6

A study of longitudinal causal models comparing gain score analysis with structural equation approaches

LESLIE HENDRICKSON AND BARNIE JONES

Introduction

Current trends in applied research have witnessed the widespread adaptation of multiple regression techniques to research projects and program evaluations. Although regression analysis is a powerful technique, it owes much of its power to highly restrictive and often unrealistic assumptions. The interpretation of regression results, especially the assessment of the relative impact or importance of independent variables, can be difficult.

This chapter compares methodological procedures for analyzing longitudinal data. It critically compares regression analysis of gain scores with structural equation approaches. The analytic techniques discussed here are applicable to any longitudinal analysis. These general techniques are exemplified by the secondary analysis of data from the *What Works in Reading?* study conducted by the School District of Philadelphia (Kean et al. 1979a,b).

Following an introduction to the data, the analysis proceeds in three steps. First, specification of the dependent variable is examined. The original report (Kean et al. 1979a) treated reading improvement as a net change or gain score. Gain scores are widely used in American schools. Results of using the gain score as a dependent variable are compared with results obtained when reading at time 1 (T_1) and reading time 2 (T_2) are treated as separate dependent variables in a longitudinal model (see Models 1 and 2 in Figures 6.1 and 6.2, respectively). Second, the model is reformulated as a latent variable structural model to relieve problems due to collinearity among the independent variables. Third, the latent variable model is subjected to a sensitivity analysis (Land & Felson 1978) with regard to random measurement error in the dependent variables and to specification error due to the omission of theoretically important independent variables. This analysis demonstrates how small changes in model specification and residual assumptions can modify results.

Sample and data collection

The original sample consisted of 1,800 fourth grade students in 25 schools drawn from a population of 190 schools. Schools were stratified on the basis of average scores in 1974 and 1975, for grades 1–4, on reading portions of the California Achievement Test (CAT). The sample excluded schools that showed major shifts in average reading score level from 1974 to 1975, selecting 10 with high, 10 with low, and 5 with medium scores in both years. Schools were selected from all eight administrative subdistricts of the city. The resulting sample is representative of the range of average school achievement levels in the district, but it purposely screens out schools in which the average ability level is changing. Student-level data were gathered from school records. In all, data on 245 variables were gathered and analyzed.

Selection of variables

Using regression analysis, the researchers (Kean et al. 1979a,b) narrowed the field from 245 to 18 variables that had statistically significant regression coefficients when predicting change in reading achievement. The selection process by which these variables were identified was evidently statistical significance alone.[1] Our secondary analysis began with these 18 variables. Seven were quickly eliminated because they accounted for less than 1 percent of the variance in the dependent variable and appeared to contribute nothing to the analysis.

Table 6.1 lists definitions, means, and standard deviations for 11 of the independent variables and for the 3 dependent variables: the gain score and the third and fourth grade reading scores. The 11 independent variables include measures of student, teacher, and school organization. These variables were selected because the Philadelphia researchers found that they had a statistically significant β weight in predicting the gain score.

Table 6.2 shows the correlation matrix of the variables listed in Table 6.1. The impression obtained from Table 6.2 is that the matrix is thin. Of the 90 correlations in it, only 19 percent are greater than .15, and only 13 percent are greater than .25. Among pairs of the 11 independent variables only 9 percent of the correlations are greater than .25. The highest correlation of any variable with CATGAIN, the gain score, is .08.

Gain score model

The regression analysis used the difference between the third and fourth grade reading achievement scores as a single dependent variable. The use

Table 6.1. *Code names, definitions, means, and standard deviations for 11 independent variables and 3 dependent variables*

Code name	Definitions	Mean	S.D.
X_1	Days students were present in grade 4	130.51	10.41
X_2	Student attended kindergarten, 1 = NO, 2 = YES	1.80	0.40
X_3	Number of nonteaching support staff per school, grade 4	—[a]	11.02
X_4	Percentage of students scoring above 84th percentile in California Achievement Test 1976 – Total Reading, measured at grade 4; the grade 3 proportion is assumed to be similar to grade 4	0.20	0.13
X_5	Percentage of classroom teachers with less than two years of experience; measured at grade 4; the grade 3 proportion is assumed to be similar to grade 4	0.20	0.14
X_6	Number of teacher pay periods with no absence	13.89	3.79
X_7	Teacher attends outside professional conference meetings, 1 = NO, 2 = YES	1.17	0.39
X_8	First year teaching grade 4, 1 = NO, 2 = YES	1.17	0.35
X_9	Minutes per week of individual independent reading	73.35	60.31
X_{10}	Teacher would select the same reading program again	1.54	0.50
X_{11}	Times per week aide in room during reading	2.55	2.31
$T_1 - T_2$	Difference between grade 3 and grade 4 scale score	28.43	52.50
T_1	California Achievement Test – Reading Comprehension Scale Score for grade 3, 1975	385.06	67.74
T_2	California Achievement Test – Reading Comprehension Scale Score for grade 4, 1976	412.50	72.56

[a] The mean for X_3 was not shown in the November 1979 technical report of *What Works in Reading?* (Kean et al. 1979b). If not indicated, variable is measured at grade 4.

of "difference," "change," or "gain" scores has been thoroughly examined (Thorndike & Hagen 1955; McNemar 1958; Thorndike 1966; Bohrnstedt 1969; Cronbach & Furby 1970; Alwin & Sullivan 1975; Kim & Mueller 1976; Kessler 1977; Pendleton, Warren & Chang 1979). As a result of these examinations the use of gain scores has been discouraged, because the difference between the two measures has lower reliability than the

Table 6.2. *Correlation matrix for 14 variables in Philadelphia achievement study (N = 1,363)*

	Days stud. present X_1	Stud. attended X_2	No. of non-teaching support staff X_3	% Studs. above 84th percentile CAT – 1976 X_4	% Classrm. teachers with less than 2 yr. exper. X_5	No. of teacher pay periods with no absence X_6	Teacher attends outside conf. X_7	1st yr. teaching gr. 4 X_8	Min./wk. individ. independ. reading X_9	Teacher select same reading program X_{10}	Aide time during reading each wk. X_{11}	Diff. betw. gr. 3 & gr. 4 scale score $T_2 - T_1$	CAT – Read. Comp. Scale Score, gr. 3–1975 T_1	CAT – Read. Comp. Scale Score, gr. 4–1976 T_2
X_1	1.000													
X_2	.115	1.000												
X_3	−.134	−.145	1.000											
X_4	.134	.141	−.626	1.000										
X_5	−.056	−.095	.327	−.154	1.000									
X_6	−.017	−.006	.135	−.040	.078	1.000								
X_7	−.033	.004	−.027	−.082	.118	.005	1.000							
X_8	−.061	.013	.104	.007	.021	.021	.021	1.000						
X_9	.086	.034	−.143	.113	.005	−.004	−.125	.021	1.000					
X_{10}	−.029	−.086	−.142	.030	−.033	.080	.023	−.011	−.119	1.000				
X_{11}	−.116	−.110	.527	−.427	.074	.010	.234	−.012	−.069	−.106	1.000			
$T_2 - T_1$	.074	.020	−.004	−.051	.071	.042	−.037	−.002	.083	−.007	.004	1.000		
T_1	.161	.123	−.290	.386	−.118	.076	−.112	−.139	.090	.187	−.362	−.292	1.000	
T_2	.197	.129	−.273	.382	−.065	.101	−.132	−.132	.144	.169	.335	N.A.	.722	1.000

measures considered separately. Consequently, their use requires low error variance and high reliability of measurement. Also, calculations of the gain score reliability tend to be untrustworthy because the calculations depend on five estimates: three correlations and two variances. Finally, the analysis of gain scores is complicated by the effects of regression toward the mean.

However, in addition to the problem of poor reliability, there is another, perhaps more serious problem with the gain score model. To illustrate, consider the following two models:

$$T_2 = \beta T_1 + \sum \lambda_i' X_i + e_1 \tag{6.1}$$

$$T_2 - T_1 = \sum \lambda_i X_i + e_2 \tag{6.2}$$

Equation (6.1), which we call the conditional model, is derived from Figure 6.2. Equation (6.2) represents the gain score model described in Figure 6.1. If T_1 is added to both sides of (6.2), the result is

$$T_2 = T_1 + \sum \lambda_i X_i + e_2 \tag{6.3}$$

Comparing (6.3) with (6.1), it can be concluded that, unless $\beta = 1$,

$$\sum \lambda_i' X_i + e_1 \neq \sum \lambda_i X_i + e_2 \tag{6.4}$$

One could say that the gain score model produces biased estimates of the effects of the independent variables by unnecessarily constraining β to equal 1.

A structural equation model

Model 6.1, a gain score model, is shown in Figure 6.1. All independent variables are assumed to influence the gain score and are assumed to be measured without error. One alternative to the gain score analysis is one that uses data from both time points rather than the difference score. The analyses reported in this chapter used the maximum likelihood exploratory factor analysis (EFAP) and structural equation programs (LISREL V) of Jöreskog and Sörbom (1981).

Figure 6.2 shows one alternative model (Model 6.2) for analyzing the Philadelphia data using both dependent variables, T_1 and T_2, together instead of analyzing their difference. Two structural equations were estimated using the 11 variables; first the third grade achievement variable was used as the dependent variable, then the fourth grade variable was used. Three of the 11 variables are hypothesized to influence both the third and fourth grade scores, whereas the other eight are hypothesized to influence only the fourth grade score. The three variables influencing scores at both times were the student's attendance in kindergarten (X_2),

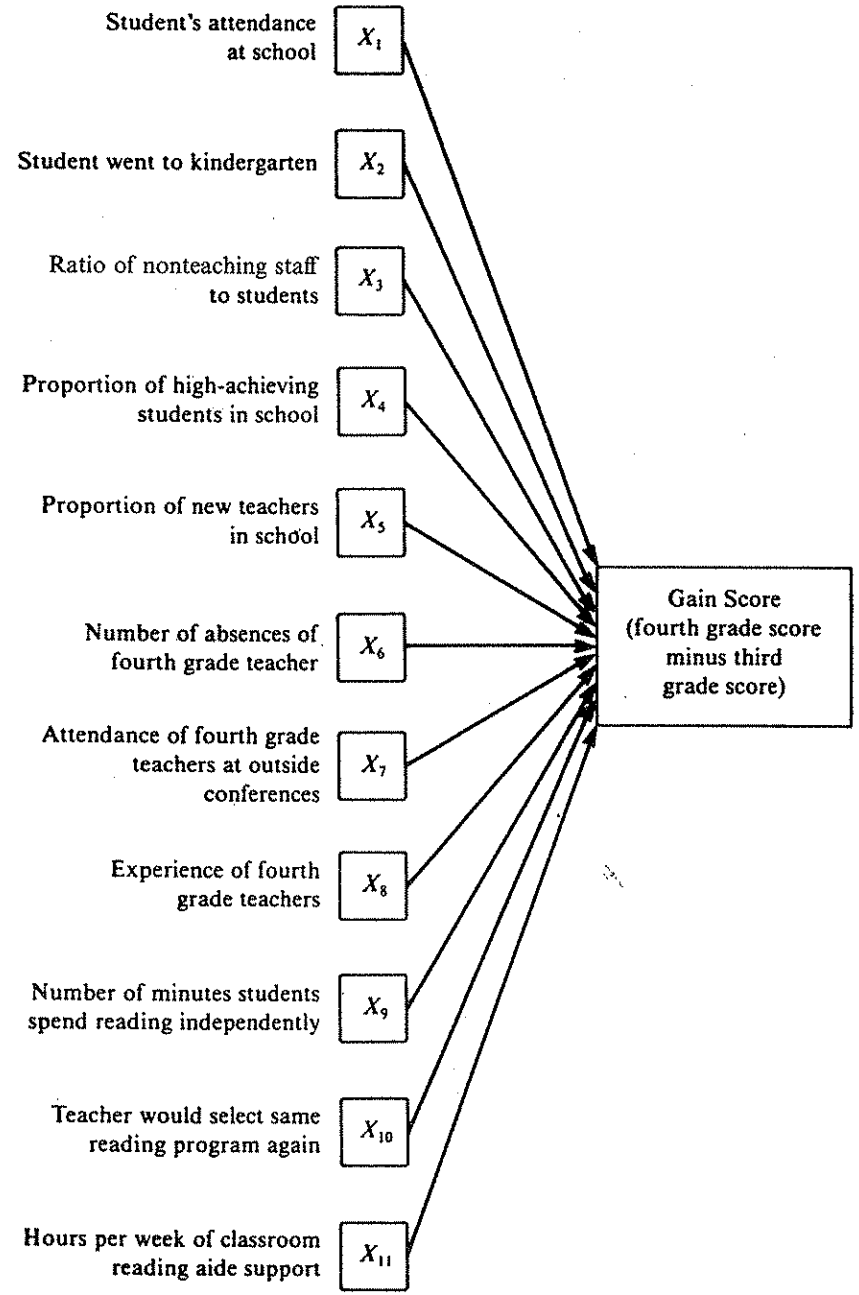

Figure 6.1. Model 6.1: a gain score model. No time assumptions are made; all independent variables are assumed to affect a single dependent variable. All error terms are assumed to be zero.

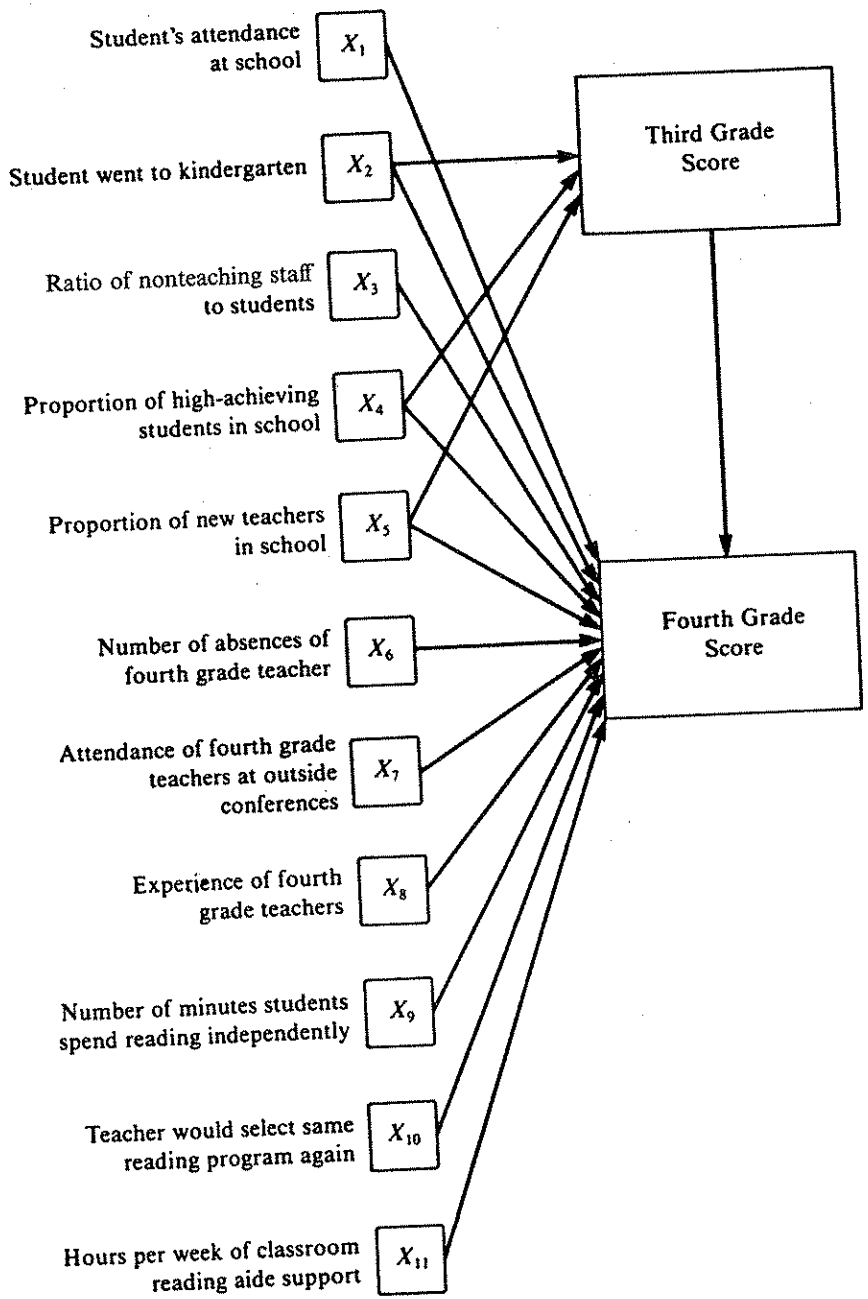

Figure 6.2. Model 6.2: a longitudinal model. Three independent variables are assumed to affect T_2, the third grade score. All variables and the third grade score are assumed to affect T_2, the fourth grade score. All error terms are assumed to be zero.

the proportion of students in the school scoring well on the achievement test (X_4), and the proportion of new teachers (X_5).

This model is recursive in that the fourth grade score is assumed to have no effect on the third grade score. The identification of recursive models is usually obtained by making particular assumptions about error terms. A common procedure for identifying Model 6.2 is to assume that the disturbance terms (residuals) are uncorrelated and that the independent variables are measured without error. Two additional modeling strategies might be proposed. One is simply to include T_1 among the X_i in the following single-equation model:

$$T_2 = \sum \beta_{i+1} X_{i+1} + e_3 \tag{6.5}$$

The difficulty here is that autocorrelation in the residuals of the serially measured variables T_1 and T_2 will affect all the estimates of the b_i.

Bornstedt (1969) has proposed a method using residualized scores that avoids this problem. The first step is to calculate the regression of T_1 on T_2, as in (6.6):

$$T_2 = \beta T_1 + e_4 \tag{6.6}$$

Although estimates of β and e_4 may be inefficient because of autocorrelation, they are unbiased (Johnston 1972, p. 246; Neter & Wasserman 1974, p. 352). Estimates for the remaining independent variables are then obtained by solving for β in (6.7):

$$e_4 = \sum \beta_i X_i + e_5 \tag{6.7}$$

However, rearranging (6.6), we see that

$$e_4 = T_2 - \beta T_1 \tag{6.8}$$

And substituting for e_4 in (6.7) it is evident that (6.9) and (6.10) are formally equivalent to (6.1); that is, the longitudinal model is formally equivalent to Bornstedt's (1969, p. 118) model for residualized scores:

$$T_2 - \beta T_1 = \sum \beta_i X_i + e_5 \tag{6.9}$$

$$T_2 = \beta T_1 + \sum \beta_i X_i + e_5 \tag{6.10}$$

This method does not resolve doubts about estimates of β when autocorrelation is present. However, it does isolate β so that more satisfactory estimates of the other slope coefficients in the equation can be obtained. To a limited extent, additional steps are taken to address autocorrelation in the sensitivity analysis presented later in this chapter.

Table 6.3. *Estimates of Model 6.1 and Model 6.2*[a]

Independent variable	Gain Score Model	Longitudinal Model	
	$T_2 - T_1$	T_1	T_2
X_1	0.373* (0.074)		0.579* (0.083)
X_2	2.587 (0.020)	10.445* (.062)	4.183 (0.024)
X_3	−0.407* (−0.085)		0.434* (0.066)
X_4	−44.753* (−0.111)	190.846* (.366)	63.114* (0.113)
X_5	31.111* (0.081)	−24.782* (−.050)	3.733 (0.007)
X_6	0.635 (0.046)		1.064* (0.056)
X_7	−3.829 (0.028)		−4.075 (−0.022)
X_8	1.803 (0.012)		14.180* (−0.068)
X_9	0.062* (0.071)		0.123* (0.102)
X_{10}	−0.625 (−0.006)		12.267* (0.085)
X_{11}	0.010 (0.000)		−3.642* (−0.116)
T_1			0.653* (0.610)
R^2	.03	.20	.47
χ^2/d.f.		278/8	

[a] Asterisked values are significant at less than .05; values in parentheses are standardized estimates.

Analysis of Models 6.1 and 6.2

Estimates were obtained for all 11 independent variables. These estimates are presented in Table 6.3. The gain score model yields quite a different picture than the longitudinal model. First, $\beta = 0.65$ in the conditional model; we demonstrated earlier that in order for the gain score model to be unbiased it is necessary for β to equal unity. This suggests substantial misspecification in the gain score model. The variables X_2 and X_7 are not significant in Model 6.1 or in equation (6.2) of Model 6.2. Teacher's attendance at outside conferences, X_7, is an ambiguous measure. It may measure level of professional interest and awareness, but it may also measure teacher absence from the classroom or a desire for upward professional mobility, that is, to get out of the classroom.

Student attendence at kindergarten, X_2, is an interesting variable since it is insignificant in Model 6.1 and in the equation involving dependent variable T_2 of Model 6.2, but significant in the equation involving dependent variable T_2 in Model 6.2. Kindergarten experience has an indirect effect on achievement, which is omitted in the specification of the gain score model.

The variable for teacher experience, X_5, is significant in Model 6.1 but not in Model 6.2. This suggests that students of experienced teachers show more improvement than students of inexperienced teachers, but when we control for reading competence at T_1, teacher experience makes no difference in reading competence at T_2. The effect found in Model 6.1 could represent a difference in assignment, since Model 6.2 suggests that the assumption that experienced teachers are more effective is false.

Four teacher and classroom variables, X_6, X_8, X_{10}, and X_{11}, are nonsignificant in Model 6.1 but are significant in Model 6.2. Again this may reflect patterns of assigning pupils with low achievement to classrooms with more available resources.

Three remaining variables, X_1, X_4, and X_9, are significant in both models. However, X_4 changes sign. It is interesting that X_1 and X_9, along with X_2, are the only independent variables measured at the student level. All others are observed at the classroom and school levels. The interpretation of X_1, student attendance, and X_9, time in the classroom spent reading independently, is straightforward. Students who come to school more often and spend more time reading while at school can read better at the end of the year.

The variables X_3 and X_4, supplementary staff and proportion of high-achieving students, have a positive sign in Model 6.2 and a negative sign in Model 6.1. Model 6.2 provides more plausible results, indicating that supplemental staff contribute to, rather than detract from, a student's ability to read.

This is a complex association. The variables X_3 and X_4 are highly correlated negatively, −.626. Considering just Model 6.2, they have opposite signed correlations with the dependent variable, but their effects in Model 6.2 have the same sign. Substantively, it seems that X_4 is measuring the level of general reading achievement in the school. It is also possible that what is being measured is the socioeconomic level of the school. Middle- and upper-middle-class students tend to have higher levels of scholastic success than working and lower-class students.[2]

In either case, Model 6.2 suggests that since supplemental staff persons are assigned on the basis of need, schools with low general levels of competence will receive more staffing resources, accounting for the high negative correlation between X_3 and X_4. Consequently, X_3 has a negative correlation with T_2, because of this allocation effect; but when X_3 and X_4 are entered in the same equation, the partial effect of X_3 is positive, suggesting that, when the allocation effect of staffing is controlled, the effect of supplement staffing on reading levels is positive.

In Model 6.1 the effects of both X_4 and X_3 on the gain score are negative, which has led users of the earlier study to conclude that supplementary staffing has a detrimental influence (Rankin 1980). However, it is likely that this is due instead to the negative association between gain and initial competence level. Low-achieving students make higher gains, perhaps

because there is more room for improvement and perhaps also because of supplementary staff, that is, more concentrated instruction.

The foregoing interpretation seems satisfactory except that it is contradicted by the behavior of X_{11} (classroom aide time) in the model. Support staff (X_3) and aide time (X_{11}) are positively correlated (.527), and it should be reasonable to expect each to measure the same underlying attribute. However, the effect of X_{11} on reading achievement is negative. One or both of the following may account for this apparent anomaly. First, since X_3 and X_{11} are correlated at a moderately high level, the effect of each may be distorted when both are included in the same equation. Second, since X_{11} is measured at the classroom level and X_3 at the school level, X_{11} may be sensitive to within-school effects that are not picked up by X_3. It seems reasonable to assume that similar considerations that lead to allocation of more staff to low-achieving schools will lead to a similar allocation among classrooms within a school. Once again, however, whatever compensatory results aides may accomplish, these may be offset by the circumstances that led to their assignment in the first place.

In any case, it seems quite clear that three correlated variables X_3, X_4, and X_{11} have both common and unique effects on reading achievement. In the next section we describe a measurement model that is intended to simplify this complex structure.

In summary, Model 6.2 tends to produce a pattern of effects that comes closer than the pattern of Model 6.1 to matching reasonable expectations about reading achievement. Reversals in signs of effects suggest that the performance of a particular student depends largely on that student's achievement at T_1. When a student's initial achievement level is taken into account, a clearer picture of the factors contributing to his or her progress is obtained.

Model 6.2 accounts for approximately 20 percent of the T_1 variance and 45 percent of the T_2 variance. This is a substantial improvement over the small (2.5%) amount of gain score variance accounted for by Model 6.1. At the same time it must be emphasized that effects are small in both models and, although statistically significant, may be substantively trivial. For example, Model 6.2 indicates that each day of absence from the classroom results in an expected loss of half a point on the CAT – Total Reading when the mean and standard deviation of that test are 412.50 and 72.56, respectively. Model 6.2 also indicates that each additional hour per week spent reading independently results in an increase of 6.12 points on the CAT, perhaps a small return for the increase in effort.

These findings must be viewed in the context of model specification. We have seen how readily the sign and magnitude of effects can be altered

when new information is added. The addition of other variables would probably alter the estimates, because the low percentage of variance explained suggests that there are other major influences on reading abilities that have not been taken into consideration.

The results demonstrate the principal advantage of a two-equation longitudinal model over the more conventional gain score model. The gain score model incorporates the assumption that the third grade score has no effect on achievement, except to define a starting point relative to which gain is measured. We have argued that decisions about the use of educational resources are based partly on the child's past performance. Consequently, the effect of past achievement on present and future achievement is much more complex than what the gain score assumes, and it is therefore advisable to estimate the effect of past on present achievement directly from data. We have shown that differences in estimates that were found between the gain score and longitudinal models could plausibly be accounted for in terms of decisions to allocate resources, based on a child's past performance and current needs, and that the gain score model presents a remarkably distorted view of the effects of school resources on reading achievement.

A measurement model with correlated errors

Model 6.2 leaves the issue of the effects of X_4 and X_{11} unresolved. Along with X_3, which we temporarily treat as a proxy measure of staff allocation, these two variables were analyzed using confirmatory factor analysis to explore a range of factor structures. The two-factor structure shown in Model 6.3 (Figure 6.3) produced the most satisfactory fit.

The measurement model was identified by fixing λ_{33} at 1. The final fit was very good ($\chi^2 = 20.79$ with 15 degrees of freedom and $p = .143$). Substantial improvement in the indicators of goodness-of-fit were obtained by allowing the indicated error terms to be correlated.

Note that variables specified to have errors correlated to that of X_4 are variables relating to the teacher's training, confidence, and experience. We suggest that these correlations may indicate that teachers in schools that have been identified as low-achieving schools may feel more pressure to exaggerate their qualifications. Not having participated in the data collection process, we undertake this discussion of measurement error with some hesitation, but we venture to say that this would not be the first time that subjects in an evaluation study felt threatened.

The Model 3 structure suggests that X_3 and X_{11} have different but overlapping structures with X_4. The correlation between the two factors is $-.75$. If the lack of similarity of effects of X_3 and X_{11} had been due to a

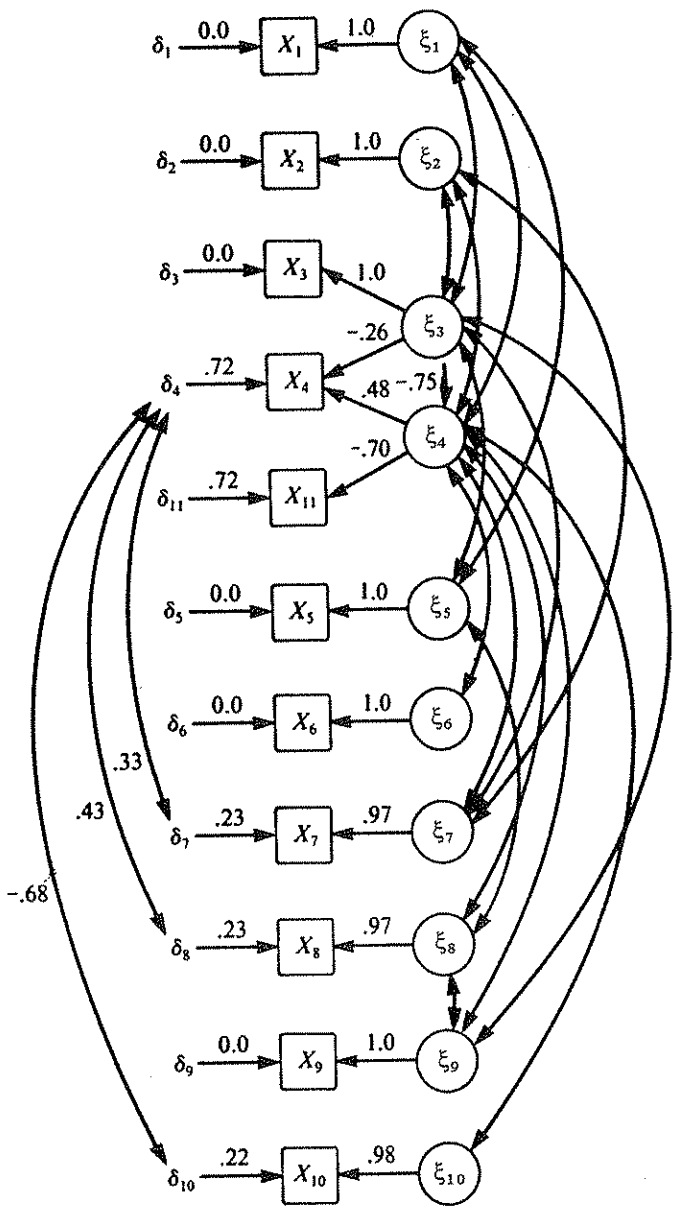

Figure 6.3. Model 6.3: independent measurement model, showing detail of factor structure. Estimates have been standardized.

distortion from collinearity, it should have been possible to load all three variables on one factor. That a single-factor structure did not fit is empirical evidence that the effect of X_{11} is indeed partly different from that of X_3.

A modified structural model

Model 6.4, shown in Figure 6.4, includes Model 6.3, the measurement model. To simplify presentation, links among the factors shown in Model 6.3 are not shown in Model 6.4. The full model links the measurement model for the independent variables to that for the dependent variables.

Model 6.4 is like Model 6.2 except that it incorporates a measurement model (Model 6.3) on the independent side. In Model 6.2, X_4 is hypothesized to influence both T_1 T_2, but X_3 and X_{11} influence only T_2. This results in a dilemma concerning the place of the two factors that replace these three variables in Model 6.4. Empirical underidentification is a possible problem here (Rindskopf 1984). Since X_4 loads on both factors, it was decided that both factors should be allowed to influence T_1.

As mentioned χ^2 for the fit of the measurement model was low (20.79 with 15 d.f.). For the full structural model, the fit was not as good (χ^2 = 212.34 with 74 d.f.). The fit improved when six Γ parameter estimates, Γ_{11}, Γ_{16}, Γ_{17}, Γ_{18}, Γ_{19}, Γ_{10}, were freed (χ^2 = 43.73 with 80 d.f.). However, it is not theoretically sensible to free these Γ elements, because they represent events in 1976, which can have no causal impact on a 1975 test score.

Comparison of Models 6.2 and 6.4

Table 6.4 compares estimates obtained from Model 6.2 with those obtained from Model 6.4. The more complex factor structure has increased the proportion of variance explained in the T_1 and T_2 variables, T_1 from .20 to .31, T_2 from .47 to .55. Of major interest are the effects of the two factors associated with X_3, X_4, and X_{11}. Factor 3, which is influenced by X_3 but not X_{11}, has a positive effect greater than the effect of either X_3 or X_4 in Model 6.2. Factor 4, influenced by X_{11} but not X_3, has a very large effect, which is considerably larger than the negative effect of X_{11} in Model 6.2. Recalling that X_{11} (aide time) loads negatively (−.698) on Factor 4, it appears that specification of a latent structure did not eliminate the negative effect of aide time on reading achievement. Indeed, that negative effect is stronger and more evident.

There are other notable changes in estimates. Teacher attendance, X_6,

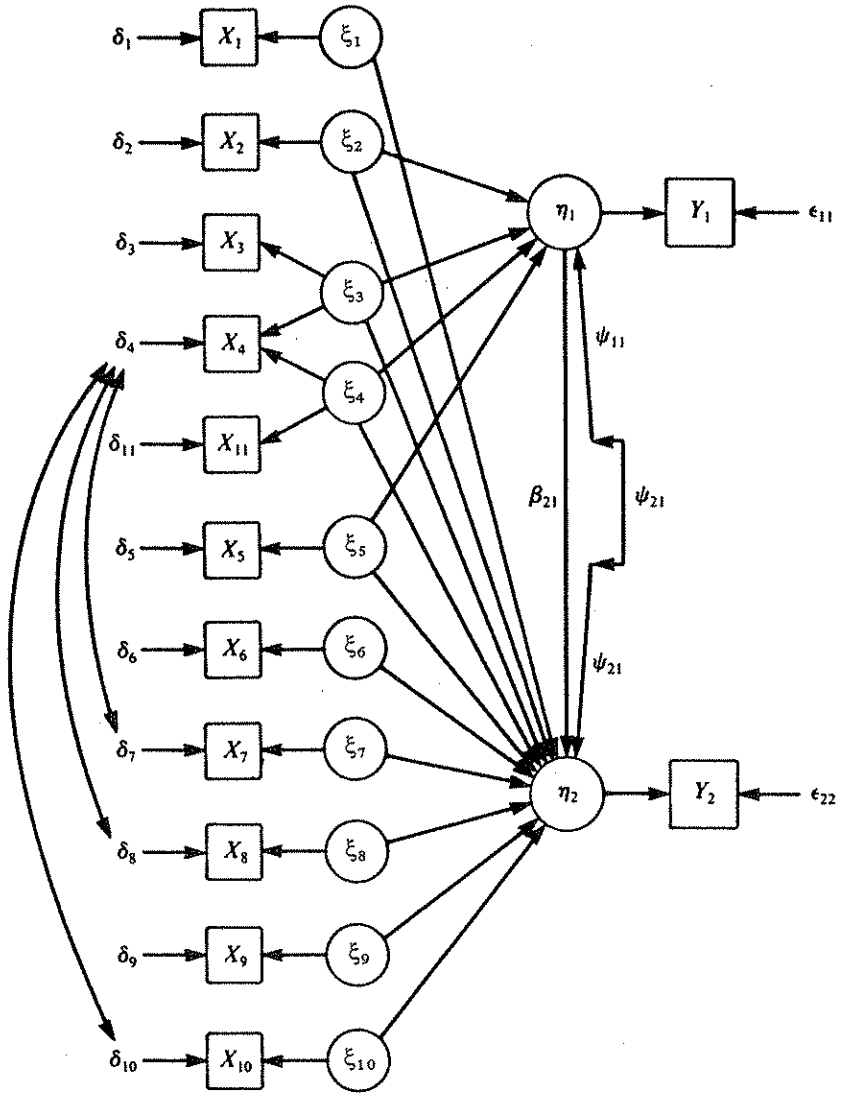

Figure 6.4. Model 6.4: the full model, detailing effects between independent factors omitted for clarity. See Model 6.3 for details.

has a small but significant effect in Model 6.2 but an insignificant effect in Model 6.4. Teacher attendance at outside conferences, X_7, is not significant in Model 2 but has a small significant effect in Model 6.4. Teacher approval of the reading program, X_{10}, is significant in Model 6.2 but not

Table 6.4. *Comparison of Model 6.2 with Model 6.4*[a]

Independent variable	Model 6.2		Model 6.4	
	T_1	T_2	T_1	T_2
X_1		0.58* (0.08)		0.30* (0.04)
X_2	10.45* (0.06)	4.183 (0.02)	3.81* (0.02)	2.13 (0.01)
X_3		0.43* (0.07)		
Factor 3			1.91* (0.31)	1.46* (0.22)
X_4	190.85* (0.37)	63.11* (0.11)		
Factor 4			−30.94* (0.74)	−17.81* (0.40)
X_{11}		−3.64* (0.12)		
X_5	−24.78 (−0.05)	3.77 (0.01)	−63.19* (−0.13)	−6.50 (−0.01)
X_6		1.06* (0.06)		0.38 (0.02)
X_7		4.07 (0.02)		12.80* (0.06)
X_8		14.18* (0.07)		−9.18* (−0.04)
X_9		0.12* (0.10)		0.13* (0.11)
X_{10}		12.27* (0.09)		3.72 (0.02)
T_1		0.65* (0.61)		0.59* (0.55)
R^2	.20	.47	.31	.55
χ^2/d.f.	277/8		212/18	

[a] Asterisked values are significant at less than .05; values in parentheses are standardized estimates.

in Model 6.4. Since the magnitude of these effects is of minor substantive importance, one hesitates to draw conclusions, but they are consistent with the suggestion that a teacher's performance is related to his or her effectiveness in the use of aides. The process of separating out the effect of

a possible overreliance on aides caused the effects of other teacher and classroom variables to shift.

The analysis of measurement and specification error

To this point we have considered three different models. Regardless of which is considered best, our modifications have resulted in changes in the magnitudes of estimated effects, leading to changes in interpretation as well. However, investigation need not, and often should not, end with a model with an acceptable fit. Variables in education and social science research are usually measured with more than negligible error. Also, it is never possible to be sure that all relevant variables have been included in a given model.

Given strong theory and excellent research design, it is sometimes possible to obtain direct estimates of error by constructing a measurement model based on multiple indicators, as suggested by Hauser and Goldberger (1971). However, although we have neither strong theory nor a particularly good design, all is not lost. Sensitivity analysis (Land & Felson 1978; Kim 1984) is a general method by which specific estimates obtained from a particular model can be scrutinized in terms of "sensitivity" to alterations of assumptions.

The sensitivity analysis described here uses alternative combinations of fixed values and compares their results. This procedure is discussed by Kim (1984, pp. 276). The LISREL framework makes it very convenient to perform sensitivity analysis. To specify assumed error in a dependent variable, for example, one need only specify a fixed value for the appropriate error term. To specify errors in equations, appropriate values are entered as fixed parameters in the Ψ matrix and, for errors in the measurement of variables, in the θ matrix.

Thirty-six alternative models were specified. Each of the two dependent variables was assumed to contain zero, 5 or 10 percent measurement error, resulting in nine possible combinations. In addition, for each combination of measurement assumptions, four levels of specification error in equations were tested (0%, 5%, 10%, 15%).

The level of specification error was to reflect variables not in the model, which should be expected to have some impact on both η_1 and η_2. Specification error may also be reflected in a correlation between the residuals $(\varepsilon_1, \varepsilon_2)$ for each equation explaining η_1 and η_2, respectively. A correlation between ε_1 and ε_2 reflects influences on both equations, which can be attributed to omitted variables. In Model 6.4, there are no measures of student background characteristics, and we would expect this omission to be reflected in such a correlation. Thus, the direct effect of η_1 and η_2 and (β_{21}) may be spuriously high.

The error assumptions investigated are, even in the worst case, fairly optimistic about Model 6.4, assuming that the test scores (dependent variables) are 90 percent reliable and that all relevant unmeasured independent variables would explain only 15 percent of additional variance in η_2.

Results of the sensitivity analysis are presented in Table 6.5. Several expected but interesting conclusions can be drawn from this analysis. First, β_{21} is not sensitive to error in Y_2.[3] Unstandardized slope estimates will be affected only when measurement error is in the independent variable. Of more interest is the impact of specification error ψ_{21}, which has a dramatic impact on the magnitude of β_{21} and its standard error. Note that, as β_{21} decreases, its standard error increases. One of the insidious aspects of this kind of autocorrelation is this double bias toward rejection of a true null hypothesis with respect to β_{21} because of underestimation of mean squared error (Neter & Wasserman 1974).[4]

Concluding comments

The original report (Kean et al. 1979a) bases conclusions on methodological practices that may be inappropriate. Among these are the procedure by which 18 "significant" independent variables (out of 245) were selected, the uncritical use of gain scores, and disregard for problems of measurement error.

Important methodological lessons can be drawn from the secondary analysis described in this chapter. The procedure by which independent variables were selected reflected a lack of theoretical guidance about the substantive model of interest, although this is essential in multivariate analysis. With a large number of variables, statistical significance is not a particularly useful criterion. With 245 variables, 12 correlations can be expected to be "significant" by chance (at the .05 level), to say nothing of the even larger number of partial regression coefficients that can be expected to be significant.

The secondary analysis reported here entailed successive refinements. This is not to say that other approaches would not be equally appropriate. For example, the gain score model could also be refined by "residualizing" the gain score variable, as recommended by Bohrnstedt (1969).

The gain score model (Model 6.1) yields results that are virtually uninterpretable. Effects contradict long-standing principles of educational practice. The longitudinal model (Model 6.2) results in large changes in the magnitude and sign of effects compared with findings in Model 6.1. Effects in Model 6.2 are also more in agreement with expectations (see Rankin 1980). Subsequent refinements, including the introduction of a measurement model among the independent variables (Model 6.3), and

Table 6.5. *Resulting parameter estimates given error assumptions in Y_1, Y_2, and ψ_{21}*

Percentage of error in:			Estimates[a] of:			Percentage of error in:			Estimates of:			Percentage of error in:			Estimates of:		
Y_1	Y_2	ψ_{21}	β_{21}	S_β	R^2	Y_1	Y_2	ψ_{21}	β_{21}	S_β	R^2	Y_1	Y_2	ψ_{21}	β_{21}	S_β	R^2
0	0	0	0.592	0.029	.581	5	0	0	0.639	0.029	.597	10	0	0	0.694	0.031	.617
0	0	5	0.515	0.029	.577	5	0	5	0.557	0.030	.594	10	0	5	0.607	0.031	.613
0	0	10	0.437	0.031	.566	5	0	10	0.476	0.031	.582	10	0	10	0.520	0.032	.601
0	0	15	0.359	0.032	.548	5	0	15	0.393	0.033	.564	10	0	15	0.434	0.034	.582
0	5	0	0.593	0.029	.611	5	5	0	0.639	0.030	.629	10	5	0	0.694	0.032	.649
0	5	5	0.516	0.030	.608	5	5	5	0.559	0.031	.625	10	5	5	0.609	0.032	.645
0	5	10	0.440	0.031	.597	5	5	10	0.479	0.032	.614	10	5	10	0.524	0.033	.634
0	5	15	0.364	0.033	.579	5	5	15	0.399	0.034	.595	10	5	15	0.440	0.035	.615
0	10	0	0.592	0.030	.645	5	10	0	0.639	0.031	.664	10	10	0	0.694	0.032	.685
0	10	5	0.519	0.031	.642	5	10	5	0.561	0.032	.660	10	10	5	0.611	0.033	.681
0	10	10	0.445	0.032	.631	5	10	10	0.483	0.033	.649	10	10	10	0.529	0.034	.670
0	10	15	0.371	0.034	.613	5	10	15	0.405	0.035	.630	10	10	15	0.447	0.036	.651

[a] S_β is the standard error of β_{21}.

an analysis of the sensitivity of the estimates to measurement error (Model 6.4), do not suggest large shifts in parameter estimates from the unrefined model, but they do illustrate techniques that can be applied as new generations of software make them not only practical but accessible.

Notes

1. This kind of analysis may clearly capitalize on chance, and if the analysis was in fact conducted as we have suggested, by an undisciplined romp through a correlation matrix, it might be said that there is no point in further consideration of variables selected in this manner. However, our point of view is that, like too much policy research, its methodological limitations have not hindered the adoption of the study's recommendations by practitioners in education who may lack sophistication in research methodology.

   We did not have the option of going back and replicating the analysis. Instead, we chose to make those refinements in method that were available to us. In this way, we showed that, by introducing a more appropriate specification of the model, support for some of the more important and controversial policy recommendations of the original study disappeared or was reversed.
2. This variable, it might be logically concluded, should have been based on data from 1975 (T_1) instead of 1976 (T_2). However, data from 1975 were not available. Conversations with School District of Philadelphia staff indicated that year-to-year changes in such school wide measures of achievement could be assumed to be negligible. Also, the sampling method described above (see also Kean et al. 1979a,b) helps to ensure that 1975–6 changes in schoolwide reading achievement are trivial.
3. Changes in β_{21} over levels of error in Y_2 occur whenever $\psi_{21} > 0$. Parameter ψ_{21} is specified as a percentage of η_{21}, whereas ε_2 is specified to be a percentage of Y_2. When ε_2 changes, so does η_2 and, therefore, the quantity $(\psi_{21})(\eta_2)$. The changes in β_{21} result from this level of specification error.
4. It is tempting to be reassured by the observation that $t = \beta/S_\beta$ is relatively constant, but in a multivariate model it is possible to have upward bias in one or more β_{21} estimates with downward bias in corresponding standard errors, resulting from poor reliability of one or more variables. It is only in the bivariate case that measurement error can be relied on to result only in simple attenuation (Won, 1982).

References

Alwin, D. F., & Sullivan, M. J. (1975). "Issues of Design and Analysis in Evaluation Research," *Sociological Methods and Research* 4 (August): 77–100.

Bohrnstedt, G. W. (1969). "Observations on the Measurement of Change," pp. 113–33 in E. F. Borgatta & G. W. Bohrnstedt (eds.), *Sociological Methodology 1969*. San Francisco: Jossey-Bass.

Cronbach, L. J., and Furby, L. (1970). "How We Should Measure 'Change' – or Should We?" *Psychological Bulletin* 74: 68–80.

Hauser, R. M., & Goldberger, A. (1971). "The Treatment of Unobservable Variables in Path Analysis," pp. 81–117 in H. Costner (ed.), *Sociological Methodology 1971*. San Francisco: Jossey-Bass.

Johnston, J. (1972). *Econometric Methods*. New York: McGraw-Hill.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods*, University of Uppsala, Department of Statistics.

Kean, M. H., Summers, A. S., Raivetz, M. J., & Farber, I. J. (1979a). *What Works in Reading? The Results of a Joint School District/Federal Reserve Bank Empirical Study in Philadelphia*. Office of Research and Evaluation, School District of Philadelphia (May).

(1979b) *What Works in Reading? Technical Supplement of a Joint School District/Federal Reserve Bank Empirical Study in Philadelphia*. Office of Research and Evaluation. School District of Philadelphia (November).

Kessler, R. C. (1977). "The Use of Change Scores as Criteria in Longitudinal Survey Research." *Quality and Quantity* 11: 43–66.

Kim, J. (1984). "An Approach to Sensitivity Analysis in Sociological Research," *American Sociological Review*, 49 (April): 272–82.

Kim, J., & Mueller, C. W. (1976). "Standardized and Unstandardized Coefficients in Causal Analysis: An Expository Note," *Sociological Methods and Research* 4 (May): 423–37.

Land, K. C., and Felson, M. (1978). "Sensitivity Analysis of Arbitrarily Identified Simultaneous-Equation Models," *Sociological Methods and Research* 6 (February): 283–307.

McNemar, Q. (1958). "On Growth Measurement," *Educational and Psychological Measurement*, 18: 47–55.

Neter, John, and Wasserman, William. (1974). *Applied Linear Statistical Models*. Homewood, IL: Irwin.

Pendleton, B. F., Warren, R. D., & Chang, H. C. (1979). "Correlated Denominators in Multiple Regression and Change Analysis," *Sociological Methods and Research* 7 (May): 451–74.

Rankin, R. (1980). *What Works in Reading*. Unpublished manuscript, University of Oregon, College of Education.

Rindskopf, D. (1984). "Using Phantom and Imaginary Latent Variables to Parameterize Constraints in Linear Structural Models," *Psychometrika*, 49 (1): 37–47.

Thorndike, R. L. (1966). "Intellectual Status and Intellectual Growth," *Journal of Educational Psychology* 57: 121–7.

Thorndike, R. L., and Hagen, E. (1955). *Measurement and Evaluation in Psychology and Education*. New York: Wiley.

Won, E. (1982). "Incomplete Corrections for Regression Unreliabilities: Effects on the Coefficient Estimates," *Sociological Methods and Research* 10 (February): 271–84.