ADJUSTING FOR CONFOUNDING FACTORS IN QUASI-EXPERIMENTS: ANOTHER REANALYSIS OF THE WESTINGHOUSE HEAD START EVALUATION[1]

Jay Magidson
Abt Associates
Cambridge, MA

Dag Sörbom
U. of Uppsala
Uppsala, Sweden

Abstract

Evaluations of social programs based upon quasi-experimental designs are typically plagued by problems of nonequivalence between the experimental and comparison group prior to the experiment. In such settings it is extremely difficult if not impossible, to isolate the effects of the program from the confounding effects associated with the relevant preexisting differences between the groups. A classic occurrence of the problem was in the 1969 large-scale quasi-experimental evaluation of the Head Start program, the negative findings from which were used to justify phasing out the summer programs. In this paper we reanalyze a portion of the data using Sörbom's (1978) statistical adjustment. The results do not support the strong inferences drawn by the original evaluators.

Introduction

One of the most challenging methodological problems facing evaluators of social programs is that of the nonequivalent comparison group. In the case that the comparison group differs from the experimental group in important ways prior to the program (or treatment) these differences are confounded with the treatment and therefore compete with the treatment in explaining any post program differences between the groups. How should an evaluator determine what portion of post program differences to attribute to the sources of preexisting differences and what portion to attribute to the program? What analytic technique should be used and under what assumptions can a reasonable mathematical model be specified? Or should an evaluator disregard quantitative approaches altogether? After all, Lord (1967) points out that in such situations "no logical or statistical procedures can be counted on to make proper allowances for uncontrolled preexisting differences between groups."

While there are no unequivocal answers to these questions, under certain situations, some approaches have important advantages over others. This paper illustrates and recommends a general methodology for a particular situation which is, unfortunately, all too common in the evaluation of social programs. The weak quasi-experimental design to be described here is the rule rather than the exception in practice.

We will be stressing throughout this paper the need to interpret our results cautiously as well as the results of any quantitative analyses of data obtained from less than a true randomized experiment. Nevertheless, it is important to recognize that policy decisions are often based upon quantitative analyses of data obtained from less than a true randomized experiment. Nevertheless, it is important to recognize that policy decisions are often based upon quantitative analyses of data obtained from weak designs. Thus, evaluators may wish to apply a variety of difference techniques to the data to determine whether the conclusions differ depending upon the different analytic assumptions underlying the techniques. While we believe that multiple analyses are necessary in this case, no strategy is sufficient to assure that all relevant confounds have been appropriately taken into account. Ultimately, one must rely upon theory to help interpret the results. The weaker the design of the study, the heavier the burden of interpretation must rest with theory.

Background

The data reanalyzed here are taken from the original evaluation of the summer Head Start program as conducted in 1969 by the Westinghouse Learning Corporation. This was an important evaluation in that it represented the first large-scale national study to evaluate the impact of Head Start on school achievement, focussing on children who attended Head Start summer programs in 1965, 1966, 1967 or 1968.

The design of the study was ex post facto and no pretest data was available. Nonattendees from the same communities as the Head Start pupils were recruited for the comparison group. The comparison children were matched to the Head Starters on race and kindergarten enrollment but not on socio-economic status because the researchers felt this would be too costly. Although the evaluators acknowledge that the design suffered a lack of internal validity (Cicirelli et al., 1969: 34), they present the results in no uncertain terms:

> "Results from the summer program are so negative that it is doubtful that any change in design would reverse the findings." (Cicirelli et al., 1969: 245)

After the completion of the study, the directors of the summer Head Start programs (involving over a million children) were given the option of shifting their funds to full-year programs (Smith and Bissell, 1970).

Some reanalyses of these data yielded different conclusions. For example, Barnow (1973) concluded that the summer programs (as well as the full-year programs) were effective for blacks and Mexican Americans but not for whites. Magidson (1977), using a different analytic

approach, concluded that the summer programs also had a small positive effect on the whites. Cicirelli, Barnow and Magidson all attempted to control for socioeconomic status in their analysis, although each measured it in a different manner. Magidson (1977) and Bentler and Woodward (1978) agree that the original ANCOVA analysis by Cicirelli inappropriately used a forecasting technique (analysis of covariance) to estimate causal parameters of effect.

It is not the purpose of this paper to argue the merits of these differing conclusions. Rather, we present an alternative analysis of the data analyzed by Magidson (1977), and argue that the assumptions made here are more realistic than those underlying the earlier analyses. More specifically, our present approach improves upon previous analyses in the following ways:

1. It recognizes that the Head Start and comparison groups are separate and distinct populations (this was acknowledged by Bentler and Woodward, 1978, and by Magidson, 1978, to be a desirable property).

2. It offers a statistical test of the null hypothesis that the two groups are equal on a latent factor we call Socio-Economic Advantage (S).

3. A goodness of fit statistic providing an overall test of the assumptions of the model indicates that the model fits the data better than any previous model.

Description of the Data

Two tests of cognitive ability were used in the original evaluation as the criterion measures of performance. The first, denoted Y_1, is the Illinois Test of Psycholiguistic Abilities (ITPA). The object of the ITPA is to aid in the diagnosis of specific abilities and disabilities and to guide in the administration of remedial work. It is comprised of ten subtests: auditory reception, visual reception, auditory-vocal association, visual-motor association, verbal expression, manual expression, grammatic closure, visual closure, auditory sequential memory, and visual sequential memory. The second outcome measure, denoted Y_2, is the Metropolitan Readiness Test (MRT), consisting of six subtests: word meaning, listening, matching, alphabet, numbers, and copying.

The two samples consist of 148 white six-year-old first graders who previously attended a Head Start summer program and 155 white six-year-old first grade comparison children from the same communities who did not attend a Head Start summer nor full-year program. The comparison children were selected (after the Head Start experience had been completed) to match the Head Start sample on age, sex and kindergarten attendance. Only children for which data on both parents was available are included in our analysis here. As mentioned above, this is the same sample analyzed earlier by Magidson, a subset of the first grade summer sample analyzed by Cicirelli et al.

Table 1 displays the correlations, means, and standard deviations for the post tests (Y_1 and Y_2) and four measures of socio-economic status, (X_1) mother's education, (X_2) father's education, (X_3) father's occupation and (X_4) Income.[2] Notice that although the comparison children outscore the Head Start children on each of the two tests, they are also higher on each indicator of socio-economic status. Thus, it seems reasonable to conclude that if pre-test data were available it would similarly show the comparison children outscoring the Head Start children, even before the Head Start experience.

Analytic Approach

We assume that there is a single relevant causative factor upon which the two groups differ. We label this factor Socio-Economic Advantage (S) and hypothesize that each of the four X-variables are measures or indicators of S, being positively correlated with it. Thus, the critical aspect of the analysis is the specification of a measurement model for S. After obtaining a reasonable measurement model, we use it to estimate the impact of socio-economic advantage on cognitive ability, and remove (adjust for) it in the analyisis. The resulting difference in cognitive ability between the Head Start and Comparison children after adjusting for the effects of S is attributed to the summer Head Start experience.

In the next section we develop the measurement model for S. In the following sections we develop the causal model for adjusting for S and estimate the effects of Head Start.

The Measurement Model for S

The measurement model relates the four observed indicators of socio-economic advantage to the unobserved latent construct S as in a single factor model in traditional factor analyses. However, there are some important differences between the single factor model presented here and the traditional model. For more details about this model, see Sörbom (1978).

First, our model is a 2-group rather than a 1-group model. We will assume that the factor loadings are the same for both groups but since we explicitly formulate the model for two groups, this assumption can be tested. The traditional single population model does not allow a direct test of equal factor loadings on selected subsets in the population.

A second difference between our model and the traditional model is that the traditional model assumes that the factor accounts for all the correlations among the variables. That is, it requires that the unexplained or residual components of each of the variables be uncorrelated with each other. For example, suppose mother's and father's education (X_1 and X_2) shared some

correlation not shared by the other S-indicators. In this case, the traditional approach would require explicitly hypothesizing a second factor to adequately account for all correlations. Our approach, on the other hand, does not require residuals to be uncorrelated. Thus, in the above example, a single factor model allowing for correlated residuals associated with mother's and father's education could be formulated without the necessity of explicitly hypothesizing a second factor. Since it is reasonable to believe that a person may marry another with a similar level of education for reasons at least partially unrelated to social class, it is convenient to allow for such residual correlation without explicitly creating a "nuisance factor" which would be irrelevant to the theory being tested.

The third difference is that the traditional model relates only to the correlations or covariances among the observed variables. Alternatively, our model includes a structure on the means. Thus, our model can test the hypothesis that a single factor not only accounts for the observed correlations among the variables in both groups but also explains the difference in means on these variables. Thus, we can test the hypothesis that the higher values for the comparison group on all four S-indicators is that amount expected under our single factor model.

The final difference is that our approach allows us to directly estimate the difference in factor scores between the groups. Thus, we can estimate the socio-economic advantage for the comparison group relative to the Head Start group and to test whether this value is significantly greater than zero.

Formally, the model is:

$$
\begin{aligned}
X_1^{(H)} &= v_1 + \lambda_1 S^{(H)} + z_1^{(H)} \\
X_2^{(H)} &= v_2 + \lambda_2 S^{(H)} + z_2^{(H)} \\
X_3^{(H)} &= v_3 + \lambda_2 S^{(H)} + z_3^{(H)} \\
X_4^{(H)} &= v_4 + \lambda_4 S^{(H)} + z_4^{(H)} \\
& \qquad\qquad (1) \\
X_1^{(H)} &= v_1 + \lambda_1 S^{(C)} + z_1^{(C)} \\
X_2^{(C)} &= v_2 + \lambda_2 S^{(C)} + z_2^{(C)} \\
X_3^{(C)} &= v_3 + \lambda_3 S^{(C)} + z_3^{(C)} \\
X_4^{(C)} &= v_4 + \lambda_4 S^{(C)} + z_4^{(C)}
\end{aligned}
$$

where the superscripts H and C refer to Head Start and the Comparison group respectively; the v's are unknown constants and the z's are random variables having zero expectation.

This model assumes that the factor loadings (λ) are the same in each group and that the S factor explains the differences in means ($\mu^{(C)} - \mu^{(H)}$) between the groups. The latter follows from the assumption that the v- constants are equal between groups.

Thus, model (1) implies

$$\mu_k^{(H)} = v_k + \lambda_k E(S|\text{Head Start}) \qquad (2)$$

and $$\mu_k^{(C)} = v_k + \lambda_k E(S|\text{comparison group}) \qquad (3)$$

and thus

$$\mu_k^{(C)} = \mu_k^{(H)} + \lambda_k [E(S|\text{Comparison}) - E(S|\text{Head Start})] \qquad (4)$$

Note from (2) that if we add any constant c to E(S|Head Start), the constant can be absorbed into the intercept by subtracting $\lambda_k^{(C)}$. This means that v_k and E(S|Head Start) cannot be identified simultaneously without imposing some restrictions on this parameter space. Thus, without loss of generality we make the following restriction:

$$E(S|\text{Head Start}) = 0 \qquad (5)$$

so that E(S|comparison group) represents the average socio-economic advantage of the comparison group relative to the Head Start group.

As shown in the appendix, model (1), as identified using restriction (5), can be estimated by the LISREL IV computer program (Jöreskog and Sörbom, 1978). We first estimate the model assuming that within each group the z's are uncorrelated. X_4 in Table 1 we get a measure of goodness of fit for the overall model of 43.6. This measure is asymptotically distributed as X^2 with 10 degrees of freedom under the assumption that within each group the X - vector has a multivariate normal distribution, thus indicating that the fit of the model is not very good.

Sörbom (1975) suggests improving the fit of a model by relaxing (allowing to be estimated) that fixed parameter that has the largest derivative. An examination of the derivative of the free parameters reveals that there might be a significant correlation between the errors z_1 and z_2. That is, when the correlation among the observed variables caused by the construct S has been accounted for, there seems to be a correlation left between mother's education (X_1) and father's education (X_2). As mentioned earlier, it is reasonable to believe that parents' education levels correlate more than can be explained by social class or by the social advantage construct. Including the z_1, z_2 correlation as an additional parameter yields a model with an acceptable fit, X^2 with 8 degrees of freedom equals 10.7 (P = .22). The difference in degrees

of freedom from the previous model is 2, since we have now added two parameters, namely the correlation of z_1 and z_2 in the two groups. The difference in X^2 between the two models is highly significant with 2 degrees of freedom indicating that the correlation between z_1 and z_2 (estimated to be .58 for the Head Start group and .68 for the controls) are significantly different from zero.

The estimates and their estimated standard errors are given in Table 2. We can see that the compariosn group is significantly advantaged in comparison with the Head Start group. The difference in the mean values of S is 0.271 with a standard error equal to 0.076, a highly significant result.

By the above analysis we have justified the use of the variables X_1 - X_4 as indicators of the same construct S for both the Head Start and comparison groups. We have called this construct socio-economic advantaged. We have also found that the two groups significantly differ on this construct, the comparison group being the more advantaged (or less disadvantaged) group.

Measurement of Cognitive Ability

As criterion Magidson (1977) used two cognitive ability tests, the Metropolitan Readiness Test (ITPA). He made separate analyses for the two tests. In this paper we follow the approach used by Bentlter and Woodward (1978), and take both tests to define the latent construct "cognitive ability" (A). The model is depicted in Figure 2 and data for both groups on variables Y_1 and Y_2 are given in Table 1. Again, to fix the origin of the latent variable A, we set the mean value of A in the Head Start group equal to zero.

This model has no degrees of freedom, so we can compute the estimates simply by equating the first and second order moments implied by the model to their observed counterparts. The estimates and their estimated standard errors are listed in Table 3. We can see that the comparison group also scores significantly higher in cognitive ability than the Head Start group since the estimated expectation for the construct is positive (it is estimated to be 0.743). However, the difference is non-significant at the 0.05 level, having a standard error equal to 0.439.

Adjusting for the Effect of S

The next model to investigate is the combined model as depicted in Figure 3 where the main focus is on the structural equation

$$(3) \qquad A^{(g)} = \alpha^{(g)} + \beta^{(g)} S^{(g)} + \zeta^{(g)}$$

where g designates the group g = H or C. As previously discussed in the case of S, there is no absolute origin for A. All we can do is to compare groups and look at differences. For example, we could fix α in the control group to be zero, and then α in the Head Start group could be interpreted as the effect of the Head Start program when social class has been controlled for (assuming $\beta^H = \beta^C$).

The χ^2 for the combined model equals 31.04 with 22 degrees of freedom so the fit of the model is acceptable (P = 0.10). An examination of the β parameters in the two groups shows that they probably are equal, since $\hat{\beta}^{(H)} = 2.698$ and $\hat{\beta}^{(C)} = 2.521$ with estimated standard errors equal to 0.870 and 0.771, respectively. Thus, the final model is a model with the β:s constrained to be equal. The χ^2 for this model is 31.08 with 23 degrees of freedom. The difference in χ^2 for the last two models can be used as a test of the hypothesis that the β:s are equal. χ^2 with 1 degree of freedom is 0.04 and thus we can treat the β:s as equal. Then it is meaningful to talk about α as a measure of the effect of Head Start.

The estimates of the model are listed in Table 4. We note no statistically significant effect for the Head Start program when controlling for socio-economic advantage (S), although the inclusion of S has changed the effect estimate from negative to positive. The estimate of α a is 0.131 with a standard error equal to 0.373.

In more general case, when we have more than two groups and/or more than one dependent variable, we can test the hypothesis of no effect by reestimation of the model with the restriction $\alpha^{(1)} = 0$ added, and then compare χ^2 :s. In the above case we obtain χ^2 equal to 31.20 with 24 degrees of freedom. The test of no effect results in a χ^2 with 1 degree of freedom equal to 0.12 which in this case is the same as what we get when we compute the squared z statistic which equals the square of the estimate of α divided by its standard error.

Appendix

Estimation with the LISREL IV computer program

Using matrix notation the model in figure 3 could be specified

$$(4) \quad \begin{matrix} y = \Lambda\eta + \varepsilon \\ B\eta \quad \Gamma 1 + \zeta \end{matrix}$$

where

$$Y = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ Y_1 \\ Y_2 \end{bmatrix} \quad \Lambda = \begin{bmatrix} 1 & 0 & v_1 \\ \lambda_2 & 0 & v_2 \\ \lambda_3 & 0 & v_3 \\ \lambda_4 & 0 & v_4 \\ 0 & 1 & v_5 \\ 0 & \lambda_6 & v_6 \end{bmatrix} \quad \eta = \begin{vmatrix} S \\ A \\ 1 \end{vmatrix} \quad \varepsilon = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 0 & 0 \\ -\beta & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \Gamma = \begin{bmatrix} \theta \\ \alpha \\ 1 \end{bmatrix} \quad \zeta = \begin{bmatrix} S-\theta \\ \zeta \\ 0 \end{bmatrix}$$

(4) is a special case of the full LISREL model (see Jöreskog and Sörbom, 1978), which is specified by the equations

$$Y = \Lambda_Y \eta + \varepsilon$$

$$x = \Lambda_x \xi + \delta$$

$$B\eta = \Gamma\xi + \zeta$$

In (4) the ξ variables in (5) are specified to be a single variable which is fixed to be equal to 1 for all observations and the third η- variable is set equal to this fixed variable. By this it is possible to incorporate the location parameters, μ_1, μ_2,..., μ_6, θ, and α into the model.

Since we are specifying a structure of the means of the observed variables we should analyze the sample second order moment matrix instead of the usually used sample covariance matrix. In order to get a ξ variable identically equal to 1 we can add a sample variable $w \equiv 1$, i.e. in the sample moment matrix we add a row and a column with entries equal to the sample means of the y - variables and a diagonal element equal to 1. Then, by letting $\Lambda_x = I$ and $\delta = 0$ in (5) we see that we get the specification in (4). As a matter of fact, in the LISREL IV computer program this kind of model is automatically generated if the so called fixed - x - case is specified. The program can handle raw data, covariance matrices, or correlation matrices as input and then compute the moment matrix if also the sample means are supplied for the latter two cases.

By the above use of the LISREL IV program it can be shown (see Jöreskod & Sörbom, 1979) that the likelihood function that the program maximizes to get the estimates will be correctly computed. Hence, also the first order derivatives and the expected second order derivatives will be correct, which in turn implies that the estimates of the standard errors for the estimated parameters will be correct.

Footnotes

1. Supported in part by NIE-G-79-0128 and by DAR 7820374.

2. The coding of all variables was that used by Barnow (1973) except that (X_4) Income is measured in thousands of dollars.

References

Bentler, P. and Woodward, J.A. A Head re-evaluation : positive effects are not yet demonstrable. Evaluation Quarterly, 1978, 2, 493-510.

Cicerelli, V. G. et al. The impact of Head Start; An evaluation of the Effects of Head Start on Children's Cognitive and Affective Development, Vols. 1 and 2. A report presented to the Office of Economic Opportunity pursuant to contract B89-4536, June, 1969. Athens : Ohio University and Westinghouse Learning Corporation.

Jöreskog, K. G. and Sörbom, D. LISREL IV: Analysis of linear structural relationships by the method of maximum likelihood. Chicago : International Educational Services, 1978.

Jöreskod, K. G. and Sörbom, D. Simultaneous Analysis of Longitudinal Data from Several Cohorts. Paper presented at the SSRC conference on "Analyzing Longitudinal Data for Age, Period and Cohort Effects", in Snowmass, Colorado, June, 1979.

Magidson, J. Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation : A general alternative to ANCOVA. Evaluation Quarterly, 1977, 1, No. 3, 399-420.

Sorbom, D. Detection of Correlated Errors in Longitudinal Data. British Journal of Mathematical and Statistical Psychology. 1975, 28, 138-151.

Sörbom, D. An Alternative to the Methodology for Analysis of Covariance. Psychometrika, 1978, 43(3).

Table 1 Correlations, standard deviations and means for the Head Start and Comparison Group Data

Head Start Group

	Correlations						Standard deviations	Means
X_1	1.000						1.032	3.210
X_2	.466	1.000					1.281	3.081
X_3	.253	.203	1.000				1.075	2.088
X_4	.361	.182	.377	1.000			2.648	5.358
Y_1	.275	.265	.208	.084	1.000		3.765	19.672
Y_2	.256	.122	.251	.198	.664	1.000	2.677	9.562

Control Group

	Correlations						Standard deviations	Means
X_1	1.000						1.022	3.387
X_2	.561	1.000					1.195	3.290
X_3	.224	.342	1.000				1.193	2.600
X_4	.306	.215	.387	1.000			3.239	6.435
Y_1	.239	.215	.196	.115	1.000		3.901	20.415
Y_2	.281	.297	.234	.162	.635	1.00	2.719	10.070

X_1 = Mother's education
X_2 = Father's education
X_3 = Father's occupation
X_4 = Income (thousands of dollars)

Y_1 = Post test ITPA
Y_2 = Post test MRT

Table 2 Estimates for the final measurement model of socio-economic advantage with standard error estimates within parenthesis.

	Head Start Group				Control Group	
v_1			3.432	(0.071)		
v_2			3.324	(0.081)		
v_3			2.558	(0.089)		
v_4			6.466	(0.238)		
λ_1			1.000†			
λ_2			1.034	(0.179)		
λ_3			1.596	(0.310)		
λ_4			4.162	(0.811)		
σ^2_{z1}	0.843	(0.110)			0.832	(0.111)
σ^2_{z2}	1.467	(0.183)			1.155	(0.149)
σ^2_{z3}	0.743	(0.132)			0.846	(0.157)
σ^2_{z4}	3.936	(0.809)			6.550	(1.133)
σ_{z1z2}	0.422	(0.109)			0.447	(0.103)
E(S)	0†				0.271	(0.076)
σ^2_S	0.172	(0.060)			0.223	(0.075)

†fixed parameters to specify the scale of S.

Table 3 Estimates for the model of cognitive ability (A) with standard error estimates within parenthesis

	Head Start Group				Control Group	
v_5			20.415	(0.313)		
v_6			10.070	(0.218)		
λ_6			0.684	(0.344)		
$\theta^2_{\varepsilon_5}$	4.348	(5.010)			5.328	(5.061)
$\theta^2_{\varepsilon_6}$	2.571	(2.348)			2.772	(2.369)
E(η)	0.0				0.743	(0.440)
θ^2_η	9.730	(5.112)			9.789	(5.147)

Table 4 Estimates for the combined model with standard error estimates within parenthesis.

	Head Start Group				Control Group	
v_1			3.444	(0.072)		
v_2			3.337	(0.082)		
v_3			2.559	(0.089)		
v_4			6.410	(0.229)		
v_5			20.357	(0.286)		
v_6			10.085	(0.216)		
λ_1			1.0†			
λ_2			1.057	(0.165)		
λ_3			1.476	(0.259)		
λ_4			3.517	(0.625)		
λ_6			0.850	(0.142)		
σ^2_1	0.807	(0.108)			0.787	(0.108)
σ^2_2	1.433	(0.182)			1.083	(0.144)
σ^2_3	0.718	(0.123)			0.856	(0.146)
σ^2_4	4.395	(0.730)			7.223	(1.052)
σ^2_5	6.267	(1.536)			7.280	(1.597)
σ^2_6	1.458	(0.970)			1.637	(1.002)
$\sigma_{z_1z_2}$	0.386	(0.108)			0.390	(0.099)
β			2.135	(0.549)		
α	0.0†				0.131	(0.373)
σ^2_S	0.209	(0.066)			0.261	(0.080)
σ_ε	6.377	(1.487)			6.187	(1.472)
E(S)	0.0†				0.295	(0.081)

† fixed

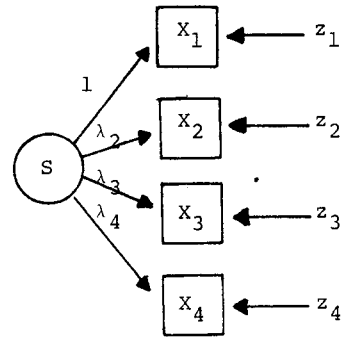

Figure 1. The initial model for the social class construct.

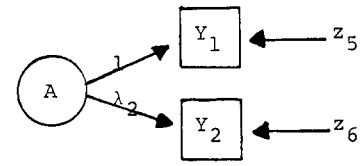

Figure 2. The model for cognitive ability.

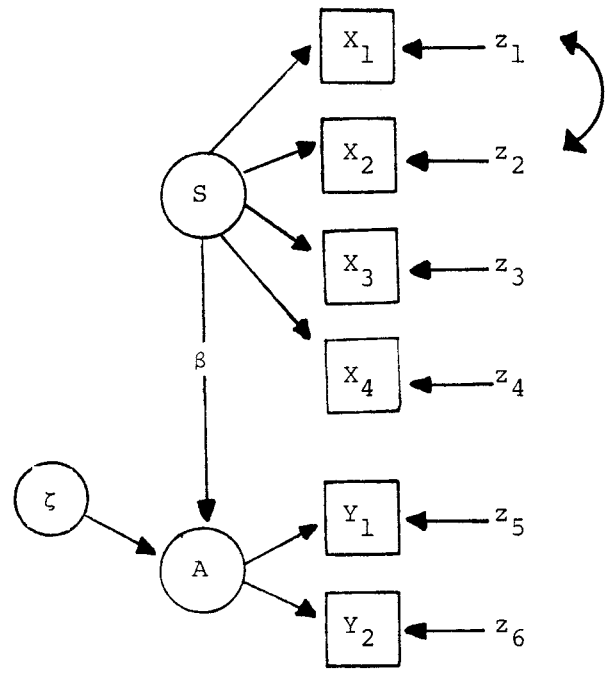

Figure 3. The combined model.