The "IQ paradox" reconceived: Visualizing the limited relevance of human heritability estimates in explaining differences between means across groups or across generations

PETER TAYLOR
Programs in Science, Technology & Values and Critical & Creative Thinking
University of Massachusetts, Boston, MA 02125, USA peter.taylor@umb.edu

1. Introduction: The IQ paradox and the persistent plausibility of genetic explanations

This working paper represents an ongoing process of conceptual clarification that began with my attempt to understand the motivation for and logic of the models of Dickens and Flynn (2001). Dickens and Flynn address the "IQ paradox," which they see as the co-existence of high estimates of heritability (Neisser et al. 1996) and large IQ test score gains between generations (the "Flynn effect") (Flynn 1994). The change in gene frequencies in a human population over one generation is negligible. Yet high heritability implies, in the conventional view of human behavioral genetics, that "observed variance in environment accounts for… little variance in adult IQ" test scores (Dickens and Flynn 2001, 346). So, Dickens and Flynn ask, how could large differences between generations be caused by environmental factors? Their answer depends on the possibility that environmental differences can be linked to and amplify genetic differences.

This idea is not new (see citations in Dickens and Flynn 2001, 347), but their "reciprocal causation" models allow systematic examination of its implications. Two features of these models contribute to the resolution of the paradox: 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).

Dickens and Flynn's models not only resolve the IQ paradox, but also challenge the conventional wisdom about differences between mean IQ test scores for racially defined groups. Many psychometricians and human behavioral geneticists believe that high heritability of IQ test scores within groups, combined with a failure of environmental hypotheses to account for the differences, lends plausibility to explanations of differences in means in terms of genetic factors (even if these factors have yet to be elucidated) (e.g., Jensen in Miele 2002, 111ff). The same logic, Dickens and Flynn note, would apply to explaining differences across generations in mean IQ test scores. However, given the negligible change in gene frequencies over one generation, genetic explanations are not plausible in the latter case and so something must be wrong with the logic. Reciprocal causation overcomes the problem, allowing us to conceive of a strong role for environmentally induced differences between means scores for generations or racial groups together with high heritability.

Dickens and Flynn's contribution has the potential to move the debate about heritability and differences between racial-group means onto fresh ground, although the response to their work has of yet been limited (Loehlin 2002, Rowe and Rodgers 2002; see Dickens and Flynn 2002). This paper challenges the conventional wisdom at deeper levels. In particular, I question two lines of thinking in which high heritability is held to bolster the plausibility of genetic explanations of differences between group means. Both lines begin by conceding that high heritability of IQ test scores within a racial group does not on its own allow us to conclude that the persistent difference between groups in mean IQ test scores also has a high heritability, but the following rejoinders are then put forward:

1. high within-group heritability suggests that it will be possible for researchers to find direct effects of genotypes on traits that influence IQ test scores (or indirect effects of environments induced by such genotypes). Researchers can expect different forms of the same genes to direct the development of traits that influence IQ test scores within other racial groups given the human biology that the groups share. The rapid advances in molecular genetics lend support for the expectation that genes and their effects can be identified.

2. high heritability of between group mean differences becomes more plausible when we note that:

a. If the differences are not caused by genetic differences, then they must be caused by environmental differences. Yet, all environment-only explanations that have been tested have been disproved (Flynn 1980, 40ff; Jensen in Miele 2002, 127ff).

b. High heritability means that the fraction of variation in IQ test scores within a group that is associated with environmental variation is low. If the IQ test score gap between group means were solely due to environmental causes and these causes varied within groups, then the number of standard deviations (SDs) of change in the environment causes that would be necessary to produce the gap is high if the heritability is high. Heritability of .75 translates to 2 or more SDs (Dickens and Flynn 2001, 348). No known environmental factor shows such wide variation between racial groups.

c. If the IQ test score gap between group means were solely due to environmental causes and these causes did not vary within groups, this could be consistent with high heritability, but no known environmental factor operates in such a fashion (sometimes called an X-factor) (Dickens and Flynn 2001; Flynn 1980, 62; Jensen 1973, 137ff; Sesardic 2000).

In questioning these lines of thinking I am not claiming that environment-only (or "culture-only") explanations account for variation between racial group means. Instead, my contention is that the concept of heritability and the statistical Analysis of Variance (AOV) on which it is based cannot logically or methodologically support the two lines of thinking above. Clearly I am entering a debate with a long and politically charged history. (For early points in the debate, see the 1969 Harvard Educational Review article by psychometrician Arthur Jensen, which elicited a critical response from, among others, the population geneticist, Richard Lewontin [1970a, b; 1974; Jensen 1970]. Jencks and Phillips [1998] reviews recent 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.) In order to contribute something new, I step away from human behavioral genetics into the analysis of agricultural crop trials in which a number of different varieties or "cultivars" of a crop are grown under different conditions. As will emerge, the two lines of thinking run into trouble even in this ideal case where discrimination among various factors is possible. In human behavioral

genetics, where heritability estimation and AOV depart from that ideal in significant ways, the lines of thinking are even less supportable.

Although I reinterpret selected contributions to the debate in light of my account of heritability and AOV (sect. 5), the scope of this paper does not extend to a detailed conceptual analysis of previous work on heritability and group-mean differences, to a sociological analysis of the production and maintenance of the conventional wisdom I am challenging, or to an alternative analysis of empirical data on IQ tests, racial classification, and generational differences. My goal here is primarily to motivate a set of themes (compiled in the appendix) that should help non-specialists as well as specialists visualize more clearly the limited relevance of human heritability estimates for explaining differences between means across groups or across generations. For readers who can keep these themes in mind, much that might have once seemed plausible will, I hope, become problematic. Along the way, the paradox that motivated Dickens and Flynn's reciprocal causation modeling will disappear, to be replaced by the challenge of deriving models of developmental pathways whose heterogeneous components differ among individuals at any one time and over generations.

2. Preamble: Changeability and Causes

Changeability is a key concern in this debate. I would be interested, for example, in knowing how white and black Americans' scores on IQ tests would change if everyone were raised in a non-racist environment, that is, without anyone having the disadvantages or advantages that are associated not with their individual abilities, but with their membership in a racial group (Flynn 2000, 142ff). Posing such a utopian question makes clear that changeability is not only a matter to be investigated through biological and social scientific research, but also a matter of whether the change in social arrangements implied by the question is within the range of movement of society that the researchers want to entertain. Even if I persisted in this line of questioning, I would not expect much insight to follow from social scientific research based in racially marked situations. On the other hand, I would hope that research could shed light on how and how much IQ test scores can be changed (absolutely or relatively) for individuals in racially defined groups. Yet, although such results would help people identify measures that are effective, replication of those measures might entail more social movement than the researchers

deem likely. In short, ideas about "social movement" changeability are entangled with extrapolation from particular or "local" conditions that researchers have analyzed through their specific questions and methods—with "local research" changeability, as we might call it. When Flynn (1980, 73) asks how whites would score on IQ tests if raised in the environments that blacks experience and vice versa, his question implies more change in society than I can imagine. At the same time he speaks to researchers studying IQ test scores, racial group differences, and heredity who think that their (local) research can reveal something meaningful about changeability (or lack thereof).

Against this backdrop, let me acknowledge the degree of changeability that underlies my discussion; it corresponds to asking what "local research" can learn about ways in which IQ test scores can be changed for individuals in racially defined groups and how much change (absolutely or relatively) is achieved. That question could, of course, be transferred to other socially marked groups (e.g., class and gender) and to attributes beyond IQ test scores. For example, how much can people's life achievements be changed absolutely or relatively for individuals in some socially defined group (racial, class, gender, etc.)? Indeed, IQ test scores would be of limited interest to educational and social policy-makers unless there were some link between them and life achievements. We could also ask how changeable that link is, but for the purposes of this essay, I assume that IQ test scores and traits that influence IQ test scores do have wider significance for people's lives.

Although the question about how much and through what means IQ test scores can be changed usually evokes educational interventions or social policy changes, it does not presume that changeability depends only on social ("environmental") factors or that unchangeability depends only on genetic factors. The ideal way to address the changeability question without such assumptions would be to have a detailed understanding of the ways different influences build on each other during the process of development of traits that influence IQ test scores (depicted schematically in Figure 1). Let me call this knowledge of "causes in developmental dynamics," or "developmental causes" for short. (See Pearl [2000] for a more elaborate and formalized approach to the analysis of causes and changeability.)

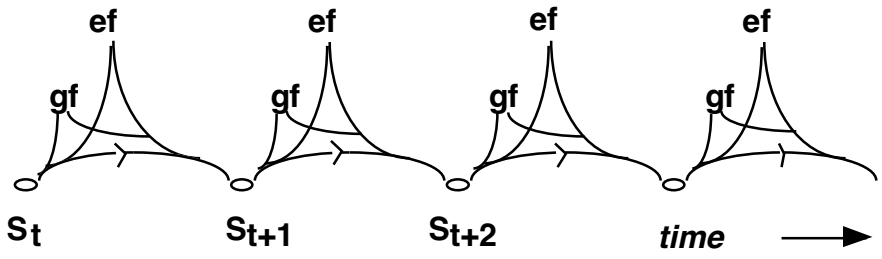


Figure 1. Schematic depiction of the process of development from time t to t+1 etc. of the state of an organism (S_t) in relation to its ability in IQ tests.
The state of organism at time t induces action by genetic factors (gf) and environmental factors (ef). gf and ef modulate the effect of each other on state of organism at time t+ 1, and so on. The nature of the genetic factors and environmental factors that are implicated in this process changes as the organism develops from a zygote to fetus to child to adult.

Debates about IQ test scores center on two other construals of causation, which I refer to as difference-in-effects causes and measurable factors. As will emerge, to move from difference-in-effects causes to measurable factors to developmental causes is to progressively loosen the assumption of control over the combinations of genetically defined varieties and locations that can be observed and, thereby, be able to provide insight about changeability.

3. Agricultural Crop Trials as the Ideal Case for Analysis of Variance, Heritability Estimation, and Regression Analysis

What can data analysis in the form of AOV, heritability estimation, and regression analysis tell researchers who conduct agricultural crop trials in which a number of different cultivated varieties or "cultivars" of a crop are grown in a number of plots (or replications) in different locations or under different conditions. Let me call these observational trials, in contrast with experimental trials in which specific environmental or genetic factors are systematically varied. (The term "location" need not be taken literally; it can refer to distinguishable conditions of many kinds, e.g., the weather in different years at the same site.)

The primary purpose of centering the paper on analysis of agricultural trials is to show that it is possible to question the lines of thinking identified in the introduction even in a case where researchers have great control over the genetic types and environmental conditions

studied. Moreover, given that my argument rests not on empirical details, but on logical and methodological grounds, it is appropriate to focus on agricultural crop trials and breeding programs because AOV and related techniques of quantitative genetics (which formed the basis of behavioral genetics) were first developed in that context. The use of agricultural terminology also allows me to avoid using the more conventional terms "genotype," "environment," and "GxE interaction," which tend to cloud the conceptual distinctions I will be making.

3.1 Linear models and Partitioning of Variance

The statistical concept of the heritability of a measurable attribute relies on partitioning into different components the variation observed among a set of individuals. This partitioning can be undertaken through an Analysis of Variance (AOV), which is always based on some specific linear model that separates into a series of "effects" the deviations of the measurements from their overall mean. For example, suppose that the yield was measured from a number of different cultivars in a number of plots (replications) in different locations. A linear model for the yield would be

Yield for cultivar i grown in plot k in location j

= the mean yield over all the measurements ("the grand mean")
+ the deviation from the grand mean of the mean of cultivar i over all locations and plots
+ the deviation from the grand mean of the mean in location j of all plots of all cultivars
+ the deviation from the sum of the above for cultivar i grown in location j of the mean over all plots of that cultivar grown in that location
+ the residual (the deviation from the sum of the above for cultivar i grown in location j of the yield in a particular plot of that cultivar grown in that location) (1a)

Conventionally, the verbal model is abbreviated as

Yield for cultivar i grown in plot k in location j
= the grand mean
+ the main effect of cultivar i
+ the main effect of location j
+ the interaction effect for cultivar i in location j

+ the residual effect (1b)

Alternatively, the model can be expressed in symbols:

$$y_{ijk} = m + c_i + l_j + cl_{ij} + r_{k:ij} \quad (1c)$$

where the notation k:ij denotes k is nested within i and j. If the residual effect is taken to be equivalent to noise or measurement error, it can be denoted by $\varepsilon_{k:ij}$.

The statistical significance of these effects is assessed in terms of the size of their variance in relation to the variation associated with noise or measurement error. Indeed, estimating the effects in the linear model is mathematically equivalent to partitioning the variation around the grand mean, i.e.:

$$\sigma^2_y = \sigma^2_c + \sigma^2_l + \sigma^2_{cl} + \sigma^2_r \quad (2)$$

where σ^2_y denotes the variance of yield measurements (i.e., sum of squared deviations from the grand mean), σ^2_c denotes the variance of cultivar effects, etc.

and the partitioning produces the lowest value of σ^2_r subject to the constraint that $\Sigma_i c_i = 0$; $\Sigma_j l_j = 0$; $\Sigma_i cl_{ij} = 0$ for each j; $\Sigma_j cl_{ij} = 0$ for each i; and $\Sigma_k r_{ijk} = 0$ for each ij combination. (These constraint means that the different kinds of effects in model 1c are independent, i.e., information about any one kind tells you nothing about any other.)

This method of partitioning can be applied to a wide range of situations. In principle, the cultivars in a trial may even be from different species, e.g., wheat, rye, oats, sorghum; the locations may be similarly heterogeneous, e.g., location 1 could be a no till cultivation site, location 2 a greenhouse site, location 3 drip irrigation, and so on.

Theme 1—Gradient-free conditions: Use of the AOV does not require that any gradient of a measurable genetic factor runs through the differences among genetically defined varieties or any gradient of a measurable environmental factor runs through the differences among locations.

Theme 2—Conditionality: All effects are conditional on the particular set of genetically defined varieties and locations observed.

(The abbreviated and symbolic models 1b and 1c hide this qualification, so we should refer back to the full wording of the linear model 1a if ever tempted to forget the conditionality.)

A simple numerical example will illustrate the conditionality of effects, variances, and heritabilities. (The figures in this paper's data sets are invented, but my arguments do not depend on correspondence between the data in the examples and actual observations.) Notice the changes in values as more locations and cultivars are included in the data analyzed.

Table 1

Data Set 1a

location	1
cultivar	
1	5.3,4.3*
2	3.1,2.1

Estimates of effects		Variance components & heritability estimates	
m	3.7	σ^2_c	1.21 (83%**)
l_1	0	σ^2_ε	0.25 (17%)
c_1	1.1	$h^2_{w/in\ location\ 1}$	0.83***
c_2	-1.1		
$\varepsilon_{k:ij}$	+/-.5		

Data Set 1b

location	1	2
cultivar		
1	5.3,4.3	0.2,1.2
2	3.1,2.1	2.4,1.4

Estimates of effects		Variance components & heritability estimates	
m	2.5	σ^2_c	0.0625 (2.5%)
l_1	1.2	σ^2_l	1.44 (58%)
l_2	-1.2	σ^2_{cl}	0.7225 (29%)
c_1	0.25	σ^2_ε	0.25 (10%)
c_2	-0.25	$h^2_{w/in\ location\ 1}$	0.83
cl_{ij}	±0.85	$h^2_{w/in\ location\ 2}$	0.59
$\varepsilon_{k:ij}$	+/-.5	$h^2_{across\ locations}$	0.25

Data Set 1c

location	1	2
cultivar		
1	5.3,4.3	0.2,1.2
2	3.1,2.1	2.4,1.4
3	4.9,5.9	1.6,2.6
4	3.7,2.7	2.8,3.8

Estimates of effects		Variance components & heritability estimates	
m	3.0	σ^2_c	0.3125 (13%)
l_1	1	σ^2_l	1 (43%)
l_2	-1	σ^2_{cl}	0.7625 (33%)
c_1	-0.25	σ^2_ε	0.25 (11%)
c_2	-0.75		
c_3	0.75	$h^2_{w/in\ location\ 1}$	0.84
c_4	0.25	$h^2_{w/in\ location\ 2}$	0.77
cl_{1j}, cl_{4j}	±1.05	$h^2_{across\ locations}$	0.13
cl_{2j}, cl_{3j}	±0.65		
$\varepsilon_{k:ij}$	+/-.5		

Notes:
* The two figures separated by a comma denote two independent replications. The order of the figures is of no significance.

** Figures in parentheses give percentages of the total variance.
*** The meaning and significance of the heritability estimates, denoted by h^2, will be discussed in sections 3.3 and 4.

If the conditionality of effects derived from partitioning of variance is kept in mind, effects can be construed as causes in a particular sense: Differences between the effects for individuals of different genetic types can be interpreted as causes of differences (on average over locations) between the yield for different cultivars. Similarly, the difference between the effects for locations can be interpreted as causes of differences (on average over cultivars) between the yield in different locations. Finally, the difference between interaction effects for cultivar-location combinations can be interpreted as causes of the differences between the cultivar yields in different locations over and above the two differences above. Let me call this a "difference in effects" sense of causality. Notice that difference-in-effects causes do not require identification of measurable genetic or environmental factors (see theme 1). (We should refer back to the full model 1a if ever tempted to read the symbolic model 1c as if the c_i's corresponded to genetic factors, l_j's corresponded to environmental factors, and cl_{ij}'s corresponded to interactions between otherwise separable genetic and environmental factors.)

Theme 3—Rerun control: The conditionality of effects means that any predictions made using difference-in-effects causes entail an assumption of control over the genetically defined varieties and locations that would allow the original combinations to be rerun and observed again.
If the control is imperfect, for example, if the weather varies between growing seasons or the cultivars are not identical, the accuracy of predictions will diminish. Nevertheless, as described in the section to follow, meaningful predictions can be made on the basis of difference-in-effects causes if the data can be appropriately rearranged and the variance repartitioned.

3.2 Clustering and repartitioning of variance

Consider the analysis of large international crop trials, e.g., the 1967-68 trial of 49 wheat varieties grown in 63 locations discussed in Byth et al. (1976). The AOV for such data sets typically finds that the cultivar-by-location-interaction variance component is as large as the cultivar component. In practical terms this means that no cultivar performs well in all locations so it would be ill-advised to make recommendations to farmers as if you can predict which cultivar would perform best; similarly, for recommendations about which cultivars to cross with

each other in subsequent breeding programs. Byth et al. (1976) addressed this problem as follows.

First, Byth's team of researchers performed cluster analysis to group cultivars by similarity in responses across all locations. Similarly, they grouped locations by similarity in responses elicited across the full range of cultivars. For cultivars in some cultivar group or locations in some location group, they found that the relative sizes of within-group variance components had changed so that the ranking of the cultivars tended to hold across locations within any location group. In any case, differences among the cultivars within a cultivar group and differences among the locations within a location group were small because the clustering method resulted in within-group variance components that were relatively small. In short, these researchers were able to make recommendations for farmers that were conditional on working within the range of cultivars and locations in the specific cultivar and location groups but were not dependent on perfect control over which cultivar and location were chosen in those groups.

Clustering resulted in variance components for among cultivar group means, among location group means, and the interaction of among cultivar group means and among location group that were much greater than the corresponding within-group components. This meant that the responses of cultivars across different locations could be efficiently summarized in terms of the patterns of cultivar group means across location groups (Figure 2). Recommendations could be made to plant breeders, for example, to cross cultivars from two specific groups so as to bring the strengths that the first group exhibited in specific location groups to overcome the weaknesses of cultivars in the other group in the same locations. It should be noted, however, that making such recommendations discounts the possibility that different conjunctions of underlying developmental causes or measurable factors might lie behind cultivars having similar patterns of responses across locations. It might seem unreasonable to worry about this when clustering takes into account response across a wide range of locations, but concern is justifiable when+ groups are formed on a limited range of locations or on some basis other than similarity in responses.

Theme 4—Questionable homogeneity: An assumption that is always open to questioning is that similar patterns of responses of different genetically defined varieties across locations (or environmental factors) have been produced by similar conjunctions of underlying developmental causes or measurable factors.

--insert Figure 2, reproduced from Byth (1976)--

As an illustration of clustering and repartitioning of variance, consider the effect of clustering cultivars for data set 1c. (For this illustration the two locations could be considered as two groups of size one.) As figure 3 shows, cultivars 1 and 3 would be grouped because they have similar responses across locations; similarly, cultivars 2 and 4 would be grouped. Within each cultivar group the ranking does not change across locations, that is, the interaction within cultivar groups has been reduced. This is shown numerically in Table 2, which presents the results of an AOV performed with the appropriate model, namely,

$$y_{ijk} = m + C_I + c_{i:I} + l_j + Cl_{Ij} + cl_{i:I,j} + \varepsilon_{k:ij} \quad (3)$$

where i:I denotes the i^{th} cultivar in cultivar group I

Figure 3. Data Set 1c

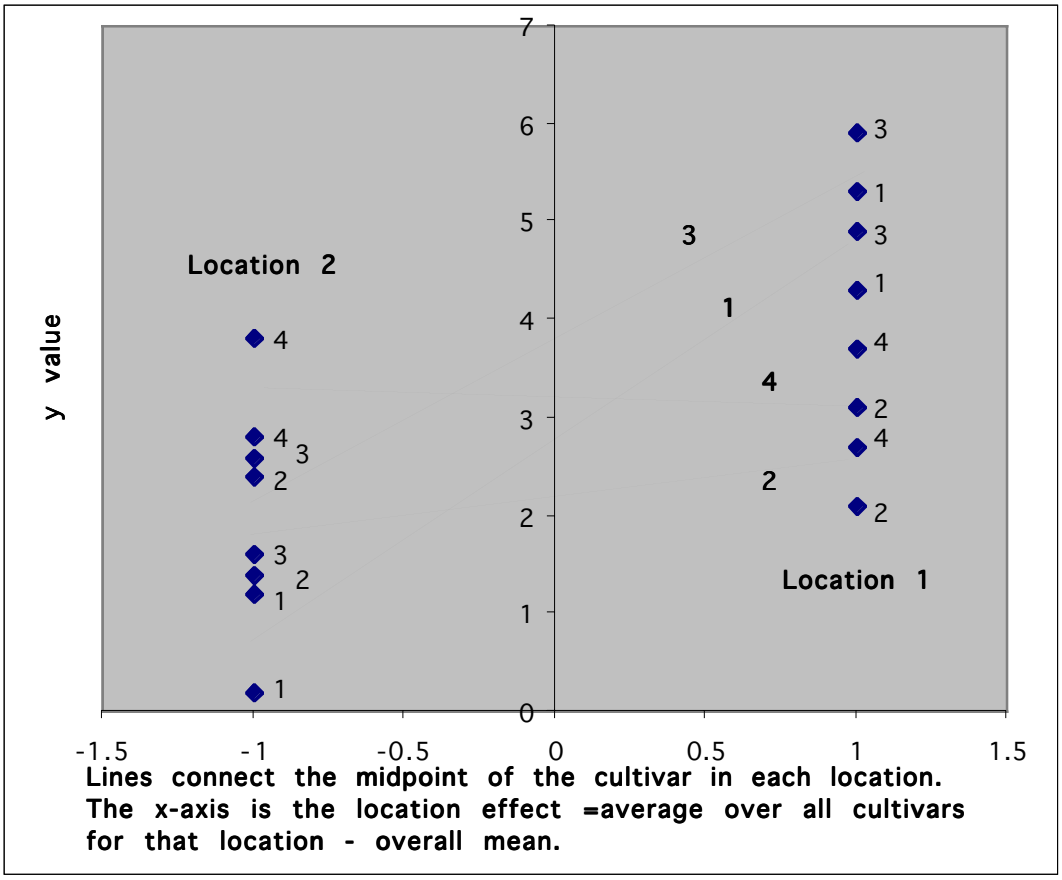


Table 2

Data Set 1c divided into two groups Estimates of effects Variance components &

by similarity of response across locations **heritability estimates**

	location	1	2
Cultivar Group	cultivar		
A	1	5.3,4.3	0.2,1.2
B	2	3.1,2.1	2.4,1.4
A	3	4.9,5.9	1.6,2.6
B	4	3.7,2.7	2.8,3.8

m	3.0
I_1	1
I_2	-1
C_A	0.25
C_B	-0.25
$c_{1:A}$	-0.5
$c_{2:B}$	-0.5
$c_{3:A}$	0.5
$c_{4:B}$	0.5
CI_{ij}	±0.85
$ci_{i:I,j}$	±0.2
$\varepsilon_{k:ij}$	+/-.5

σ^2_C	0.0625 (2.7%)
$\sigma^2_{c:C}$	0.25 (11%)
σ^2_I	1 (43%)
σ^2_{CI}	0.7225 (31%)
$\sigma^2_{c:C,I}$	0.04 (1.7%)
σ^2_ε	0.25 (11%)
h^2 within cultivar group A or B within location 1	0.26
h^2 within cultivar group A or B within location 2	0.66
h^2 within cultivar group A across both locations	0.06
h^2 within cultivar group B across both locations	0.44

When researchers in Byth's team formulated recommendations after creating groups and repartitioning variance, they were not relying on any knowledge about the biophysical pathways of the plant growth and development (developmental causes) and how these pathways were affected by the different genetic makeup of cultivars and the different environmental factors in the locations. The recommendations assumed that farmers and breeders would be able to re-grow the cultivars again in subsequent years and confine themselves to sites that were similar to the locations within a specific group of locations. This meant that difference-in-effects causes could provide meaningful predictions (theme 3). Subject to the variation in weather from year to year, farmers could expect yields from the cultivars they grew that matched those achieved in the trial in locations like their farm; plant breeders could expect that, subject to the complexities of genetics and development, crosses between cultivars would have predictable results.

3.3 Rerun predictability and heritability

Predictions made on the basis of an AOV are not identical to the original observations because of the residual variation or error that corresponds to replications within any given cultivar-location combination. We can generate formulas to estimate the correlation between observations and predictions if we use the linear model on which the AOV was based to make

the prediction and if we specify how the original situation and predictions are to be matched—Will the cultivar in the observed case be paired with the predicted values in the rerun, or the location? Will all cultivars be matched, or only those in one group? Will all locations be considered or one only?

Let me call such correlations "rerun predictability." For example, the rerun predictability for the first location and cultivar group A when each cultivar is matched in the current case and the rerun is the correlation between the all the possible permutations depicted in Figure 4. The rerun predictability across both locations is the correlation depicted in Figure 5.

Figure 4. Cultivar group A in location 1: Correlation between observed values & predicted rerun values

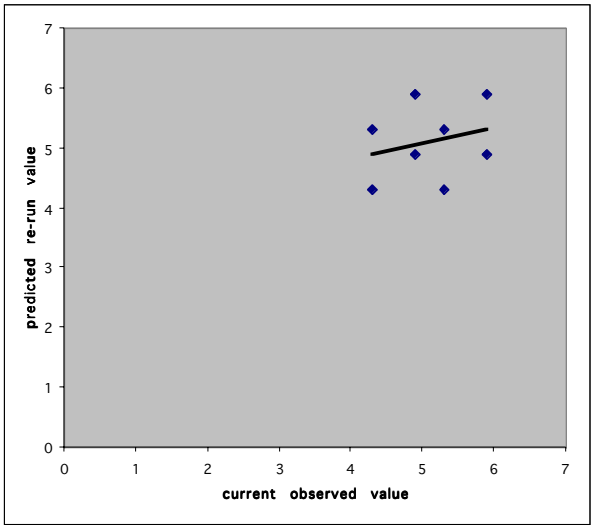
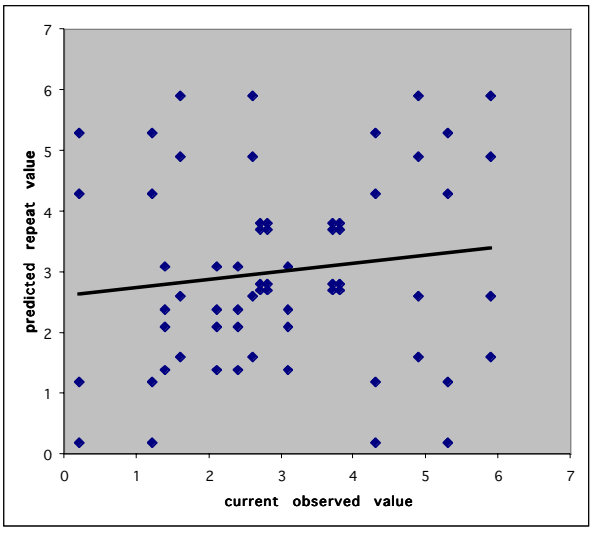


Figure 5. Cultivar group A across both locations: Correlation between observed values & predicted rerun values



Although rerun predictability is a measure of correlation, the general formula turns out to consist of variance terms, namely:

Variance of components in the model that are constrained to be the same in observed case and the rerun in the defined situation /

Variance of all components in the model that are not constant in the defined situation (which includes appropriate error variance) (4)

(Note that rerun predictability, like the variance components that enter into this formula, is conditional on the particular set of cultivars and locations observed; see theme 2.) For example,

for the situation depicted in figure 4, the formula for rerun predictability can be derived as follows:

Correlation (observed and predicted)

= Covariance (observed, predicted)/ [Variance (observed)* Variance (predicted)] $^{1/2}$

= Cov ($m + c_i + l_j + cl_{ij} + \varepsilon_{k:ij,}$, $m + c'_i + l_j + cl'_{ij} + \varepsilon'_{k:ij}$)

/ [Var ($m + c_i + l_j + cl_{ij} + \varepsilon_{k:ij}$) * Var ($m + c'_i + l_j + cl'_{ij} + \varepsilon'_{k:ij}$)] $^{1/2}$

where j is fixed, but i and k can vary & ' denotes the rerun

which can be estimated by Cov (c_i+ cl_{ij}, c'_i+ cl'_{ij}) / [(σ^2_c +σ^2_{cl} +σ^2_ε) * (σ^2_c +σ^2_{cl} +σ^2_ε)]$^{1/2}$,

given that error effects are independent and $m + l_j$ is a constant

which can be estimated by ($\sigma^2_c + \sigma^2_{cl}$) / ($\sigma^2_c + \sigma^2_{cl} + \sigma^2_\varepsilon$), given that the cultivar is constrained to be the same in the observed situation and the rerun (i.e., i = i') (5)

Quantitative and behavioral geneticists will recognize eqn. 5 as a formula for heritability in the case where cultivars are confined to a single location (Lynch & Walsh 1998, 669).

Theme 5—Heritability as rerun predictability: In its technical meaning, heritability is a special case of rerun predictability (a concept that encompasses a more general class of correlations) in which the cultivar is matched in the observed data and rerun (i.e., i = i').

As a specific form of rerun predictability, heritability is based on conditional estimates from an AOV and provides meaningful predictions to the extent that cultivars and locations can be controlled and replicated (see themes 2 and 3). The difference between heritability within one location and heritability across locations (e.g., .26 in Figure 4 versus .06 in Figure 5) can be understood in terms of rerun predictability and reflects nothing about the underlying gene-based dynamics of reproduction or their differences between locations.

Like heritability, rerun predictability can be applied more generally to linear functions of effects derived from the measurements, such as differences between the means of the cultivar groups or predicted outcomes of crossing or mating between cultivars. The latter predictions have to factor in the relatedness of parents—e.g., are they sibs or cousins or unrelated?—and require assumptions about the result of crosses—e.g., will the offspring tend to come out half way between the parents or will they show the effects of Mendelian combinations of a small number of genetic alleles? More complicated formulas for heritability are required (see Lynch and Walsh 1998), but they all involve partitioning of variance and can be construed as variants of

the basic idea of rerun predictability. (Similarly for formulas that allow estimation of heritability when there is only one replicate because there is genetically relatedness among cultivars and replicates are correlated, not random within the location; Lynch and Walsh 1998, 153ff.)

A formula for heritability that is commonly cited—genetic variance divided by the total variance of the trait ("phenotypic variance")—is too loose to help us conceptualize the variants under different conditions (e.g., within a location vs. across all locations). Moreover, the term "genetic variance," which in this context refers to the variance of the cultivar effect in a specific model, could be misread as the variance of some measurable genetic factor.

Theme 6. Heritability vs. "heritable." Ambiguity in the term "genetic variance" invites the technical term heritability to be misidentified with the colloquial idea that a trait is "heritable" or "genetic" if differences in a trait are associated with differences in specific genetic factors in the gene-based dynamics of organisms' reproduction.

The neutral term rerun predictability has the virtue of avoiding the connotations that heritability has with gene-based dynamics of organisms' reproduction.

3.4 Causes and measurable factors derived from observational and experimental trials

Plant breeders do not have to unravel the biophysical processes of plant growth and development (i.e, the developmental causes) before they make recommendations to farmers on the basis of clustering and partitioning variation from observational trials, in which multiple cultivars are each grown in multiple locations. It is possible, however, to build on such analysis to learn more about a kind of cause intermediate between difference-in-effects and developmental causes, namely, measurable genetic and environmental factors that differ among cultivars and locations in association with differences among cultivars and locations in the yield (or other trait) that has been observed. (As an analogy, without knowing how eyes develop in fruit flies, geneticists long ago identified a mutant gene associated with the flies' eyes being white instead of the normal red color. Reciprocally, researchers have identified environmental conditions that induce traits or "phenocopies" that resemble those of flies that have mutant genes; e.g., Mitchell and Lipps 1978.) Let me distinguish three ways that measurable factors associated with differences, or "measurable factors" for short, can be elucidated.

a. Natural history hypothesizing. Researchers, drawing on sources of knowledge—"natural history"—other than the data from the crop trials, can formulate hypotheses about what aspects of the locations in any particular location group elicited basically the same response from the cultivars in a particular cultivar group that distinguished them from other groups. That is, they can make hypotheses about measurable factor causes. To invent an example, if rainfall occurred in concentrated periods on poorly drained soils, then cultivars whose genes originated from particular parental stock that was more susceptible to plant rusts may have yielded badly. (See Byth et al. 1976, 224ff for actual hypotheses after analysis of the international wheat cultivar trial referred to earlier.)

b. Regression analysis. If data are available on some environmental factors (e.g., rainfall, soil type, day lengths) in the different locations within a location group, statisticians can estimate the parameters ("regression coefficients") that best fit models such as the following to the observed data:

$$y_{ijk} = m + C_I + c_{i:I} + \Sigma_q b_{iq} e_{jkq} + \varepsilon_{k:ij} \qquad (6)$$

where e_{jkq} denotes the value of environmental factor q in replication k in location j;
b_{iq} denotes the regression coefficient that conveys the (partial) correlation of cultivar I's yield with environmental factor q (across locations and replicates, j and k); and
$\varepsilon_{k:ij}$ is a residual or error term (but not the same one as in the AOV models; this residual reflects the particular range of environmental factors for which measurements are available and included in the model).

To increase the number of observations on which each regression analysis is based, the models for separate cultivars (eqn. 6) in a cultivar group may be consolidated into a smaller set of analyses. This requires that we assume that all b_{iq}'s are the same across cultivars in the same cultivar group (= "b_{Iq}").

$$y_{ijk} = m + C_I + \Sigma_q b_{Iq} e_{jkq} + \varepsilon_{k:ij} \qquad (6a)$$

The homogeneity assumption involved in such consolidations is subject to questioning (theme 4), especially since such consolidation shifts the c_i and $cl_{i:I,j}$ effects into the

residual term as if they were negligible. If data are available for measurable genetic factors, such as presence or absence of an allele at a genetic locus, eqn. 6a could be elaborated so as to bring those effects back out of the residual:

$$y_{ijk} = m + C_I + \Sigma_p b_p g_{ip} + \Sigma_q b_{Iq} e_{jkq} + \Sigma_p \Sigma_q \, b_{pq} g_{ip} e_{jkq} + \varepsilon_{k:ij} \qquad (6b)$$

where g_{ip} denotes the value of genetic factor p in location i, and
b_p and b_{pq} denote the regression coefficients that convey the (partial) correlation of cultivar yield with, respectively, genetic factor p and the product of genetic factor p and environmental factor q (as these factors vary across cultivars, locations, and replicates i, j, and k).

c. Experiments. Hypotheses drawn from natural history can be examined through experimental crop trials in which the cultivars are grown subject, for example, to differing watering schedules and soils of varying degrees of drainage. Similarly, the results of regression analysis can be validated if experimental crop trials are conducted in which the genetic and environmental factors are systematically varied. Results of such trials can be analyzed using AOV under models such as:

$$y_{ijpqk} = m + c_i + l_j + w_p + d_q + cl_{ij} + cw_{ip} + cd_{iq} + lw_{jp} + ld_{jq} + wd_{pq} + clw_{ijp} + cld_{ijq} + cwd_{ipq} + lwd_{jpq} + clwd_{ijpq} + \varepsilon_{k:ijpq} \qquad (7)$$

where w denotes the "watering schedule" effect, d the "degrees of drainage" effect, and the longer terms denote interaction effects,
and i and j are restricted to the particular cultivar and location groups under consideration. (Strictly, the notation should be specify i:I and j:J at appropriate points.)

If the watering schedule and degrees of drainage effects were significantly different from zero we would gain the insight that differences in these environmental factors made a difference (conditional again on the particular groups of cultivars and locations). (In

principle, since the advent of genetic engineering, we can also conduct experimental trials varying the degrees of expression of specific genes in the cultivars.)

To the extent that other researchers are unraveling the pathways of plant growth and development, insights from crop trials in which environmental and genetic factors are varied may contribute to understanding (or hypothesizing about) the ways that pathways of growth and development are affected by the genetic makeup of cultivars and the environmental factors in the locations. Let me articulate a theme about the contribution of regression analysis and experiments that build on observational trials, then draw out several implications or sub-themes.

Theme 7—From Observation to Hypothesis to Experiment to Understanding of Developmental Causes: After the appropriate simplification of the data set by clustering, the AOV of observations and subsequent regression analysis can contribute to the formulation of hypotheses that may be subject to experimental trials designed to examine the effect of varying specific genetic or environmental factors. In such trials the number of factors that can be considered simultaneously is constrained in practice (an extension of theme 3 on control) and the effects exposed by AOV are conditional (theme 2), but such experimental trials may contribute to a larger project of unraveling the biophysical pathways of development.

The distinctions between observation of multiple cultivars in multiple locations, natural history hypotheses, regression analyses, and experimental crop trials that vary specific factors have a number of implications for thinking about causes and factors:

Sub-theme 7a. Hypothesis generation following observational trials is enhanced by simplifying the large data set into groups for which the response of a cultivar group member (or response elicited by a location group member) is similar to the average for the group as a whole. Hypotheses become more difficult to formulate when groups are defined not by clustering, but by using some other criteria that does not minimize the variance within groups, including the within-group interaction variance (see theme 4).

As illustration of this difficulty, compare being asked to generate hypotheses in the following two situations: first with data set 1c where cultivars 1 and 3 formed group A and 2 and 4 formed B, then with the same data set where cultivars 1 and 2 formed group A and 3 and 4

formed B. (Table 3 shows the repartitioning of variance for the second grouping, again using model 3.) It is easy to imagine something about cultivars 1 and 3 that makes them respond more positively to some environmental condition in location 1 than do cultivars 2 and 4. For the second grouping, however, factors have to be hypothesized that could be associated with different differences between locations for the cultivar groups on average while allowing the different differences for cultivars within both groups (see Fig. 2). This problem becomes even greater if researchers made no groups and tried to hypothesize about differences between individual cultivars' responses in a location or across a group of locations.

Table 3

Data Set 1c divided into two arbitrary groups				**Estimates of effects**		**Variance components & heritability estimates**	
	location	1	2	m	3.0	σ^2_C	0.25 (11%)
Cultivar Group	cultivar			I_1	1	$\sigma^2_{c:C}$	0.0625 (2.7%)
A	1	5.3,4.3	0.2,1.2	I_2	-1	σ^2_I	1 (43%)
A	2	3.1,2.1	2.4,1.4	C_A	-0.5	σ^2_{CI}	0.04 (1.7%)
B	3	4.9,5.9	1.6,2.6	C_B	0.5	$\sigma^2_{c:C,I}$	0.7225 (31%)
B	4	3.7,2.7	2.8,3.8	$c_{1:A}$,$c_{3:B}$	0.25	σ^2_ε	0.25 (11%)
				$c_{2:A}$, $c_{4:B}$	-0.25		
				CI_{A1}, CI_{B2}	0.2	$h^2_{\text{within cultivar group A or B within location 1}}$	0.83
				CI_{A2}, CI_{B1}	-0.2	$h^2_{\text{within cultivar group A or B within location 2}}$	0.59
				$cI_{i:I,j}$	±0.85	$h^2_{\text{within cultivar group A across both locations}}$	0.025
				$\varepsilon_{k:ij}$	+/-.5	$h^2_{\text{within cultivar group B across both locations}}$	0.037

Although hypothesis generation based on observational crop trials is enhanced by appropriate clustering, it tends, in practice, to revolve around one or two major environmental factors whose effect on the cultivars in a group makes sense given their shared genealogical origins. It discounts the possibility that behind a similar response across locations of cultivars that cluster together may lie different conjunctions of measurable factors (theme 4). To use data analysis to expose more heterogeneity in measurable factors, researchers need, ironically, to know already quite a lot about the biophysical pathways of growth and development (developmental causes).

If data are available on environmental factors, hypotheses can be formulated and tested (statistically) through regression analysis. In order to produce sufficient observations for each model to be fitted to the data consolidation of analyses across cultivars within groups is usually needed (see 3.4b, above). Clustering that minimizes within-group variances makes such consolidation seem reasonable with respect to the homogeneity assumption; arbitrary groups or consolidation across all cultivars invites attention to heterogeneity (theme 4). If regression analyses are performed that consolidate heterogeneous models, their results are less likely to be validated by experimental trials. Statistical testing of such models become less a step towards understanding developmental causes and more akin to AOV that identifies difference-in-effect causes. In this latter context, themes 2 and 3 about conditionality and control, and the concept of rerun predictability are apt. (Appendix 2 considers a special case in which there is one measurable genetic factor and one measurable environmental factor.)

Sub-theme 7b. High heritability has a mixed relationship with researchers' ability to formulate hypotheses.

Notice that under the second grouping within-cultivar-group heritability within one location tends to be larger (an average of .71 versus .46). This higher heritability has come at the expense of homogeneity within groups; the heterogeneity makes it harder for researchers to formulate hypotheses about what aspects of the location (or location group) elicit basically the same response from the cultivars in a particular cultivar group. At the same time, under the second grouping the heritability across both locations decreases (an average of .031 versus .25). This decline simply reflects the sizeable amount of cultivar by location interaction that remains within cultivar groups under the second grouping. As noted above, any hypothesis from the second grouping has to accommodate differences between cultivars as well as different differences from one location to the other.

Sub-theme 7c. The results of experimental trials in which environmental factors are systematically varied are conditional on the levels of other factors not subject to experimental variation and on the combinations of factors varied, as well as on the combinations of cultivars and locations in the trial.

To invent an example, we might find that concentrating the same seasonal total of water into short periods was better than constant moisture if the cultivars in the trial were susceptible to rust or if the mid-season in that location is cool, but not otherwise.

Sub-theme 7d. In the absence of experiments, estimates derived from regression analysis are not only conditional, but are like difference-in-effects causes in that predictions made using them assume rerun control (themes 3& 7a).

Although it is possible mathematically to extrapolate a regression equation involving measurable environmental factors beyond the cultivar or cultivar group in which it was derived, we would need more knowledge about the developmental causes to ascertain whether this was justified.

Sub-theme 7e. There are important conceptual distinctions among the different quantities referred to as "environmental variance." The variance of a location effect in an AOV of an observational trial is not the same as the variance of the effect in an AOV of a trial in which an environmental factor is systematically varied. The latter is not the same as the variance of such an environmental factor itself. The variance associated with an environmental factor in a regression equation is yet another different quantity.

Recall that, in non-experimental trials a location effect has a variance even when there is no continuous gradient running through the differences among locations (theme 1). Yet, even if there are such gradients, the different environmental variances do not have a clear relationship. For example, if the factor varied were kg/ha of nitrogen in fertilizer applied to the crop, low levels of nitrogen would be positively associated with yield, but in trials using higher levels of nitrogen the influence would diminish or even turn negative. In such trials the AOV may still show a significant effect for nitrogen application, but it would be smaller than if nitrogen had been restricted to a smaller range and the variance in nitrogen had been smaller. It is a confusion of categories to equate a treatment effect in a model to a measurable environmental factor. They may turn out to be correlated, but the expected degree of correlation is not something that can be determined without reference to a model of the underlying biophysical processes of development (see sect. 3.7, example 2 and appendix 2).

Another way to visualize the conceptual difference is to recognize that the subscript j in model 1c applied to the fertilizer trial would refer to one of the fertilizer treatments—the order is not important—while l_j refers to an amount of the measured trait, e.g., tonnes/ha of yield, not to kg/ha of fertilizer.

Sub-theme 7f. Practical considerations place limits on experimentation that stems from observational trials.

Some factors are more difficult than others to vary experimentally. For example, alterations in day length are harder to implement than additions of fertilizer. The more factors to be varied experimentally, the more cross-combinations that should be tested, and the more difficult it is in practice to implement the necessary crop trials and encompass mentally the many interaction effects. In practice, there is a tendency for hypothesis generation and experiments to revolve around one or two measurable factors that can be readily manipulated (see theme 7a).

Sub-theme 7g. If measurable genetic factors are identified, then the previous sub-themes also apply to generation of hypotheses about genetic factors and to different meanings of the term "genetic variance" (see also theme 6).

3.5 A thought experiment to illustrate the difficulty of exposing complex causes through AOV

Let me introduce a simple model and thought experiment to accentuate the difficulties of exposing the complexity of biophysical processes of growth and development using AOV. Consider the generic model of development in figure 1, but simplify it by not allowing the state of the organism to induce actions by genetic and environmental factors. That leaves the attribute in question being produced by a sequence of gene actions, each one modulated by a corresponding environmental factor and subject to noise. One way this could be modeled is as follows:

$$y'_{ijk} = \Pi_r (g_{ir}{}^{e_{jr}}) f_{jrk} \qquad (8)$$

where g, e, f denote genetic factors, environmental factors, and random noise respectively,

g_{ir} = 1 or $1+\gamma$ with equal probability,

e_{jr} = ±β with equal probability, and

f_{jrk} = 1 for the 1st replicate, 1+random number in interval(-δ,δ) for the 2nd replicate

To facilitate comparison with data set 1c, I will scale any data generated from model 8 so it has the same mean and SD:

$$y_{ijk} = \text{constant}_1 + \text{constant}_2 * y'_{ijk} \quad (8a)$$

One data set generated by model 5 with the values γ = .8, β = .5, δ = .25, r=1,...,5, constant_1 = -0.88, and constant_2 = 3.53 is given in the bottom right hand corner of Table 4. The AOV of these data using linear model 3 is given in Table 5 and Figure 6.

Table 4

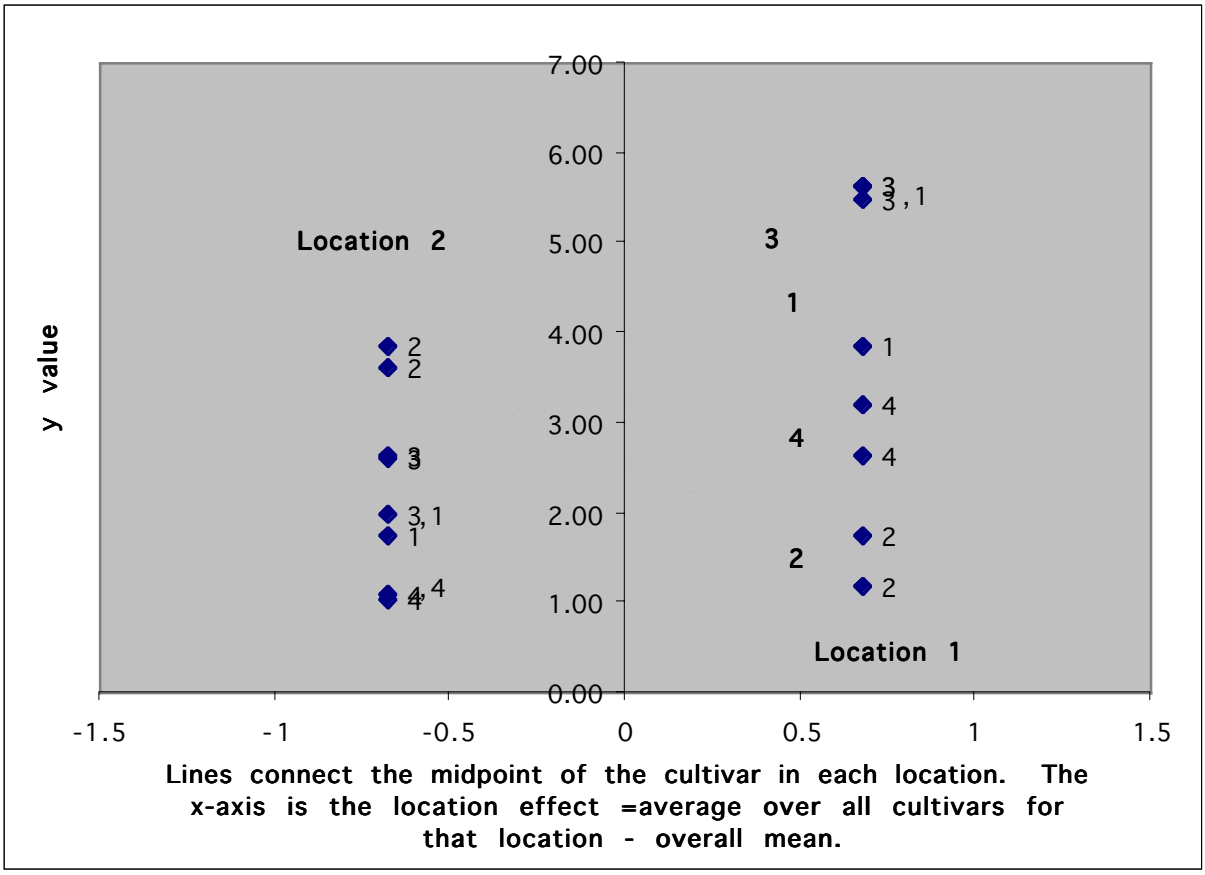
						locations	
					envtl factors	1	2
					1	0.5	0.5
					2	0.5	-0.5
					3	-0.5	0.5
	genes				4	-0.5	-0.5
cultivars	1	2	3	4	5	0.5	0.5
1	1.8	1.8	1	1.8	1	3.9, 5.6	1.8, 2.6
2	1	1	1.8	1.8	1.8	1.8, 1.2	3.9, 3.6
3	1	1.8	1	1	1.8	5.5, 5.6	2.7, 2.0
4	1	1.8	1	1.8	1	2.7, 3.2	1.1, 1.0

Table 5

Data Set 2 divided into two groups by similarity of response across locations				**Estimates of effects**		**Variance components & heritability estimates**	
	location	1	2	m	3.0	σ^2_C	0.49 (21%)
Cultivar Group	cultivar			l_1	0.67	$\sigma^2_{c:C}$	0.07 (3%)
A	1	3.9, 5.6	1.8, 2.6	l_2	-0.67	σ^2_l	0.45 (20 %)
B	2	1.8, 1.2	3.9, 3.6	C_A	0.70	σ^2_{Cl}	0.60 (26 %)
A	3	5.5, 5.6	2.7, 2.0	C_B	-0.70	$\sigma^2_{c:C,l}$	0.55 (24%)
B	4	2.7, 3.2	1.1, 1.0	$c_{1:A,}c_{3:A}$	±0.24	σ^2_ε	0.16 (7 %)
				$C_{2:B}$, $c_{4:B}$	±0.30		
				Cl_{A1}, Cl_{B2}	-0.77	$h^2_{\text{within cultivar group A within location 1, 2}}$	0.29, 0.029
				Cl_{A2}, Cl_{B1}	0.77	$h^2_{\text{within cultivar group B within location 1, 2}}$	0.87, 0.996
				$cl_{i:A,j}$	±0.17	$h^2_{\text{within cultivar group A across both locations}}$	0.02

$cl_{i:B,j}$	±1.04	h^2 within cultivar group
$\varepsilon_{k:ij}$	varied	B across both locations 0.08

Figure 6. Data set 2.



The first part of the thought experiment is to consider what a comparison of the AOV of data sets 1c and 2 would suggest to researchers about the processes that generated the data. A comparison of figures 3 and 6 shows that, although cultivars 4 and 2 only converge in data set 1c but cross in data set 2, the overall trends are very similar. The similar AOVs would not suggest that the data sets 1c and 2 were generated by radically different kinds of models. However, this is the case. Although I have not stated this before now, data set 1c was simply (and unrealistically) generated by the linear model 3 using the parameter values shown in Table 3.

The second part of the thought experiment is to consider what the researchers following the natural history approach would hypothesize about measurable factors on the basis of the AOV in Table 5. Suppose that their prior knowledge about the cultivar genetics and the locations led them to suspect that genetic factors 1-5 and environmental factors 1-5 could be

important in explaining differences between cultivar groups and between locations for any cultivar group. The best hypotheses they could make are given in the second column of Table 6. Compare this with the third column, which summarizes the actual factors underlying the data in Table 4.

Table 6 A comparison of hypotheses and actual factors

To be explained	Best natural history hypothesis	Actual factors
Factors underlying separation of cultivar groups A and B	Uniform within each cultivar group for all genetic factors and different between groups on at least some of the factors.	Heterogeneous for 3 of 5 genetic factors within group A and for a different 3 of 5 within group B. There is no genetic factor for which A and B are both uniform within the group and different from one group to the other.
Factors underlying different responses of cultivar group A to location1 vs. location 2	Environmental factors 2 and 3 differ from location 1 to 2.	Partially as hypothesized, but some of the difference is due to the three non-uniform genetic factors modulated by environmental factors 1, 4 & 5.
Factors underlying different responses of cultivar group B to location1 vs. 2	Same as above: Environmental factors 2 and 3 differ from location 1 to 2.	The factors are not the same as above: Some of the difference is due to two non-uniform genetic factors modulated by environmental factors 2 and 3; some to the other non-uniform genetic factor 5 modulated by environmental factor 5.

The last part of the thought experiment is to confirm that the discrepancy between hypotheses and actual factors would be worse if cultivars were not grouped by similarity over locations, say, 1 and 2 in group A and 3 and 4 in group B. I leave this as "an exercise for the reader," but want to note that this second grouping yields very high heritability estimates for each cultivar group within any one location (ranging from .86 to .98).

If the actual values of the genetic and environmental factors were known, a second thought experiment could be envisaged in which regression analyses was compared with the actual factors. There are, however, too few observations in data set 2 for meaningful estimation of more than a few regression coefficients to be made.

3.6 A special case in which heritability, AOV, and measurable factors can be related

The variability of measurable genetics factors is often assumed to be related to heritability, although the latter concept is related to difference-in-effects causes (theme 6). To bring attention to the difference between the concepts, let me consider the special conditions in which systematic relationships between heritability and measurable factors could be derived. (This section is somewhat technical and is not essential for understanding the sections to follow; some readers may choose to move ahead to the sections that review the plausibility of genetic explanations of differences between group means.)

Let us imagine that genetic as well as environmental factors can be measured and vary among replicates. The appropriate regression model for a cultivar-location combination would be an extension of eqn. 6:

$$y_{ijk} = m_{ij} + \Sigma_p b_{jp} g_{ikp} + \Sigma_q b_{iq} e_{jkq} + \Sigma_p \Sigma_q \, b_{ijpq} g_{ikp} e_{jkq} + \varepsilon_{k:ij} \qquad (9a)$$

To increase the number of observations on which the regression analyses are based, the models for separate cultivars and locations could be consolidated into one analysis for each cultivar group -location group combination—provided that the cultivar and location groups are defined on the basis of clustering so the homogeneity assumption is reasonable (theme 4):

$$y_{ijk} = m_{IJ} + \Sigma_p b_{Jp} g_{ikp} + \Sigma_q b_{Iq} e_{jkq} + \Sigma_p \Sigma_q \, b_{IJpq} g_{ikp} e_{jkq} + \varepsilon_{k:ij} \qquad (9b)$$

Suppose that a large number of individuals take one of two values on a genetic factor, p, and similarly for a dichotomous environmental factor, q, but otherwise are treated as identical genetically and environmentally. (Although this is not a standard situation in agricultural research, I consider it because of its relevance to recent research in human behavioral genetics in which the effect of an environmental exposure is conditional on a person's genotype at a specific locus; Moffitt et al. 2005.) If we equate these values with belonging to one of two cultivars in a cultivar group and one of two locations in a location group, we can simplify model 9b:

$$y_{ijk} = m_{IJ} + b_{Jp} g_{ikp} + b_{Iq} e_{jkq} + b_{IJpq} g_{ikp} e_{jkq} + \varepsilon_{k:ij} \qquad (10)$$

If each i,j (or p,q) combination is assumed to have identical numbers of observations and the values of the genetic and environmental factors are scaled to be ±1, the regression analysis using model 10 produces estimates of b_{Jp}, b_{Iq}, b_{IJpq} identical to estimates for c_1, l_1, cl_{11} derived from an

AOV (within the IJ cultivar group-location group combination). Heritability within one location can be estimated from the regression estimates:

$$(\sigma^2_c + \sigma^2_{cl}) / (\sigma^2_c + \sigma^2_{cl} + \sigma^2_\varepsilon)$$
$$= (c_1^2 + cl_{11}^2) / (c_1^2 + cl_{11}^2 + \sigma^2_\varepsilon)$$
$$= (b_{Jp}^2 + b_{IJpq}^2) / (b_{Jp}^2 + b_{IJpq}^2 + \sigma^2_\varepsilon) \quad (11a)$$

Heritability across both locations can be estimated by:

$$\sigma^2_c / (\sigma^2_c + \sigma^2_l + \sigma^2_{cl} + \sigma^2_\varepsilon)$$
$$= c_1^2 / (c_1^2 + l_1^2 + cl_{11}^2 + \sigma^2_\varepsilon)$$
$$= b_{Jp}^2 / (b_{Jp}^2 + b_{Iq}^2 + b_{IJpq}^2 + \sigma^2_\varepsilon) \quad (11b)$$

If $b_{Jp}^2 + b_{IJpq}^2$ is negligible, the within-location and across-location heritability estimates from the AOV will be similar. If the coefficients are not negligible, the across-location heritability will be smaller than the within-location estimate.

This equivalence suggests an analog of heritability for measurable factors defined by the last terms in eqns. 11a and 11b—"mf-heritability" we might call such estimates. The same formulas might be extended to cases in which the genetic and environmental factors are continuous, rather than dichotomous. The downside of the mf-heritability formulation is that the requisite assumptions might be accepted without question, namely, the homogeneity required to consolidate the separate regressions (from eqn. 9a to 9b) and the negligible variation in genetic and environmental factors other than p and q. Moreover, the mf-heritability estimates might be used beyond the cultivar group-location group combination (IJ) in which they were estimated—"After all," it might be presumed, "if the same factors are measurable in a different population and location, surely they will function in similar ways in that population too, so heritabilities and mf-heritabilities should be similar." This, in turn, might predispose people to think that heritability estimated in one location can be extended to apply across locations or used to explain differences between means of cultivar groups grown in separate locations. The sections to follow will show that such extensions are not justified.

3.7 Plausibility reviewed

Based on the discussion of agricultural crop trials, it is now possible to question the two lines of thinking presented in the introduction in which high heritability is held to bolster the

plausibility of genetic explanations of differences between group means. These lines of thinking will be translated into agricultural analogs for the purposes of this review.

Common theme in both lines of thinking: High heritability of measurements within a cultivar group does not on its own allow us to conclude that the difference between mean measurements for the groups grown in different locations also has a high heritability.
Review: This theme can be understood in terms of the independence of the variance components, σ^2_c and $\sigma^2_{c:C,l}$, which enter the formula for within-group, within-location heritability, and the variance components, σ^2_C , σ^2_l and σ^2_{Cl} , which enter the formula for the heritability of the difference between mean measurements for the cultivar groups in different locations. (It should be noted that the latter heritability is very close to 1 unless σ^2_ε is very large.)

Line of thinking 1. High within-group heritability suggests that it will be possible for researchers to find direct effects on measurements of genotypes or effects of environments induced by such genotypes.
Review: High heritability of a measured attribute within a cultivar group in one location does not indicate that researchers will be able to find direct effects of genotypes on the attribute or effects of environments induced by such genotypes. Heritability is a measure, derived from a given data set, of predictability of outcomes if the range of cultivars and other conditions that produced the data set were repeated or rerun. As such, heritability does not reflect underlying causes related to the dynamics of gene-based reproduction or to the organisms's subsequent development responding to environmental factors (themes 5 and 6).

Consider the high within-location heritabilities for cultivar group B in Table 5 and for both cultivar groups under the second grouping (as noted at the end of section 3.5). In the model producing these data (eqn. 8) the effects of genetic factors cannot be meaningfully separated from the effects of environmental factors that modulate them, and the effects of individual genetic or environmental factors on cultivar-group differences cannot be isolated from each other. It should not be surprising that heritability decreases when estimated over more than one location—this follows from the formulas used to calculate heritability—or that heritability changes with the mix of cultivars and locations—such conditionality is to be expected for a measure of rerun predictability, which is derived from an AOV.

Heritability is related to the generation of hypotheses about causes of differences in crop trials, but not in a positive way. The heritability within a cultivar group and within a location group (or location) is lower if the cultivars are grouped so that the response of a cultivar group member (or response elicited by a location group member) is similar to the average for the group as a whole. When groupings are made arbitrarily or are based on criteria not derived from the data, the heritability will be higher but it will be harder for researchers to generate hypotheses about what environmental factors present in the locations in any particular location group elicited the responses from the cultivars in a particular cultivar group. In any case, hypotheses formulated after observational crop trials need to be subject to testing in experimental crop trials and the results become one component of the larger project of unraveling the biophysical pathways of the plant growth and development (developmental causes) and exposing the ways these pathways are affected by the different genetic makeup of cultivars and the different environmental factors in the locations (measurable factors).

Finally, it is only in special conditions that systematic relationships can be derived between heritability and regression analyses in relation to measurable factors. The cultivars and locations need to have been grouped by clustering and the relations should be viewed as conditional, holding within the specific combination of cultivars and locations where the relations are derived, not across into other groups of cultivars or locations (sect. 3.6).

Line of thinking 2a. If the differences are not caused by genetic differences, then they must be caused by environmental differences. Yet, all environment-only explanations that have been tested have been disproved.

Review. There is a false dichotomy in the first sentence. Differences in cultivar group mean yields could be caused by a combination of genetic and environmental differences, whether we are thinking about causes at the level of difference-in-effects (model 3), measurable factors (model 6a), or developmental causes (model 8 and the generic model in Fig. 1). This would, I believe, be readily acknowledged by anyone promoting this line of thinking. There is, however, a deeper conceptual problem: the second sentence refers to explanations in terms of measurable factors, while heritability, which is derived from AOV of observations, relates to difference-in-effects causes. The AOV of observations is not very helpful in exposing causes at the level of measurable factors (let alone in helping to stimulate models of the dynamics of development).

After the appropriate simplification of the data set by clustering, the AOV of observations in crop trials can contribute to the formulation of hypotheses that may be subject to experimental trials to examine the effect of varying specific environmental (and perhaps genetic) factors. If data are available for measurable genetic and environmental factors, regression analysis can provide support for models in the form of measurable factors (e.g., model 6a) or a mix of effects and measurable factors (e.g., model 6). Notice, however, that regression analysis of environment-only models entails consolidation of the analyses for separate cultivars. Recalling theme 4, such consolidation could limit the fit of these models even if environmental factors had a direct influence on the yields for individual cultivars.

Lines of thinking 2b & 2c. High heritability means that the fraction of variation in measurements within a group that is associated with environmental variation is low. Therefore, the number of SDs of change in the environment that would be necessary to produce a 1 SD gap between the means for the groups grown in different locations is too large to be solely due to any known environmental causes. Any environmental "X-factor" that explains cultivar group differences must vary hardly at all within cultivar groups. No known environmental factor operates in such a fashion.

Review of 2b & 2c. High heritability of a measured attribute within a cultivar group in one location means that the residual variance component (σ^2_{ε}) is low in relation to within-group, within-location cultivar variance component ($\sigma^2_{c:C} + \sigma^2_{c:C,l}$). Suppose that the size of the residual variance component when the replications are randomly assigned (i.e., model 3) is taken as an upper limit to the expected size of the sub-location variance component in a hypothetical trial where replications are systematically assigned to multiple sub-locations. This sub-location variance component is not, however, the same as the variance of the effect in an AOV of a trial in which some environmental factor is systematically varied, which is not the same as the variance of such an environmental factor itself (theme 7e).

Let me drive the point home perhaps further than is necessary. Imagine that the distinctions among different construals of "environmental variance" (theme 7e) were ignored. Even in this thought experiment it would not be justified to extend results from within a group to between means of groups in different locations. This follows from noting that within-group

variance components are independent of between-group-mean variance components in the appropriate AOV based on model 3. Similarly, for the "X-factor" argument.

3.8 Numerical examples to reinforce the argument against plausibility

Despite the preceding logical and methodological arguments, I can anticipate requests to be shown a realistic case in which, in the absence of an unrealistic environmental factors, the values of the treatment effect variances from an AOV (and heritability estimates based on them) and variances of measurable genetic and environmental factors were disparate. Let me provide two numerical examples that reinforce the preceding critical review of the second line of thinking.

1) Consider Table 3, in which Data Set 1c is divided into two arbitrary groups. The within-cultivar-group by location interaction variance component ($\sigma^2_{c:C,l}$) is comparable to the location variance component (σ^2_l) at the same time as key features for IQ test scores are found, namely, within-cultivar-groups, within-location heritabilities are high and the difference between the means across cultivar groups and locations is substantial in relation to the standard deviation (SD) for the data set as a whole. For data set 1c, the difference between means of 1.0 is 65% of the SD for the data set as a whole; for data set 2 the difference between means is 91% of the SD.

2) The first example does not refer to measurable environmental factors for which variances could be calculated. Consider then following variant of model 8:

$$y'_{ijk} = \Pi_r (g_{ir}{}^{e_{jrk}}) \qquad (12)$$

where g, and e denote genetic and environmental factors (with random noise built into the latter),

g_{ir} = 1 or $1+\gamma$ with equal probability,

e_{jrk} = $\pm\beta * (1+$random number in interval$(-\delta,\delta))$ with $\pm\beta$ having equal probability

Table 7 gives an AOV using linear model 3 of one data set generated with $\gamma = .8$, $\beta = .5$, $\delta = 1.25$, r=1,...,5, again scaled using eqn. 8a so the data the same mean and SD as data set 1c ($constant_1 = -2.19$, and $constant_2 = 3.87$). Table 8 presents the variances of the genetic and environmental factors. The square root of the average within-group or within-location variances isgiven to allow comparison to the corresponding gap between cultivar-group and location means.

Table 7

Data Set 3 divided into two arbitrary groups				**Estimates of effects**		**Variance components & heritability estimates**	
	location	1	2	m	3.0	σ^2_C	0.01 (1%)
Cultivar Group	cultivar			l_1	0.78	$\sigma^2_{c:C}$	1.12 (48%)
A	1	2.1, 1.6	0.8, 2.2	l_2	-0.78	σ^2_l	0.61 (26 %)
A	2	6.3, 5.9	2.7, 3.4	C_A	0.12	σ^2_{Cl}	0.00 (0 %)
B	3	5.3, 3.4	1.8, 2.6	C_B	-0.12	$\sigma^2_{c:C,l}$	0.29 (13%)
B	4	2.2, 3.6	2.1, 2.1	$c_{1:A,}$ $c_{2:A,}$	±0.37	σ^2_ε	0.28 (12 %)
				$c_{4:B}$, $c_{3:B}$	±1.45		
				Cl_{A1}, Cl_{B2}	-0.07	$h^2_{\text{within cultivar group within location}}$	0.84
				Cl_{A2}, Cl_{B1}	0.07	$h^2_{\text{within cultivar group across both locations}}$	0.49
				$cl_{i:A,j}$	±0.34		
				$cl_{i:B,j}$	-/+0.69		
				$\varepsilon_{k:ij}$	varied		

Table 8

						envtl fact-ors	Location, replication 1,1	1,2	2,1	2,2	square root of mean b/w replicat-ion var.	differ-ence b/w location means
						1	0.03	0.77	0.00	0.01	0.26	0.40
						2	-0.28	-0.89	-0.61	-0.37	0.23	-0.09
						3	-0.73	-0.44	-0.92	-0.55	0.17	0.15
		genetic factors				4	0.91	0.45	0.64	-0.12	0.29	0.38
		1	2	3	4	5	0.97	0.91	0.30	0.95	0.23	0.32
cultivars	1	1	1.8	1.8	1	1.8	2.1	1.6	0.8	2.2		
	2	1.8	1	1	1.8	1.8	6.3	5.9	2.7	3.4		
	3	1.8	1.8	1	1.8	1.8	5.3	3.4	1.8	2.6		
	4	1.8	1	1	1	1	2.2	3.6	2.1	2.1		
square root of mean b/w cultivar var.		0.28	0.40	0.28	0.40	0.28						
difference b/w cultivar group means		-0.40	0.00	0.40	0.00	0.40						

This example may not be typical—there is considerable variation among data sets generated by model 12 even with the same values of the parameters. Yet, there is nothing about model 12 that renders the example unrealistic. This case shows that it is possible to produce high within-cultivar group, within-location heritability values without any systematic difference between the two cultivar groups in the values of the genetic factors (see bottom two rows on the left in table 8). Moreover, the between-location gap for the environmental factors is comparable to the within-location (between replicates) standard deviation (top right two columns in table 8).

3.9 Partial data sets and their implications, I. AOV

Let me now consider a less-than-ideal scenario that will become important when translating from the agricultural situation to human behavioral genetics. Suppose that there can be only two replicates of any cultivar in an observational trial. If cultivars are measured in two locations, there is no within-location replication and thus heritability within any one location is always 1 (see eqn. 5). Nothing interesting can be inferred from that. If cultivars are measured in one location only, say, cultivars 1 and 2 are grown in location 1 and cultivars 3 and 4 in location 2, they cannot be grouped by similarity across locations. This limits the formulation of hypotheses about environmental factors (themes 7a & b).

Theme 8—Limited replication: limited hypothesizing. With only two replicates, an AOV of observations provides a limited basis for hypothesizing about measurable factors; any conclusions made on the basis of the AOV must center on difference-in-effects causes.

What could be learned about difference-in-effects causes from AOV of the limited-replication data set? This question can be addressed in two ways, according to two linear models.

a. Nested model.

Consider linear model 13 and the effects estimated under this model (Table 9)

$$y_{ijk} = m + l_j + c_{i:j} + \varepsilon_{k:ij} \qquad (13)$$

where i:j denotes that i is nested within j

and estimates are constrained so that $\Sigma_j l_j = 0$; for each l, $\Sigma_{i:l} c_{i:l} = 0$; and for each i, $\Sigma_k \varepsilon_{ik} = 0$

Table 9

Subset of data Set 1c analyzed using Estimates of effects Variance components &

nested model **heritability estimates**

	location	1	2
Cultivar Group	cultivar		
A	1	5.3,4.3	
A	2	3.1,2.1	
B	3		1.6,2.6
B	4		2.8,3.8

m	3.2
l_1	0.5
l_2	-0.5
$C_{1;1}$	1.1
$C_{2;1}$	-1.1
$c_{3:2}$	-0.6
$c_{4:2}$	0.6
$\varepsilon_{i;j,k}$	+/-.5

σ^2_l	0.25 (19%)
$\sigma^2_{c:l}$	0.785 (61%)
σ^2_ε	0.25 (19%)
$h^2_{\text{within cultivar group A within location 1}}$	0.83
$h^2_{\text{within cultivar group B within location 2}}$	0.59
$h^2_{\text{within cultivar group A across both locations}}$	not applicable
$h^2_{\text{within cultivar group B across both locations}}$	not applicable

The interpretation of these effects, however, is not the same as for the effect with the same symbol in models 1 or 3.

Theme 9—Nested effects. The existence of a significant cultivar-nested-within-location effect does not tell us whether the average for a cultivar is generally higher than another or whether it is only higher when paired with the particular location. Similarly, a significant location ("l") effect does not tell us whether one location is superior to the other, only that the cultivars grown in one location are different, on average, from the cultivars grown in the other location.

The superiority of the average of cultivars 1 and 2 in location 1 over the average of cultivars 3 and 4 in location 2 does not address any more general question. To put this in other words, the model allows no estimation of effects of the cultivars in the location in which they were not grown, or of location effects over all cultivars.

A caution about interpretation of nested analyses along the lines of theme 9 is given in Lindman's textbook (1992) and is illustrated with the example of assessing the dependence of high school students' test scores in algebra on their teacher and school. The students within a school were randomly assigned to a teacher in their usual school. Lindman notes that a significant location (school) effect "is likely to be interpreted as due to differences in physical facilities, administration, and other factors that are independent of the teaching abilities of the teachers themselves… [However, d]ifferences between teachers in different schools are part of the [location or school] effect, and the observed differences between schools could be due

entirely to the fact that some schools have better teachers [or] some schools have smarter children attending them" (Lindman 1992, 194).

If regression analysis is performed, equations 6 or 6a can still be used, but it should be remembered that the regression coefficients have a different meaning, now referring to a cultivar nested in a location. Extrapolation to other locations is mathematically straightforward, but not justified on the basis of underlying developmental causes or measurable factors, especially since the groups cannot have been formed on the basis of similarity across locations (sect. 3.2).

b. Grouped model.

A second way to analyze the partial data set is to use linear model 3. Some of the effects can still be estimated but in composite form (Table 10).

Table 10

Subset of data Set 1c divided into two arbitrary groups				**Estimates of effects**		**Variance components & heritability estimates**	
	location	1	2	m	3.2	$\sigma^2_{(C+I+Cl)}$	0.25 (19%)
Cultivar Group	cultivar			$I_1 + C_A + Cl_{A1}$	0.5	$\sigma^2_{(c:C + c:C, I)}$	0.785 (61%)
A	1	5.3,4.3		$I_2 + C_B + Cl_{B2}$	-0.5	σ^2_{ε}	0.25 (19%)
A	2	3.1,2.1		$C_{1:A} + cl_{1:A,1}$	1.1		
B	3		1.6,2.6	$C_{2:A} + cl_{2:A,1}$	-1.1	$h^2_{\text{within cultivar group A within location 1}}$	0.83
B	4		2.8,3.8	$c_{3:B} + cl_{3:B,2}$	-0.6	$h^2_{\text{within cultivar group B within location 2}}$	0.59
				$c_{4:B} + cl_{4:B,2}$	0.6	$h^2_{\text{within cultivar group A across both locations}}$	not estimable
				$\varepsilon_{k:ij}$	+/-.5	$h^2_{\text{within cultivar group B across both locations}}$	not estimable

The values correspond directly to effects from the nested model (Table 7) and Lindman's caution applies equally well to these composite effects. The composite estimates are not, however, directly translatable to the effects derived from AOV of the full data set 1c (Table 3).

Let me make some other observations about the comparison of the analyses of the partial and full data sets. The partial data set is a subset of the full data set, which means that nothing about the underlying causal structure of the world that produced full data set has changed. Quantitative and qualitative changes in the effects are simply further illustration of the conditionality of effects in an AOV (theme 2). Consider, in particular, the effects that make up

the difference between the mean for cultivar group A in location 1 and the mean for cultivar group B in location 2. According to the nested analysis

$$Mean_{A1} - Mean_{B2} = l_1 - l_2 \quad (14a)$$

while in the grouped analysis (where l has a different value)

$$Mean_{A1} - Mean_{B2} = C_A - C_B + l_1 - l_2 + Cl_{A1} - Cl_{B2} \quad (14b)$$

In this case there are only two groups of cultivars and two locations, so many effects come in equal and opposite pairs, e.g., $C_A = -C_B$, etc., which means eqns. 14a & 14b simplify to

$$Mean_{A1} - Mean_{B2} = 2l_1 \quad (15a)$$

$$Mean_{A1} - Mean_{B2} = 2C_A + 2l_1 + 0 \quad (15b)$$

In the nested analysis, whatever difference there may be between group A cultivars and group B cultivars is subsumed in the location effects. Equivalently, in the grouped analysis, there is no way that data comparing A1 and B2 can disentangle the two effects remaining in eqn. 15b – the same difference could result from a positive C_A and zero l_j, or from negative C_A and positive l_j, and so on.

Furthermore, there is no way that data from comparing A1 and B2 can indicate the size of Cl_{A1}, which could be very large or very small or somewhere in between. An estimate of this effect is needed if we wanted to predict the difference between the means of the two cultivar groups in the same location, e.g.,

$$Mean_{A1} - Mean_{B1} = C_A - C_B + l_1 - l_1 + Cl_{A1} - Cl_{B1} \quad (16a)$$

which simplifies in the case of two locations to

$$Mean_{A1} - Mean_{B1} = 2\,C_A + 0 + 2Cl_{A1} \quad (16b)$$

Equations 14-16 do not include effects related to within-group variation (i.e, $c_{i:j}$ and $\varepsilon_{k:ij}$ in the nested analysis or $c_{i:I}$ and $cl_{i:I,j}$ and $\varepsilon_{k:ij}$ in the grouped analysis), so <u>within-group</u> variance components and heritability estimates can have no relationship to the quantities needed to estimate the changes in cultivar-group mean, location, and cultivar-group-mean-by-location-interaction effects that would be needed to shift one mean to the level of the other.

The preceding observations have further implications for the two lines of thinking in which high heritability is held to bolster the plausibility of genetic explanations of differences between group means.

Theme 10—Partial data sets and inseparable effects. The within-group estimates that can be derived from the partial data set cannot speak to the relative contribution to differences between cultivar-group means in different locations of differences in cultivar-group effects versus location effects versus cultivar-group-mean-by-location-interaction effects.

If some differences in cultivars and/or locations were hypothesized to explain the difference between cultivar group means in different locations, these differences should encompass all these differences-in-effects. In any case, the logic of the arguments made in section 3.7 against lines of thinking 2b & c still holds (i.e., concerning the different construals of "environmental variance" and the independence of within-group variance components and between-group-mean variance components). Note also that, in the case of two locations, because no estimate at all can be made of the cultivar-group-mean-by-location-interaction effect, no answer could ever be derived from AOV of the partial data set to the agricultural analog of Flynn's reversal of fortunes question (sect. 2), i.e., how would cultivar group B's mean change if grown in location 1, not location 2.

3.10 Partial data sets and their implications, II. Regression

The problem of inseparable effects (theme 10) cannot be overcome by regression analyses in relation to measurable environmental factors. To show this, let me examine standard regression analyses from the perspective of this paper, namely, one in which we envisage multiple cultivars grown in multiple locations, but observe a nested subset in which any cultivar is grown only in one of two locations.

When environmental factors can be measured for each location and replicate, the standard regression analysis that allows comparison of cultivars in two cultivar groups involves a model using a dummy variable to designate the average effect of cultivar group membership:

$$y_{ijk} = m + d_I + \Sigma_q b_q e_{jkq} + \varepsilon_{k:ij} \qquad (17a)$$

where d_I denotes the dummy-coefficient for group I (set to zero for the reference group) Alternatively, separate regression analyses can be performed:

$$y_{ijk} = m_I + \Sigma_q b_{Iq} e_{jkq} + \varepsilon_{k:ij} \qquad (17b)$$

Now consider the full observational trial, but again imagine that genetic factors can be measured and vary among replicates, so that the appropriate regression model is eqn. 9a. or, after

consolidation to increase the number of observations, eqn. 9b. Eqn. 9b can be used to compare a cultivar (i) from group A grown in location 1 with a different cultivar (i') from group B grown in location 2 (i.e., location groups are of size 1). The difference is given by:

$$y_{i1k} - y_{i'2k'} = (m_{A1} - m_{B2}) + \Sigma_p b_{1p} g_{ikp} - b_{2p} g_{i'k'p} + \Sigma_q [e_{1kq}(b_{Aq} + \Sigma_p b_{A1pq} g_{ikp}) - e_{2k'q}(b_{Bq} + \Sigma_p b_{B2pq} g_{i'k'p})] + \varepsilon \quad (18)$$

Clearly, genetic and environmental factors are intertwined. Table 13 compares the difference in eqn. 18 with those derived from regression analyses 17a and 17b, respectively, namely:

$$y_{i1k} - y_{i'2k'} = - d_B + \Sigma_q b_q (e_{1kq} - e_{2k'q}) + \varepsilon \quad (19a)$$

$$y_{i1k} - y_{i'2k'} = m_A - m_B + \Sigma_q (b_{1q} e_{1kq} - b_{2q} e_{2k'q}) + \varepsilon \quad (19b)$$

Table 13

Regression analysis based on model			**Comparison**
18	19a	19b	
$m_{A1} - m_{B2}$	$- d_B$	$m_A - m_B$	coefficient specific to cultivar group-location combinations *subsumed in* a coefficient related only to cultivar group membership
$\Sigma_p b_{1p} g_{ikp} - b_{2p} g_{i'k'p}$	part of ε	part of ε	location-specific coefficient related to differences in genetic factors *subsumed in* the residual
b_{Aq}	b_{1q}	b_q	coefficient specific to cultivar group together with a coefficient related to differences in genetic factors (but specific to cultivar group-location combinations) *subsumed in* a coefficient related to location only or a coefficient independent of cultivar and location
$\Sigma_p b_{A1pq} g_{ikp}$			
Similarly for cultivar group B and location 2			

This comparison complicates the objection that environmental factors may be, in part, surrogates for genetic factors (see end of sect. 5.2). From the perspective of this paper, misattribution is more than a matter of correlations between measurable genetic and environmental factors. It is thoroughgoing and this is to be expected in light of the inseparability of effects in AOV of nested and partial data sets (themes 9 & 10). Moreover, if homogeneity is questionable, the

consolidation from eqn. 9a to 9b cannot be justified and the coefficients in a standard regression analysis become even harder to interpret in terms of distinct genetic and environmental factors.

Of course, the methodology of regression analysis allows us to use models 17a or 17b without attention to nesting or partiality of the data, and in cases where there is no replication. The results may stimulate experimental trials (sect. 3.4c), but cannot be linked to heritability estimates within or across locations.

4. Human behavioral genetics and differences in IQ test scores

Human behavioral genetic analysis of IQ test scores and other traits departs from the ideal agricultural case in significant ways. Agricultural researchers have the kind of control that allows them in observational trials to replicate cultivars over many locations (and over time) and to test many cultivars in the same location. In experimental trials they can also vary specific environmental factors. Human behavioral genetics does not have such control over the genetic types or environmental factors. The maximum number of replicates is two (and this figure assumes that there are no departures from randomness when identical twins are separated and assigned to sub-locations).

I have called into question, even for the ideal, agricultural case, the lines of thinking in which high heritability is held to bolster the plausibility of genetic explanations of differences between group means. The use of partial data sets in human behavioral genetics has the further effect of restricting heritability estimation to within-location heritability and of limiting the formulation of hypotheses to test about measurable environmental or genetic factors. Given those limitations, discussion should be limited to difference-in-effects causes and as such can shed no light on the question of how whites would score on IQ tests if raised in the environments that blacks experience and vice versa (even if we assume that this reversal of fortunes could be brought about without changing any other social conditions and without knowing the developmental causes of or measurable factors associated with IQ test scores for individuals in different racial groups). Regression analysis cannot escape the limitations of partial data sets for separating the contribution of measurable genetic or environmental factors.

My arguments depend on analyzing data from racial groups as if they grew up in different locations.

Theme 11—Racial groups as separate locations in an AOV. The appropriate model for an AOV when replications—the identical twins raised separately—are not assigned independently with respect to membership in a racial group is one in which the experience of racial group membership is modeled as a separate location.

(In any case, if identical twins were ever to be randomly assigned across racial groups, the experience of trans-racial twins would not be simply that of the racial group in which they were raised.) Any AOV using a model such as 3 (or 13) means that statistics, such as heritability, that are calculated from data in one or the other racial group can provide no insight about the causes of the gap between the two racial groups in the averages over the genetic types within that location.

(My case is based on the logic and methodology of AOV and regression analysis of observational trails. If any reader needs numerical examples to help them visualize the arguments, they should subtract 3.0 from each data point in section 3, multiply by $15/SD_y$, i.e., 9.84, and add 100—the resulting values for the genetic types or cultivars will look very much like values for IQ test scores. The difference between the mean for genetic types 1 and 2 grouped together and genetic types 3 and 4 grouped together would be 15 IQ points for data set 1c and 21 IQ points for data set 2. For the same groupings, the within-cultivar-group, within-location heritabilities are high, ranging from .59 to .83 for data set 1c and from .86 to .88 for data set 2.)

Finally, all the arguments and themes I have introduced apply to apply to differences between mean IQ test scores not only in different racial groups, but also in different generations. Because separated identical twins cannot be assigned across generations, model 3 (or 13) is appropriate for the AOV. Any AOV using that model produces statistics, such as heritability, that are calculated from data in one or the other generation and can provide no insight about the causes of the gap between the two generations in the averages over the genetic types within that generation. The coexistence of high intergenerational gains and with within-generation heritability is not a genuine paradox, even though the Flynn effect remains intriguing and warrants explanation (see sect. 5.3).

The psychologist and behavioral geneticist Arthur Jensen is well known for promoting the second line of thinking identified in the introduction. He has stated that "we are left with… various lines of evidence, no one of which is definitive alone, but which, viewed all together,

make it a not unreasonable hypothesis that genetic factors are strongly implicated in the average Negro-white intelligence difference" (1969, 80). However, according to my account there are no grounds for the reasonableness of this hypothesis that can be found in the concept of heritability and the statistical Analysis of Variance on which it is based. Perhaps of more importance, heritability and AOV cannot shed light on the changeability of differences across racial groups or generations.

It is beyond the scope of this essay to assess the power of newer techniques in human behavioral genetics, such as mapping Quantitative Trait Loci (Plomin et al. 2003) and Moffitt et al's (2005) "investigations of measured genes in measured environments." Such work should, I believe, be reviewed in relation to the distinctions and themes of this essay (especially themes 9-11) before anyone presumes that the new behavioral genetics can do better than the old in explaining differences between group means and addressing questions about changeability.

5. Critiques reviewed and questions reconceived

The goal of this essay has been to introduce distinctions and themes that, I hope, help readers visualize more clearly the limited relevance of human heritability estimates for explaining differences between means across groups or across generations. In this section I want to revisit three critiques related to the conventional wisdom about differences between mean IQ test scores for racially defined groups to show how important distinctions can be unintentionally obscured.

5.1 The Analysis of Variance and the Analysis of Causes Revisited

Lewontin (1982, 132-3) introduces two agricultural thought-experiments to help readers visualize why heritability within groups is not relevant to explanation of differences between groups. In one example, two set of seeds sampled from one open-pollinated cultivar are planted in two pots of washed sand. Both pots are fed with plant-growth solution, but the solution in the second pot lacks nitrogen. Reading from the diagram in Lewontin's book and inventing arbitrary units for height of the plants gives us Data set 4, which is analyzed using a variant of eqn.1c (Table 11):

$$y_{jk} = m + l_j + \varepsilon_{k:j} \qquad (20)$$

Table 11

Data Set 4			Estimates of effects		Variance components & heritability estimates	
location	1	2	m	3.15	σ^2_l	1.9525 (75 %)
cultivar			l_1	1.35	σ^2_ε	0.655 (25 %)
1	5, 4, 6, 3	2, 1.8, 2, 1.4	l_2	-1.35		
			ε_{jk}	varied	h^2 within location or across both locations	0

Lewontin observes that, because each location (pot) is uniform, variation within them will be associated with genetic differences among the sampled seeds. Lewontin calls this a heritability of 1, but the correct value is 0 because a new sample of seeds grown in any one location would have no correlation with the first sample. Although all the within-location variance is associated with genetic variability, this variability is error variance in the AOV. (If we thought of the sample of seeds as a set of <u>different</u> cultivars and if we could clone each seed and replant it, then the appropriate model would be eqn. 1c and, in the absence of replications, heritability within locations would be 1.)

Similarly, the heritability across both locations is 0. This value seems consistent with Lewontin's observation that the difference between (the average measurements for) locations is "totally environmental" (i.e., entirely associated with the environmental difference of nitrogen vs. no nitrogen). Consider, however, the rerun predictability when the location is constrained to be the same in both cases—the correlation between the current data and the predicted results is .75. The only way to increase this correlation, that is, for the difference between the average measurements for each location is more strongly associated with the nitrogen difference, is to reduce the within-location or error variance that is associated with genetic differences among those seeds.

Precision about these technical issues helps us to avoid linking heritability, which depends on AOV of observational trials and thus difference-in-effects causes, with the idea of measurable genetic factors that differ among individuals in association with differences among individuals in the trait in question (theme 6; but note the special case in sect. 3.6). As shown in sect. 3, the partial data sets of human behavioral genetics cannot fulfill the conditions under which the AOV of observational trials stimulates hypothesizing about measurable environmental or genetic factors or under which it is reasonable to extrapolate regression coefficients to locations other than those for which they were estimated.

When Lewontin (and many others following him) use his example to make conceptual points, they overlook the control the example presumes over which varieties to interbreed or plant and the ability to replicate environmental conditions. Such control characterizes observational and experimental crop trials (theme 3), but is not available to human behavioral genetics. Moreover, the example blurs the distinction between observational and experimental crop trials (theme 7) as if human behavioral genetics could use experiments to generate knowledge of measurable factors. These oversights, together with the implicit linkage of heritability with measurable genetic factors (theme 6), lend unintended plausibility to lines of thinking in which genetic factors are separable from environmental factors and in which insight about those factors could follow from learning that heritability estimates are high.

In Lewontin's other example (1982, 132), one seed from each of two inbred cultivars is planted in a series of pots of soils taken from different locations. Again, suppose we read from the diagram in Lewontin's book and invent arbitrary units for height of the plants. This can give us Data set 5, which is analyzed (Table 12) using a different variant of eqn.1c:

$$y_{ij} = m + c_i + l_j + cl_{ij} \qquad (21)$$

Table 12

Data Set 5					Estimates of effects		Variance components & heritability estimates	
location	1	2	3	4	m	2.75	σ^2_l	2.5625 (69 %)
cultivar					c_1	.75	σ^2_c	0.5625 (15%)
1	4	7	2	1	c_2	-.75	σ^2_{cl}	0.5625 (15%)
2	3	3	1	1	l_j	.75, 2.25, -1.25, -1.75		
					cl_{1j}	-.25, 1.25, -.25, -.75	$h^2_{\text{within a location}}$	0
					cl_{2j}	$-cl_{1j}$	$h^2_{\text{across locations}}$	.15

Lewontin observes that for each of the inbred cultivars there are no genetic differences across the locations (pots) and all the location-to-location (pot-to-pot) variation must be "environmental," corresponding to differences in the soils. At the same time, noting that cultivar 1 does better than (or as well as) cultivar 2 in each location, Lewontin asserts that this gap is entirely "genetic" because the cultivars experienced identical sets of locations.

The meaning of the terms "genetic" and "environmental" are ambiguous (see themes 6 and 7d) in this context. Again, it is instructive to be precise. First, notice that the advantage of cultivar 1 over cultivar 2 varies from one location to the next. In terms of AOV and difference-in-effect causes, there is a cultivar effect (i.e, a non-zero gap between the means across all locations of cultivar 1 and 2), and there are cultivar-by-location-interaction effects. In terms of measurable genetic and environmental factors, no hypotheses are obvious. What we can infer, however, is that the varying within-location, between-cultivar differences are associated either with a different mix of genetic factors in different locations, or with the same genetic factors having a different influence. The within-cultivar differences across locations may correspond with one or many environmental factors and these factors need not be the same from one cultivar to the next. In short, the best we can say about the gap between cultivars in each location and the gap between locations for each cultivar is that they are associated with a combination of environmental and genetic factors.

In terms of rerun predictability, the absence of replication means that within-cultivar heritability is not a very interesting quantity in this case. Within one location, this heritability must be 1; across locations it must be 0—even if the cultivars are not inbred. On the other hand, taking both cultivars into account, the heritability across locations is .15 and the rerun predictability when the location is constrained to be the same in both cases is .69—not 0 and 1 as might be naively expected if these heritability and rerun predictability measures corresponded to the terms "genetic" and "environmental" as used by Lewontin and others following him. In summary, this example, like Lewontin's other one, blurs distinctions in ways that lend plausibility to lines of thinking in which measurable genetic and environmental factors are separable and in which insight about those factors could follow from learning that heritability estimates (and difference-in-effect causes) are high or low.

In an earlier, much-cited essay Lewontin (1974) argues, in effect, that, because any AOV is conditional (theme 2), it cannot shed light on causes (in the terms of this paper, all three kinds of causes) beyond the local combination of genetic types and locations observed. He supports this argument with diagrams of "norms of reaction" that summarize the response of a cultivar or genetic type when some environmental factor is varied. Norms of reaction for different cultivars that vary in slope and position can confound any attempt to extrapolate the relative ranking of genetic types (or cultivars) observed over part of the range of the environmental factor to the full

range. Some of Lewontin's diagrams are schematic, with the measured trait plotted against an unspecific environmental factor E; one plots real data on viability of strains of fruit flies against temperature. Having a single continuous environmental factor as the horizontal axis in both the schematic and real cases reinforces the idea that location effects in an AOV can be readily translated into environmental factors with continuous gradients (contra themes 1 and 7e). Compared with agricultural breeders, fruit fly breeders have even greater control over genetic types and environmental conditions (theme 3), so they can readily envisage generating such plots. However, using diagrams of norms of reaction to make conceptual points about the AOV and human behavioral genetics steers us away from visualizing the difficulties in using AOV to expose equivalent environmental factors in humans (theme 8).

5.2 Sesardic's critique of critics of genetic determinism

Philosopher of science Neven Sesardic (2000) has argued pointedly that his colleagues need to delve more deeply into the science on which they make their arguments. In particular, they should recognize that the contribution of behavioral genetics to explanation of differences between racial group means in IQ and other test scores rests not on an invalid extension of within group heritability, but on something like the second line of thinking I presented in the introduction. In this light, the validity of extending within group heritability to explanation of differences between racial group means cannot be resolved on logical or methodological grounds, but is an empirical issue—Are there environment-only explanations that have been tested and not disproved? (He thinks not, but see Fryer and Levitt 2004.) Are there environmental factors that show wide variation between racial groups, but narrow variation within groups? Unfortunately for Sesardic's argument, although he may be justified in noting the inattention of critics of genetic determinism to the second line of thinking, that thinking does not stand up to scrutiny (sects. 3 and 4). The shortcomings I describe involve matters of logic and methodology; empirical considerations are beside the point.

Sesardic (2003, 1004) also argues against many critics of genetic determinism in asserting that the concept of heritability, "when properly understood, actually accords well with our common-sense etiological ascriptions." In his account, "heritable" and "genetic" are synonymous, and "genetic" means differences in a trait are associated with genetic differences. This formulation obscures the technical meaning of heritability, which refers not to differences in

measurable genetic factors, but to difference-in-effects causes as derived from AOV of observational data (theme 6; but note the special case in sect. 3.6).

However, for the sake of argument, let me put aside the distinction between heritability and "genetic" in order to follow Sesardic's rebuttal of the criticism that "genetic" is improperly ascribed to causes that involve both measurable genetic and environmental factors. Following Plomin (1977), Sesardic distinguishes three forms of association between genetic and environmental factors:

> Passive, in which parents contribute both genetic and complementary environmental factors to their children;
>
> Reactive, in which people provide environmental factors in response to the genetic factors of the children so as to amplify the effects of the genetic factors; and
>
> Active, in which children seek environmental factors that amplify the effects of their genetic factors.

Sesardic notes that behavioral geneticists do not refer to the second case as genetic and can use adoption studies to separate the contributions of genetic and complementary environmental factors in the first case. This leaves the third case—active association; to call this "genetic" accords well, Sesardic contends, with our common-sense.

I see two kinds of problem with Sesardic's treatment of associations between genetic and environmental factors. Although it is easy to identify processes through which people respond to observed traits or through which children's traits lead them to seek out certain environmental factors, it is more difficult to envisage mechanisms through which people link environmental factors to the genetic factors, rather than to traits that the people can observe. What evidence is there for assuming that such a mechanism could exist? More importantly given the thrust of my paper, researchers need methods of data analysis that can discriminate among competing models. It is difficult, in the absence of knowledge about the pathways of growth and development of traits that influence IQ test scores (developmental causes), to design a method that discriminates between reactive and active associations between genetic and environmental factors. This difficulty only increases if the data are partial, i.e., genetic types are observed in different locations (sect. 4).

Theme 12— When alternative models based on different dynamics are proposed, a method is needed that not only compares the models' fit with observations but also assesses the support for their assumptions independent of that fit.

Overall, Sesardic endorses Jensen's conclusion (cited in sect. 4), which leads me to note an asymmetry in their rhetoric and conceptualization of the debate over genes, race, and IQ test scores. Behavioral geneticists point to the failure of environment-only explanations to fully account for the racial-mean gap, but insist that they are not wedded to gene-only explanations. Instead, they "hypothesize" that genes are "strongly implicated" and claim that an environmentalist orthodoxy has held social scientists from considering this possibility (e.g., Pinker 2002). This formulation invites a symmetrical rejoinder: Opponents could propose that environmental factors are strongly implicated in the gap and point to the striking failure of gene-only explanations (in the sense of significant associations of between-group-mean differences in IQ test scores and measurable genetic factors related to degree of African ancestry; see summary in Nisbett 1998, 89-90). Would behavioral geneticists entertain an ideological interpretation of their tendency to discount the standard environmental factors of social scientists? This seems a fair question given that behavioral geneticists have made little headway in connecting high heritability to measurable genetic factors within groups and no success in connecting measurable genetic factors to differences on average between racial groups. Why have they continued to highlight genetic contributions to behavior and the role of environments induced by genes, which includes their hypothesizing that sociological variables are surrogates for genetic factors (e.g., socioeconomic status that parents confer on their children may reflect abilities or limitations that are influenced by genetic factors related to, say, mental illness or intellectual abilities) (Plomin and Bergeman 1991; Plomin et al. 2003)?

A symmetrical conclusion would be that sociological regression models and accounts of changes over time cannot show the equality on average of genetic factors across races (Flynn 2005), and the methods of behavioral genetics cannot show the inequality on average of genetic factors across races. It will be interesting to observe, then, which side takes up the challenge of deriving empirical models of developmental pathways whose heterogeneous components differ among individuals at any one time and over generations.

5.3 Dickens and Flynn's Reciprocal Causation Models Revisited

Dickens and Flynn's (2001) reciprocal causation models are elaborations of a two-part linear model that differs in many ways from the linear models used in AOV:

$$y_{it} = G_i + E_{i,t-1} \quad (22a)$$

where

y_{it} denotes IQ test score at time t (measured in steps within a lifespan),

G_i an unchangeable genetic endowment, and

$E_{i,t}$ the environment experienced by individual genetic type i at time step t, which matches the observed value of the trait as follows:

$$E_{i,t-1} = \text{constant}_1 * y_{i,t-1} + \text{constant}_2 + \varepsilon_{i,t-1} \quad (22b)$$

Some data sets exist in which traits are measured at various steps in the lifespan (e.g., Medical Research Council 2004) so in theory estimating the G_i effects and the constants in this model would be possible, even though computing their values would be more difficult than a standard AOV. Data analysis is not, however, the focus of Dickens and Flynn's exposition. Instead, their argument is that, if there were environmental factors that changed over the lifetime in response to the trait of IQ test scores, then their model could produce results that mimicked observed features of those scores, especially the high within-group heritability and large gains between generations. According to my arguments in the previous sections, the coexistence of these observed features is not paradoxical and many other models, including eqn. 8, can produce both the observed features. The challenge then is to discriminate among alternative models based on a broad array of possible dynamics. For this, we would need to compare not only the models' fit with observations but also the support for their assumptions independent of that fit (theme 12).

Economists—Dickens is an economist by training—often take fit with a number of key features of the observations as confirmation of a model (Friedmann 1953), but I am convinced by philosophers of science who insist on independent support for the assumptions built into any model (Taylor 2000). Consider in contrast to model 22:

$$y_{it} = E_{i,t-1} \quad (23a)$$

$$E_{i,t-1} = \text{constant}_1 * y_{i,t-1} + \text{constant}_2 + \varepsilon_{i,t-1} \quad (23b)$$

where $y_{i0} = G_i$

It could be that model 22 fits certain patterns in observed data better than model 23, but model 23 does not require the assumption of something unchangeable that enters at every time step into the development of traits that influence IQ test scores. Following my questioning of Sesardic, what evidence can Dickens and Flynn provide for a mechanism that links every point in current development directly to genetic endowment?

Similarly, we could question the assumption that environmental factors can be packaged into one continuous, measurable gradient, E. In the context of widespread discussion referring to "genes versus environment" or "genes interacting with the environment" there is a certain conceptual ease in envisaging such a gradient. However, such an assumption reinforces, as noted in section 5.1 for Lewontin's norms of reaction diagrams, the idea that location effects in an AOV can be readily translated into environmental factors with continuous gradients (contra themes 1 and 7e). It steers us away from visualizing the difficulties in using AOV to expose environmental factors in humans (themes 8-11). Collapsing the conceptual distinctions among the different quantities referred to as "environmental variance" (theme 7e) (and correspondingly, the distinction between difference-in-effect causes and measurable factors) is key to the second line of thinking in which high heritability is held to bolster the plausibility of genetic explanations of differences between group means. I hope that some readers now see this line of thinking as problematic.

In the introduction I remarked that Dickens and Flynn's contribution has the potential to move the debate about heritability and differences between racial-group means onto fresh ground. Let me now qualify this remark. Such discussion should take into account the themes I have introduced in this essay and distinguish between developmental causes, measurable factors, and difference-in-effects causes, which means that there is no heritability paradox that reciprocal causation modeling resolves. Yet, even if human heritability estimation is put to the side, the Flynn effect still needs explanation. I believe that investigation of differences between means of different generations that incorporates reciprocal causation in its models has the potential to move the debate about differences between racial-group means onto fresh ground.

Flynn (2005) has begun to consider how heterogeneity complicates the interpretation of data about change and difference:

> I can suggest no common metric to measure the black environment of today against the white environment of some 50 years ago… [remainder of passage omitted until author gives go ahead for quoting it]

Once it is recognized that the potency of social multipliers depends on different groups' capacity to capitalize on historical changes in society, there is no reason to assume that they apply uniformly across individuals, given their differences in age, gender, geographical location, culture, and so on, or even that they move different individuals in the same direction but at different speeds. To adapt Dickens and Flynn's basketball analogy, TV coverage of basketball elicited greater participation in basketball at the same time as it elicited more "couch potato" spectatorship.

The next step is to envisage social multipliers operating heterogeneously across social groups within any generation and heterogeneously across individuals within any social grouping. It is beyond the scope of this paper to review research already advancing along these lines. Woodhead's (1988) review, however, sets the scene by summarizing studies explaining how the IQ test score increases produced by Head Start preschool programs tend to be transient, but in the long term, through social support systems initiated or enhanced during the Head Start years, the children end up with significantly higher high school graduation rates, employment, and many other socially valued measures.

If pathways of reciprocal causation are heterogeneous, this weighs against analyses that employ gross categories, such as racial group genes and environment, but it does not rule out quantitative analysis. We might take the lead from the innovative attempt of Kendler et al. (2002) to produce a comprehensive developmental model for major depression in women. They recognized that major depression is an etiologically complex disorder, which required "consideration of a broad array of risk factors from multiple domains," but their model was able to account for 52% of the variance in liability to episodes of major depression and to characterize different pathways to the outcome to be explained, namely, depression. I find it plausible likewise that various traits influence IQ test scores and people develop the combinations of traits they have in different ways.

Kendler et al. (2002, 1133) show admirable reserve in concluding that their "results, while plausible, should be treated with caution because of problems with causal inference,

retrospective recall bias, and the limitations of a purely additive statistical model." At the same time, they did not remark on the absence of variables that correspond to therapeutic interventions (as if to suggest that these had no effect on the etiology of depression or its preceding risk factors) or to social changes that have led to the rising incidence of depression. Such omissions would seem important to rectify in any analogous, reciprocal causation modeling of IQ test scores. This brings us back to the issue of causes and changeability. In this paper I asked only what "local research" can learn about ways in which IQ test scores can be changed, a question which allowed me to show the limited relevance of human heritability estimates in developing explanations of differences between means across groups or across generations (or in exposing developmental causes). Insights in this area will only be delayed by the persistence in the thinking of researchers and other commentators of the conflation of ideas of heritability, genetic, and unchangeable.

6. **Coda: the "IQ paradox" reconceived**

Suppose that researchers want to develop empirical models of developmental pathways whose heterogeneous components differ among individuals at any given point of time. If genetic factors are to be included in the models, there are good methodological reasons for not categorizing individuals according to racial group membership. (This grouping is not based on clustering across a range of locations [theme 7a] and no measurable genetic factor admits a clean subdivision between whites and African-Americans.) 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. Yet, 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, even if all those pathways factor in the effects of racial group membership. The challenge for researchers is to shift the focus from group membership to heterogeneous pathways without bolstering the fiction that racial group membership no longer brings social/environmental benefits and costs. Can this be achieved? But, conversely, can researchers continue to track average differences among racial groups without bolstering the ubiquitous stereotyping that employs group membership

when deciding how to treat an individual? In short, a genuine paradox that applies to the use of IQ test scores in U.S. 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.

References

Byth, D. E., R. L. Eisemann, et al. (1976). "Two-way pattern analysis of a large data set to evaluate genotypic adaptation." Heredity **37**(2): 215-230.

Dickens, W. T. and J. R. Flynn (2001). "Heritability estimates versus large environmental effects: The IQ paradox resolved." Psychological Review **108**(2): 346-369.

Dickens, W. T. and J. R. Flynn (2002). "The IQ paradox is still resolved: Reply to Loehlin (2002) and Rowe and Rodgers (2002)." Psychological Review **109**(4): 764-771.

Flynn, J. R. (1980). Race, IQ and Jensen. London, Routledge and Kegan Paul.

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

Flynn, J. R. (2000). How to Defend Humane Ideals: Substitutes for Objectivity. Lincoln, NE, University of Nebraska Press.

Flynn, J. R. (2005). Are blacks genetically superior for IQ and GQ? What Germany did that America could not. (unpublished manuscript).

Friedmann, M. (1953). Essays in Positive Economics. Chicago, University of Chicago Press.

Fryer, R. and S. Levitt (2004). "Understanding the black-white test score gap in the first two years of school." The Review of Economics and Statistics **86**(2): 447-464.

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

Jensen, A. R. (1969). "How much can we boost IQ and scholastic achievement?" Harvard Educational Review **39**: 1-123.

Jensen, A. R. (1970). "Race and the genetics of intelligence: A reply to Lewontin." Bulletin of the Atomic Scientists **26**: 17-23.

Jensen, A. R. (1973). Educability & Group Differences. New York, Harper & Row.

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

Lewontin, R. C. (1970a). "Race and intelligence." Bulletin of the Atomic Scientists **26**: 2-8.

Lewontin, R. C. (1970b). "Further remarks on race and the genetics of intelligence." Bulletin of the Atomic Scientists: 23-25.

Lewontin, R. C. (1974). "The analysis of variance and the analysis of causes." American Journal of Human Genetics **26**: 400-411.

Lewontin, R. C. (1982). Human Diversity. New York, Freeman Press.

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

Loehlin, J. C. (2002). "The IQ Paradox: Resolved? Still an Open Question." Psychological Review **109**(4): 754-758.

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

Medical Research Council (2004). "National Survey of Health and Development (The British 1946 birth cohort study)." http://www.nshd.mrc.ac.uk/(viewed 4 March 2005).

Miele, F. (2002). Intelligence, Race, and Genetics: Conversations with Arthur Jensen. Boulder, CO, Westview Press.

Mitchell, H. K. and L. S. Lipps (1978). "Heat shock and phenocopy induction in Drosophila." Cell **15**(3): 907-918.

Moffitt, T. E., A. Caspi, M. Rutter. (in press). "Strategy for investigating interactions between measured genes and measured environments." Archives of General Psychiatry.

Neisser, U., G. Boodoo, et al. (1996). "Intelligence: Knowns and unknowns." American Psychologist **51**: 77-101.

Nisbett, R. E. (1998). Race, genetics, and IQ. The Black-White Test Score Gap. C. Jencks and M. Phillips. Washington, DC, Brookings Institution Press**:** 86-102.

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

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

Pinker, S. (2002). The Blank Slate: The Modern Denial of Human Nature. New York, Viking.

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

Plomin, R. and C. S. Bergeman (1991). "The nature of nurture: genetic influence on "environmental" measures." Behavioral and Brain Sciences **14**: 373-427.

Plomin, R., J. C. Defries, I. W. Craig, P. McGuffi, Eds. (2003). Behavioral Genetics in the Postgenomic Era. Washington, DC, American Psychological Association.

Rowe, D. C. and J. L. Rodgers (2002). "Expanding variance and the case of historical changes in IQ means: A critique of Dickens and Flynn (2001)." Psychological Review **109**(4): 759-763.

Sesardic, N. (2000). "Philosophy of science that ignores science: Race, IQ and heritability." Philosophy of Science **67**: 580 - 602.

Sesardic, N. (2003). "Heritability and indirect causation." Philosophy of Science **70**: 1002-1014.

Taylor, P. J. (2000). "Socio-ecological webs and sites of sociality: Levins' strategy of model building revisited." Biology and Philosophy **15**(2): 197-210.

Taylor, P. J. (2004). "What can we do? -- Moving debates over genetic determinism in new directions." Science as Culture **13**(3): 331-355.

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

Woodhead, M. (1988). "When psychology informs public policy." American Psychologist **43**(6): 443-454.

Appendix. Compilation of themes introduced in this paper

Difference-in-effects causes and their relation to measurable factors and developmental causes

1. Gradient-free conditions: Use of the AOV does not require that any gradient of a measurable genetic factor runs through the differences among genetically defined varieties or any gradient of a measurable environmental factor runs through the differences among locations.

2. Conditionality: All effects are conditional on the particular set of genetically defined varieties and locations observed.

3. Rerun control: The conditionality of effects means that any predictions made using difference-in-effects causes entail an assumption of control over the genetically defined varieties and locations that would allow the original combinations to be rerun and observed again.

4. Questionable homogeneity: An assumption that is always open to questioning is that similar patterns of responses of different genetically defined varieties across locations (or environmental factors) have been produced by similar conjunctions of underlying developmental causes or measurable factors.

5. Heritability as rerun predictability: In its technical meaning, heritability is a special case of rerun predictability (a concept that encompasses a more general class of correlations) in which the cultivar is matched in the observed data and rerun (i.e., i = i').

6. Heritability vs. "heritable." Ambiguity in the term "genetic variance" invites the technical term heritability to be misidentified with the colloquial idea that a trait is "heritable" or "genetic" if differences in a trait are associated with differences in specific genetic factors in the gene-based dynamics of organisms' reproduction.

7. From Observation to Hypothesis to Experiment to Understanding of Developmental Causes: After the appropriate simplification of the data set by clustering, the AOV of observations and subsequent regression analysis can contribute to the formulation of hypotheses that may be subject to experimental trials designed to examine the effect of varying specific genetic or environmental factors. In such trials the number of factors that can be considered simultaneously is constrained in practice (an extension of theme 3 on control) and the effects exposed by AOV are conditional (theme 2), but such experimental trials may contribute to a larger project of unraveling the biophysical pathways of development.

> 7a. Hypothesis generation following observational trials is enhanced by simplifying the large data set into groups for which the response of a cultivar group member (or response elicited by a location group member) is similar to the average for the group as a whole. Hypotheses become more difficult to formulate when groups are defined not by clustering, but by using some other criteria that does not minimize the variance within groups, including the within-group interaction variance (see theme 4).
>
> 7b. High heritability has a mixed relationship with researchers' ability to formulate hypotheses.

7c. The results of experimental trials in which environmental factors are systematically varied are conditional on the levels of other factors not subject to experimental variation and on the combinations of factors varied, as well as on the combinations of cultivars and locations in the trial.

7d. In the absence of experiments, estimates derived from regression analysis are not only conditional, but are like difference-in-effects causes in that predictions made using them assume rerun control (themes 3& 7a).

7e. There are important conceptual distinctions among the different quantities referred to as "environmental variance." The variance of a location effect in an AOV of an observational trial is not the same as the variance of the effect in an AOV of a trial in which an environmental factor is systematically varied. The latter is not the same as the variance of such an environmental factor itself. The variance associated with an environmental factor in a regression equation is yet another different quantity.

7f. Practical considerations place limits on experimentation that stems from observational trials.

7g. If measurable genetic factors are identified, then the previous sub-themes also apply to generation of hypotheses about genetic factors and to different meanings of the term "genetic variance" (see also theme 6).

Limitations of analyses based on partial data sets

8. Limited replication: limited hypothesizing. With only two replicates, an AOV of observations provides a limited basis for hypothesizing about measurable factors; any conclusions made on the basis of the AOV must center on difference-in-effects causes.

9. Nested effects. The existence of a significant cultivar-nested-within-location effect does not tell us whether the average for a cultivar is generally higher than another or whether it is only higher when paired with the particular location. Similarly, a significant location ("l") effect does not tell us whether one location is superior to the other, only that the cultivars grown in one location are different, on average, from the cultivars grown in the other location.

10. Partial data sets and inseparable effects. The within-group estimates that can be derived from the partial data set cannot speak to the relative contribution to differences between cultivar-group means in different locations of differences in cultivar-group effects versus location effects versus cultivar-group-mean-by-location-interaction effects.

11. Racial groups as separate locations in an AOV. The appropriate model for an AOV when replications—the identical twins raised separately—are not assigned independently with respect to membership in a racial group is one in which the experience of racial group membership is modeled as a separate location.

Methods to discriminate among more elaborate models

12. Methods to discriminate among models. When alternative models based on different dynamics are proposed, a method is needed that not only compares the models' fit with observations but also assesses the support for their assumptions independent of that fit.