Biol Philos
DOI 10.1007/s10539-009-9174-x

Three puzzles and eight gaps: what heritability studies and critical commentaries have not paid enough attention to

Peter Taylor

Received: 19 August 2008 / Accepted: 11 June 2009

Abstract This article examines eight "gaps" in order to clarify why the quantitative genetics methods of partitioning variation of a trait into heritability and other components has very limited power to show anything clear and useful about genetic and environmental influences, especially for human behaviors and other traits. The first two gaps should be kept open; the others should be bridged or the difficulty of doing so should be acknowledged: 1. Key terms have multiple meanings that are distinct; 2. Statistical patterns are distinct from measurable underlying factors; 3. Translation from statistical analyses to hypotheses about measurable factors is difficult; 4. Predictions based on extrapolations from existing patterns of variation may not match outcomes; 5. The partitioning of variation in human studies does not reliably estimate the intended quantities; 6. Translation from statistical analyses to hypotheses about the measurable factors is even more difficult in light of the possible heterogeneity of underlying genetic or environmental factors; 7. Many steps lie between the analysis of observed traits and interventions based on well-founded claims about the causal influence of genetic or environmental factors; 8. Explanation of variation within groups does not translate to explanation of differences among groups. At the start, I engage readers' attention with three puzzles that have not been resolved by past debates. The puzzles concern generational increases in IQ test scores, the possibility of underlying heterogeneity, and the translation of methods from selective breeding into human genetics. After discussing the gaps, I present each puzzle in a new light and point to several new puzzles that invite attention from analysts of variation in quantitative genetics and in social science more generally. The article's critical perspectives on agricultural, laboratory, and human heritability studies are intended to elicit further contributions from readers

P. Taylor (✉)
Programs in Science, Technology & Values and Critical & Creative Thinking,
University of Massachusetts, Boston, MA 02125, USA
e-mail: peter.taylor@umb.edu

Published online: 26 June 2009

across the fields of history, philosophy, sociology, and politics of biology and in the sciences.

Keywords Data analysis · Environmental · Genetic · Group differences · Heritability · Heterogeneity · IQ paradox · Measurable factors · Quantitative genetics · Selective breeding · Twin studies

Introduction

In a 2006 review article, Plomin and Asbury present quantitative genetic findings about heritability of human traits and shared versus non-shared environmental influences. They point to the "flood of molecular genetic research whose goal is to identify the specific DNA sequences responsible for genetic influence on common behavioral disorders such as mental illness and on complex behavioral dimensions such as personality" (2006, 87). They conclude, moreover, that those who resist genetic explanations of behavior are "sticking [their] heads in the sand" and that [t]here is nothing to be gained… by pretending that genetic differences do not exist" (2006, 96). The issue taken up in this article is not whether genetic differences exist, but whether the quantitative genetics methods of partitioning variation in a trait into heritability and other components[1] can show anything clear and useful about genetic and environmental influences. This article argues that by and large these methods (hereon: "heritability studies") cannot, especially in the case of human behaviors and other traits. Heritability studies are therefore not a reliable basis or point of departure for new molecular genetic research on hereditary variation among humans, for judging that the search for environmental influences and corresponding social policies is not warranted, or for sociological research that focuses on differences of experiences within families.

At first sight, this position might be pegged as another contribution on the skeptical side of the long history of scientific and policy debates around heritability, which includes the politically charged discussions of the heritability of IQ test scores (and other human traits) and genetic explanations of the differences between the mean scores for racial groups.[2] As will become clear, however, the article aims to challenge quantitative genetic researchers, commentators on nature-nurture debates, and readers from various fields who believe that this arena has been exhaustively (exhaustingly?) covered. The centerpiece of what is new here is the

[1] The "classical" quantitative genetics methods discussed in this article includes partitioning variation into heritability and other components, but not the technique of mapping of quantitative trait loci (QTL). QTL are regions of the genome containing genetic factors that influence a continuously variable trait. QTL mapping has had most success in animal and plant varieties that can be replicated and raised in controlled conditions. Reliable QTL results for human populations have been few (Majumder and Ghosh 2005), but Genome-wide association (GWA) studies may be changing this picture (Khoury et al. 2007).

[2] For key points in the debate, see the Harvard Educational Review article by psychometrician Arthur Jensen (1969), which elicited a critical response from, among others, the population geneticist, Richard Lewontin (1970a, b, 1974); see also Jensen (1970). Jencks and Phillips (1998) reviews research on the black-white test score gap and Parens (2004) provides an even-handed overview of past and potential contributions of human behavioral genetics to discussions of social importance well beyond IQ tests.

distillation of issues into eight conceptual and methodological "gaps" that need attention: some should be kept open; others should be bridged or the difficulty of doing so should be conceded (Table 1). Previous researchers and commentators have not acknowledged some or all of the gaps, have not taken the appropriate responses, or have not consistently sustained them.

The meaning and significance of the gaps and appropriate responses summarized in Table 1 are not meant to be self-evident at the outset; they have to emerge as the gaps are introduced in sequence. Nevertheless it may help to preview five features of the exposition thereby orienting readers to what lies ahead and distinguishing it from previous contributions.

First, the picture that results from laying out the gaps incorporates, but is not limited to, critical points that have been made before, namely, "genetic" does not mean unchangeable; estimates in quantitative genetic studies are contingent on the specific population and environments from which the data were collected; estimates of the degree of genetic and environmental influence are not helpful in identifying the specific factors that make up that influence; and explanations applying within groups cannot be extrapolated to apply to differences among groups. Past discussion of these points, by critics and supporters of heritability studies alike, does not address or leaves unresolved a significant puzzle: By what steps or assumptions can heritability and other quantities that summarize the variation among measurements made at one point of time shed light on the influence of underlying measurable factors involved in the processes of reproductive transmission and development of the trait? (The obvious quick response, that some authors, such as Francis Galton, associate heritability with the correlation between parents and offspring, does not

Table 1 Eight gaps in analyzing and interpreting heritability

Gap	Appropriate response
1. Key terms have multiple meanings that are distinct	Meanings need to be kept distinct, for which terminological moves can help
2. Statistical patterns are distinct from measurable underlying factors	Needs to be highlighted and kept open
3. Translation from statistical analyses to hypotheses about measurable factors is difficult	The steps and conditions to bridge this gap—or to circumvent it—warrant attention
4. Predictions based on extrapolations from existing patterns of variation may not match outcomes	Compensate for the discrepancies (if any actions depend on the predictions)
5. The partitioning of variation in human studies does not reliably estimate the intended quantities	Gap is difficult to remedy; this needs to be acknowledged
6. Translation from statistical analyses to hypotheses about the measurable factors is even more difficult in light of the possible heterogeneity of underlying genetic or environmental factors	The steps and conditions to bridge this gap—or to circumvent it—warrant attention
7. Many steps lie between the analysis of observed traits and interventions based on well-founded claims about the causal influence of genetic or environmental factors	Recognize that estimates of heritability and other fractions of the variation are of even more limited utility than gaps 3–6 indicate
8. Explanation of variation within groups does not translate to explanation of differences among groups	Recognize that this gap is firm and its implications are deep

resolve the puzzle because heritability can be derived from similarities among members of the same generation.) The four other features of the exposition can be seen in the way this puzzle was stated.

Second, data analysis is the point of entry and central concern. Readers need to be willing to put aside ideas and speculation about what genes can do or what genes can make organisms (humans included) do unless (or until) those ideas can be assessed by some reliable method of data analysis. The idea, for example, that genes might elicit matching environments sounds plausible at first, but one has to ask how an association with genes would be shown through analysis of observations of traits (see Gap 5, #g).

Third, the gaps apply to heritability studies for traits in all species, not only in humans (except, as will be obvious, for Gap 5). This said, the article predominantly makes reference to human heritability studies because the debate has been most active in that arena. Yet, when there is not an everyday human example to illustrate the point, which can be the case when examining relevant issues about data analysis, it will be helpful to consider agricultural studies of animals or plants. We can then contrast what can be known through those studies with what can be known through analyses of data from humans.

Fourth, the issues are conceptual, not technical or empirical. Data analysis may be the central concern, but no data, equations, or mathematical symbols are needed and the issues can be presented in terms accessible to non-specialists. At the same time, the sequence of gaps is meant to challenge researchers in the social and life sciences and other specialist commentators: How do they address the first gap? When this is clarified, how do they address the second gap? And so on. When the answers are not clear, the ball lies in their court, not mine. It cannot be very fruitful to participate in an extended back-and-forth on technical or empirical issues if such an exchange presupposes understandings about the gaps that are not, in fact, shared. For example, to take a look ahead, I have been asked how heterogeneity-centered problems—discussed under Gap 6—relate to other problems for heritability analyses, say, genotype–environment interaction. The question cannot be answered, however, unless I know whether the questioner's conceptualization of genotype–environment interaction keeps summaries of variation among traits distinct from measurable factors that underlie these traits—as discussed under Gap 2. Indeed, because this last distinction is one that many researchers and commentators do not make, or do not consistently maintain, there are many potential side-discussions of the kind "author X say this, but under my account that would look like…" that can be put aside until it is established whether author *X* shares the trait-underlying factor distinction. At the same time, other authors do have responses, sometimes implicit, to the gaps and these need to be acknowledged. I do so by spelling out various connections between concept, method, and application related to each gap. In that task I try to minimize technical details or include them in notes, but non-specialists might choose to skip or skim some of the discussion once the character of the gap is clear to them.

The last feature of this article evident in the puzzle about how to connect patterns at a single point of time with processes over time is the very use of puzzles. Three key puzzles are introduced in the section to follow, before the main discussion of the gaps. (The third puzzle subsumes the initial one above.) These puzzles, which have not been resolved by past debates, are meant to engage those who think either that everything significant must have been said already or that any further issues require technical expertise to appreciate and resolve. After discussion of the eight gaps, I revisit the three puzzles to present each in a new light. In the process, I point to several new puzzles that invite attention from analysts of variation in quantitative genetics and social science. As the article's very last puzzle will highlight, the critical perspectives on heritability studies developed here are intended to elicit further contributions from Biology and Philosophy readers across the fields of history, philosophy, sociology, and politics of biology, as well as in the sciences (Taylor 2008a, b). Appropriate responses to each of the eight gaps are, I believe, a precondition for meaningful discussion and progress on the puzzles. Neither well-rehearsed skepticism nor qualified acceptance of heritability studies offers much help in this regard—thus the need for this article.

A terminological preliminary: "Factor" is used throughout this article in a non-technical sense, referring simply to something whose presence or absence can be observed or whose level can be measured (a quality that is emphasized in some places in the text by adding the adjective "measurable"). For any given trait, the factors of interest are those that influence the trait's development, but the causal quality of a factor is a secondary matter. 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. Measurable environmental factors can range widely, say, from average daily intake of calories to maltreatment as a child.

Three puzzles not resolved by past debates about heritability

Puzzle 1. The two-part argument and the IQ paradox

Flynn (1994) has pointed to large gains in average IQ test score between generations (the "Flynn effect"). No environmental factor, or composite of factors, such as diet or years of education, has been shown to be associated strongly with the generational differences. At the same time, according to the current consensus, heritability of IQ test scores is high (Neisser et al. 1996, but see Turkheimer et al. 2003). Persistent large differences in average IQ test score also exist between racial groups. In parallel with the generational difference, no environmental factor, or composite of factors, has been associated strongly with the racial group average differences.[3] This has led many psychometricians and human behavioral geneticists to make a two-part argument: the high heritability of IQ test scores within racial groups coupled with a failure of environmental hypotheses to account for the group differences supports—or lends plausibility to—explanations of mean differences in terms of genetic factors (even if these factors have yet to be elucidated) (e.g., Jensen in Miele 2002, 111ff). However, the same logic would lead us towards explanations of generational

[3] There has been some success recently in using regression analysis to identify associations between environmental factors and differences between the mean test scores for racial groups (Fryer and Levitt 2004).

differences in terms of genetic factors, but the change in gene frequencies in a human population over one generation is negligible. Where is the hole in the logic of the two-part argument? Once we have an answer, does it help explain how large differences between generations in this highly heritable trait can be explained in terms of environmental factors? These questions constitute an “IQ paradox” to which Dickens and Flynn (2001) draw our attention.

Puzzle 2. The possibility of underlying heterogeneity

Claims that some human trait, say, IQ test score at age 18, shows 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 (see Fig. 2 later on in the text). If the genetic and environmental factors underlying the observed trait are heterogeneous, what can researchers do on the basis of knowing a trait’s heritability? If the method of data analysis does not allow researchers to rule out underlying heterogeneity, what can researchers do on the basis of knowing a trait’s heritability?

Puzzle 3. Confusing terms and methods borrowed from breeding

The term “heritability” connotes a connection between parent and offspring through transmission of genes, but, as mentioned above in Puzzle 2, its technical meaning and its statistical estimation involve no reference to measurable, transmissible genetic factors. Moreover, no connection between one generation and the next need be entailed—heritability for a given trait is defined as a quantity that summarizes observations made of a specific set of varieties and locations at one point of time.[4] Equivalent potential for confusion accompanies the related term, genetic variation. This might seem to refer to variation among different individuals in who possesses some measurable genetic factor(s). However, the genetic variation that enters into

[4] Heritability can be related to correlations between parents and offspring, but to do so requires models of hypothetical genes that determine the trait and a suite of assumptions (Lynch and Walsh 1998, 48–50; 142; see also Gaps 3–6).

estimation of heritability actually refers to variation across groups of related individuals (“varieties” or “genotypes”) in the average value of the trait for each group. The potential for confusion increases still further when researchers who are technically proficient in the statistical analysis of measurements on a trait interpret their results in terms of about the influence of the (unknown) measurable genetic and environmental factors that underlie the development of the trait. Are the researchers simply making unfounded interpretations? How has the potential for confusion played out over long history of application of statistical analyses of trait variation in agricultural trials and selective breeding? (This, after all, is the context in which the methods of analysis arose.) In what ways have agricultural and laboratory breeders made the translation from a pattern in data (from a specific set of individuals in a specific range of situations at one point of time) to factors in the dynamics of development and reproduction under selective mating? Do their methods of translation depend on the agricultural or laboratory context in which varieties and locations—”genotypes” and “environments”—can be controlled and replicated? What does this say about the interpretation of statistical analyses of human variation, where control and replication are difficult?

Gaining more insight concerning these three puzzles requires us, I contend, to acknowledge and make appropriate responses to the eight conceptual and methodological gaps laid out in the section to follow.

Eight gaps in analyzing and interpreting heritability

Gap 1. Key terms have multiple meanings that are distinct

Key terms have multiple meanings that are distinct and should not be conflated; in other words, this first gap is one that needs to be kept open. Terminological moves can reduce the potential for confusion. Consider, especially, the term “genetic.” Two different meanings of that term were identified in the third puzzle above. To these we can add the common language uses of the term for inherited or something that recurs in a family. To reduce the potential for confusion, let us reserve genetic as an adjective in reference to entities or factors that are transmitted from parents to offspring and whose presence or absence can, in principle, be observed. Similarly, “environmental” can also be reserved as an adjective to refer to measurable factors.

Additional terminological moves to reduce potential confusion include avoidance of the term “phenotype” to refer to traits and of “genotype” and “environment” to refer to groups of identical or related individuals and the locations or situations in which they are raised or grown. Those terms are liable to obscure the fact that analysis of variation in traits neither requires knowledge nor, on its own, produces knowledge about the genetic or environmental factors that influence the trait (“phenotype”) in the various “genotype-environment” combinations (see Gap 2). The agricultural terms “variety” and “location” can serve as suitably neutral replacements. A variety is a group of individuals whose relatedness by genealogy can be characterized, such as offspring of a given pair of parents, or a group of individuals whose mix of genetic factors can be replicated, as in an open pollinated plant variety. A location is the situation or place in which the variety is raised, such as a family or a specific experimental research station.

Note that classical quantitative genetics (see note 1) is not the analysis of genes or genetic factors, but the analysis of continuous variation in traits of humans, other animals, or plants in ways that take account of the genealogical relatedness of the varieties whose traits are observed. However, for want of a recognizable alternative, the term quantitative genetics will be preserved in this article, as will “heritability.” We should stay mindful that using those terms risks perpetuating misleading connotations (see puzzles 2 and 3). The potential for confusion is reduced, but not eliminated, by acknowledging the following gap.

Gap 2. Statistical patterns are distinct from measurable underlying factors

Estimates of heritability derive from statistical analysis of variation in traits among related and unrelated individuals. They involve no reference to measurable genetic or environmental factors involved in the development of those traits. It follows that any estimates or patterns detected by the analysis, such as the size of heritability relative to other fractions of the overall variation, must also be distinct from those factors. This is not a new point, but the distinction is not always preserved, even by critical commentators (e.g., Turkheimer 2000; see also Taylor 2006a [online appendix] on Lewontin’s much-cited agricultural thought experiments). The gap between statistical patterns and measurable underlying factors is one that needs to be highlighted and, like Gap 1, to be kept open.

To accentuate the distinction and, at the same time, make the meaning of heritability clear, consider the simplest case for analysis, 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 (Fig. 1). The total variation among replicates for any observed trait can be partitioned into four components: the variation among the variety means (v_A, v_B, etc.), the variation among the location means (l_1, l_2, etc.), the variation among means for variety-location combinations after subtracting the variety and location means for each combination (not indicated in Fig. 1), and the variation among replicates within variety-location combinations (indicated by the curly brackets). (This partitioning is called the Analysis of Variance or ANOVA.) Heritability for the trait is simply the first of the four components expressed as a fraction of the total variation.[5]

[5] Strictly this defines “across-location” heritability. An alternative, “within-location” heritability, is relevant where researchers envisage that the variety will continue to be raised in the same location. In effect, this quantity takes the heritability estimated in each location separately and averages these estimates over all locations. This estimation means that differences among the averages for the trait from one location to the next are not taken into account. Across-location heritability, which always has a smaller value than the within-location heritability, is relevant when the varieties could be raised or grown again in any of the original set of locations. Strictly it also defines “broad” heritability. “Narrow” heritability, which is used to predict change under selective breeding, is a construct that depends on assumptions about the action of hypothetical genes in the standard models of quantitative genetics (see note 10). 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 and Walsh 1998, 823). The reliability

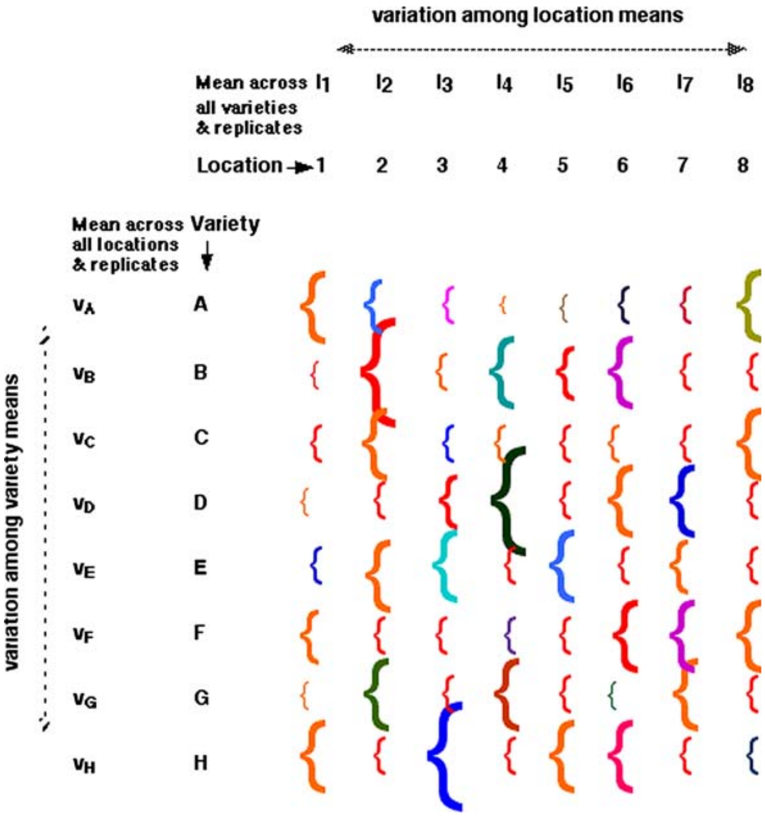
variation among location means
Mean across all varieties & replicates
l1
l2
l3
l4
l5
l6
l7
l8
Location
1
2
3
4
5
6
7
8
Mean across all locations & replicates
Variety
vA
vB
vC
vD
vE
vF
vG
vH
A
B
C
D
E
F
G
H
variation among variety means

Fig. 1 Partitioning of variation 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 variation among replicates within variety-location combinations is indicated by the size of the *curly brackets*

In the spirit of spelling out the various connections between concept, method, and application related to each gap, let us review several ways in which this distinction between statistical patterns and measurable underlying factors can be obscured:

a. The components of variation are given shorthand names—variety variance, location variance, variety-location interaction variance, and error (or residual) variance.[6] Heritability becomes the ratio of the variety (or "genetic") variance to the total variance, which can be reexpressed ambiguously as the

Footnote 5 continued
of the estimates depends on the assumptions built into the networks, such as the similarity of relatives of different degrees and the inclusion/exclusion of coefficients for variety-location interaction (see Gap 5 and note 10). When the same assumptions are used, ANOVA and path analysis estimate the same quantities.

[6] Variance is the common measure of variation. The variety variance, for example, assesses the size of the deviations of the variety means from the overall mean for the trait by averaging the square of those deviations.

"proportion… that is attributable to genetic variation among individuals" (Wikipedia 2008). This formulation can, in turn, lead to statements that obscure the distinction altogether: "Heritability analyses estimate the relative contributions of differences in genetic and non-genetic factors to the total phenotypic variance in a population" (Wikipedia 2008). Similarly, location (or "shared environmental") variance becomes interpreted as a measure of the effect of experiences or environmental factors shared by replicates in a variety-location combination (e.g., the members of a family growing up together).[7]

b. When estimating heritability from datasets where varieties have varying degrees of genealogical relatedness (e.g., human monozygotic and dizygotic twins), models are used that refer to hypothetical genes each adding a small contribution to the trait. The analyses built around those models are, however, analyses of observations of the traits, so there must always be alternative formulations that make no reference to genes (Taylor 2007, 2009; see also Gaps 4 and 5).

c. A gradient of a measurable genetic factor (or composite of factors) is assumed to run through the differences among variety means.[8] If the varieties were drawn from different species we would not assume such a gradient. Yet there is nothing in the method of data analysis that changes if we shift from multi-species situation to one in which that the varieties are all drawn from a single species—perhaps even made up of inbred lines. There is no point in such a shift at which we have to assume that an underlying gradient is present. Similarly, for a gradient of environmental factors running through differences among location means.

d. The conditionality of patterns derived from statistical analyses of traits on the specific sets of varieties and locations in the data is discounted. A variety mean does serve as a single value of the trait for each variety that best conveys its average difference from other varieties; this might seem to make the assumption in #c plausible. However, this is an average calculated over the particular range of locations in which varieties are observed, so the variety mean is not simply a property of the variety. Similarly, for location means.

e. Heritability is given causal significance without translating the quantity into the terms of measurable genetic and environmental factors. In the broad sense of causes as differences that make a difference, a difference between two variety means makes a difference among the observed traits. Heritability quantifies the average of these differences (squaring them so the direction of the difference is not important). Similarly, differences between location means can be thought of as causes. However, given the conditionality of the variety and location means

[7] In an ANOVA the components of variation are conventionally labeled "effects," a term that misleadingly connotes the influence of some causal factor. This connotation tends to be especially confusing in the case in discussion of shared versus non-shared environmental effects (Turkheimer 2000).

[8] The existence of such gradients is suggested by the symbols often used in equations, e.g., $P = G + E$, where P is the measurement of the trait ("phenotype"), G a contribution from the variety ("genotype") and E a contribution from the location ("environment," or environment plus error). However, such contributions are derived using statistical methods, such as ANOVA and path analysis, that partition the variation in traits across some specific set of varieties and locations into components. The results are thus conditional on that set of varieties and locations (see #d to follow in the text, as well as the other points under Gap 2).

on the particular sets of varieties and locations, this difference-between-means form of causality corresponds to a situation in which the only thing that can vary from the original to any "rerun" is the "noise" (i.e., unsystematic or "error" variation in an ANOVA) (Taylor 2006a).

f. Replicability of varieties and locations is imagined for human traits. In agriculture, a variety refers to a group of individuals whose relatedness by genealogy can be characterized, such as offspring from repeatable mating of a certain sire and dam, and 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, but when methods are used to analyze variation across "genotypes" and "environments"—here: across varieties and locations—, this implies a thought-experiment in which replication of varieties and locations is indeed possible. Without this thought-experiment, it is not obvious why we would analyze variation among human genotypes/varieties and environments/locations. (Recall that the definitions of variety and location [see Gap 1] do not assume that researchers know or can specify the genetic or environmental factors that influence the trait for any variety-location combination.) Eventually the analysis of variation in a trait may help researchers identify the measurable genetic or environmental factors that influence that trait and lead to data on those factors being brought into the analysis (see Gap 3 and Table 2 in Gap 7), but heritability studies can, and usually do, proceed from data about the trait alone.

g. The idea of "genotype-environment correlation" (or "gene-environment correlation"),[9] which has a well-defined technical meaning in quantitative genetics (Jacquard 1983; Lynch and Walsh 1998, 47), is used in more colloquial discussions that assume that genes can elicit particular environmental factors or vice versa (Plomin et al. 1977; Sesardic 2005). For example, "other people react to children with genotypically higher IQ by… imposing on them more intellectually demanding conversation and otherwise challenging their ability even further" (Sesardic 2005, 93). Now, it is easy to identify processes for humans through which people respond to observed traits or through which children's traits lead them to seek out certain environmental factors, but it is more difficult to envisage mechanisms through which people could match environmental factors to the genetic factors (Taylor 2006b). Evidence is needed

[9] In the terminology of this article, genotype-environment correlation is "variety-location correlation" (or covariance), which is distinct from "variety-location-interaction variance." The correlation is readily explained by referring to the general case of an agricultural evaluation trial, in which the means of varieties, locations, and variety-location combinations can be estimated. For the full data set, the usual method of estimation ensures that the variety and location means (averaged across, respectively, all locations and all varieties) are uncorrelated. Within a subset of the full data, however, those same means may be positively or negatively correlated—this is the variety-location correlation. In human studies, varieties are raised in at most two locations (identical twins raised apart), so the variety and location means across all locations and varieties are unknown; variety-location correlation is thus difficult to estimate (Jacquard 1983), and, if estimated (e.g., Otto et al. 1995), requires many additional assumptions (Lynch and Walsh 1998, 142).

to support any assumption of a matching mechanism that does not involve traits that people can observe. Even if that evidence could be secured, researchers would still need methods of data analysis that could discriminate among competing models of the association between traits and genetic and environmental factors (e.g., reactive versus active; Plomin et al. 1977). Analogous reservations apply to species other than humans whenever proposed explanations of variation in a given trait refer to genes eliciting environmental factors.

If the gap between statistical patterns in traits and underlying measurable factors is highlighted, not obscured, we can ask how it can be bridged, which leads to the third gap.

Gap 3. Translation from statistical analyses to hypotheses about measurable factors is difficult

It is difficult to translate from statistical analyses of data on traits to hypotheses about the measurable genetic or environmental factors involved in the development of the traits. The steps and conditions to bridge this gap—or to circumvent it—warrant attention. To this end, consider three directions that researchers might pursue:

a. 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). Discussion of this direction of research lies beyond the scope of an article on heritability studies.
b. 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). This assumes that a gradient of a measurable genetic factor (or composite of factors) runs through the differences among variety means (contra Gap 2 #c, above). There may well be certain traits for which such a gradient exists. These might be worth finding even if, in the course of doing so, researchers end up conducting fruitless investigations of other heritable traits for which it turns out there is no such gradient. (This search for the traits with a genetic gradient is not for traits, such as presence of extra digits (or polydactyly), that are largely determined by genes at a single locus whose effect is more or less independent of the individuals' upbringing. Such "high penetrance major genes" can be detected through examination of family trees; heritability analysis need not be involved.) In pursuing this direction, there is a risk that the proportion of fruitful investigations will be low compared to those confounded by the lack of an underlying gradient. In any case, additional knowledge, not derived from the statistical analysis, is always required.
c. Restrict attention to variation within a set of relatives. Even if the underlying factors are not yet known, 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, given that this scenario assumes that the timing of getting the condition differs from the first twin to the second, researchers can also ask: What factors influence the timing? How changeable are these? How much reduction in risk comes from changing them? To address these issues researchers have to identify the genetic and environmental factors involved in the development of the trait and to secure larger sample sizes than any single set of relatives allows. The question then arises whether the results can be extrapolated from one set of relatives to others. This issue is an empirical one; there is a risk, as before, that the proportion of fruitful investigations will be low compared to those confounded by factors not extrapolating well from the initial set of relatives.

It should be noted that the last two directions seem to rest on an intuition that is not reliable, namely, high heritability indicates that measurable genetic factors have more influence on variation in the trait than measurable environmental factors (even though the specific factors are unknown); similarly for the ratio of variation among location means ("shared environmental effects") and other fractions of variance (Turkheimer 2000). However, support is lacking, even in the ideal case of agricultural evaluation trials, for this intuition. To gather such support would require us, in the absence of prior knowledge of how genetic and environmental factors influence the development of the trait, to take a number of steps: consider a range of models of factors influencing development; justify the assumptions on which the models are built; calculate heritability for a representative range of values of each model's parameters; and discover associations between the heritabilities and the corresponding genetic and environmental factors that are robust across models (Taylor 2006a, Sect. 4.2). The intuition arises without these steps having been taken; in short, it is problematic.

Now, to speak of considering a range of models is to imply that alternatives exist to the standard models presented in quantitative genetic texts. The standard models are constructed through a sequence of steps and assumptions beginning with a trait governed by a pair of alleles of a single hypothetical gene (i.e., at a single locus) and raised in a single location (Falconer and Mackay 1996; Lynch and Walsh 1998).[10]

[10] The first step in the construction of the standard models of quantitative genetics is to consider the case of a trait governed by a pair of alleles of a single gene (i.e., at a single locus) and where all individuals are raised in a single location. In that location, the presence or level of such a trait depends only on whether the individual has two copies of one allele (i.e., is "homozygote" for that allele), two of the other, or one of each ("heterozygote"). For example, phenylketonuria (PKU) in humans is associated with having two copies of a non-functioning allele for the enzyme phenylalanine hydroxylase (PAH). The development of such individuals is extremely impaired by phenylalanine at the level present in normal diets. In this "normal-diet" location relatives will resemble each other more than unrelated individuals because if, say, a twin has PKU, both parents have at least one copy of the non-functioning PAH allele so the other twin is more likely to have two non-functioning PAH alleles than an unrelated individual (i.e., one chosen at random from the population). This seems straightforward, but few traits are dictated only by alleles at a single locus. The standard models of quantitative genetics envisage the influence of alleles at many loci adding up to shape the traits to be analyzed, allowing also for the effect of one allele to eclipse that of the other ("dominance") at any given locus and some degree of interaction ("epistasis") among alleles at different loci. Next, the models allow for some noise (from measurement error or unsystematic variation among the replicates of the variety). Finally, to allow for the variety to be raised in a number of locations, the standard models of quantitative genetics incorporate a term for variance across locations of the mean

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 2006a [online appendix] for a more complex example). Alternatives to the standard quantitative genetic models and the assumptions built into them warrant more attention (see Gap 5).[11]

Employing models of hypothetical genes and giving advice to relatives are examples of trying to circumvent Gap 3. Other ways of doing something in the absence of knowledge of underlying factors are addressed under Gap 4.

Gap 4. Predictions based on extrapolations from existing patterns of variation may not match outcomes

It is possible, even without knowledge of the underlying measurable factors involved in the processes of reproductive transmission and development of the trait, to extrapolate from existing patterns of variation (as captured, for example, by heritability estimates), but we can expect that imperfect predictions of the actual outcomes. If any actions depend on such predictions, we need to be able to compensate for the discrepancies.

If we do not have knowledge of the measurable underlying factors we can focus on heritability as a fraction of the variation among measurements. This is evident in agricultural and laboratory breeding, where selection of parents of the next generation can proceed on the basis of observed traits and without knowledge of the underlying factors. High (or low) heritability is used to indicate whether (or not) to expect selective breeding to produce the desired improvement in the average value of the trait across the population. (Selective breeding, is not, of course, an acceptable option for humans.) Moreover, simple models of multiple, hypothetical genes underlying the traits (note 10) allow breeders to make more refined

Footnote 10 continued
value of the trait in each location (i.e., "location variance" or "shared environmental variance"). Application of the models to the analysis of data from related and unrelated individuals, such as human twins, requires additional assumptions for which plausible alternatives exist (Taylor 2007, 2009 and Gap 5). Most notably, it is conventional to assume that, all other things being equal, fraternal twins are half as similar as identical twins because fraternal twins share half the genes that vary in the species or population, while monozygotic twins share them all (e.g., Kendler and Prescott 2006, 42). However, it is straightforward to invent plausible models of the contributions of multiple genes to a trait that do not result in this ratio (see example to follow in the text). Ratios other than .5 should not be surprising because measures of similarity (such as, "intraclass correlations") are based on observed traits and, as such, are not directly given by the number of shared genes involved in the development of those traits (Taylor 2007, 2009).

[11] Genealogical relatedness can be taken into account without the models of hypothetical multiple genes and additional assumptions sketched in note 10. Analyses without those assumptions may, however, require data collected under special conditions, such as, replicates of a variety raised in separate, randomly chosen locations (e.g., twins raised apart) and replicates from different varieties raised in the same location (e.g., unrelated individuals raised in the same family) (Taylor 2007, 2009). Whether these special conditions obtain for humans in any actual cases remains under debate (Richardson and Norgate 2005).

predictions of the advance under different breeding plans (e.g., mating of half-sibs versus mating of cousins), which 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 (Gap 2, #d), it should 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, in agricultural and laboratory settings, researchers can replicate varieties and locations (see Gap 2, #f) and they can select among varieties for the next generation on the assumption that the environmental factors will remain unchanged. With such control the predictions can be made with more confidence. Moreover, if the actual advance under selective breeding is less than predicted, breeders can always compensate for discrepancies: they discard the undesired offspring, breed the desired ones, and continue.

Neither control of conditions nor selective breeding is acceptable or achievable for humans. Nevertheless, when models of genes (albeit hypothetical ones) are fitted to the observed variation in the trait, they appear to show the relative degree of influence on the trait of genetic and other factors, even if the identity of those factors remains unknown. (In other words, Gaps 1 and 2 seem not to matter, and the intuition behind the two directions under Gap 3 seems to be justified.) By extension, high heritability leads us to expect only small changes following shifts in environmental factors. This last idea would seem to apply to humans as well, that is, implications can be drawn from patterns in variation even outside of the realm of selective breeding. Yet, in the absence of evidence for the assumptions of the gene-based models or, at least, without comparing a range of alternative models (see discussion under Gap 3), the reliability of the predictions and claims about the relative degree of influence is uncertain. In other words, it is not so easy to sidestep the difficulty of translating from statistical analyses of data on traits to hypotheses about the measurable genetic or environmental factors involved in the development of the traits (Gap 3).

The situation for human studies is further limited by the following gap.

Gap 5. The partitioning of variation in human studies does not reliably estimate the intended quantities

The discussion under Gaps 3 and 4 implies that for humans, where selective breeding and control of conditions are not acceptable or achievable, high heritability for a trait is relevant only for directions #b and c under Gap 3 and rests on an intuition about relative degree of influence by genetic and environmental factors that is difficult to justify. Even then, the conventional estimation of heritability (and of other fractions of the trait's variation) is not reliable because it depends on certain fundamental assumptions listed in the next paragraph. The results and interpretation change profoundly if these assumptions are not made (Taylor 2009). The gap between the actual estimation and reliable results is difficult to remedy, but needs to be acknowledged. To put an exclamation point on this, Taylor (2007) shows that a simple adjustment for assumption #c below results in most human heritability estimates falling to values below the fractions for variance among-location-averages (“shared environment effects”).

Assumptions in the standard quantitative genetic analysis of human twins include: (a) the analysis requires models of genes with simple Mendelian inheritance and direct contributions to the trait;[12] (b) All other things being equal, fraternal or dizygotic twins are half as similar as identical or monozygotic twins;[13] (c) Variance among variety-location-combination means (“genotype–environment-interaction”) can be discounted or can be incorporated into the heritability estimates;[14] (d) Residual variance is a within-family environmental contribution (“non-shared environmental effects”);[15] (e) When similarity among a set of close relatives (such as twin pairs) is associated with similarity of (yet-to-be-identified and measured) genes or genetic factors, those factors are the same from one set of relatives to the next. These assumptions and alternative analyses without those assumptions are discussed in Taylor (2007, 2009). The last assumption, which was raised by puzzle 2, leads to the next gap.

Gap 6. Translation from statistical analyses to hypotheses about the measurable factors is even more difficult in light of the possible heterogeneity of underlying genetic or environmental factors

Puzzle 2 noted that, even when similarity among a set of close relatives (such as twin pairs) is associated with similarity of (yet-to-be-identified and measured) genetic or environmental factors underlying the development of the trait, it is

[12] This assumption runs through all quantitative genetics, not only in human studies; see note 10. Models based on this assumption can be fitted to observations (and the fit of different models compared with each other), but support has not been shown for the models’ assumptions independent of that fit. This practice runs counter to the idea in philosophy of science that confirmation of a model requires both aspects (Taylor 2005, 35ff).

[13] See next to last paragraph under Gap 3 for an example of a model in which DZ twins are almost always more than half as similar as DZ twins. (Simulation available from author on request.)

[14] The standard methods of human quantitative genetics cannot demonstrate that the variety-location-interaction variance is negligible, so the resulting estimates of heritability, which incorporate the interaction fraction, may be systematically inflated (Taylor 2007). It is important to be able to separate out the interaction fraction because the claim that the effect of family members growing up in the same location (family) is of small importance (e.g., Turkheimer 2000) requires showing not only that the location variance is a small component of the total variance, but so also is the variety-location-interaction variance. In agricultural plant evaluation trials, variety-location-interaction variance is typically as large a fraction of the total as variety variance (heritability), but it is not known whether this is also the case for animal or human populations observed in a typical range of locations. To estimate interaction variance separately from heritability requires data collected under the special conditions mentioned in note 11.

[15] Residual variance is a “non-shared” component in the sense of not being variation among location averages, variety averages, or additional contributions from averages for variety-location combinations. However, this component should not be labeled an environmental component given the two sources of residual variance, namely: measurement error (after subtracting any systematic differences in measurement error across varieties or across locations); and differences among replicates within variety-location combinations in the ways that the (unknown) genetic and environmental factors possessed or experienced by the replicates influence the trait. Greater accuracy in measurement can reduce the first source of residual variation. The second source can be reduced if replicates of a variety are more uniform and positioned randomly within the location. The neutral terms “noise” or “unsystematic” for this fraction of the variation is more appropriate than “non-shared environmental effects.”

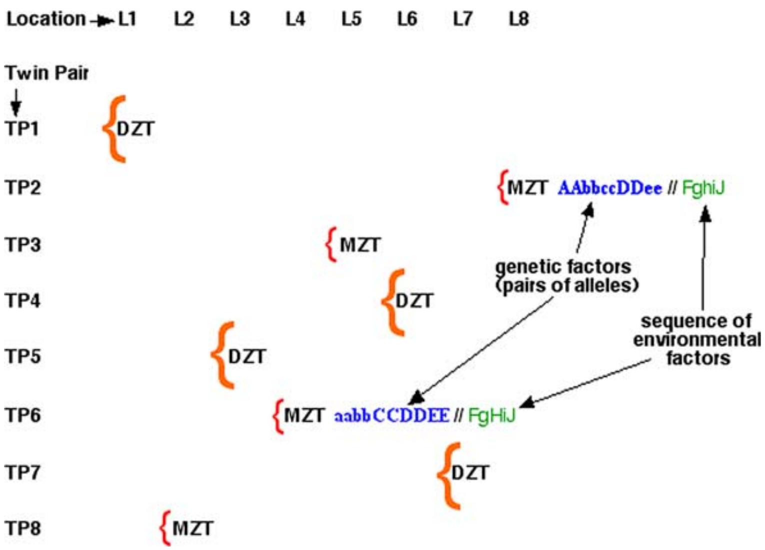
Location → L1 L2 L3 L4 L5 L6 L7 L8
Twin Pair
TP1
TP2
TP3
TP4
TP5
TP6
TP7
TP8
DZT
MZT AAbbccDDee // FghiJ
MZT
DZT
DZT
MZT aabbCCDDEE // FgHiJ
DZT
MZT
genetic factors (pairs of alleles)
sequence of environmental factors

Fig. 2 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*). In contrast to the agricultural evaluation trial, the replicates of any variety (twins) are raised in only one location per variety

possible (contra assumption #e under Gap 5) that those factors are not the same from one set of relatives to the next (Fig. 2).[16] The questions then, as raised in puzzle 2, are: what can researchers do on the basis of knowing a trait's heritability if the genetic and environmental factors underlying the observed trait are heterogeneous, or if the method of data analysis does not allow researchers to rule out the possibility of underlying heterogeneity? What steps and conditions are needed for researchers to bridge or circumvent the knowledge that underlying factors may be heterogeneous? Difficulties in translation over and above—and more serious than—those discussed under Gaps 3 and 4 are evident in the following six directions that researchers might pursue.

The first four directions parallel those under Gaps 3 and 4:

a. 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. Again, discussion of this direction of research lies beyond

[16] This sense of "heterogeneity" should be distinguished from three other uses of the term 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 the underlying heterogeneity referred to here, 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.

the scope of this article (but see the new puzzle that emerges in the concluding section's discussion of "Puzzle 2").

b. Use high heritability (perhaps adjusted downwards; see Gap 5) to guide molecular research to identify the specific genetic factors involved. There may be traits for which the underlying factors are not heterogeneous. These might be worth finding even if researchers do not know in advance the proportion of fruitful investigations compared with those confounded by the underlying heterogeneity. Again, the search is not for high penetrance major genes. Researchers need to find 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 within some defined population.

c. Restrict attention to within a set of relatives. The same thinking as given under Gap 3, #c, means that the differential timing of getting the condition becomes an issue. Again, researchers have to identify the genetic and environmental factors involved in the development of the trait and to employ larger sample sizes than any single set of relatives. The question of what to do about the possibility of underlying heterogeneity thus persists.

d. Put aside the search for measurable factors. Instead, focus on heritability as a fraction of the variation among measurements, a focus that is useful in agricultural and laboratory breeding. If the actual advance under selective breeding is less than predicted, one source of the discrepancy might be the underlying heterogeneity of genetic factors and their reassortment through mating. Again, this matters little because breeders can always compensate for discrepancies: they discard the undesired offspring, breed the desired ones, and continue. Selective breeding is not an acceptable option for humans. The intuition discussed under Gap 3, that genetic factors have a larger influence than environmental factors for high heritability traits, is even more problematic when researchers consider models that allow for heterogeneous factors to underlie the trait.

Researchers can also address the possibility of underlying heterogeneity in two ways that were not discussed under Gaps 3 and 4:

e. Reduce the possibility of underlying heterogeneity by restricting the range of varieties or locations. Agricultural researchers can reduce the possibility of underlying heterogeneity by restricting the range of locations in which a variety is raised or grown. They can also control environmental conditions, such as, for animals, the regimes of feeding and husbandry or, for plants, the application of fertilizer and irrigated water. Agricultural breeders can also produce inbred lines and thereby eliminate the heterogeneity of genetic factors that exists within outbred varieties. However, to envisage taking action on the basis of research conducted under restrictive conditions is to presume that the restrictive conditions can be replicated. This presumption is most apparent when plant breeders recommend varieties to be grown only in defined regions and under prescribed techniques of cultivation, or when animal breeders specify the optimal feeding and husbandry for each variety. In the study of human traits, however, it is not feasible to control the full range of relevant environmental

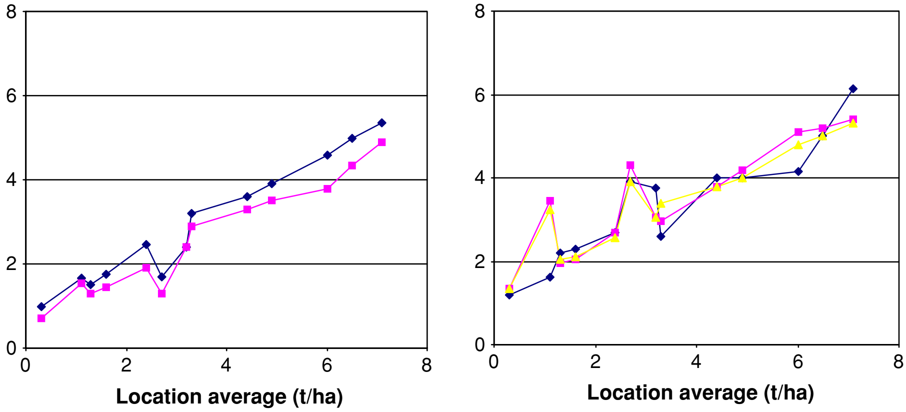

Fig. 3 Yields for 5 groups of wheat varieties grown in 13 groups of locations (from Byth et al. 1976). The *x*-axis is the average over all varieties for that location. The individual varieties (not shown) were clustered into these 5 groups by similarity of response across locations. These groups were then clustered into two groups as shown in the two *plots*

conditions or to breed for genetic uniformity. It may be possible to restrict the locations included in a human study (e.g., to include only families of low socioeconomic status; Turkheimer et al. 2003). The heritability estimates would be reliable (again perhaps adjusted downwards; see Gap 5) to the extent that these restrictions were replicated in subsequent research or policy. That is, the research could be applied even though the environmental factors underlying those locations had not been identified.

f. Reduce the possibility of underlying heterogeneity by grouping varieties that are similar in responses across locations. In agricultural trials, where a number of varieties or animals or plants can be raised or grown in multiple replicates in many locations, varieties can be grouped by similarity in responses across all locations (using techniques of cluster analysis; Byth et al. 1976). (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 (Fig. 3). 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. If the underlying factors are assumed to be homogeneous within each of the groups, researchers can hypothesize about the group averages—about what factors in the locations elicited basically the same response from varieties in a particular variety group that distinguishes them from other groups. (Again, it should be noted that 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 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, and to use that knowledge in subsequent research or in planting recommendations (Taylor 2006a). In general, any hypotheses that researchers generate need to be validated before making proposals for action. If hypotheses are not forthcoming or they fail to be validated, it is possible to shift back to approach #d above, this time making use of one of the groups, not the whole data set, in decisions about selection and breeding.

What role does heritability play in research that groups varieties to help reduce underlying heterogeneity? Clustering ensures that the variation among the means for variety groups 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 within that group is the best across locations. In other words, heritability within the variety group is not so important. This is also the case for among-group heritability. On the other hand, even if variation among variety group means is smaller than variation among location means, researchers can still hypothesize about the group averages.

It becomes more difficult to distinguish groups of varieties by similarity of responses across locations when varieties are observed in only a few locations or when the locations are not the same from one variety to the next. Clustering becomes infeasible when analyzing measurements from studies of human twins because such studies have only two replicates (twins) in one or at most two locations (families). In other words, grouping varieties in this way is not a direction by which research on human variation can bridge or circumvent Gap 6.

Gap 7. Many steps lie between the analysis of observed traits and interventions based on well-founded claims about the causal influence of genetic or environmental factors

The utility of estimates of heritability and other fractions of the variation is even more limited than conveyed in the discussion thus far. Suppose that hypotheses have been derived about measurable factors (even if, as Gaps 3–6 suggest is the case for human traits, statistical analyses, such as ANOVA, provide little guidance). The next step for researchers is then to use regression analysis and conduct experimental trials to investigate associations with measurable factors. In both cases, conditionality (Gap 2, #d) applies, now extended to conditionality on the set of factors measured as well. By choosing significant factors from 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 had been 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 2.

Table 2 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 | v	ANOVA + Cluster analysis + knowledge from sources outside data | v hypotheses about measurable factors	ANOVA (& path analysis) not helpful in generating hypotheses about measurable factors[a] (hypotheses about factors drawn from other sources)
Observed associations with measurable factors	Factors significant according to regression analysis | v	Factors significant according to regression analysis | v
| v	factors for testing through experimental trials	(Same as on the left[b])
Experiments that vary measurable factors | v	Significant factors | v insights for investigation of dynamics of development	(Rare) ?
Factors observed over the course of development [rarely-realized ideal] | v	Significant factors in development under controlled research conditions | 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 the structure of dynamics

Gap 8. Explanation of variation within groups does not translate to explanation of differences among groups

It is widely acknowledged that accounting for "within-group" variation does not explain "between-group" differences. Yet, in the contentious debates about differences among the averages for racial and other groups (see note 2), high heritability often seems to confer plausibility to hypotheses about the role of genetic factors in explaining those differences (see Puzzle 1). The following considerations counter that plausibility; the within-group/between-group gap is firm and its implications are deep.

a. Statistical analysis of variation among traits and heritability estimates provide little or no guidance in hypothesizing about measurable factors behind observations of human traits within one group of varieties (Gaps 3–6), so they can provide little or no guidance about measurable factors associated with differences between two groups. This point alone discounts the relevance of heritability to discussions of group differences (Taylor 2006b). Referring back to Puzzle 1, the two-part argument about IQ test scores dissolves into a symmetry: There may be no environmental factor associated strongly with the group or generational average differences, but there is no such genetic factor either.[17] These average differences still need explanation, but high heritability (if it is truly high; see Gap 5) poses no paradox.

b. Consider the earlier case of agricultural evaluation trials in which the observations of the trait are used to cluster varieties by similar responses across locations (Gap 6, #f). By minimizing the possibility of underlying 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, responses that distinguish one group from another. Figure 4 conveys schematically the relationship between factors and patterns in data that underlies such hypothesizing. Notice that hypothesizing involves both genetic and environmental factors and that insights about one group in one location are related through contrasts to insights about other groups in that location and about the same group in another location. (The plant example under Gap 6, #f 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.)

However, if varieties are not grouped by similarity of responses across locations, the possibility of heterogeneity of underlying factors should be considered (Gap 6). The relationship between factors and patterns in data that underlies any hypothesizing in this situation may be very difficult to disentangle (Fig. 5). To undertake such hypothesizing is akin to hypothesizing about the difference between group averages as if the spread (variance) of values (in ANOVA: within-variety-group variety means) were noise (Fig. 6).

[17] Associations have not been found in the few instances where genetic factors have been examined, e.g., genes that mark degree of African ancestry (see summary in Nisbett 1998, 89–90).

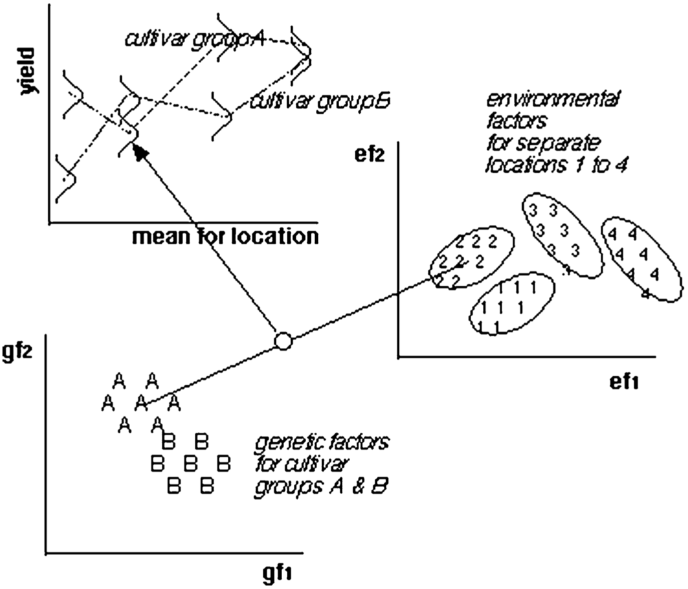
yield
cultivar groupA
cultivar groupB
mean for location
environmental factors for separate locations 1 to 4
ef2
ef1
gf2
gf1
genetic factors for cultivar groups A & B
A
B
1
2
3
4

Fig. 4 Generation of hypotheses about genetic and environmental factors underlying patterns in data when those factors are homogeneous within groups

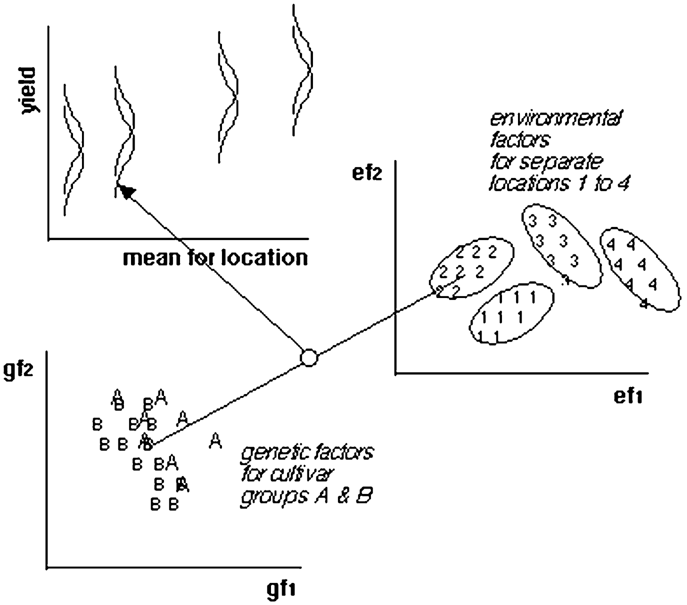
yield
mean for location
environmental factors for separate locations 1 to 4
ef2
ef1
gf2
gf1
genetic factors for cultivar groups A & B
A
B
1
2
3
4

Fig. 5 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 Fig. 4

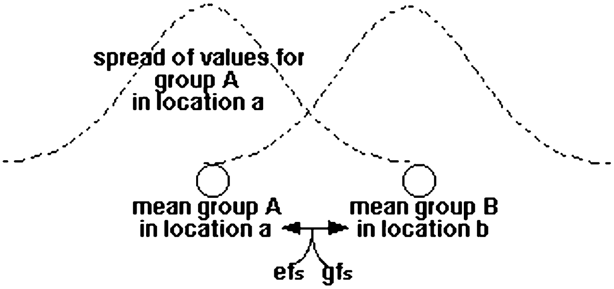

Fig. 6 Generation of hypotheses about genetic and environmental factors underlying differences among groups when the spread of values within groups is not taken into account (*gfs* and *efs* refer to measurable genetic and environmental factors)

If the possibility of heterogeneity of underlying factors has not been minimized, then, by extension, it must 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 varieties in one group are limited to a subset of 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 Fig. 5 would have to serve as the basis for any hypothesizing about the genetic and environmental factors. This is unlike the example of plants susceptible or resistant to rust in damp locations, for it means that contrasts among varieties within a location are not available to guide (or constrain) the researchers' hypothesis generation and subsequent inquiries. The same is true for contrasts for a single group of varieties across locations. The limitations just described can also be expressed in terms of "nested" analyses, to follow.

c. Lindman's (1992) textbook illustrates a cautionary note about nested ANOVA (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 difference among location (school) means "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 [average location or school difference], 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. In any case, nested ANOVA cannot help researchers hypothesize about the difference in the average scores from one school to the next when the teachers are replicated (in their students' test scores) only within schools. To translate this into the concerns of this article, nested ANOVA cannot help researchers hypothesize about the difference in the average scores from one location (or subset of locations) to the next when the varieties are replicated only within locations (or subsets of locations). Researchers might just as well conduct a separate ANOVA for each subset of varieties and locations—or, in the context of racial differences, for each combination of group of individuals and experience of membership in different racial groups.[18] (To respect this methodological limitation of nested data analysis is not to make the claim that disjunct kinds of causes must be operating in the different racial groups.)[19]

d. Another consideration relevant to the within-group/between-group gap is the relationship between lack of attention to the possibility of heterogeneity of underlying factors (see Gap 6) 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, perpetuates the typological worldview in referring to "the observed differences between schools" when he means the observed differences between averages for schools. 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 dispute the label typological, saying that the implicit variation is well appreciated and 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 variation 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 help keep in view 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. The alternatives are more likely to steer us from away thinking that there is something essentially of each group that leads to differences in their averages. Moreover, we might ask just who is empowered to do something as a result of analysis of differences in group averages (or who is given license not to have to do anything)? These larger sociological questions are touched on in the concluding discussion that follows.

[18] The limitations of nested analysis for comparing groups can be overcome using Multi-level or Hierarchical regression analysis when data are available on measurable factors within groups or at the group level (Gelman and Hill 2007). Such data are not available in conventional heritability studies, but, if they were, interpretations of the resulting regression coefficients might still be confounded by heterogeneity in the underlying factors.

[19] Although Lindman's note and the preceding discussion and diagrams in the text center on ANOVA, the points about the possible heterogeneity of underlying factors, about membership in different groups being analyzed as different locations, and about the limitations of nested analysis might also apply to drawing hypotheses and insights from regression analysis and experimental trials (Gap 7; see Table 2). This idea and its implications warrant further inquiry.

Resolving and raising puzzles

The eight conceptual-methodological gaps (Table 1), taken together, mean that the classical methods of quantitative genetics can show very little that is reliable and useful about the genetic and environmental factors underlying traits, especially human behaviors and other traits. Even in agricultural and laboratory breeding, where varieties and locations—"genotypes" and "environments"—can be controlled and replicated, the translation from statistical analyses to hypotheses about measurable factors is difficult. The translation is easier to achieve when the range of varieties or locations or varieties that are similar in responses across locations are grouped (see Gap 6, #e and f), but, even then, knowledge from sources other than the data analysis is always needed to help researchers generate any hypotheses. Hypotheses, in turn, are only one step toward interventions based on well-founded claims about the causal influence of genetic or environmental factors (see Gap 7). Agricultural and laboratory breeders have a way to circumvent the difficulty of exposing the genetic and environmental factors underlying traits. The standard quantitative genetic models, which refer to hypothetical genes each adding a small contribution to the trait, can be used to make predictions of advance under selective breeding. Then, even if the assumptions behind the models are impossible to verify or unreliable (see Gaps 3–5 and notes 10–12), breeders can always compensate for discrepancies: they discard the undesired offspring, breed the desired ones, and continue. This option is unavailable to researchers studying human variation.

In summary, the resolution to Puzzle 3 is that the methods of quantitative genetics do not translate well from agricultural and laboratory breeding to statistical analyses of human variation. This points to a new puzzle for historians, sociologists, and philosophers of biology: How were restrictive conditions—the control and replicability that can be achieved in agricultural and laboratory breeding—discounted or forgotten when methods of heritability estimation were adapted to human genetics? (Taylor 2008b).

The summary of the previous paragraph also speaks to Puzzle 2 about what researchers can do without knowing whether or not the genetic or environmental factors underlying traits are heterogeneous. In agricultural and laboratory trials, researchers can pursue approaches that make that possibility less disturbing—they have the ability to replicate varieties and locations; to control the variability in those varieties and locations; to reduce heterogeneity through grouping varieties by similarity of responses across locations; and to compensate for shortcomings in predictions of advances under selective breeding. In human studies, however, high heritability may be used primarily as a guide to decide whether to pursue molecular research to identify the specific genetic factors involved for the trait (see Gap 6, approach #a) or not to search for environmental influences and promote social policies based on them. Following this guide in molecular or social research is likely to be fruitful only if three conditions apply: the heritability is truly high (but see Gap 5); a gradient of a measurable genetic factor (or composite of factors) runs through the differences among variety means (but see Gap 2, #c); and the underlying factors are not heterogeneous. When these conditions cannot be assured, it would be prudent for researchers not to place too much stock in heritability as a guide (or the problematic intuition that may underlie it; see Gap 3). Instead, researchers could explore methods that attempt to identify the specific, measurable genetic and environmental factors without reference to the trait's heritability or the other fractions of the total variance (see #a under Gaps 3 and 6).

In light of this circumscribed answer to puzzle 2, it would be interesting to revisit studies that interpret heritability and "genetic variance" as measuring the contribution of the genetic factors in influencing the process through which the trait develops. What can be learned from the data and analyses if we highlight the gap between hypotheses about the underlying genetic and environmental factors and the statistical analysis of measurements on a trait for a specific set of individuals in a specific range of situations (Gap 2)? It would also be interesting to extend the concern about underlying heterogeneity (Gap 6) to human sciences more generally: What shortcomings of current methods of data analysis and interpretation might emerge if researchers question the methodological assumption that, when similar responses of different individual types are observed, similar conjunctions of genetic and environmental factors (or, in epidemiology, risk and protective factors) have been involved in producing those responses?

A final follow-up to puzzle 2 stems from observing that, although some prominent geneticists have noted that heritability estimates are not helpful in identifying specific genetic factors (e.g., Rutter 2002, 4), the possible heterogeneity of factors that underlie patterns in observed traits has not been recognized as a significant issue, either by quantitative geneticists or by critical commentators on heritability research (e.g., Downes 2004 and references therein; but see Taylor 2006a, b). What conceptual and sociological considerations have obscured the issue over decades of debate?

Having revisited puzzles 3 and 2, this brings us to puzzle 1. This puzzle is not resolved here, but the paradox is dissolved. Recall the two-part argument that the strong role of genetic factors within a group, coupled with the failure of environmental factors to explain differences among the average IQ test score for racial groups, lends plausibility to the idea that genetic factors are needed in an explanation. This plausibility depends, however, on interpreting high heritability within groups as evidence that genetic factors are more significant than environmental factors. Once we keep statistical patterns distinct from measurable underlying factors (Gaps 1 and 2) and acknowledge the difficulties in translating from the patterns to hypotheses about the factors (Gaps 3 and 6), this interpretation of heritability becomes problematic. The two-part argument can be put to the side and there is no paradox. (This can be said even without invoking the unreliability of heritability estimates for humans [Gap 5]; the many steps between hypotheses and intervention based on well-founded claims [Gap 7]; and the within-group/among-group disconnect [Gap 8].)

Yet, the large average differences between groups and between generations on IQ test scores remain to be explained. The puzzle becomes: how do we expose the mix of genetic and environmental factors associated with those differences. Dickens and Flynn (2001) propose "reciprocal causation" models, which involve two key features: a matching of environments to differences that may initially be small (e.g., children who show an earlier interest in reading will be more likely to be given books and receive encouragement for their reading and book-learning); and a social multiplier through which society's average level for the attribute in question influences the environment of the individual (e.g., if people grow up and are educated with others who, on average, have higher IQ test scores, this will stimulate their own development).

However, once it is recognized that the potency of social multipliers depends on different groups' capacities to capitalize on historical changes in society, there is no reason to assume that the multipliers apply uniformly across individuals despite their differences in age, gender, geographical location, culture, and so on, or even that the multipliers move different individuals in the same direction but at different speeds. To adapt a basketball analogy that Dickens and Flynn use to illustrate their reciprocal causation model, the onset of TV coverage of basketball acted as a social multiplier by eliciting greater participation in basketball, but, at the same time, it elicited more "couch potato" spectatorship. In more general terms, if researchers envisage developmental pathways whose heterogeneous components differ among individuals at any given point of time, the challenge is to develop methods to collect and analyze the data so as to discriminate among possible models.

The possibility that heterogeneous pathways underlie the variation in any given human trait leads, in turn, to a puzzle for socially engaged researchers (Taylor 2006b; Flynn 2000). If genetic factors are to be included in the models of development of traits, there are good methodological reasons for not categorizing individuals according to racial group membership (e.g., this grouping is not based on clustering across a range of locations [see #f under Gap 6]; and no measurable genetic factor admits a clean subdivision between whites and African-Americans; see also Taylor 2008b). On the other hand, racial group membership continues to bring disadvantages to African-American individuals and, reciprocally, to bring benefits to white individuals (Flynn 2000, 142ff)—moderated somewhat, but in a decreasing set of circumstances, by affirmative action for African-Americans. So, while exposing the best way to ameliorate the effects of racial group membership for any individual may depend on having empirical models of the heterogeneous pathways of development, all those pathways may have to factor in the effects of racial group membership. Yet, even if we allow for some common factors underlying diverse pathways, a shift in focus from group membership to heterogeneous pathways comes at the risk of bolstering a fiction that has gained currency in the USA, namely, that racial group membership no longer brings social (environmental) benefits and costs. Conversely, if researchers continue to track differences between averages for racial groups, how can they avoid bolstering the ubiquitous stereotyping in which group membership is employed when deciding how to treat an individual? In short, a genuine paradox that applies to the use of IQ test scores in US society seems to be that researchers and policy-makers who want to move beyond explanations and policies based on racial group membership cannot escape taking into account the disadvantages and benefits individuals experience because of their group membership.

New molecular genetic techniques may appear to circumvent the limitations of the classical methods of analysis of hereditary variation that are the focus of this article. However, is it wise to move ahead without understanding the conceptual and methodological gaps in the classical QG methods, as well as the social and historical context in which the gaps were obscured? Most especially, what is gained—and by whom—by pretending that the persistent interest in explaining differences among the averages for human groups defined on racial grounds has not been a salient part of that context?

A concluding puzzle

Suppose that philosophers of biology take up the suggestion above of revisiting studies that interpret heritability and "genetic variance" as measuring the size of the influence of genetic factors on the process through which the trait develops. Suppose they conclude, in light of the eight gaps identified in this article, that key results and interpretations from 90 years of quantitative genetics are not justified or, at best, are unreliable. What should they then do, for, as sociology and history of science reminds us, critique is rarely sufficient for a dominant paradigm to be abandoned? Perhaps human quantitative genetics could be viewed, contra Kendler (2005, 10), as akin to alchemy.[20] This was a field of inquiry that provided observations, questions, tools, debates, careers, and institutions which modern chemistry built on, but ultimately had to break away to make further progress. The shifts that led alchemy to be abandoned in the eighteenth century and to be depicted today as an exemplar of pseudoscience might be informative for thinking about the status of quantitative genetics. Whether or not this angle of interpretation is taken up, we might puzzle over what critical philosophers of science should do if they think that the scientists need to break away from fundamental and long-held assumptions and interpretations.

Acknowledgments This article is based on research supported by the National Science Foundation under grant SES–0634744. The comments of Hamish Spencer and anonymous reviewers of this manuscript and a related one helped in the revision process.

References

Byth DE, Eisemann RL, DeLacy IH (1976) Two-way pattern analysis of a large data set to evaluate genotypic adaptation. Heredity 37(2):215–230

Davey-Smith G, Ebrahim S (2007) Mendelian randomization: genetic variants as instruments for strengthening causal influences in observational studies. In: Weinstein M, Vaupel JW, Wachter KW (eds) Biosocial surveys. National Academies Press, Washington DC, pp 336–366

Dickens WT, Flynn JR (2001) Heritability estimates versus large environmental effects: the IQ paradox resolved. Psychol Rev 108(2):346–369

Downes SM (2004) Heredity and heritability. In: Zalta EN (ed) The stanford encyclopedia of philosophy. (http://plato.stanford.edu/entries/heredity/ (viewed 11 May 2006)

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman, Harlow

[20] Kendler (2005, 10) responded confidently to a trenchant criticism of some key assumptions of twin studies as follows: 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. This might have been warranted for some pseudoscientific systems, such as astrology, alchemy, and the Ptolemaic astronomic system. It is highly unlikely that modern psychiatric genetics will be judged by future historians of science to be in such company.

Flynn JR (1994) IQ gains over time. In: Sternberg RJ (ed) Encyclopedia of human intelligence. Macmillan, New York, pp 617–623

Flynn JR (2000) How to defend humane ideals: substitutes for objectivity. University of Nebraska Press, Lincoln, NE

Freedman DA (2005) Linear statistical models for causation: a critical review. In: Everitt B, Howell D (eds) Encyclopedia of statistics in the behavioral sciences. Wiley, Chichester

Fryer R, Levitt S (2004) Understanding the black-white test score gap in the first two years of school. Rev Econ Stat 86(2):447–464

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York

Jacquard A (1983) Heritability: one word, three concepts. Biometrics 39:465–477

Jencks C, Phillips M (eds) (1998) The black-white test score gap. Brookings Institution Press, Washington, DC

Jensen AR (1969) How much can we boost IQ and scholastic achievement? Harv Educ Rev 39:1–123

Jensen AR (1970) Race and the genetics of intelligence: a reply to Lewontin. Bull At Sci 26:17–23

Kaplan JM (2000) The limits and lies of human genetic research. Routledge, New York

Kendler KS (2005) Reply to J. Joseph, research paradigms of psychiatric genetics. Am J Psychiatry 162:1985–1986

Kendler KS, Prescott CA (2006) Genes, environment, and psychopathology: understanding the causes of psychiatric and substance abuse disorders. The Guilford Press, New York

Khoury MJ, Little J, Gwinn M, Ioannidis JP (2007) On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. Int J Epidemiol 36:439–445

Lewontin RC (1970a) Race and intelligence. Bull At Sci 26:2–8

Lewontin RC (1970b) Further remarks on race and the genetics of intelligence. Bull At Sci 26:23–25

Lewontin RC (1974) The analysis of variance and the analysis of causes. Am J Hum Genet 26:400–411

Lindman HR (1992) Analysis of variance in experimental design. Springer, New York

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer, Sunderland

Majumder PP, Ghosh S (2005) Mapping quantitative trait loci in humans: achievements and limitations. J Clin Investig 115(6):1419–1424

McLaughlin P (1998) Rethinking the agrarian question: the limits of essentialism and the promise of evolutionism. Hum Ecol Rev 5:25–39

Miele F (2002) Intelligence, race, and genetics: conversations with Arthur Jensen. Westview Press, Boulder

Moffitt TE, Caspi A, Rutter M (2005) Strategy for investigating interactions between measured genes and measured environments. Arch Gen Psychiatry 62(5):473–481

Neisser U, Boodoo G, Bouchard TJ, Boykin AW, Brody N, Ceci SJ, Halpern DF, Loehlin JC, Perloff R, Sternberg RJ, Urbina S (1996) Intelligence: knowns and unknowns. Am Psychol 51:77–101

Nisbett RE (1998) Race, genetics, and IQ. In: Jencks C, Phillips M (eds) The black-white test score gap. Brookings Institution Press, Washington, DC, pp 86–102

Nuffield Council on Bioethics (2002) Genetics and human behavior: the ethical context. http://www.nuffieldbioethics.org (viewed 22 Jun. 2007)

Otto SP, Christiansen FB, Feldman MW (1995) Genetic and cultural inheritance of continuous traits. Stanford University Morrison Institute for Population and Resource Studies Working Paper Series No. 64. http://www.stanford.edu/group/morrinst/pdf/64.pdf (viewed 24 March 2009)

Parens E (2004) Genetic differences and human identities: on why talking about behavioral genetics is important and difficult, Hastings center report (January-February). pp S1–S36

Plomin R, Asbury K (2006) Nature and nurture: genetic and environmental influences on behavior. Ann Am Acad Political Soc Sci 600(1):86–98

Plomin R, DeFries JC, Loehlin JC (1977) Genotype-environment interaction correlation in analysis of human behavior. Psychol Bull 84:309–322

Richardson K, Norgate S (2005) The equal environments assumption of classical twin studies may not hold. Br J Educ Psychol 75(3):339–350

Rutter M (2002) Nature, nurture, and development: from evangelism through science toward policy and practice. Child Dev 73(1):1–21

Sesardic N (2005) Making sense of heritability. Cambridge University Press, Cambridge

Taylor PJ (2005) Unruly complexity: ecology, interpretation, engagement. University of Chicago Press, Chicago

Taylor PJ (2006a) Heritability and heterogeneity: on the limited relevance of heritability in investigating genetic and environmental factors. Biol Theory Integr Dev Evol Cognit 1(2):150–164

Taylor PJ (2006b) Heritability and heterogeneity: on the irrelevance of heritability in explaining differences between means for different human groups or generations. Biol Theory Integr Dev Evol Cognit 1(4):392–401

Taylor PJ (2007) The unreliability of high human heritability estimates and small shared effects of growing up in the same family. Biol Theory Integr Dev Evol Cognit 2(4):387–397

Taylor PJ (2008a) Puzzles in the history and philosophy of heredity that warrant more attention. http://sicw.wikispaces.com/HeredityVariationPuzzles (viewed 12 Aug. 2008)

Taylor PJ (2008b) The under-recognized implications of heterogeneity: opportunities for fresh views on scientific, philosophical, and social debates about heritability. Hist Philos Life Sci 30:423–448

Taylor PJ (2009) Critical assumptions of classical quantitative genetics and twin studies that warrant more attention (manuscript)

Turkheimer E (2000) Three laws of behavior genetics and what they mean. Curr Dir in Psychol Sci 9(5):160–164

Turkheimer E, Haley A, Waldron M, D'Onofrio B, Gottesman II (2003) Socioeconomic status modifies heritability of IQ in young children. Psychol Sci 16(6):623–628

Wikipedia (2008) Heritability. http://en.wikipedia.org/wiki/Heritability (viewed 14 Mar 2008)