diabetes. The integration of fetal origins and subsequent influences is now taking place under the label of “life course epidemiology” (Kuh and Ben-Shlomo 2004). Gilthorpe and colleagues (among others) have highlighted the statistical challenges in interpreting associations between early life influences and diseases of later life (Head et al. 2005). West and Gilthorpe are developing alternative statistical analyses that enable them to characterize different pathways of growth over the lifecourse (which, in my terms, makes it easier to visualize the possible heterogeneity of factors underlying responses) (see also Croudace et al. 2003, DeStavola et al. 2006).

2.4.3 Life events and difficulties

Another line of research from England, initiated by the sociologists Brown and Harris in the late 1960s, investigates how severe events and difficulties during people's life course influence the onset of mental and physical illnesses (Harris 2000). Brown and Harris use wide-ranging interviews, ratings of transcripts for the significance of past events in their context (with the rating done blind, that is, without knowledge of whether the person became ill), and statistical analyses. Recognizing that the “same” event, e.g, death of a spouse, might have very different meanings and significance for different subjects according to the context, Brown and

Harris's methods accommodate events with diverse meanings. The approach allows apparently heterogeneous events to be subsumed under one factor, such as, in explanation of depression, a severe, adverse event in the year prior to onset.

Conclusion

If the condensed critique of heritability studies in section 1 has achieved its aim, readers should see that the possible heterogeneity of factors that underlie patterns in observed traits is worth more attention. For readers intrigued by this consideration or by its absence from previous debates, the many open questions laid out in section 2 point to diverse inquiries that can be taken up in various areas of interpretive or scientific research. I have begun to pursue some of these inquiries and report on them (e.g., Taylor 2008). I look forward to participating in a widened community of researchers developing and sharing of ideas, arguments, narratives, and new lines of inquiry. It remains to be seen, of course, whether attention to the possibility of underlying heterogeneity makes an impact on the quantitative analysis in the study of heredity, policy-making based on such research, and popular discussion.

Acknowledgements

This paper is based on research supported by the National Science Foundation under grant SES–0634744.

References

Block N. J., Dworkin A. (eds.), 1976, *The IQ Controversy: Critical Readings*, New York: Pantheon.

Burks B. S., 1928, "The Relative Influence of Nature and Nurture Upon Mental Development: A Comparative Study of Foster Parent-Foster Child Resemblance and True Parent-True Child Resemblance", *The Twenty-Seventh Yearbook of the National Society for the Study of Education*, 27**:** 219-316.

Byth D. E., Eisemann R. L., DeLacy I. H., 1976, "Two-way pattern analysis of a large data set to evaluate genotypic adaptation", *Heredity*, 37**:** 215-230.

Croudace T. J., Jarvelin M.-R., Wadsworth M. E. J., Jones P. B., 2003, "Developmental Typology of Trajectories to Nighttime Bladder Control: Epidemiologic Application of Longitudinal Latent

Class Analysis", *American Journal of Epidemiology*, 157**:** 834–842.

De Stavola B. L., Nitsch D., dos Santos Silva I., McCormack V., Hardy R. J., Mann V., Cole T. J., Morton S., Leon D. A., 2006, "Statistical Issues in Life Course Epidemiology", *International Journal of Epidemiology*, 163**:** 84-96.

Downes S. M., 2004, "Heredity and Heritability". In: Zalta E. N., (ed.) *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/heredity/, accessed 11 May 2006.

Fisher R. A., 1918, "The Correlation Between Relatives on the Supposition of Mendelian Inheritance", *Philosophical Transactions of the Royal Society of Edinburgh*, 52**:** 399–433.

Flynn J. R., 1994, "IQ gains over time". In: Sternberg R. J., (ed.) *Encyclopedia of Human Intelligence*, New York: Macmillan, 617-623.

Freedman D. A., 2005, "Linear statistical models for causation: A critical review". In: Everitt B., Howell D. (eds.), *Encyclopedia of Statistics in the Behavioral Sciences*, Chichester: Wiley

Harris T. (ed.) 2000, *Where Inner and Outer Worlds Meet*, London: Routledge.

Head R. F., Ellison G. T. F., Gilthorpe M. S. (eds.), 2005, Statistical challenges facing the foetal origins of adult disease hypothesis, 3rd International Congress on the Developmental Origins of Health and Disease, Toronto.

Jencks C., Phillips M. (eds.), 1998, *The Black-White Test Score Gap*, Washington, DC: Brookings Institution Press.

Kaplan J. M., 2000, *The Limits and Lies of Human Genetic Research*, New York: Routledge.

Kendler K. S., Gardner C. O., Prescott C. A., 2002, "Towards a comprehensive developmental model for major depression in women", *American Journal of Psychiatry*, 159**:** 1133-1145.

Kuh D., Ben-Shlomo Y. (eds.), 2004, *A Life Course Approach to Chronic Disease Epidemiology*, Oxford: Oxford University Press.

Lindman H. R., 1992, *Analysis of Variance in Experimental Design*, New York: Springer-Verlag.

Lush J. L., 1945, *Animal Breeding Plans*, Ames, Iowa: Iowa State College Press.

Lush J. L., 1947, "Family merit and individual merit as bases for selection", *American Naturalist*, 81**:** 241-261; 362-379.

Lynch M., Walsh B., 1998, *Genetics and Analysis of Quantitative Traits*, Sunderland, MA: Sinauer.

McLaughlin P., 1998, "Rethinking the Agrarian Question: The Limits of Essentialism and the Promise of Evolutionism", *Human Ecology Review*, 5**:** 25-39.

Nuffield Council on Bioethics, 2002, "Genetics and Human Behavior: The Ethical Context", http://www.nuffieldbioethics.org, accessed 22 June 2007.

Ou S.-R., 2005, "Pathways of long-term effects of an early intervention program on educational attainment: Findings from the Chicago longituidinal study", *Applied Developmental Psychology*, 26**:** 478-611.

Parens E., 2004, "Genetic differences and human identities: On why talking about behavioral genetics is important and difficult", Hastings Center Report**:** S1-S36.

Pearl J., 2000, Causality: Models, Reasoning, and Inference, Cambridge: Cambridge University Press.

Plomin R., Asbury K., 2006, "Nature and Nurture: Genetic and Environmental Influences on Behavior", *The Annals of the American Academy of Political and Social Science*, 600**:** 86-98.

Rutter M., 2002, "Nature, nurture, and development: From evangelism through science toward policy and practice", *Child Development*, 73**:** 1-21.

Sarkar S., 1998, *Genetics and Reductionism*, Cambridge: Cambridge University Press.

Sesardic N., 2005, *Making Sense of Heritability*, Cambridge: Cambridge University Press.

Taylor P. J., 2005, *Unruly Complexity: Ecology, Interpretation, Engagement*, Chicago: University of Chicago Press.

Taylor P. J., 2006a, "Heritability and heterogeneity: On the limited relevance of heritability in investigating genetic and environmental factors", *Biological Theory: Integrating Development, Evolution and Cognition*, 1**:** 150-164.

Taylor P. J., 2006b, "Heritability and heterogeneity: On the irrelevance of heritability in explaining differences between means for different human groups or generations", *Biological Theory: Integrating Development, Evolution and Cognition*, 1**:** 392-401.

Taylor P. J., 2006c, "The analysis of variance is an analysis of causes (of a very circumscribed kind)", *International Journal of Epidemiology*, 35**:** 527-531.

Taylor P. J., 2006d, "Exchange around the knowledge claim, ‘Intelligence is 75% genetic’", http://sicw.wikispaces.com/Intelligence75percentGenetic, accessed 21 August 2006.

Taylor, P.J., 2007, "The Unreliability of High Human Heritability Estimates and Small Shared Effects of Growing Up in the Same Family". *Biological Theory: Integrating Development, Evolution and Cognition*, 2(4): 387-397.

Taylor P. J., 2008, "Underlying heterogeneity: A problem for biological, philosophical, and other analyses of heritability?", *Biology and Philosophy*, 23**:** 587-589.

Turkheimer E., 2000, "Three laws of behavior genetics and what they mean", *Current Directions in*

Psychological Science, 9**:** 160-164.

Turkheimer E., Haley A., Waldron M., D'Onofrio B., Gottesman I. I., 2003, "Socioeconomic Status Modifies Heritability of IQ in Young Children", *Psychological Science*, 16**:** 623-628.

Woodhead M., 1988, "When psychology informs public policy", *American Psychologist*, 43**:** 443-454.

Woodward J., 2003, *Making Things Happen: A Theory of Causal Explanation*, New York: Oxford University Press.

Wright S., 1920, "The relative importance of heredity and environment in determining the piebald pattern of guinea pigs", *Proceedings of the National Academy of Science*, 6**:** 320-332.