Spring 3-26-2014

Heterogeneity, not randomness, sets challenges for quantitative genetics and epidemiology: A response to Davey Smith's "gloomy prospect"

Peter J. Taylor
University of Massachusetts Boston, peter.taylor@umb.edu

WORKING PAPER on Science in a Changing World

Paper # 2 -2014

Heterogeneity, not randomness, sets challenges for quantitative genetics and epidemiology

A response to Davey Smith's "gloomy prospect"

PETER J. TAYLOR

GRADUATE PROGRAM IN CRITICAL AND CREATIVE THINKING
University of Massachusetts, Boston, MA 02125, USA
http://scholarworks.umb.edu/cct_sicw/2

Heterogeneity, not randomness, sets challenges for quantitative genetics and epidemiology: A response to Davey Smith's "gloomy prospect"

Peter J. Taylor
Science in a Changing World graduate track
University of Massachusetts, Boston, MA 02125, USA
peter.taylor@umb.edu

Abstract

Social epidemiologist Davey Smith (2011) argues that epidemiologists should accept a gloomy prospect: considerable randomness at the individual level means that they should keep their focus on modifiable causes of disease at the population level. The difficulty epidemiology has had in moving from significant population-level risk factors to improved prediction of cases at an individual level is analogous to the lack of success in the search for systematic aspects of the non-shared environmental influences that human quantitative genetics claims overshadow common environmental influences (e.g., the family's socioeconomic status which siblings have in common). This article responds to the argument and analogy, aiming to draw three audiences—social epidemiologists, human quantitative geneticists, and philosophers of science—into a shared discussion that centers not on randomness, but on heterogeneity in various forms. The first half undercuts the analogy, providing a critical account of human quantitative genetics that explains why its estimates are unreliable and typical interpretations, such as that of non-shared environmental influences, unjustified. In the process, attention is drawn to the possibility of *underlying* heterogeneity—when similar responses of different types in a species or of different groups of people are observed, it need not be assumed that similar conjunctions of genetic and environmental factors (or, in epidemiology, risk and protective factors) have been involved in producing those responses. The second half introduces several additional forms of heterogeneity and captures their potential significance for epidemiology in four conceptual themes. The

themes complicate Davey Smith's advice for epidemiologists not to seek improved prediction of cases at an individual level and his view of the limited prospects for personalized medicine.

1. Introduction

In the John Snow Lecture to the World Congress of Epidemiology, George Davey Smith (2011) addressed the difficulty epidemiology has had moving from significant population-level risk factors to improved prediction of cases at an individual level. The *gloomy prospect* of his account is that epidemiology is concerned with the mean of the group and has to discount random variation around that mean. Epidemiologists should, he advises, accept considerable randomness at the individual level and keep their focus on modifiable causes of disease at the population level.

In making his case, Davey Smith draws an analogy from human quantitative genetics, a field that revolves around partitioning variation in a trait into a heritability fraction and other components. The field claims that non-shared environmental influences overshadow common environmental influences—the former corresponding to differences among siblings raised together; the latter to influences such as socioeconomic status that the siblings in a family usually have in common. Davey Smith notes that the search for systematic aspects of the non-shared environmental influences has been unsuccessful, which he sees in the same terms as the lack of success by epidemiologists who seek individual-level risk prediction.

This article takes the use of this analogy by an epidemiologist as an opportunity to address three audiences: social epidemiologists, human quantitative geneticists, and philosophers of science—especially those who have examined debates about heritability of human traits. My goal is to draw the different audiences into a shared discussion that centers not on randomness, but on heterogeneity in various forms. The first half (sections 2-5) undercuts the analogy by providing a critical account of human quantitative genetics that explains why its estimates are unreliable and typical interpretations, such as that of non-shared environmental influences, are unjustified. In the process, attention is drawn to the possibility of *underlying* heterogeneity—when similar responses of different types in a species or of different groups of people are observed, it need not be assumed that similar conjunctions of genetic and environmental factors have been involved in producing those responses. The second half (sections 6-7) extends underlying heterogeneity to include risk and protective factors in epidemiology, introduces several additional forms of heterogeneity, and

captures their potential significance for epidemiology in four conceptual themes. The themes complicate Davey Smith's advice for epidemiologists not to seek improved prediction of cases at an individual level and his view of the limited prospects for personalized medicine.

2. Genetic is not genetic is not genetic

The first step in developing a critical account of quantitative genetics is to identify three conceptually and empirically distinct senses of "genetic," pertaining to: statistical partitioning of variation in measurements on a trait; relatedness in terms of the fraction of variable part of genome shared; and the statistical association of a trait with measurable genetic factors that underlie the development of that trait. (There are other senses of genetic, such as "runs in the family," but these will not be the focus here.) Before elaborating on the significance of the different senses of genetic, two terminological notes are in order:

1. "Factor" is used throughout the article in a non-technical sense to refer simply to something whose presence or absence can be observed or whose level can be measured. Measurable genetic factors include the presence or absence of alleles at a specific locus on a chromosome, repeated DNA sequences, reversed sections of chromosomes, and so on. Measurable environmental factors can range widely, say, from mean daily intake of calories to degree of maltreatment that a person experienced as a child.

2. An agricultural variety or breed is a group of individuals whose mix of genetic factors can be replicated, as in an open pollinated plant variety, or any group of individuals whose relatedness by genealogy can be characterized, such as offspring of a given pair of parents. A location is the situation or place in which the variety is raised, such as a specific experimental research station.

In regards to *statistical partitioning of variation in measurements on a trait*, Figure 1 depicts the ideal case, namely, an agricultural evaluation trial where each of a set of varieties is raised in each of set of locations, and there are two or more replicates in each variety-location combination. The means for each variety across all locations and replicates can be estimated; the variance of these variety means is the so-called *genetic*

variance of classical quantitative genetics. The term genetic variance here is a contraction of *genotypic* variance, where genotype is a synonym for an agricultural variety and does not refer to pairs of alleles. Genotypic variance is shorthand, in turn, for variance of the genotypic values, i.e., the variety means just described. The figure makes clear that partitioning or analysis of variation for a given trait neither *requires* nor *produces* knowledge about the genetic or environmental factors that underlie the development of that trait in the different variety-location combinations.

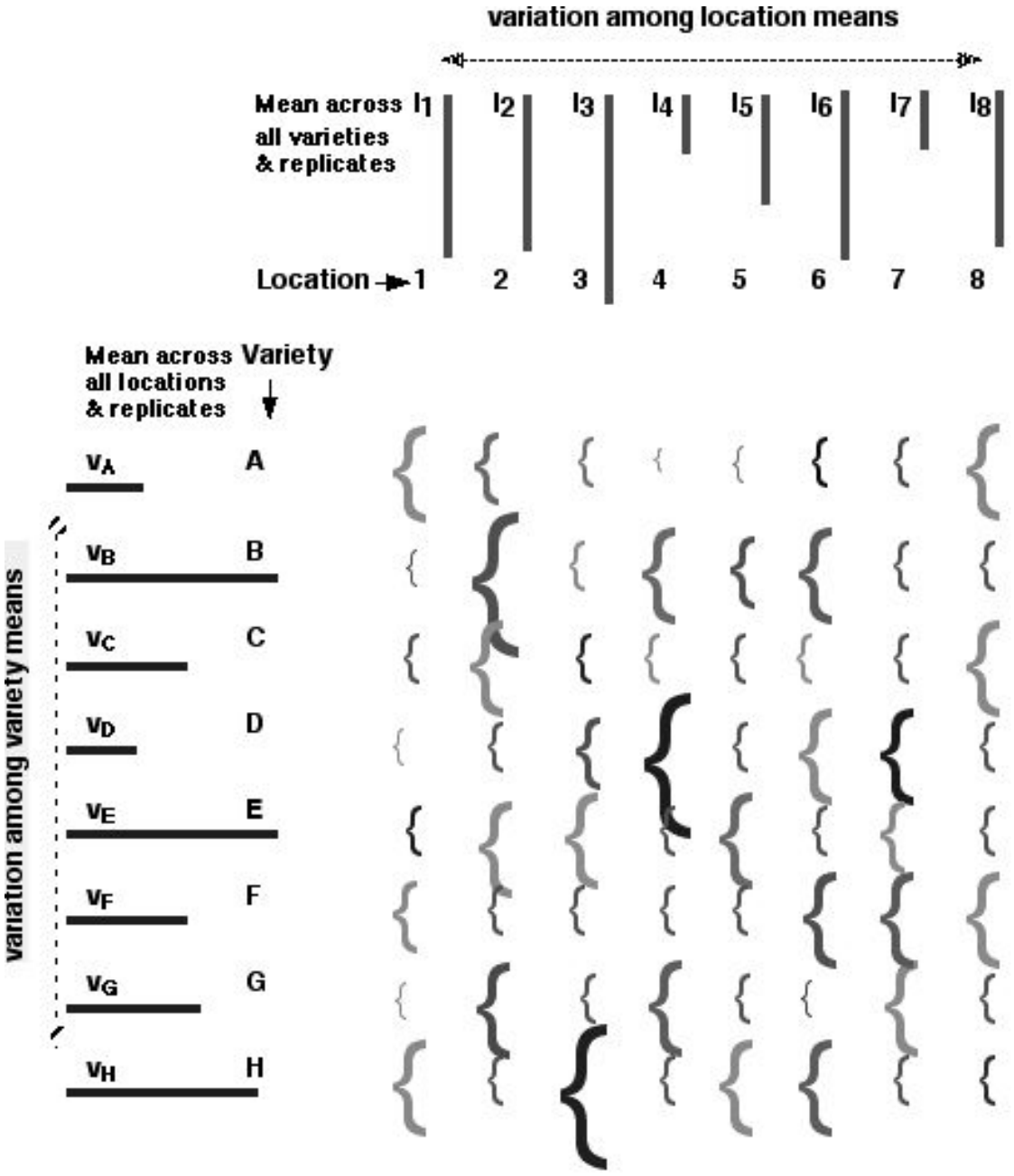

Figure 1. Partitioning of variation for a given trait in the ideal agricultural evaluation trial where each of a set of varieties is raised in each of set of locations, and there are two or more replicates in each variety-location combination. The bars next to the varieties and locations indicate the average value of the trait for, respectively, each variety and location. The variation between replicates within variety-location combinations is indicated by the size of the curly brackets. The non-systematic shading of the brackets indicates that the variation between replicates is not correlated from one variety-location combination to another. (For graphic clarity, the average values for each variety-location combination after allowing for the variety and location averages are not shown.)

Broad-sense heritability (hereon: heritability) is *by definition* is the ratio of the variance among variety means (i.e., genotypic values) to the variance of the trait across the whole data set. In the agricultural evaluation trial of Figure 1, this quantity can be readily estimated. Similarly, for estimation of the fraction of variance due to differences among location means, otherwise called the environmental variance (where environment here is a synonym for location), and for the fraction due to variety-location interaction, otherwise called genotype-environment interaction. (This last term, confusingly, has no conceptual or empirical relationship to gene-environment interaction, a concept introduced in section 6).

Relatedness in terms of the fraction of variable part of genome shared comes into play when the data set is not as complete as in the ideal agricultural trial. In the field of quantitative genetics the estimation of heritability and other fractions of the trait's overall variance typically makes use of the genealogical relatedness of the varieties or genotypes. For example, as is common in studies of humans, the similarity of pairs of monozygotic twins, which share all their genes, can be compared with the similarity of pairs of dizygotic twins, which share most of their genes, but only around half of the genes that vary in the population. An estimate of heritability can be derived using a formula or a Structural Equation Model that takes into account the extent to which the mean similarity of monozygotic twins raised in the same family exceeds the mean similarity of dizygotic twins raised in the same family (Rijsdijk and Sham 2002) (Figure 2). (The estimation of the location or environmental fraction of the variation, its division

into shared and non-shared components, and other details are taken up in section 5.)

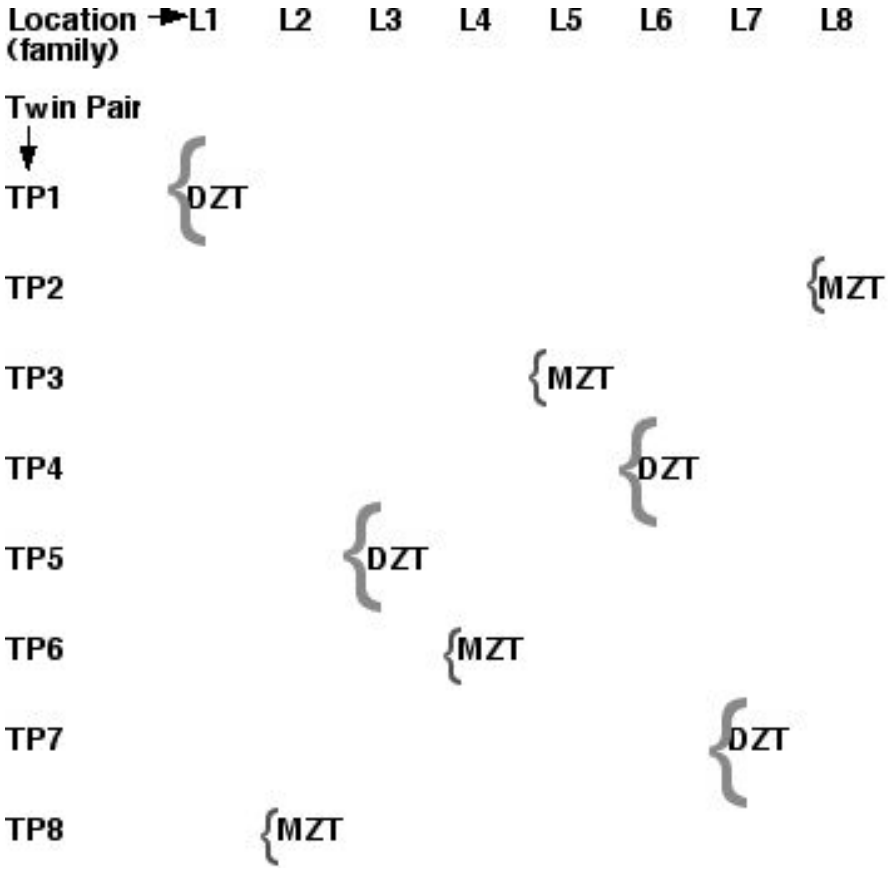

Figure 2. The basis for a comparison of the similarity of monozygotic twins raised together in the same locations or families (MZT) with the similarity of dizygotic twin pairs (DZT). The variation between twins in a pair is indicated by the size of the curly brackets. In contrast to the agricultural evaluation trial in Figure 1, the replicates of any twin pair are raised in only one household per twin pair.

Another terminological note is in order: From hereon, the agricultural terms "variety" and "location" are used even when referring to human studies whenever the entity or situation can be identified *without* reference, respectively, to measurable genetic and environmental factors. For humans, a location is typically the family of upbringing. The more common terms for variety and location, genotype and environment, invite conflation with genetic and environmental factors and confusion about what is and is not entailed (Keller 2010).

Classical quantitative genetics, like the statistical partitioning of variation from the agricultural trials, makes reference neither to the *measurable genetic factors that underlie the development of that trait* nor to the underlying environmental factors. (Models of *hypothetical* genes are required to generate the traditional formulas of quantitative genetics [Falconer and Mackay (1996), Lynch and Walsh (1998)], but this is a conceptually distinct matter—one that is taken up in section 5.) In contrast, the statistical association of a trait with measurable genetic factors that underlie the development of that trait has received increasing attention in the era of genomics (even though detecting such associations has been possible since the rediscovery of Mendelian genetics at the turn of the last century). The implications of such associations are discussed in sections 6-7.

3. Underlying heterogeneity

As described in the previous section, the estimation of heritability is associated with analysis of variation in *traits*, not transmission of heritable factors. Moreover, the analysis of variation for a given trait does not produce knowledge about the genetic or environmental factors that that underlie the development of that trait. This last point is widely recognized, yet it is commonplace for researchers and commentators to describe estimates of heritability for a trait as the fraction of variation "due to genetic differences" or "due to differences in genetic factors" (albeit factors yet to be identified). By implication, the genetic and environmental variances are held to capture the relative influence of, respectively, genetic and environmental factors.

The unwarranted construal of heritability in terms of genetic factors seems to have two sources: conflation with the commonplace term heritable as meaning transmitted through genes from parent to offspring; and use of the shorthand "genetic variance" to refer to the variance among genotypic values or, in the terms of this article, variance among variety means. More recently, the construal of heritability in terms of genetic factors has been reinforced by use of the term *missing heritability* when a conceptually distinct quantity—the fraction of variation for a trait associated with genomic variants identified in GWA studies—is much lower than the heritability estimated by classical quantitative genetics (Manolio et al. 2009).

Three observations about the agricultural trial and the quantitative genetics study of relatives help counteract the common muddying of the distinction between heritability as derived from analysis of variation in traits and the transmission of heritable factors. The third observation—about underlying heterogeneity—is the most important for the subsequent sections of the article.

1. The variety means (or genotypic values) are means over the trait as it has developed in a specific range of variety-location (genotype-environment) combinations. The variance of those means thus reflects differences in the ways those variety-location combinations—and not simply the varieties in themselves—influence the trait development.

2. The partitioning of variation could, in principle, be undertaken even if the trait measurements were of varieties from different species, classes, or even kingdoms. This thought experiment implies that the variation among the variety means need not correspond to some gradient in genetic factors. Nothing in the method of partitioning variation, even *within* a species, requires that such a gradient exists. This observation also applies to the variation among location means or environmental variance.

3. 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 location (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 outcome as alleles aabbCCDDEE subject to a different sequence of environmental factors FgHiJ (Figure 3).

The possibility of heterogeneity of underlying genetic and environmental factors has not been widely acknowledged by quantitative geneticists. Heterogeneity in the genetic factors underlying traits might explain the limited success GWA studies have had identifying causally relevant genetic variants behind variation in human traits with moderate to high heritability values. The possibility of underlying heterogeneity is consistent with, but more nuanced than, the standard view that most medically significant traits are associated with many genes, each of quite small effect (McCarthy

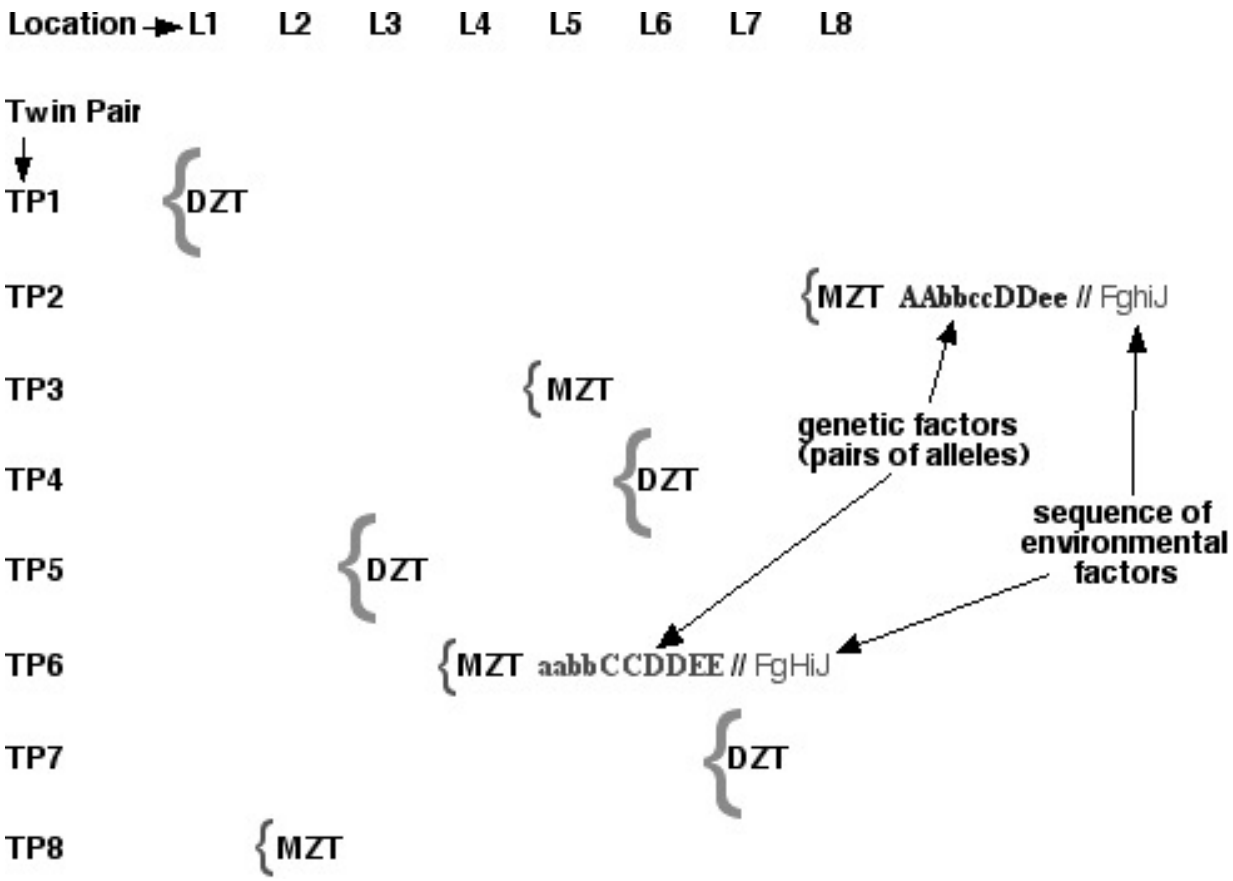

Figure 3. Factors underlying a trait may be heterogeneous even when monozygotic twins raised together (MZT) are more similar than dizygotic twins raised together (DZT). The greater similarity is indicated by the smaller size of the curly brackets. The underlying factors for two MZT pairs are indicated by upper and lower case letters for pairs of alleles (A-E) and environmental factors to which they are subject (F-J).

---

et al. 2008, Couzin-Frankel 2010). Moreover, considering heterogeneity of underlying factors is a matter of attention to environmental as well as genetic factors.

4. What is to be done?

What can researchers do on the basis of knowing that a trait's heritability or another fraction of variance is high when the factors underlying that trait might be heterogeneous? As will become evident, the range of options for human studies becomes very limited. Perspective on this can be gained by examining what is and is

not possible in an agricultural trial, in which control over varieties and locations is far greater than with human variation. This example, in turn, provides food for thought if we move beyond Davey Smith's advice that epidemiology is concerned with the mean of the group and has to discount random variation around that mean (see section 7).

Typically, in an agricultural trial there is variety by location interaction (i.e., genotype by environment interaction) in the analysis of trait variation, which means that for the trait in question the ranking of varieties varies across locations; the best variety in one location is not the best in other locations. The possibility of heterogeneity underlying the variation in trait means can, however, be reduced by *grouping varieties* that are similar in responses across locations using techniques of cluster analysis (Byth et al. 1976), or in other words, by grouping according to similarity in variety means together with variety by location interaction means. (Similarly, locations can be grouped by similarity in responses elicited from varieties grown across those locations.) Varieties in any resulting group tend to be above average for a location in the same locations and below average in the same location (Figure 4).

Now, the wider the range of locations in the measurements on which the grouping is based, the more likely it is that the ups and downs shared by varieties in a group are produced by the same conjunctions of underlying measurable factors. This gives researchers more license to discount the possibility of underlying heterogeneity within a group. They can then hypothesize about the group means—about what factors in the locations elicited basically the same response from varieties in a particular variety group that distinguishes them from other groups. Of course, knowledge from sources other than the data analysis is always needed to help researchers generate any hypotheses about genetic and environmental factors.

For example, imagine a group of plant varieties that originated from particular parental stock more susceptible to plant rusts, a form of parasitic fungi, and that these varieties yielded poorly in locations where rainfall occurred in concentrated periods on poorly drained soils. The obvious hypothesis about genetic factors modulated by environmental factors is that these varieties share genes from the parental stock that are related to rust susceptibility and this susceptibility is evident in the measurements of yield in locations where the rainfall pattern enhances rusts. Through additional research

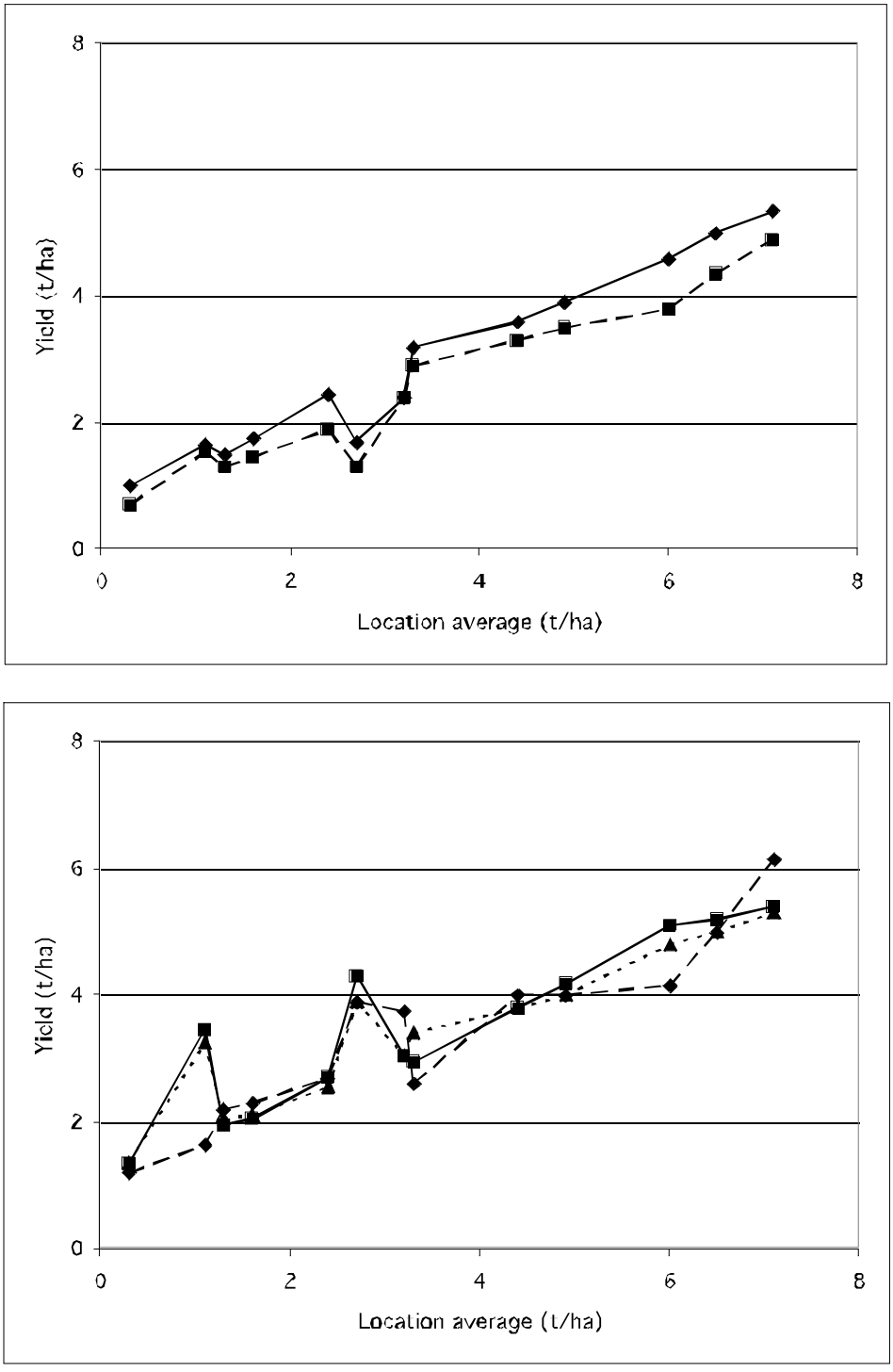

Figure 4. Yields for 5 groups of wheat varieties grown in 13 groups of locations (from Byth et al. 1976; reproduced with permission). The x-axis is the mean over all varieties for that location. The individual varieties (not shown) were clustered into these 5 groups

by similarity of response across locations. These groups were then clustered into two groups as shown in the two plots.

---

comparing the variety and parental genomes, it may be possible to identify specific sets of genes that are shared, to investigate whether and how each one contributes to rust susceptibility in certain environmental conditions, and to use that knowledge in subsequent research or in planting recommendations (Taylor 2006). (Of course, plant breeders can make informed decisions about what varieties to cross even in the absence of knowledge about the specific genetic factors and environmental conditions involved.)

What role does heritability play in research that groups varieties, thus reducing underlying heterogeneity and facilitating the generation of hypotheses about underlying genetic and environmental factors? Answer: Very little. Clustering ensures that the variation among the means for groups of varieties is much higher than the average variation among variety means within groups. The low within-group variation allows the selective breeder to select from the variety group without being very concerned about whether any one variety is the best within that group across all the locations. In other words, heritability *within* the variety group is not so important. At the same time, even if variation among variety group means is smaller than variation among location means or variation among variety by location interaction means (that is, *between-group* heritability is small), researchers can still hypothesize about the group means. In short, the size of the fractions of variance is not key to hypothesizing about the underlying genetic and environmental factors.

Grouping varieties by similarity of responses across locations becomes more difficult when varieties are observed in only a few locations or when the locations are not the same from one variety to the next. This is the case, of course, for variation among traits in humans. It should be noted, however, that reduction of underlying heterogeneity, although helpful, is not essential to agricultural and laboratory breeding. Breeders know, by the very definition of heritability, that a high value means that differences among the mean values for the varieties make up much of the total variation for the trait in question. So the breeders can mate or cross individuals with the desired

values for the trait, expecting that this will lead to offspring with similar, desired values, and thus to improvement in the overall average compared with the previous generation. Now, if there is underlying heterogeneity, the reassortment of genes from parents may well lead to some far-from-expectation offspring, even when the offspring are raised in the locations of their parents (in the example given in Figure 3, AabbCcDDEe subject to the factors FghiJ or FgHiJ). Yet, such outcomes are not troubling to breeders because they can compensate when the results do not meet their expectations. They discard the far-from-expectation offspring and select only those offspring that do have the desired values.

In the study of human traits, selective mating and discarding of defectives is obviously not acceptable. Nor is it possible to replicate human genetic material together with the environmental conditions of many defined locations. It is impossible to have the level of control needed to obtain data for grouping varieties that are similar in responses across locations. In short, as noted earlier, human studies involve substantially less control over materials and conditions than agricultural and laboratory trials. Researchers should expect, therefore, to gain correspondingly less insight about genetic and environmental factors from partitioning of variation for a human trait into the various fractions.

Returning then to the question with which this section began: What then can researchers do once they acknowledge that similarity among human twins or a set of close relatives may well be associated with similarity of yet-to-be-identified genetic factors but the factors may not be the same from one set of relatives to the next, or from one location to the next? Researchers have three options for proceeding on the basis of knowing that a trait's heritability is high when the factors underlying that trait might be heterogeneous:

1. Undertake research to identify the specific, measurable genetic and environmental factors *without reference to the trait's heritability* or the other fractions of the total variance. (e.g., Moffitt et al. 2005, Davey Smith and Ebrahim 2007, Khoury et al. 2007).

2. Restrict attention to variation *within a set of relatives*. For example, even if the underlying factors are unknown, high heritability still means that if one twin develops the

trait, e.g., type 1 diabetes, the other twin is more likely to as well. This information might stimulate the second twin to take measures to reduce the health impact if and when the disease starts to appear. However, notice that this scenario assumes that the timing of getting the condition differs from the first twin to the second. Researchers might well then ask: What factors influence the timing? How changeable are these? How much reduction in risk comes from changing them? To address these issues they would have to identify the genetic and environmental factors involved in the development of the trait, but to do so requires larger sample sizes than any single set of relatives allows. In any case, the question would then arise as to whether the initial results would carry over from one set of relatives to others. Resolving this question is an empirical matter, but it is obvious that carry over becomes less likely the more heterogeneity there is in the factors underlying the development of the trait. Researchers have to be prepared, therefore, for the possibility that the proportion of fruitful investigations will be low compared to those confounded by factors not carrying over well from the initial set of relatives.

3. Use high heritability as an indicator that "the trait [is] a potentially worthwhile candidate for molecular research" to identify the specific genetic factors involved (Nuffield Council on Bioethics 2002, chap. 11) and hope that for *some* traits a gradient of a measurable genetic factor (or composite of factors) runs through the differences among variety means. Such traits might be worth finding even if, in the course of doing so, researchers end up conducting many fruitless investigations of other high-heritability traits for which it turns out there is no such gradient.

Equivalent options apply to proceeding on the basis of knowing another fraction of a trait's variance is high but the factors underlying that trait might be heterogeneous. In particular, when a high value for the among-location-means (or "shared environmental") fraction of variance is used to decide when to search for the specific environmental factors involved, the hope is that for some traits a gradient of a measurable environmental factor (or composite of factors) runs through the differences among location means. Again, this option involves the possibility of fruitless investigations of traits for which there is no gradient in underlying factors. Similarly, if

the researcher chooses to look for specific measurable factors underlying high fractions of non-shared environmental variance and variety-location interaction variance.

The possibility of fruitless investigations of traits for which there is no gradient in underlying factors is not the only concern researchers should have about being guided by high estimates of heritability or other fractions of variance in the traits. As the next section explains, there are more serious problems about estimates and interpretation of fractions of variance for human traits. (Appendix A is provided to counter the response that, given decades of debate among methodologically sophisticated scholars, it is implausible that some fundamental problems in quantitative genetic estimation could have been overlooked [Kendler 2005]. The appendix describes some of the background that has allowed the author to see the study of heredity and variation differently from most researchers and philosophers of science who have contributed to debates about classical quantitative genetics.)

5. Unreliable estimates of fractions of variance for human traits

The reliability of the established methods of estimation of fractions of variance for human traits can be checked by considering a hypothetical agricultural trial where every variety is raised in every location and replicated twice and where the replicates for each variety-location combination are either monozygotic or dizygotic twins (Taylor 2012). The analysis of variance allows one to estimate the trait's actual heritability in that trial, that is, the fraction of the trait's variance associated with differences among the means for the varieties, where these means are taken over all locations and replicates. Similarly, the actual fractions can be estimated for the among-location-means variance, the among-variety-location-interaction-means fraction, and the residual or "error" variance, that is, the average variance within variety-location-combinations. At the same time, given that the replicates are twins, one can also estimate the fractions of variance using the standard formulas of human quantitative genetics (Rijsdijk and Sham 2002). If the standard formulas do not yield estimates that match the correct values from the analysis of variance, something must be wrong with those formulas.

That the standard formulas of human quantitative genetics do not, indeed, match the correct values can be demonstrated theoretically and through analysis of simulated

agricultural trials (Taylor 2007, 2012). To appreciate why the standard formulas, as well as their generalization as Structural Equation Models, do not yield reliable values, it is helpful to note the following features of those formulas:

1. The standard formulas and Structural Equation Models do not separate a variety-location-interaction variance fraction, but subsume it in the among-variety-means variance fraction, that is, in the so-called heritability estimate.

2. Empirical estimation of the variety-location-interaction variance fraction is nevertheless possible, but only if the appropriate classes of data are available; that is not often the case. For example, it is not common to have data about monozygotic twins raised in separate locations in which there is no correlation from one twin's location to the other.

3. Partitioning of trait variation into components rests on models of hypothetical, idealized genes with simple Mendelian inheritance and direct contributions to the trait. Given that the data are about *traits*, an analysis of trait variation that makes no reference to theoretical, idealized genes must also be possible (Taylor 2012). Such a *gene-free* analysis affirms the distinction made in section 2, that estimation of heritability is associated with analysis of variation in traits, not transmission of heritable factors.

4. The derivation of the standard formulas and models assumes that, all other things being equal, similarity in traits for relatives is proportional to the fraction shared by the relatives of all the genes that vary in the population. For example, fraternal or dizygotic twins share half of the variable genes that identical or monozygotic twins share. However, plausible models of the contributions of multiple genes to a trait can be shown to result in, all other things being equal, ratios of dizygotic similarity to monozygotic similarity that are not .5 and that vary considerably around their average (Taylor 2007). This point here does not depend on the validity of any particular hypothetical model of multiple genes contributing to the trait. The assumption is unreliable because the relevant correlations need to be based on observed *traits* and, as such, cannot be directly given by the proportion of shared *genes* involved in the development of those traits. For the same reason, heuristic values of the similarity of relatives of other degrees, which are ubiquitous, are also unreliable. (Identifying the

exact fraction of genes shared by relatives [Visscher et al. 2006] does not address the problem that similarity in traits for relatives need not be proportional to this fraction.)

5. Empirical estimation of a parameter to take degree of relatedness into account depends on the appropriate classes of data being available. Data about unrelated individuals raised in the same location are especially valuable.

6. The residual variance, or the sub-fraction of that variance not related to measurement error, has been interpreted as a “non-shared *environmental* effect," where a large value means that within-family or "non-shared environmental differences" are large relative to the effects due to the members of a family growing up in the same location or "shared environmental differences" (Plomin 1999; see critical review by Turkheimer and Waldron 2000). The more careful interpretation is that residual variance corresponds to the variation of trait differences among replicates within variety-location combinations that has no systematic relation to variation among variety means, among location means, or among variety-location combinations. Now, to interpret the unsystematic variation of a trait in terms of differences in the underlying factors is not warranted in the first place (see first paragraph of section 4). However, if one were to go ahead and attempt to do just that, the interaction of genetic as well as environmental factors in specific variety-location combinations would need to be included in the picture.

When Davey Smith, the epidemiologist, invokes human behavioral genetics, he cites the work of Turkheimer, a researcher on the more critical wing of his field. Turkheimer (2004) emphasizes, as this article does, that a heterogeneity of pathways can lead to the traits measured. The implications for human sciences that he and Davey Smith attach to heterogeneous pathways are, however, colored by an insufficiently critical interpretation of human quantitative genetic estimates. In particular, following point 6 above, those estimates can provide no warrant for the claim they make that the non-shared environmental influences overshadow common environmental influences. Moreover, a corollary of the discussion so far is that *none* of Turkheimer’s (2000) three laws of behavioral genetics is reliable; nor is an oft-cited addition (labeled 4 in Table 1).

Table 1. Laws of behavioral genetics, recast

Laws of behavioral genetics	**Revised statement**
1. All human behavioral traits are heritable.	1. All human behavioral traits show heritability, but: a. estimation methods are not reliable; and b. heritability does not mean that the "genetic" (among-variety-means) variance translates into differences in genetic factors across the population studied.
2. The effect of being raised in the same family is smaller than the effect of the genes.	2. This effect has not been established, because a. estimation methods are not reliable; and b. fractions of variance do not translate into differences in genetic and environmental factors across the population studied.
3. A substantial portion of the variation in complex human behavioral traits is not accounted for by the effects of genes or families.	3. The residual fraction of the variation in complex human behavioral traits is often substantial, but the interaction of unknown genetic as well as environmental factors specific to the particular variety-location combinations underlies this unsystematic variation.
4. Heritability tends to increase over people's lifetimes, that is, genetic differences come to eclipse environmental differences (Plomin 1999).	4. Unless the estimation method separates the interaction fraction from heritability and it is shown to be negligible, this trend could equally well indicate that the *interaction* component increases over time.

The critical account provided above and in previous sections calls into question the interpretation from human quantitative genetics that "non-shared environmental" influences among siblings overshadow shared environmental influences. (Appendix B can be referred to for responses to the critical account and counter-responses.) Yet, if the analogy that Davey Smith draws to the lack of success in the search for systematic aspects of those "non-shared environmental influences" has been undercut, we are still

left with the difficulty epidemiology has had in moving from significant population-level risk factors to improved prediction of cases at an individual level. To address that issue, the next section extends the possibility of underlying heterogeneity raised in sections 3 and 4, and reviews several additional forms of heterogeneity. The final section captures in four conceptual themes the potential significance for epidemiology of various forms of heterogeneity.

6. Heterogeneity of heterogeneities

It is quite simple to extend the possibility of underlying heterogeneity to epidemiology: when similar responses of different groups of people are observed, it need not be assumed that similar conjunctions of risk and protective factors have been involved in producing those responses. One difference, however, is that, whereas in quantitative genetics the different varieties or locations are not distinguished by some measurable genetic or environmental factors, in epidemiology the different groups of people are typically characterized by reference to measurable factors, such as socio-economic status, gender, or even genetic factors. The following example, in which researchers detected associations of a trait with a combination of measurable genetic and environmental factors, serves to illustrate this distinction and, at the same time, point to various other forms of heterogeneity. (The example is also used to illustrate conceptual points in section 7, but no endorsement of the importance or generality of the results should be read into this choice; see Wasserman 2004 and Morris et al. 2007 for critical commentaries.)

Figure 5 depicts a reported association of behavior with enzymatic activity related to a single-locus genetic variant, MAOA, and with the social or environmental factor, childhood maltreatment (Caspi et al. 2002). The association of MAOA and childhood maltreatment with anti-social behavior is an instance of *gene-environment interaction*, a term that refers to the statistical interaction between measured genetic factors and measured environmental factors (Moffitt 2005). In other words, the association of the trait with the genetic factor is modified by the value of the environmental factor, and vice versa. (Such interaction, confusingly, has no conceptual or empirical relationship to

"genotype-environment interaction" or variety-location interaction, as discussed in sections 2-6.) The example involves a single genetic variant, but Genome-Wide Association (GWA) studies are now able to quantify the mean effect of each of a wide array of variants across the human genome on a trait, such as mean height differential (Weedon et al. 2008), and relative risk for type 1 diabetes (Barrett et al. 2009).

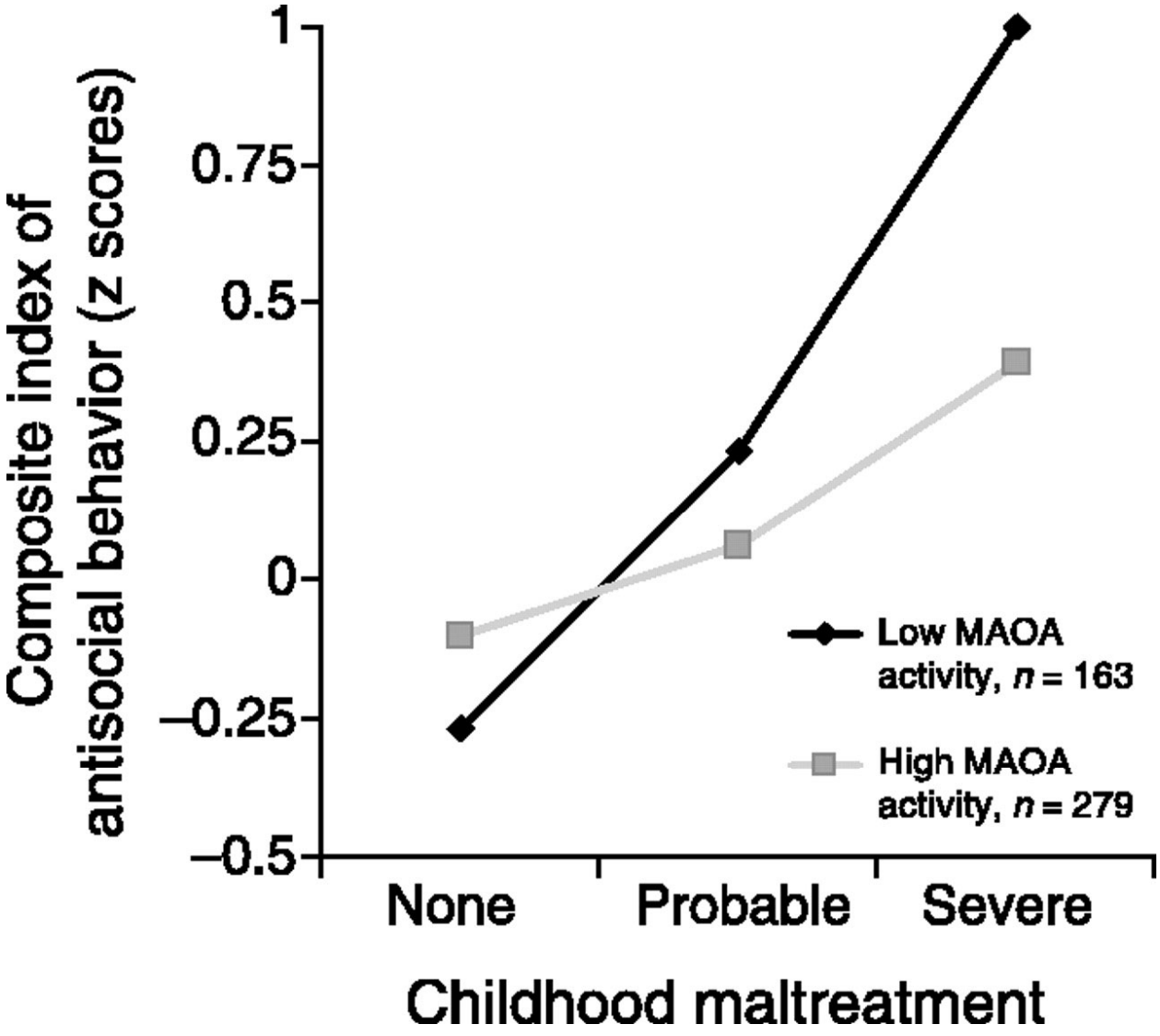

Figure 5. Mean adult composite anti-social behavior score in relation to levels of MonoAmineOxidaseA and Level of Childhood Maltreatment for a sample from Dunedin, New Zealand (Caspi et al. 2002, 852; reproduced with permission).

Epidemiologists should be familiar with heterogeneity of various forms that complicate the statistical association of a trait with measurable factors. Consider the MAOA-maltreatment example:

1. The points plotted in Figure 5 represent a mean value, around which there is variation (which is the simplest form of heterogeneity) in the measure of antisocial behavior.

2. The binary of high-versus-low MAOA activity is a simplification of the many variants in the MAOA gene and related regulatory regions (Di Giovanni et al. 2008, 84ff).

3. Similarly, childhood maltreatment varies in nature, degree, and timing, as well as in its meaning to the child.

4. The trait in question, anti-social behavior, can be assessed in different ways (indeed, Figure 3 plots a composite index derived from various assessments).

5. Even for any one assessment of anti-social behavior, different individuals may have arrived at similar values through different pathways of development (in the same way that is clearly the case for development of human height).

Equivalent considerations pertain to the effects associated with multiple genetic variants exposed through GWA studies (although the number of GWA studies of interactions between genetic and environmental factors is still small). In short, there is a heterogeneity of heterogeneities.

7. What is to be done?—Part II

Should epidemiologists follow Davey Smith's advice to accept considerable randomness at the individual level and keep their focus on modifiable causes of disease at the population level? Or are there other courses of action? In particular, when researchers find a statistical association of a trait with measurable risk or protective factors, what can they do given the potential for various forms of heterogeneity to complicate the association? These questions are viewed in relation to four conceptual themes in the sub-sections to follow. The cases used are presented with a view to motivating the conceptual themes, not to surveying the literature on any of the cases.

7.1 *The push for individualized or personalized medicine can lead to more population-level treatment or treatment according to the mean of one's group.*

Consider population-level approaches in the MAOA-maltreatment case. As the authors conclude, their results "could inform the development of future pharmacological treatments" (Caspi et al. 2002, 853). By implication, if low-MAOA children could be identified, prophylatic drug treatment could reduce their propensity to anti-social behavior as adults, or, more strictly, their vulnerability to childhood maltreatment insofar as that increases their propensity to anti-social behavior as adults. Reciprocally, if severe childhood maltreatment could be identified and reduced or stopped early, this action could nullify the influence of a child's MAOA level on undesired adult outcomes. The risk reduction for anti-social behavior produced by each population-level approach would depend on whether the threshold for unacceptable anti-social behavior is set high or low.

Of course, with only 3% of the population in the low-MAOA plus severe-maltreatment category, epidemiologists might not recommend any population-level action based on the reported association. Moreover, they may wait to see if the results in this case apply to populations other than that of the original study (Wasserman 2004, Morris et al. 2007). Indeed, some meta-analyses have cast doubt on the generality of a similar study by Caspi, Moffitt and collaborators (Risch et al. 2009). Suppose, however, *for discussion purposes*, that something analogous to the MAOA-maltreatment result is replicated and the proportion in the high-risk category is sufficient to stimulate action. The population-level approaches could still run into troubles, in the following manner.

Notice that Figure 3 presents the means; around any mean there will be variation. From other figures in the study (Caspi et al. 2002) it can be seen that some of the high-MAOA individuals end up with higher anti-social behavior scores than some of the low-MAOA individuals. Moreover, depending on the threshold, a substantial fraction of the low-MAOA plus severe-maltreatment category does not end up as anti-social adults. Yet, in practice, once the resources were invested to screen children for MAOA levels, the attention of parents, teachers, social workers, and so on would be focused on *all* low-MAOA children. Indeed, *how could treatment on the basis of group*

membership be avoided if such adults do not know from a childhood MAOA assessment whether any particular individual is one who would go on, after childhood maltreatment, to become an antisocial adult? (Of course, adults would take other traits into account in how they treated any child, but what is relevant here is how information about MAOA status influences treatment.)

Now, some of the parents of low-MAOA children might resist their children being treated according to the mean of the MAOA group. They might also balk at years of prophylatic drug treatment or of maltreatment monitoring by social workers. These parents—together with others concerned about the same issues—could push for additional research to identify other characteristics that differentiate among the low-MAOA children (and perhaps also help predict who among the high-MAOA children are vulnerable). It may be that no systematic characteristics would be found; variation within the low- and high-MAOA subpopulations would then fit Davey Smith's view of unavoidable randomness. Nevertheless, it would have been understandable that researchers had sought a more refined account of risk factors than given by the population-level approaches.

To understand the need for more refined risk accounts is not, however, to endorse the pursuit of personalized medicine customized to the individual. Indeed, the scenario played out above points to a serious shortcoming of that very endeavor. In its simplest form, personalized medicine involves the use of genetic information to predict which patients with a given condition (e.g., heart aryhthmia) will benefit from a particular drug treatment (e.g., beta blockers). More ambitiously, personalized medicine promises to inform people of their heightened vulnerability or resistance to specific environmental, dietary, therapeutic, and other factors early enough so they can adjust their exposure and risky behaviors accordingly.

If the MAOA analogy holds, the path to personalized medicine would often involve a phase in which large numbers of people are treated *according to their group membership*. Consider, however, which kinds of medical conditions would receive the necessary investment in pharmaceutical and sociological research, screening, and preventative treatment or monitoring to address the conjunction of genetic and environmental factors involved. Some well-organized parental advocacy groups may

secure funding to address the prenatal or neonatal diagnosis and post-natal treatment of rare debilitating genetic disorders (Panofsky 2011, Terry et al. 2007), as happens with phenylketonuria. However, public and corporate policy would more likely focus on conditions with a large value for the *average benefit* of ameliorating the effect of the genetic difference multiplied by *number* of people considered vulnerable.

For high-average-benefit conditions, taking the MAOA case as a guide, *if* the effect of the genetic difference depends on identified social or environmental factors, and *if* variability within the groups that have on-average high and low vulnerability produces a problem of misclassification, then pressure would arise for researchers to *differentiate among individuals within the groups*. However, until distinguishing characteristics were found, parents, teachers, doctors, social workers, insurance companies, policy makers, friends, and the individuals themselves could make no better use of the genetic information than to treat individuals according to which genetic group they belonged to. (Again, some people listed in the previous sentence would take other traits into account in how they treated any individual, but what is relevant here is how the genetic information influences that treatment.) Moreover, *if* the additional research were not conducted or were not successful, or *if* the cost of differentiating among individuals were too high, we might *never get beyond treating individuals according to their genetic-group membership*. In short, an under-acknowledged danger in the pursuit of personalized medicine lies, ironically, in genetic information being used to treat people according to the mean of their genetic group—a population-level approach—leaving variation around that mean accepted as unavoidable noise.

7.2 *It can be quite reasonable to try to differentiate among individuals so as to improve risk prediction, even though finding ways to do so may not be straightforward.*

Consider the modifiability in practice of any given population-level risk factor and take the MAOA-maltreatment scenario as a thought experiment. Population-level measures would require more than finding a safe and effective prophylatic drug treatment. MAOA screening would need to become routine, compliance with the treatment achieved, and so-called side-effects addressed—including the effects of considering all low-MAOA children as incipiently anti-social. On the maltreatment side,

the detection and prevention of childhood maltreatment might entail intrusions into many households, surveillance or mandatory reporting, monitoring and intervention by state agencies, diversion of government budgets from other needs, and so on. In short, even if reduced childhood maltreatment were seen as a positive outcome, the means might not be unconditionally positive to all (Wasserman 2004).

What the example illustrates is that, for the population-level approach to make sense, the population-level risk factor has to be modifiable *in practice*. This depends on the political, economic, and cultural circumstances around public health measures, as is well illustrated by the uneven implementation of laws in the United States requiring seat belt use in cars (Wikipedia 2011). Moreover, the effects of shifting the distribution of that factor should not disadvantageous to individuals who are not high risk. Unless both of these conditions hold, it makes sense to search for risk factors that differentiate individuals within a population. For some traits such a search might reveal no systematic characteristics; variation within the populations would then fit Davey Smith's view of unavoidable randomness. However, for other traits the within-population risk factors might be identifiable; some of these factors might be easier to modify than the population-level factors.

The population-level approach, advocated most famously by Rose (1985), is well illustrated by the case of smoking and lung cancer. The relative and absolute risks of smoking are high; there is a dose-response curve (Bjartveit and Tverdal 2005); and the historical trends after allowing for latency match smoking rates. Moreover, population-level policies, such as cigarette taxes and indoor smoking bans, are not disadvantageous for the health of those smokers who turn out to be less susceptible to lung cancer. Nor are the policies disadvantageous to others outside the population in question, i.e., non-smokers. In short, there seems to be little incentive now that smoking-reduction policies are gained traction, to find ways to screen for high-risk individuals and focus resources on treating them.

Most cases are, however, not like smoking and lung cancer; the justification for de-emphasizing within-population risk factors is usually far less clear. Consider cardiovascular disease. Lynch and colleagues reported that for a population of Finnish men, "94.6% of [coronary heart disease] events occurred among men exposed to at

least one conventional risk factor" (smoking, hypertension, dyslipidaemia, and diabetes). They concluded that the focus of efforts to reduce coronary heart disease should be on reducing these factors, not on further refinement of psychosocial or other risk factors (Lynch et al. 2006). Yet, focusing on the conventional risk factors could translate variously to population-level efforts for smoking reduction, to screening for and treatment of individuals at high risk for hypertension, dyslipidaemia, and diabetes, or to some hybrid effort in which anti-smoking efforts highlighted the special risk, say, to people with high blood pressure. In short, it would be very reasonable for anyone promoting or subject to the screening and treatment approach to be interested in improved prediction of who is at high risk.

In this vein, Ridker et al. (2007) noted that the conventional risk factors for heart disease in women (as combined in the Framingham score) misclassify many women as of intermediate risk who were actually higher or lower risk. Ridker et al's refined Reynolds Risk Score performs better primarily, it seems, by including the risk marker cReactive Protein (CRP). Health policymakers might examine the ratio of benefits from improved risk determination to the increased cost from measuring CRP levels. Biomedical researchers might see inclusion of CRP in the Reynolds Risk Score as grounds for examining mechanisms around a rise in CRP levels for an individual. They might hope for eventual improvement in therapy and prevention—perhaps targeted at high-risk individuals, but perhaps also yielding population-level interventions.

It turns out that CRP is almost surely a marker, not a direct cause or valid candidate for therapeutic reduction (C Reactive Protein Coronary Heart Disease Genetics Collaboration 2011). Yet the point remains: Focusing on modifiable causes of disease at the population level is a choice for epidemiologists that is not dictated solely or primarily by considerable randomness at the individual level. There are situations in which it is quite reasonable to be interested in improved prediction of cases at an individual level. The value of any such improvements, however, depends on the economics of health policy as well as on whether a richer picture of risk factors translates through biomedical research into improvements in therapy and possibly prevention.

Suppose researchers decide neither to discount nor to give blanket acceptance

to randomness at the individual level. What they then face is not unbridled exploration of risk factors that might differentiate individuals or sub-groups within the population. Rather, it is the challenge of exposing possible risk factors and assessing which ones are worth considering for insertion into refined risk equations for screening and treatment. And for this challenge there are no off-the-shelf guidelines. Consider the following contrasting cases. The pathway seems straightforward in the "changing view of pathophysiology" underlying atherothrombosis, which led Ridker et al. (2007) to assay a large number of biomarkers then compare a range of new risk prediction equations that included CRP and other markers. The role of refined within-population distinctions has, however, been more subtle in the research and discussions of Lustig and his colleagues around the association of sugar (i.e., fructose-glucose) with metabolic syndrome of diabetes, hypertension, lipid problems, and cardiovascular disease (Lustig et al. 2012). They end up advocating a population-level policy of reducing availability of sugar, but the path to that position is more complex. Arguments over epidemiological associations—after all, countries with high sugar consumption also have high fat consumption—are combined with physiological research to unravel mechanisms of metabolism under various conditions. The latter research takes note, among other things, that "20% of obese people have normal metabolism and… up to 40% of normal-weight people develop the diseases that constitute the metabolic syndrome" (Lustig et al. 2012). This observation led to epidemiological comparisons among countries that controlled for physical activity and being overweight or obese, which have shown that "[d]ifferences in sugar availability statistically explain variations in diabetes prevalence rates at a population level" (Basu et al. 2013).

Now, even if it turned out that approaches other than limiting sugar consumption for the population were more feasible or effective in reducing coronary heart disease, juxtaposing the Ridker and Lustig examples suffices to make the point that no simple guidelines can be expected for research and policy for risk factors that differentiate among individuals. That, however, is no argument against trying to differentiate among individuals so as to improve risk prediction.

7.3 *When researchers think about the causal dynamics underlying patterns in data, such as associations with risk factors, it may be helpful not to see deviations from patterns as noise and to pay attention to heterogeneity in forms such as there being multiple paths to the "same" trait.*

The observation that CRP improves risk prediction even though it is not a causal factor points to a further challenge. The conventional way to think about risk factors is to distinguish the modifiable (e.g., blood pressure) from the non-modifiable (e.g., gender). We envisage modifying the former so as to change the risk; for the latter clinicians screen more vigilantly or researchers investigate further with a view to exposing what modifiable risk factors underlie them. There is, however, an alternative way of thinking about both kinds of risk factors, namely to see them as patterns detected in data generated by some unknown causal dynamics. Then, when researchers envisage modifying a risk factor, they are *hoping* that the causal dynamics generating past data persist into the current situation and that modification of a factor does not restructure those dynamics—or, at least, not very much.

To modify factors without restructuring dynamics would mean, for example, that, if low income is a risk factor for a condition (e.g., stress as measured by some biomarker), an experiment in which some families get a lump sum increase in income would be expected to lead to a lower rate of that condition. If this expectation seems crude, it is because we are not surprised by results that indicate greater complexity in the causal dynamics that generated the patterns as well as in the modified dynamics emerging under the experiment. Ludwig et al. (2012), for example, showed that for families joining an experiment in which some were randomly chosen to move "from a high-poverty to lower-poverty neighborhood," moving "leads to long-term (10- to 15-year) improvements in adult physical and mental health and subjective well-being, despite not affecting economic self-sufficiency." The challenge that such examples pose for epidemiologists is to expose the causal dynamics generating patterns in data, moreover to do so in ways that provide expectations for how the dynamics might be modified by interventions. Such interventions might be at a population level or they might be only among high-risk individuals.

Exposing underlying causal dynamics may strike data analysts as an elusive

goal, but it is instructive to at least entertain the quest. First, to avoid setting off in the wrong direction, researchers might find avoid using the term “risk factor” because its causal connotations invite confusion—perhaps “risk-associated variable” instead? They should also put aside the contrast between modifiable causes of conditions at the population level and randomness at the individual level. That framing leads us to see deviation from a risk equation as noise whereas it could indicate the gap between the fitted equation and the causal dynamics generating the data. Then researchers might pick up the possibility of heterogeneity in genetic and environmental factors underlying heritability and other fractions of the variation for a trait. Attention to this possibility can be readily extended so that, when similar responses of different individual types are observed, researchers do not assume that similar conjunctions of risk or protective factors have been involved in producing those responses. There may be multiple pathways to the same trait value. In this vein, the data analyses of Kendler et al. (2002) and Ou (2005) in psychology and education are notable for exposing alternative pathways—as well as admirable for their reserve about how to interpret their fitted models causally.

Attention to heterogeneity may lead us to reconceive the trait being analyzed as a heterogeneous mix of traits. Tilley (2000) provides a striking exemplar of this in his discussion of evaluations of the effect of Closed Circuit television (CCTV) on crime in parking lots. Such evaluations could mix together studies of situations in which different mechanisms (or a mix of mechanisms) and different contexts apply. The different CCTV-crime studies might be subject to a meta-analysis, but no meaningful recommendation would be likely, even if all results were in the same direction (see also Cartwright and Hardie 2012). Even though not all phenomena are like CCTV and crime, key variables in epidemiology, such as socioeconomic status, admit to substantial heterogeneity. As Davey-Smith et al. (2000) note: “socioeconomic categories—whether based on education, employment, occupational social class or housing tenure—may have neither the same nor consistent meanings in different ethnic groups.” Moreover, especially in relation to implications for action, different kinds of heterogeneity can be distinguished, ranging from a "cabinet of curiosities" to “participatory restructuring through multiple points of engagement” (Taylor 2011). Discussion of different senses of

heterogeneity would require a separate article, but the issue of *meaning* is picked up in the final angle to follow, which brings this article to a close on a speculative note.

7.4 *Attention to the meaning of life events and other factors can allow for heterogeneous events to be subsumed under single factors. However, given the labor-intensive nature of research into meaning, epidemiologists may need the subjects to show how they connect knowledge with action to change their lives and communities in response to ongoing social changes in specific situations.*

A line of research initiated by the sociologists Brown and Harris in the late 1960s has investigated 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. Because what might be recorded as 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. At the same time, apparently heterogeneous events can be subsumed under one factor, such as, in explanation of depression, a severe, adverse event in the year prior to onset. In sum, their Life Events and Difficulties methodology integrates "the quantitative analyses of epidemiology and the [in] depth understanding of the case history approach" (Brown and Harris 1989, x).

The methodology of Brown and Harris is labor intensive; many who have been trained in it have tried to streamline the case-history component and eventually shifted to different lines of research. Suppose, however, that researchers took as starting points the labor intensity of methodology that exposes diversity of meanings and the possibility of underlying heterogeneity. One way forward might be to allow the subjects in specific situations that continue to change to show how they connect knowledge with action. That is, researchers could depart from the traditional emphasis on exposures impinging on subjects and, instead, observe communities where people are resilient and reorganize their health, lives and communities in response to social changes (Sampson et al. 1997). Although the patterns and variation among people exposed by

those studies might not extrapolate readily over time, place, and scale, they could provide a point of departure for research and policy engagements in subsequent situations the researchers study.

Such *agent-oriented* epidemiologists would need to be conversant with studies of resilience and reorganization in communities. They would need to train in participant observation and qualitative methods for research on population health changes that arise through grassroots and professional initiatives and then grow into loosely knit social movements, e.g., around innovations in short-term therapy for depression (e.g., Griffin and Tyrrell 2007, White 2007). In return, they—or anyone following the other three conceptual themes in this section—need not accept Davey Smith's conclusion that considerable randomness at the individual level requires epidemiologists to focus on modifiable causes of disease at the population level. There are pathways to explore that take us beyond the gloomy prospect of epidemiology limited in its power to generate reliable and useful associations of multiple factors with a disease.

Acknowledgements

This article is based on research supported by the National Science Foundation under grant no. SES–0634744. Some sentences and diagrams have been drawn from Taylor (2006, 2010, 2012). I acknowledge thought-provoking comments on an earlier version of this article contributed by Lindsay Morgia, George Davey Smith, an editor, and anonymous reviewers.

Appendix A

Getting beyond the implausibility response

This article's primary goal is to promote wider discussion of heterogeneity, in various forms, among genetic and epidemiological researchers and philosophers of science. Yet, to undercut the analogy invoked by Davey Smith I take a radical path, explaining how the estimates made in human quantitative genetics are unreliable and typical interpretations, including interpretation of non-shared environmental influences, are unjustified. Some readers may deem it implausible, given decades of debate among methodologically sophisticated scholars, that some fundamental problems in

quantitative genetic estimation have been overlooked (Kendler 2005). With a view to moving readers beyond the implausibility response, let me describe some of the background that has allowed me to see the study of heredity and variation differently from most researchers and philosophers of science who have addressed quantitative genetics.

My initial research work in the mid-1970s involved the statistical analysis of large plant breeding trials, in which many cultivated varieties would be tested in each of many locations around the world. A first step in the analysis was to partition the variation in a given trait, say, yield of wheat plants, into components related to the averages or means of the varieties (across all locations), the means of the locations (across all varieties), and so on (see Figure 1). (Indeed, agricultural breeding was where partitioning of variation and measuring heritability originated.) The challenge for the plant breeders with whom I worked was to go beyond the partitioning and hypothesize what it was about any variety that led to its pattern of response across locations and what it was about any location that led to the varieties' responses in that location compared to others. Knowing what aspects of, say, the pedigree of the variety or of the environment conditions in the location could inform subsequent breeding or cultivation decisions. Yet, hypothesis generation was not easy even though we had large and complete data sets to work from. A lesson I took from that research was that the limits to hypothesizing about genetic and environmental factors must be even greater when researchers partition variation for *human* traits. In human studies any genetically-defined type is, at best, replicated twice—as identical twins—and different genetic types cannot be systematically raised across the same range of "locations"—families, socio-economic conditions, and so on.

Fast forward to a decade ago: I was learning about three disparate areas of quantitative research that attempt to make sense of the complexity of biological and social factors that build on each other in the development of the given trait over the life course (Taylor 2004). I was impressed by what had been accomplished, but had some reservations about the models used in one of the areas, namely, Dickens and Flynn's (2001) attempt to resolve the IQ paradox, in which researchers find large generation-to-generation advances in IQ test scores even though the trait is held to have high

heritability. I explained my reservations to Dickens, digested his responses, and explained my reservations about his subsequent responses. In the course of this I found myself digger deeper into the conceptual foundations of heritability estimation and partitioning of variation. In order to present a picture that differed from what Dickens, Flynn, and others accepted without second thought, I was explicating first principles, not disputing specialized models or mathematics or pointing to new data. Making extensive use of perspectives and examples from the earlier plant breeding research, my exposition took a pedagogical, build-from-the-foundations style (Taylor 2006, 2007, 2010).

Meanwhile, my investigation continued of the other two areas—life events and difficulties research (Brown and Harris 1989) and developmental origins of chronic diseases (Barker 1998). Barker's work led me to life-course epidemiology (Kuh and Ben-Shlomo 2004), so I spent time with Ben-Shlomo and the active social epidemiology research group at Bristol University. Davey Smith is a leading figure in that group and co-edits the *International Journal of Epidemiology* based at Bristol. While visiting in 2007 I gave a talk on "new and old debates about genes and environment," which touched on some of the questions about heterogeneity raised in this article. Davey Smith's spoken response was along the lines of his subsequent "gloomy prospect" article: epidemiologists have to accept considerable randomness at the individual level and keep the focus on modifiable causes of disease at the population level. In his ensuing article, Davey Smith links this perspective to claims from quantitative genetics, thus providing me an opportunity to address social epidemiologists and human quantitative geneticists at the same time as I responded to his account. In bringing my interest in heterogeneity to the attention of those audiences, this article extends the first-principles emphasis of the other recent work and thereby speaks to philosophers of science. My contribution to philosophy takes the form, however, of articulating conceptual themes, not dissecting specific cases on empirical, analytical, bioethical or policy grounds. The expository approach reflects the background, with its roots in plant breeding trials, that I have sketched in this Appendix, as well as the idea that contributing to the conceptual toolbox of readers will prepare them to make their own contributions to wider discussion of heredity, variation, and heterogeneity.

Appendix B

Table of objections to critiques of heritability studies and counter-responses

Response	**Counter-Response**
There are many definitions or formulas for heritability other than the ratio of the variance among variety means (genotypic values) to the variance of the trait across the whole data set.	The justification for calling the other definitions *heritability* is that they approximate the original definition or the predictions it provides of advance under selection.
In particular, narrow-sense heritability is favored over the broad-sense heritability above.	Narrow-sense heritability can only be defined when quantitative genetics employs models of hypothetical, idealized genes with simple Mendelian inheritance and direct contributions to the trait. However, a gene-free analysis of trait variation must also be possible (Taylor 2012).
Genotype-environment interaction variance (here: variety-location interaction variance) has been shown to be negligible for humans.	From Taylor (2012): "Often cited in this vein, Plomin et al. (1977) consider a proxy for variety-location interaction, namely, the interaction of some variable, e.g., average educational attainment, averaged for biological parents and for adoptive parents. How well such proxy results reflect actual variety-location interaction and the generality of the low values found by Plomin et al." has not been established.
Gene-environment interaction is not assumed to be negligible—Indeed, there is an institutionalized field of research on the topic.	This use of the term interaction in this response refers to the statistical interaction between measured genetic factors and measured environmental factors. This is conceptually and empirically distinct from

	variety-location (genotype-environment) interaction that means that for the trait in question the ranking of varieties varies across locations.
The intraclass correlation formulas in heritability studies have been superceded by calculations that use Structural Equation Models.	Structural Equation Modeling shares the features that render unreliable estimates made using the formulas (see points #1-6 under Unreliable estimates in section 6).
"Research into the genetics of complex traits has moved from the estimation of genetic variance in populations to the detection and identification of variants that are associated with or directly cause variation" (Visscher, et al. 2007)	Classical quantitative genetics… remains relevant in several ways, such as indicating that "the trait [is] a potentially worthwhile candidate for molecular research" to identify the specific genetic factors involved (Nuffield Council on Bioethics 2002, chap. 11) and, in turn, the difficulty of identifying such genetic factors has led to concerns about "missing heritability" (e.g., Manolio et al. 2009). [Text adapted from Taylor (2012)]
It is past time to move beyond debates on the existence of genetic influences.	The problems summarized in this account are worth recognizing: they call most estimates and interpretations from human quantitative genetics into question; and the possibility of underlying heterogeneity that arises from the more critical account helps in clarifying the challenging paths ahead.
"It is one thing to criticize the methodology of specific studies. It is quite another to suggest… that we reject the results of an entire field of scientific inquiry… It is highly unlikely that modern psychiatric genetics will be judged by	"Human quantitative genetics could be viewed, *contra* Kendler, as akin to alchemy… a field of inquiry that provided observations, questions, tools, debates, careers, and institutions which modern chemistry built on, but ultimately had to

future historians of science to be in such company [as astrology and alchemy]." (Kendler 2005, 10)	break away from to make further progress" (Taylor 2010).
Some assumption has to be made that connects similarity in traits for relatives to the fraction shared by the relatives of all the genes that vary in the population; Direct proportionality is as good as any other assumption.	No assumption has to be made because, in a gene-free analysis of trait variation, empirical estimation of a parameter to take degree of relatedness into account is possible (Taylor 2012). In any case, estimates should indicate the sensitivity to any such assumption.
It is well known that high heritability does not show where to look for the variants.	It is not widely acknowledged that high heritability is not a good indicator of which traits are worthwhile candidates for molecular research.
Human quantitative genetics is subject to various well-recognized, albeit contested objections, for example, has zygosity of twins been correctly ascertained and representatively sampled?; is the treatment or experience of the twins or unrelated individuals within a location unaffected by whether they are monzygotic twins, dizygotic twins, or unrelated? (Taylor 2012).	A focus on these technical issues runs the risk of missing or even reinforcing the more fundamental problems raised in this account.

References

Barker, D. J. P. (1998). *Mothers, Babies, and Health in Later Life*. Edinburgh: Churchill Livingstone.

Barrett, J. C., Clayton, D., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., Plagnol, V., Pociot, F., Schuilenburg, H., Smyth, D. J., Stevens, H., Todd, J. A., Walker, N. M., Rich, S. S., & Type 1 Diabetes Genetics Consortium (2009). Genome-wide association study and meta-analysis finds over 40 loci affect risk of type 1 diabetes. *Nature Genetics, 41,* 703-707.

Basu, S., Yoffe, P., Hills, N., & Lustig, R. H. (2013). The Relationship of Sugar to Population-Level Diabetes Prevalence: An Econometric Analysis of Repeated Cross-Sectional Data. *PLoS ONE, 8,* e57873.

Bjartveit, K., & Tverdal, A. (2005). Health consequences of smoking 1-4 cigarettes per day. *Tobacco Control, 14,* 315-320.

Brown, G. W., & Harris, T. O. (Eds.) (1989). *Life Events and Illness*. New York: Guilford Press.

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.

C Reactive Protein Coronary Heart Disease Genetics Collaboration (2011). Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *British Medical Journal, 342,* d548.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.

Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I. W., Taylor, A., & Poulton, R. (2002). Role of Genotype in the Cycle of Violence in Maltreated Children. *Science, 297,* 851-854.

Couzin-Frankel, J. (2010). Major Heart Disease Genes Prove Elusive. *Science, 328,* 1220-1221.

Davey Smith, G. (2011). Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *International Journal of Epidemiology, 40,* 537-562.

Davey Smith, G., Charsley, K., Lambert, H., Paul, S., Fenton, S., & Ahmad, W. (2000). Ethnicity, health and the meaning of socio-economic position. In H. Graham (Ed.), *Understanding health inequalities* pp. 25-37. Buckingham [England]: Open University Press.

Davey-Smith, G., & Ebrahim, S. (2007). Mendelian randomization: Genetic variants as instruments for strengthening causal influences in observational studies. In M. Weinstein, J. W. Vaupel, & K. W. Wachter (Eds.), *Biosocial Surveys* pp. 336-366. Washington, DC: National Academies Press.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological review, 108,* 346-369.

Di Giovanni, G., Di Matteo, V., & Esposito, E. (2008). *Serotonin-dopamine interaction: experimental evidence and therapeutic relevance*. Amsterdam: Elsevier.

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics* (4th ed.). Harlow: Longman.

Griffin, J., & Tyrrell, I. (Eds.) (2007). *An idea in practice: Using the human givens approach*. Chalvington, UK: Human Givens Publishing.

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

Keller, E. F. (2010). *The Mirage of a Space between Nature and Nurture*. Durham: Duke University Press.

Kendler, K. S. (2005). Reply to J. Joseph, Research Paradigms of Psychiatric Genetics. *American Journal of Psychiatry, 162,* 1985-1986.

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.

Khoury, M. J., Little, J., Gwinn, M., & Ioannidis, J. P. (2007). On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *International Journal of Epidemiology, 36,* 439-445.

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

Ludwig, J., Duncan, G. J., Gennetian, L. A., Katz, L. F., Kessler, R. C., Kling, J. R., & Sanbonmatsu, L. (2012). Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults. *Science, 337,* 1505-1510.

Lustig, R. H., Schmidt, L. A., & Brindis, C. D. (2012). The toxic truth about sugar. *Nature, 482,* 27–29.

Lynch, J., Davey-Smith, G., Harper, S., & Bainbridge, K. (2006). Explaining the social gradient in coronary heart disease: comparing relative and absolute risk approaches. *Journal of Epidemiology and Community Health, 60,* 436-441

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

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A., & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461,* 747-753.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics, 9,* 356-369.

Moffitt, T. E. (2005). The New Look of Behavioral Genetics in Developmental Psychopathology: Gene-Environment Interplay in Antisocial Behaviors. *Psychological Bulletin, 131,* 533-554.

Moffitt, T. E., Caspi, A., & Rutter, M. (2005). Strategy for investigating interactions between measured genes and measured environments. *Archives of General Psychiatry, 62,* 473-481.

Morris, C., Shen, A., Pierce, K., & Beckwith, J. (2007). Deconstructing Violence. *GeneWatch, 20*.

Nuffield Council on Bioethics (2002). Genetics and Human Behavior: The Ethical Context. *http://www.nuffieldbioethics.org, viewed 22 June 2007*.

Ou, S.-R. (2005). Pathways of long-term effects of an early intervention program on educational attainment: Findings from the Chicago longitudinal study. *Applied Developmental Psychology, 26,* 578-611.

Panofsky, A. (2011). Generating sociability to drive science: Patient advocacy organizations and genetics research. *Social Studies of Science, 41,* 31–57.

Plomin, R. (1999). Genetics and general cognitive ability. *Nature, 402,* C25-C29.

Plomin, R., DeFries, J. C., & Loehlin, J. C. (1977). Genotype-environment interaction correlation in analysis of human behavior. *Psychological Bulletin, 84,* 309-322.

Ridker, P. M., Buring, J. E., Rifai, N., & Cook, N. R. (2007). Development and Validation of Improved Algorithms for the Assessment of Global Cardiovascular Risk in Women: The Reynolds Risk Score. *Journal of the American Medical Association, 297,* 611-619.

Rijsdijk, F. V., & Sham, P. C. (2002). Analytic approaches to twin data using structural equation models. *Briefings In Bioinformatics, 3,* 119–133.

Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., Griem, A., Kovacs, M., Ott, J., & Merikangas, K. R. (2009). Interaction Between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression: A Meta-analysis. *Journal of the American Medical Association, 301,* 2462-2471.

Rose, G. (1985). Sick individuals and sick populations. *International Journal of Epidemiology, 14,* 32-38.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy. *Science, 277,* 918-924.

Taylor, P. J. (2004). What can we do? -- Moving debates over genetic determinism in new directions. *Science as Culture, 13,* 331-355.

Taylor, P. J. (2006). 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. (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,* 387-397.

Taylor, P. J. (2010). Three puzzles and eight gaps: What heritability studies and critical commentaries have not paid enough attention to. *Biology & Philosophy, 25,* 1-31.

Taylor, P. J. (2011). Heterogeneity and Data Analysis. Paper presented in the Statistics Department , Iowa State University, 29 September. *http://www.faculty.umb.edu/pjt/11b.pdf*, viewed 29 Sept. 2011.

Taylor, P. J. (2012). A gene-free formulation of classical quantitative genetics used to examine results and interpretations under three standard assumptions. *Acta Biotheoretica, 60,* 357-378.

Terry, S. F., Terry, P. F., Rauen, K. A., Uitto, J., & Bercovitch, L. G. (2007). Advocacy groups as research organizations: the PXE International example. *Nat Rev Genet, 8,* 157-164.

Tilley, N. (2000). Realistic Evaluation: An Overview (Presented at the Founding Conference of the Danish Evaluation Society, September 2000). *http://www.evidence-basedmanagement.com/wp-content/uploads/2011/11/nick_tilley.pdf*, viewed 1 Nov. 2013.

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science, 9,* 160-164.

Turkheimer, E. (2004). Spinach and Ice Cream: Why Social Science Is So Difficult. In L. DiLalla (Ed.), *Behavior genetics principles: Perspectives in development, personality, and psychopathology* (pp. 161-189). Washington, DC: American Psychological Association.

Turkheimer, E., & Waldron, M. (2000). Nonshared Environment: A Theoretical, Methodological, and Quantitative Review. *Psychological Bulletin, 126,* 78-108.

Visscher, P. M., Macgregor, S., Benyamin, B., Zhu, G., Gordon, S., Medland, S., Hill, W. G., Hottenga, J.-J., Willemsen, G., Boomsma, D. I., Liu, Y.-Z., Deng, H.-W., Montgomery, G. W., & Martin, N. G. (2007). Genome Partitioning of Genetic Variation for Height from 11,214 Sibling Pairs. *American Journal of Human Genetics, 81,* 1104-1110.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., & Martin, N. G. (2006). Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. *PLoS Genetics, 2,* e41.

Wasserman, D. (2004). Is There Value in Identifying Individual Genetic Predispositions to Violence? *Journal of Law, Medicine and Ethics, 32,* 24-33.

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., Perry, J. R. B., Stevens, S., Hall, A. S., Samani, N. J., Shields, B., Prokopenko, I., Farrall, M., Dominiczak, A., Diabetes Genetics Initiative, The Wellcome Trust Case Control Consortium, Johnson, T., Bergmann, S., Beckmann, J. S., Vollenweider, P., Waterworth, D. M., Mooser, V., Palmer, C. N. A., Morris, A. D., Ouwehand, W. H., Cambridge GEM Consortium, Caulfield, M., Munroe, P. B., Hattersley, A. T., McCarthy, M. I., & Frayling, T. M. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics, 40,* 575 - 583.

White, M. (2007). *Maps of Narrative Practice*. New York: Norton.

Wikipedia (2011). Seat belt legislation in the United States. *http://en.wikipedia.org/wiki/Seat_belt_legislation_in_the_United_States*.