Supplementary Electronic Materials for
Heritability and Heterogeneity
Peter Taylor
Biological Theory (2) 2006, 150-164.

Online Appendix 1: Analyses of Data Sets to Illustrate the Paper's Conceptual Steps and Themes

The simple numerical examples in this Appendix illustrate the conceptual steps and themes presented in this article. The data sets are fabricated, but the arguments in the body of the text do not depend on any correspondence between the data in these examples and actual observations.

1. Theme 1—Conditionality: "All Effects are Conditional on the Particular Set of Varieties and Locations Observed"

Model 5 can be expressed in symbols:

$$y_{ijk} - m = c_i + l_j + cl_{ij} + r_{ijk} \qquad (5')$$

Just as model 5 is equivalent to model 7, model 5' is equivalent to:

$$\sigma^2_y = \sigma^2_c + \sigma^2_l + \sigma^2_{cl} + \sigma^2_r \qquad (7')$$

where σ^2_y denotes the variance of yield measurements, σ^2_c denotes the variance of cultivar effects, etc.

The estimates of effects and variances in Table 1 are the ones that correspond to the minimum value of σ^2_r subject to the constraint that $\sum_i c_i = 0$; $\sum_j l_j = 0$; $\sum_i cl_{ij} = 0$ for each j; $\sum_j cl_{ij} = 0$ for each i; and $\sum_k r_{ijk} = 0$ for each ij combination. (Similarly for subsequent tables.)

Theme 1 is illustrated by the changes in values of effects and variances in Table 1 as more locations and cultivars are excluded from data set 1.

Table 1
Analysis of Variance (AOV) of Data Set 1.

Data Set 1			**Estimates of effects**		**Variance & heritability estimates**	
			m	3.0	σ^2_c	0.31 (13%)
	location		l_1	1	σ^2_l	1 (43%)
cultivar	1	2	l_2	-1	σ^2_{cl}	0.76 (33%)
1	5.3,4.3	0.2,1.2	c_1	-0.25	σ^2_r	0.25 (11%)
2	3.1,2.1	2.4,1.4	c_2	-0.75		
3	4.9,5.9	1.6,2.6	c_3	0.75	$h^2_{w/in\ location\ 1}$	0.84
4	3.7,2.7	2.8,3.8	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		
			r_{ijk}	+/-.5		
Data Set 1a						
			m	2.5	σ^2_c	0.06 (2.5%)
	location		l_1	1.2	σ^2_l	1.44 (58%)
cultivar	1	2	l_2	-1.2	σ^2_{cl}	0.72 (29%)
1	5.3,4.3	0.2,1.2	c_1	0.25	σ^2_r	0.25 (10%)
2	3.1,2.1	2.4,1.4	c_2	-0.25	$h^2_{w/in\ location\ 1}$	0.83
			cl_{ij}	±0.85	$h^2_{w/in\ location\ 2}$	0.59
			r_{ijk}	+/-.5	$h^2_{across\ locations}$	0.25
Data Set 1b						
			m	3.7	σ^2_c	1.21 (83%**)
	location		l_1	0	σ^2_r	0.25 (17%)
cultivar	1					
1	5.3,4.3*		c_1	1.1	$h^2_{w/in\ location\ 1}$	0.83***
2	3.1,2.1		c_2	-1.1		
			r_{ijk}	+/-.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, are discussed in section 4.

2. Grouping of Cultivars by Similarity in Responses Across All Locations

Consider the effect of grouping cultivars from data set 1 by similarity in responses across all locations. As Figure 1 in the article 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, model 8, or, in symbols:

$$y_{ijk} = m + C_I + c_{i:I} + l_j + Cl_{Ij} + cl_{i:I,j} + r_{ijk} \qquad (8')$$

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

Table 2
AOV of Data Set 1 divided into two groups by similarity of response across locations.

Cultivar Group	cultivar	location 1	location 2	Estimates of effects		Variance & heritability estimates	
				m	3.0	σ^2_C	0.06 (2.7%)
				l_1	1	$\sigma^2_{c:C}$	0.25 (11%)
				l_2	-1	σ^2_l	1 (43%)
A	1	5.3,4.3	0.2,1.2	C_A	0.25	σ^2_{Cl}	0.72 (31%)
B	2	3.1,2.1	2.4,1.4	C_B	-0.25	$\sigma^2_{c:C,l}$	0.04 (1.7%)
A	3	4.9,5.9	1.6,2.6	$c_{1:A}$	-0.5	σ^2_r	0.25 (11%)
B	4	3.7,2.7	2.8,3.8	$c_{2:B}$	-0.5		
				$c_{3:A}$	0.5	$h^2_{\text{within cultivar group A or B within location 1}}$	0.26
				$c_{4:B}$	0.5	$h^2_{\text{within cultivar group A or B within location 2}}$	0.66
				Cl_{Ij}	±0.85	$h^2_{\text{within cultivar group A across both locations}}$	0.06
				$cl_{i:I,j}$	±0.2	$h^2_{\text{within cultivar group B across both locations}}$	0.44
				r_{ijk}	+/-.5		

Now compare this with an arbitrary grouping, analyzed in Table 3.

Table 3
AOV of Data Set 1 divided into two arbitrary groups.

Cultivar Group	cultivar	location 1	location 2	Estimates of effects		Variance & heritability estimates	
				m	3.0	σ^2_C	0.25 (11%)
				l_1	l_1	1 $\sigma^2_{c:C}$	0.25 (11%)
				l_2	-1	σ^2_l	1 (43%)
A	1	5.3,4.3	0.2,1.2	C_A	-0.5	σ^2_{Cl}	0.04 (1.7%)
A	2	3.1,2.1	2.4,1.4	C_B	0.5	$\sigma^2_{c:C,l}$	0.72 (31%)
B	3	4.9,5.9	1.6,2.6	$c_{1:A}$,$c_{3:B}$	0.25	σ^2_r	0.25 (11%)
B	4	3.7,2.7	2.8,3.8	$c_{2:A}$, $c_{4:B}$	-0.25		
				Cl_{A1}, Cl_{B2}	0.2	$h^2_{\text{within cultivar group A or B within location 1}}$	0.83
				Cl_{A2}, Cl_{B1}	-0.2	$h^2_{\text{within cultivar group A or B within location 2}}$	0.59
				$cl_{i:I,j}$	±0.85	$h^2_{\text{within cultivar group A across both locations}}$	0.025
				r_{ijk}	+/-.5	$h^2_{\text{within cultivar group B across both locations}}$	0.037

Notice that 33.7% of the total variance in yield is associated with within group effects when the grouping is arbitrary, but only 14.7% is when the groups have been formed by similarity of responses across locations. This latter figure is comparable to the residual error.

3. Consolidation of Regressions

Consider the regression model 9 in symbolic form, with the cultivar effect partitioned into a cultivar group effect and a cultivar-within-group effect:

$$y_{ijk} = m + C_I + c_{i:I} + \sum_q b_{iq}e_{jkq} + r_{ijk} \qquad (9')$$

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

r_{ijk} is a residual 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 (equation 9') in a cultivar group may be consolidated into a smaller set of analyses. This requires that that all b_{iq}'s are assumed to be the same across cultivars in the same cultivar group (= "b_{Iq}").

$$y_{ijk} = m + C_I + \sum_q b_{Iq}e_{jkq} + r_{ijk} \qquad (9a')$$

The homogeneity assumption involved in such consolidations is subject to questioning (theme 3), especially since such consolidation shifts the c_i and $cl_{i:I,j}$ effects (from model 8) into the residual term as if they were negligible. This shift also increases the residual variance and reduces the strength of the association. (If data were available for measurable genetic factors equation 9a' could be elaborated so as to bring those effects back out of the residual.)

4. Focusing Hypothesis Generation on Cases of High Heritability is Not Equivalent to Focusing on Groups with Low Within-group Effects

The non-equivalence of the two approaches discussed in section 4.1 can be seen by considering the simple case in Figure 1 in the article. Suppose that researchers, using some criterion other than similarity of responses across locations, had chosen to group cultivar 1 with 2, and cultivar 3 with 4. In each location, the size of the cultivar-in-location effect relative to the residual can be estimated by eye by comparing the difference between the cultivar means (i.e, the distance between the lines for the cultivars in a group) with the average difference between replicates (i.e, the average of the distances between the two diamonds for each cultivar). In location 1, the ratio of cultivar to residual is clearly much greater for both groups than it would be if groups chosen had happened to be those that cluster analysis of the full data set would have produced, i.e. (1,3) and (2,4). However, the homogeneity of

each of the latter groups is less questionable, so hypotheses formulated about measurable factors operating those groups, not the groups (1,2) and (3,4), are more likely to be validated by subsequent investigation. (For this example, nothing can be said about what those hypotheses might be, because no other knowledge is available concerning the cultivars or locations.)

5. Derivation of Formula for Rerun Predictability

For the situation depicted in Figure 8, the formula for rerun predictability can be derived as follows:

Correlation (observed and predicted)

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

= Covariance ($m + c_i + l_j + cl_{ij} + r_{ijk}$, $m + c'_i + l_j + cl'_{ij} + r'_{ijk}$)

/ [Variance ($m + c_i + l_j + cl_{ij} + r_{ijk}$) * Variance ($m + c'_i + l_j + cl'_{ij} + r'_{ijk}$)] $^{1/2}$

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

which can be estimated by Covariance ($c_i + cl_{ij}$, $c'_i + cl'_{ij}$) / [($\sigma^2_c + \sigma^2_{cl} + \sigma^2_r$) * ($\sigma^2_c + \sigma^2_{cl} + \sigma^2_r$)]$^{1/2}$
given that residual effects (noise) are uncorrelated and $m + l_j$ is a constant

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

Formulas for rerun predictability can also be derived using path analysis, a data analysis technique that quantifies the relative contributions of variables ("path coefficients") to the variation in a focal variable once a certain network of interrelated variables has been accepted (Lynch and Walsh 1998: 823). The usual starting point for path analysis is a regression model that associates the focal variable (here, the yield) with several other measured variables, but it is still possible to employ the technique when there are no measured variables except the observed focal variable. This can be done by formulating an additive model of constructed variables that take the values of the effects from an AOV. The path coefficients are then set to equal the square root of the ratio of the variance of the effect to the total variance for the trait. The "equation of complete determination" that lies at the heart of path analysis becomes

$$1 = \sum \sigma^2_x / \sigma^2_y \qquad (16)$$

where x denotes the corresponding effect or variable.

When the same trait is observed in parent and offspring, their separate path analyses can be linked and the correlation between the parent and offspring calculated (Lynch and Walsh 1998: 826), provided it is assumed that the effects (and path coefficients) are constant across generations and the residuals are uncorrelated, that is, the rerun conditions apply (sections 2.1 and 4.1). For the network as defined by model 5 (figure 1), the predicted correlation between parent and offspring over all locations is the corresponding formula for heritability:

$$\sigma^2_c / \sigma^2_y = \sigma^2_c / (\sigma^2_c + \sigma^2_l + \sigma^2_{cl} + \sigma^2_r) \qquad (11')$$

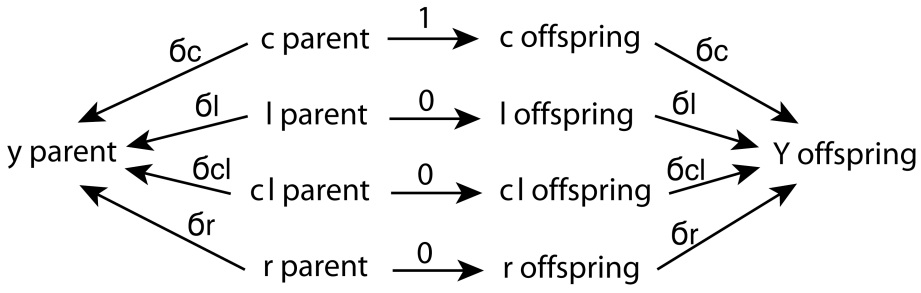

Figure 1.
Path diagram linking trait values of parent and offspring, both modeled by equation 5.

More complicated networks of interrelated variables can be analyzed. Typically, these incorporate diploidy and biparental inheritance, degrees of relatedness of different cultivars (not only parent-offspring pairs), and correlations among replicates (e.g., when plots are assigned non-randomly within a single location). (In the analysis of human traits, non-random replication is usually attributed to siblings in a family all sharing some unspecified environmental factors or to factors that differ among siblings.) In addition to the assumption that effects and path coefficients are constant across generations, all such path analyses, like the AOV and rerun predictability, a) are based on observed traits and do not require reference to measurable genetic factors that are transmitted from parent to offspring; and b) are conditional on the particular set of genetically defined varieties and locations observed (from which effects/path coefficients are estimated). (Pearl [2000, 135 & 344-5] interprets path analysis as an analysis of causes, but does not acknowledge these conditions-conditions that render path diagrams used in heritability estimation quite different from engineering circuit diagrams.)

6. Heterogeneity May Remain Even Within Cultivar Groups Formed by Cluster Analysis

Consider the generic model of development in online Appendix 2, 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:

$$y'_{ijk} = \prod_r (g_{ir}{}^{e_{jr}})\, f_{jrk} \qquad (17)$$

where g, e, f denote genetic factors, environmental factors, and random noise, respectively, active at time r in the sequence,

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

$e_{jr} = \pm ß$ with equal probability, and

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

To facilitate comparison with data set 1, I scale any data generated from model 17 so it has the same mean and SD:

$$y_{ijk} = \text{constant}_1 + \text{constant}_2 * y'_{ijk} \qquad (17a)$$

One data set generated by model 17 with the values $\gamma = .8$, $ß = .5$, $\partial = .25$, 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 additive model 5 is given in Table 5 and Figure 2. Cultivar groups are assigned on the basis of similarity of responses across locations. Note the high within-location heritabilities for cultivar group B (cultivars 2 and 4) and the low values for cultivar group A (cultivars 1 and 3).

Table 4
Parameter values in Model 17 that generate Data Set 2.

						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
AOV of Data Set 2 divided into two groups by similarity of response across locations.

				Estimates of effects		**Variance & heritability estimates**	
				m	3.0	σ^2_C	0.49 (21%)
		location		l_1	0.67	$\sigma^2_{c:C}$	0.25 (11%)
Cultivar Group	cultivar	1	2	l_2	-0.67	σ^2_l	0.45 (20 %)
A	1	3.9, 5.6	1.8, 2.6	C_A	0.70	σ^2_{Cl}	0.60 (26 %)
B	2	1.8, 1.2	3.9, 3.6	C_B	-0.70	$\sigma^2_{c:C,l}$	0.55 (24%)
A	3	5.5, 5.6	2.7, 2.0	$c_{1:A}$,$c_{3:A}$	±0.24	σ^2_r	0.16 (7 %)
B	4	2.7, 3.2	1.1, 1.0	$C_{2:B}$, $c_{4:B}$	±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	$h^2_{\text{within cultivar group B across both locations}}$	0.08
				r_{ijk}	varied		

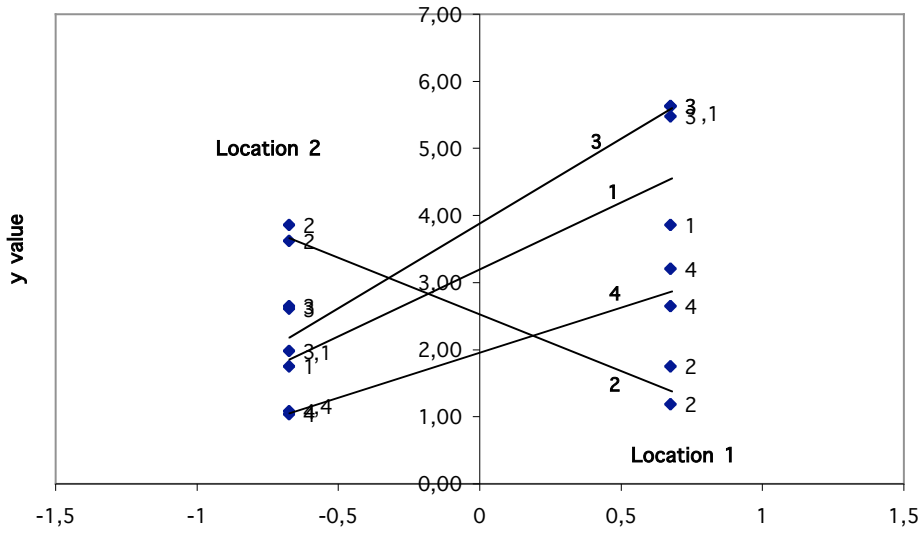

Figure 2.
Data set 2. Lines connect the midpoint of the cultivar in each location. The x-axis is the location effect = average over all cultivars for that location - overall mean.

(This simple model also accentuates the difficulties of exposing the complexity of biophysical processes of growth and development using AOV. As a thought experiment, consider what a comparison of the AOV of data sets 1 and 2 would suggest to researchers about the processes that generated the observed data. A comparison of Figure 1 in the article and Figure 2 above shows that although cultivars 4 and 2 only converge in data set 1 but cross in data set 2, the overall trends are very similar. The similar AOVs would not suggest that the data sets 1 and 2 were generated by radically different kinds of models. However, this is the case. Although I did not state this, data set 1 was simply generated by the additive model 5 using the parameter values shown in Table 3.

This simple model can also be used to illustrate a point to be made in section 4.2, namely, if the contributions of genetic and environmental factors modulate each other, the relationship of the factors to heritability is complex. The same values of the parameters γ , ß , ∂ in model 17 can result in widely varying heritability estimates. A spreadsheet that allows exploration of this feature is available from the author on request.)

7. Example of Genetically Correlated Relatives Without Any Common Genetic Factor(s) Accounting for the Similarity of Outcomes Within the Larger Group of Cultivars

Table 6
Genetically correlated relatives without any common genetic factor(s) accounting for the similarity of outcomes within the larger group of cultivars. Data generated by a variant of Model 17 (available from the author by request).

Group of relatives	Sequence of genes		Yield when genes are modulated by a sequence of environmental factors FGHij	
	Twin 1	Dizygotic twin 2	Twin 1	Dizygotic twin 2
1	ABcDE	AbcdE	1.409	1.264
2	AbcDE	AbCDE	0.975	1.144
3	Abcde	ABCde	0.685	0.715

References

Taylor PJ (2006) Heritability and heterogeneity: The irrelevance of heritability in explaining differences between means for different human groups or generations, Forthcoming in Biological Theory 1(4).

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