Analysis of Variance in the Service of Interactionism

By: Douglas Wahlsten

Wahlsten, D. (2000). Analysis of variance in the service of interactionism. *Human Development*, 43: 46-50.

Key Words: Behavior genetics, Effect size, Heritability, Multiplicative model, Sample size, Statistical power

Article:

Vreeke (this issue) asserts: 'If you hold that reality is interactive, but use methods which presuppose the separateness of genes and the environment, your are stuck with a problem of interpretation.' A similar point was made recently by a colleague in behavior genetics who chastised me, a critic of heritability analysis [Wahlsten, 1990, 1994, 1999a, 1999b], for employing the analysis of variance (ANOVA) in a recent study [Crabbe, Wahlsten and Dudek, 1999]. In reply, I argue that (a) ANOVA does not presuppose the separateness of genes and environment; (b) devotées of interactionism often employ ANOVA to examine results of factorial experiments precisely because the method can reveal the presence of interactions; and (c) all users of ANOVA must struggle with problems of interpretation, but these problems are especially severe for those who apply correlational methods to study human populations.

As I see it, the main difficulty arises when a user of statistics claims that a certain percentage of total variance among individuals can be attributed to variation in the causal influence of one factor, whereas the remaining percentage can therefore be chalked up to the other factor, so that the relative causal importance of the two sources of individual differences can be compared. This is commonly done when a heritability ratio claims that, for example, 60% of variance is attributable to genetic variation and the other 40% is environmental. This procedure is based on the assumption that genetic and environmental effects act separately in development and consequently are statistically additive. Vreeke correctly points out that we already know the factors do not act separately during development. Consequently, heritability analysis is based on obsolete theory and lacks credibility.

The controversy over the relative influence of two effects is usually not at issue in genetic research with laboratory animals. For example, Crabbe et al. [1999] chose to study 8 strains of mice that were already known to differ greatly on several simple tests of behavior, and they tried to equate many aspects of the test situation and the rearing environment in three labs. Obviously, the main effect of strain in the ANOVA was expected to be quite large. Two questions were of interest: Would there be systematic differences among the three labs, despite efforts to minimize the lab main effect, and would the pattern of strain differences depend on the specific lab, as indicated by the strain x lab interaction? Both were answered in the affirmative for several measures of behavior. As in almost all laboratory experiments, the relative strengths of the main effects were dictated by the specific values of each factor deliberately chosen by the researchers. By assigning equal numbers of animals to each cell of a complete factorial design, covariance between heredity (H) and lab environment (E) was eliminated. The F ratio for the interaction term in the ANOVA was used to evaluate the null hypothesis that effects of H and E were additive, and for several measures this ratio indicated that H and E were definitely not additive in the statistical sense. Hence they did not act separately during development.

Statistical analysis often involves partitioning variance, but, as Vreeke suggests, the chief problem occurs with interpretation of the results rather than the mechanics of the calculations. Consider an experiment with only two groups that are elegantly arranged to differ in only one factor [see Wahlsten, 1999b]. The t test asks whether

their means could differ merely because of sampling (Type I) error, whereas the CO2 coefficient, an indicator of effect size, estimates the proportion of total variance attributable to the difference between group means. That is, the ω^2 coefficient partitions total variance into two components, one between and the other within groups. This is perfectly appropriate and often informative, and many avowed interactionists do it without incurring the wrath of their sagacious gods. Here the devil is not lurking in the details of the algebra but rather casts a shadow over the larger issue of interpretation.

Suppose two inbred strains of mice yield an estimated $\omega^2 = 0.35$. Does this tell us that 35% of variance arises purely from their genetic difference, whereas 65% is environmental? Not necessarily. The additive model asserts that the group mean is a simple sum of two components ($M = G + E$), and when the strains are reared in the same environment, algebra avers that the difference in group means is solely a result of the difference in genotypic value ($\Delta M = \Delta G$), whereas it has nothing to do with the rearing environment. That is, additivity requires that the group difference should be the same in all environments. Now, suppose instead that the factors are multiplicative, such that $M = G \cdot E$. For two strains reared in environment E_1, the difference in means will be $\Delta M = E_1 \Delta G$, and ΔM will have a different value when rearing occurs in some other environment. Thus, a study with two groups can prove that a factor is a noteworthy part of a system responsible for development of behavior, but it does not justify a conclusion that the precise numerical magnitude of a group difference arises from only one factor. We can partition variance in many situations, even though the antecedent causes do not act separately during development, and there is no harm in doing so, provided we do not reify the statistical ploy.

Only an experimental design with two or more factors can assess whether effects of heredity and environment are nonadditive. Interaction can only be seen clearly when individuals with the same genotype are reared in different environments, and when the experiment involves two or more genotypes. The variance attributable to interaction is *defined* as the deviations of observed group means from values expected on the basis of strictly additive main effects. This definition then leads to an algebraically perfect partitioning of the sums of squares into three nonoverlapping portions: $SS_{Total} = SS_G + SS_E + SS_{GxE}$. This relationship must hold for any factorial array of numbers, no matter what biological or psychological phenomenon is being investigated. Contrary to the claim of Plomin [1990, p. 144], however, it does not tell us that the three factors are independent. Instead, if the relation between the two variables is multiplicative, then the two main effects and the interaction effect are completely intertwined.

$$\sigma_H = \frac{(K+1)\,he}{4}\sqrt{\frac{(J+1)(J-1)}{3}}$$

$$\sigma_E = \frac{(J+1)\,he}{4}\sqrt{\frac{(K+1)(K-1)}{3}}$$

$$\sigma_{HxE} = \frac{he}{12}\sqrt{(J+1)(J-1)(K+1)(K-1)}$$

I previously investigated a situation where *J* strains had genetic values (G) separated by *h* units and were reared in *K* environments with values separated by *e* units [Wahlsten, 1990]. The theoretical expectations for the effect sizes for the two main effects and the interaction effect are shown in three equations in terms of the standard deviation among group means, which is related to Cohen's effect size ($f = \sigma_M/\sigma$). If the separation between the strains (the value of *h*) is increased by 50% or the number of strains (*J*) is increased, this of course increases the main effect of strain, but under the multiplicative model it also increases the main effect of environment as well as the interaction effect. Increasing the number of strains also changes the relative magnitude of the main effects and the interaction. Biologically or psychologically, the three statistical effects are not independent when their functional relationship is multiplicative, even though algebra can separate the total sums of squares into three neat little piles. When two causes are genuinely interactive, the interaction effect in an ANOVA is not independent of the main effects and the two main effects are not independent of each other. When the interaction effect in the ANOVA accounts for an appreciable fraction of the total sums of squares, partitioning variance between two factors is nonsensical. The interaction effect is not separate from the main effects, and there is no magic device that can tell us whether or how much of it belongs with one term or the other. Statistical interaction tells us that partitioning variance between two sources is an invalid exercise because the

consequences of variation in one factor depend on the specific levels of the other factor. Only when the factors really do act separately in development will the two main effects be independent and separable in the statistical sense expressed by a heritability ratio.

Those who take interaction effects seriously often employ the ANOVA method to analyse data, but they do this only after careful consideration of the sample sizes needed to make the method sufficiently sensitive to interaction effects. For many common kinds of interaction, the power to detect main effects tends to be substantially greater than the power to detect interaction [Wahlsten, 1990, 1991, 1999b], and as a result the research ought to study larger samples than those required simply to detect a main effect. How much larger the sample size needs to be depends strongly on the kind of interaction effect that one would like to be able to detect. The discrepancy in sample sizes needed to detect main effects versus interaction is often surprisingly large, and studies in psychology and neuroscience typically employ sample sizes that are woefully inadequate to reveal interesting interactions.

Table 1. Two hypothetical models of strain and lab effects on average test scores

	Strain A	Strain B
Model 1: Additive Effects		
Lab L1	20	30
Lab L2	30	40
Model 2: Interaction		
Lab L1	20	30
Lab L2	30	50

Let us consider a simple 2 × 2 design with four independent groups of subjects, two inbred strains reared in two different labs. Suppose strain B scores 10 units higher than strain A in lab L1, whereas strain A reared in lab L2 scores 10 units higher than in lab L1. That is, suppose the effects of strain and lab are about the same size. If the factors are strictly additive, then strain B in lab L2 must score 20 points higher than strain A in lab L1 (see table 1). If there is interaction, the mean for strain B in lab L2 will deviate from this value. There is no universal standard for what constitutes noteworthy interaction, but I believe the following situation would be of interest to most investigators. Suppose that the effect of lab L2 environment is *twice as great* for strain B as for strain A, which would give strain B in lab L2 a mean score of 50 units.

Sample size per group (*n*) to yield power of 1-β when a one-tailed test with Type I error probability α can be approximated with a simple formula based on the method of contrasts [Wahlsten, 199 1]. For the four groups in a 2 × 2 design, the true value of the contrast is $\psi = c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4$, where $\Sigma c_j = 0$. In this example, the standard deviation within a group is set at $\sigma = 15$. Under Model 2, the contrast for the strain main effect with be $\psi_{Strain} = -20 - 30 + 30 + 50 = 30$ and $\psi_{Lab} = 30$ as well. The interaction effect will be$\psi_{Strain \times Lab} = 20 - 30 - 30 + 50 = 10$. For all three effects, $\Sigma c^2 = 4$, $z_\beta = 1.645$ and $z_\alpha = 1.645$ when power is to be 95% and Type I error probability is 0.05. Thus, $n = 43.3/(\psi/15)^2 + 2$. For the two main effects, $n_{Strain} = n_{Lab} = 13$, whereas to detect the interaction effect, the sample size must be $n_{Strain \times Lab} = 100$ mice per group, about 8 times the number required to detect the main effects with the same degree of power!

$$N == \frac{(z_\alpha + z_\beta)^2 \Sigma c_j^2}{\left(\frac{\psi}{\sigma}\right)^2} + 2$$

The analysis of variance applied to factorial designs with laboratory animals is a good method for evaluating the occurrence of statistical interaction, provided sufficiently large samples are used. If small samples of 10 to 20 animals per group are studied, as is very commonly done in physiological psychology, the ANOVA will often give a false impression of additivity, despite genuine interaction between the factors. Sample size is inversely related to the square of the effect size, which means that even substantial interactions will require relatively large samples to be detectable. It seems to me that when the strain difference is twice as large in one lab as in the other, we really ought to be able to detect the interaction; it is neither trivially small nor hugely obvious.

This magnitude of interaction warrants serious attention in any field of psychology where theories pass or fail on the basis of significant interaction effects, and it therefore serves as a convenient standard. In many instances, if only 10 animals per group are adequate to detect main effects, then 50 to 100 per group will be needed to detect an interaction. Thus, the avowed interactionist employs large samples using the ANOVA method and does not take tests of interaction very seriously when small samples are studied.

The power problem is fundamentally the same when working with correlational methods in uncontrolled human populations, although statistical models will typically entail random rather than fixed effects. Correlational studies are additionally beset by vexing problems of research design because heredity and environment are often confounded or covary [Kempthorne, 1990], and poorly controlled environmental effects can artifactually inflate heritability estimates to a surprising degree [Guo, 1999]. There is one final difficulty that douses enthusiasm for additive models. To test interaction between genotype and environment, there must be many individuals with the same genotype who are reared in different environments. This is easily achieved with standard laboratory strains but not with humans. For our species, there is no valid test of gene x environment interaction, no matter what the sample size, unless distinct alleles of a specific gene in question can be identified. For research on heredity in general that relies on twins and adopted children, effects of heredity cannot be cleanly separated from those of environment, and there is no valid test of the presence of interaction. Because the additivity assumption cannot be tested empirically, the whole edifice of path models must be accepted on faith, if it is to be accepted at all. As pointed out by Vreeke, this unverified faith conflicts with established facts about animal biology. Heredity and environment do not act separately in development.

References

Crabbe, J.C., Wahlsten, D., and Dudek, B.C. (1999). Genetics of mouse behavior: Interactions with lab environment. *Science, 284*, 1670–1672.

Guo, S.-W. (1999). The behaviors of some heritability estimators in the complete absence of genetic factors. *Human Heredity*, *49*, 215–228.

Kempthorne, O. (1990). How does one apply statistical analysis to our understanding of the development of human relationships? *Behavioral and Brain Sciences*, *13*, 138–139.

Plomin, R. (1990). Trying to shoot the messenger for his message. *Behavioral and Brain Sciences*, *13*, 144–145.

Wahlsten, D. (1990). Insensitivity of the analysis of variance to heredity-environment interaction. *Behavioral and Brain Sciences*, *13*, 109–16 1.

Wahlsten, D. (199 1). Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. *Psycho-logical Bulletin*, *110*, 587–595.

Wahlsten, D. (1994). The intelligence of heritability. *Canadian Psychology*, *35*, 244–258.

Wahlsten, D. (1999a). Single-gene influences on brain and behavior. *Annual Review of Psychology, 50*, 599–624. Wahlsten, D. (1 999b). Experimental design and statistical inference. In W.E. Crusio and R.T. Gerlai (Eds.), *Handbook of molecular-genetic-techniques for brain and behavior research* (pp. 40–57). Amsterdam: Elsevier.