Sample Size Requirements for the Capron and Duyme Balanced Fostering Study of IQ

By: Douglas Wahlsten

Wahlsten, D. Sample size requirements for the Capron and Duyme balanced fostering study of I.Q. International Journal of Psychology, 1993, 28, 509-516.

Abstract:

A simplified method is used to estimate the appropriate sample sizes needed to detect main effects and an interaction effect in analysis of variance, using the IQ data from the Capron and Duyme (1991) adoption study as an example. To achieve power of 80% to reject an hypothesis of no interaction when there is in reality a modest interaction requires about 215 children in each of four groups in a 2 × 2 design, whereas only 9 to 10 children per group are needed to detect main effects. Only a transnational collaborative study could hope to find this many children in the condition where a child from high socioeconomic status background is adopted into a low status family.

Article:

The adoption study by Capron and Duyme (1989, 1991) with its balanced fostering design provides a good opportunity to evaluate possible interactions between factors present before and after adoption. They reported that for full scale IQ there is no statistically significant interaction and that the increase in IQ from improving the postnatal family environment is similar to the decrease occasioned by adoption into a poorer environment. In the 1989 report they argued that the absence of a significant interaction term is "convincing evidence" that environmentally induced changes in IQ "exhibit the same general trend" for children from both low and high socioeconomic status (SES) birth mothers. In the 1991 report they noted that the power of their test of interaction was quite low (about 5%), but they argued that the observed size of the interaction effect was so small that it is reasonable to conclude no interaction exists.

In a long commentary published in the same issue of *Nature* as the original Capron and Duyme (1989) report, McGue (1989) went even further by claiming the adoption data demonstrate "an absence of an interaction between biological background and rearing circumstance," which he suggested means there was no genotype-environment interaction. However, as Capron and Duyme (1989) clearly and correctly stated, adoption dissociates "pooled effects of genetic and prenatal factors from factors related to the postnatal environment. But is not equipped to differentiate prenatal from genetic factors." Hence an adoption study cannot possibly test for genotype-environment interaction, which requires replicated genotypes reared in different environments. Elegant techniques are available for separating genetic and prenatal environmental factors in laboratory animals (Carlier, Nosten-Bertrand, & Michard-Vanhee, 1992), and in this realm a valid test of genotype-environment interaction is possible. Maternal environment can be evaluated in humans (Rose, Uchida, & Christian, 1981), and numerous interactions of single genetic loci with environmental conditions have been documented (Desnick & Gabrowski, 1981). Nevertheless, an adoption study can provide no such test.

The question remains whether the Capron and Duyme study did indeed prove the independence of factors acting before and after adoption on childhood IQ score. I suggest that proof cannot be provided by a nonsignificant interaction term, especially when the power of the test of interaction is low. To confer adequate power on a test of interaction, an adequate sample size must be used, but the sample size employed in practice

is usually sufficient to detect main effects while being far short of requirements for a powerful test of several realistic kinds of interaction (Wahlsten, 1990, 1991). According to a simplified formula for calculating sample size to detect one degree-of-freedom effects (Wahlsten, 1991), a much larger sample than that used by C and D would be needed but not one so large as the more than 1,000 subjects proposed in their 1991 article (Table 7, p. 337).

C and D reported a minuscule $F = 0.011$ for their test of interaction. When the true interaction effect is exactly zero, the expected value of the F ratio is 1.0; that is, the estimated variance for the interaction should tend to equal the variance within groups. By virtue of sampling error, the observed value of the F ratio could be somewhat below or above 1.0. $F = .011$ is much below 1.0 but not enough to raise the eyebrows; in fact, with degrees of freedom 1 and 34 the probability of F being this small or smaller when the true interaction effect is zero is .083. Such a result is no more surprising than $F = 3.19$; the probability of F being this large or larger when there is no interaction is also .083. However, $F = 3.19$ corresponds to a substantial interaction effect size, even though it is below the criterion for statistical significance ($F = 5.50$ when $\alpha = .05$, two-tailed). An overestimate (Glass & Hakstian, 1969) of effect size is *est* $\eta^2 = SS_{int}/(SS_{int} + SS_{error})$, which is a partial correlation ratio (Maxwell, Camp, & Arvey, 1981). For $F = 3.19$ and $df_{error} = 34$ this gives *est* $\eta^2 = .086$. A somewhat better estimate of effect size for a one degree-of-freedom effect in analysis of variance is (Hays, 1973)

$$est\ \omega^2 = \frac{F-1}{F+df_{Within}+1},$$

which yields *est* $\omega^2 = .084$. Cohen (1988) gives the relation between his effect size f and the population value of the partial correlation ratio as

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}}.$$

Cohen's effect size f would be about .30 for *est* $\eta^2 = .086$ or .28 for *est* $\omega^2 = .084$ for a true effect yielding an expected $F = 3.19$. Thus, the C and D data are equally consistent with a true interaction effect accounting for either 0% of total variance or 8 to 9% of total variance. This great uncertainty about the true interaction effect size arises from the relatively small sample size.

Because the estimated effect size in one relatively small, albeit well controlled study is subject to large sampling error, using the obtained value to calculate required sample size is not the best way to proceed. If the sample means in C and D's (1991) Table 5 are used as population means, Cohen's formula 8.3.6 yields f = .034, which is within rounding errors of the value of f = .04 cited by C and D. From Cohen's Table 8.4.4 and formulae 8.4.1 and 8.4.4, power of 80% with a one degree-of-freedom interaction effect size off = .034 requires *1700 subjects in each of the four groups!* This is considerably larger than the sample size I propose using a different method.

To determine sample size to achieve desired power, a null hypothesis of no interaction is tested against an alternative hypothesis about the true value of group means. This ought to be done before the data are collected, using available information to make an informed guess at a plausible pattern of results. Let us determine the appropriate sample size if a comparable 2×2 study is to be done with sufficient numbers of children to detect an interaction effect with power of 80%. As shown previously (Wahlsten, 1991), a 2×2 design can be analyzed as a set of three linear contrasts of the form $\Psi = c1\mu1 + c2\mu2 + c3\mu3 + c4\mu4$, and the set of means specifying the kind of interaction leads to the required sample size via the formula

$$N(pergroup) = \frac{(z_{\alpha/2} - z_{1-\beta})^2 \Sigma c_j^{\ 2}}{\left(\frac{\Psi}{\sigma}\right)^2} + 2.$$

This is a normal approximation that provides an answer very close to the methods of Cohen (1988) and

Kraemer and Thiemann (1987), although these two sources do not present a general method for complex contrasts.

The first step is to estimate the standard deviation (cr) within a group, which is assumed to be the same for all groups. For a study using a restricted range of socioeconomic status (SES) scores within a group, the true standard deviation of IQ scores must be less than the value of 15.0 for the entire population. From the MS_{within} = 174.7 in the C and D study, a reasonable estimate is $\sigma = 13$. Any greater accuracy would be illusory.

Next, the likely increase in IQ for French children adopted from a poor family into a much higher SES family can be estimated from the Schiff, Duyme, Dumaret, and Tomkiewicz (1982) study to be about 14 IQ points, which is close to the result in C and D.

The most difficult thing to estimate is the loss of potential IQ points for adoption from high to low SES families, which happens rarely. Poverty can greatly suppress the achievement of a child, but there can also be substantial recovery from early deprivation (Clarke & Clarke, 1976). How much should an opposite change in environment harm a child from a superior background? A child from poverty might as a fetus have suffered from poor nutrition or inadequate medical care, which in turn might impair its capacity to prosper in better circumstances. On the other hand, poor postadoption environment might exert relatively greater influence on the child from a low SES background, whereas the child adopted from high into low SES after birth might change less. Available evidence from studies of prenatal malnutrition provide support for a sensitive period but the timing of maximum and minimal sensitivity remains obscure (Dobbing, 1985; Morgane et al., 1992: Stein & Susser, 1985). Either scenario seems reasonable, given current knowledge, and it is noteworthy that C and D did not propose what kind of interaction should be expected. Hence, a two-tailed test of the null hypothesis should be used. If the interaction effect is the major interest of the study, it would be reasonable to test it with $\alpha = .05$, two-tailed. However, if all three contrasts are of interest, it is preferable to use $\alpha = .05/3$ for each test, known as the Dunn or Bonferroni approximation of the Sidak (1967) inequality.

If children from a low SES background improve by 14 IQ points when adopted into a high SES family, the crucial question is what change in the opposite direction would be a psychologically interesting interaction. If the true high to low SES effect is 13 points, this one point difference in adoption effect might be seen as trivial and unworthy of study with a sample of any size. A high to low effect of 9 points would constitute a more interesting interaction and 7 points would be quite a dramatic attenuation of the environmental effect. Let us contrast the 14 point low to high SES effect with the 9 point high to low SES effect.

TABLE 1
Group Means, Contrast Coefficients (± 1), Contrast Sizes and Sample Sizes (italics) Needed to Detect the Effects with Power of 80% When σ = 13 for Each Group

Pre: *Post:*	*Low-* *Low*	*Low-* *High*	*High-* *Low*	*High* *High*	ψ	α = .05		α = .017	
Mean	92	106	107	116		1 tail	2 tail	1 tail	2 tail
Pre Effect	-1	-1	1	1	25	*9*	—	*12*	—
Post Effect	-1	1	-1	1	23	*10*	—	*14*	—
Int. Effect	-1	1	1	-1	5	*170*	*215*	*240*	*284*

Note: Pre = family condition prior to adoption; Post = family condition after adoption; Int. = interaction between Pre and Post conditions; Low = Low family socioeconomic status; High = High family socioeconomic status; ψ = true interaction contrast value; α = probability of Type I error.

The overall mean IQ score in a study depends on the year it is done and the children's ages (Flynn, 1987), but these will have no impact on required sample sizes when a narrow range of ages is used, as was done by C and D. Using results in C and D as a guide, the means in Table 1 provide an alternative to the hypothesis of no interaction. The required sample size for each test in Table 1 depends strongly on the contrast effect size and the appropriate Type I error probability. The use of 10 children per group by C and D is shown to be quite appropriate to detect the main effects with a one-tailed test using $\alpha = .05$. For the test of interaction, on the

other hand, there are probably not 215 children in the high to low SES condition in all of France. If less extreme values of SES are used, many more children will be available but the effect size will also be reduced substantially and advantages of a 2 × 2 design with homogeneous groups will be lost. Pooling data from several countries might achieve an adequate sample.

If 215 children per group are studied and there still is no significant interaction effect, this would not prove the true effect is exactly zero, but it would suggest it must be very small. The main problem with the C and D findings is that they are consistent with a true interaction effect accounting for a substantial 9% of the variance. A larger sample would impose narrower limits on the likely size of the interaction.

Alternatively, a hybrid approach might be fruitful. Two large groups of birth parents could be identified, those with very low or high SES (B- and B+). Then among their adopted away children, a wide range of adopting family SES could be chosen, and the IQ data could be analysed with multiple regression using effect coding (B- = -0.5, B+ = +0.5) for biological parents, treating adopting family SES as a continuous variable. The test of interaction becomes a test of difference in slopes of the regression lines (Marascuilo & Serlin, 1988).

REFERENCES

Capron. C.. & Duyme, M. (1989). Assessment of effects of socio-economic status on IQ in a full cross-fostering study. *Nature, 340.* 552-554.

Capron, C., & Duyme, M. (1991). Children's IQs and SES of biological and adoptive parents in a balanced cross-fostering study. *Cahiers de Psychologie Cognitive/European Bulletin of Cognitive Psychology. 11.* 323-348.

Carlier. M., Nosten-Bertrand, M.. & Michard- Vanhee, *(* 1992). Separating genetic effects from maternal environmental effects. In D. Goldowitz, D. Wahlsten. & R. E. Wimer (Eds.), *Techniques for the genetic analysis of brain and behavior* (pp. 111-126). Amsterdam: Elsevier.

Clarke, A. M., & Clarke, A. D. B. *(* 1976). *Early experience: Myth and evidence.* New York: Free Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Desnick, R. J., & Grabowski, G. A. (1981). Advances in the treatment of inherited metabolic diseases. In H. Harris & K. Hirschom (Eds.). *Advances in human genetics* (Vol. II, pp. 281-369). New York, NY: Plenum.

Dobbing, J. (1985). Maternal nutrition in pregnancy and later achievement of offspring: a personal interpretation. *Early Human Development. 12.* 1-8.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: what IQ tests really measure. *Psychological Bulletin, 101,* 171-191.

Glass, G. V., & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal, 6,* 403-414.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd edition). New York, NY: Holt, Rinehart & Winston.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage.

Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences.* San Francisco, CA: Freeman.

Maxwell, S. E., Camp, C. J., & Arvey, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology. 66.* 525-534.

McGue, M. (1989). Nature-nurture and intelligence. *Nature, 340,* 507-508.

Morgane, P. J., Austin-LaFrance, R. J .. Bronzino, J. D., Tonkiss, J., & Galler, J. R. (1992). Malnutrition and the developing central nervous system. In R. L. Isaacson & K. F. Jensen (Eds.), *The vulnerable brain and environmental risks. Vol. 1: Malnutrition and hazard assessment* (pp. 3-44). New York. NY: Plenum.

Rose, R. J., Uchida, I. A., & Christian, J. C. (1981). Placentation effects on cognitive resemblance of adult monozygotes. In *Twin research 3: Intelligence, personality and development* (pp. 35-41). New York, NY: Liss.

Schiff. M., Duyme, M., Dumaret, A., & Tomkiewicz. S. (1982). How much *could* we boost scholastic achievement and IQ scores? A direct answer from a French adoption study. *Cognition, 12,* 165-196.
Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association. 62,* 626-633.
Stein, Z., & Susser. M. (1985). Effects of early nutrition on neurological and mental competence in human beings. *Psychological Medicine, 15,* 717-726.
Wahlsten, D. (1990). Insensitivity of the analysis of variance to heredity-environment interaction. *Behavioral and Brain Sciences. 13.* 109-161.
Wahlsten. D. (1991). Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. *Psychological Bulletin, 110,* 587-595.