CHAPTER 4

ISSUES IN THE INTERPRETATION OF THE RESULTS OF SCHOOL EFFECTIVENESS RESEARCH

ROEL J. BOSKER* and JAAP SCHEERENS†‡

*Institute for Educational Research, University of Groningen, The Netherlands
†University of Twente, Enschede, The Netherlands

Abstract

In this chapter three issues in the interpretation of the results of school effectiveness research are discussed: criterion choice, effect size and stability of effects. With respect to the first issue the overall conclusion is, that criterion choice and definition depend on the effectiveness perspective and the particular theory one wishes to corroborate. The issues of effect size and stability of school effects are treated both from the angle of a synthesis of available empirical results and from the angle of conceptual analysis. An overall evaluation of the available data on effect size and stability leads to the conclusion that school effectiveness models are not as shaky as certain critics would have it, but at the same time not established as firmly as enthusiastic school improvers treat them. Various suggestions as to the improvement of future school effectiveness research are offered, notably more refined research designs and more elaborate theory development.

Introduction

Three points of criticism concerning the validity of the current empirical school effectiveness models are that effects are insignificantly small, that the effect criterion (mostly achievement) is not well chosen and that school effects are unstable. In this chapter these issues are reconsidered.

Criterion-Definition

The major challenges to the choice of a particular effectiveness criterion are critical questions concerning the ultimateness, fairness, validity and economy of the actual output measures. We assume that school effectiveness is to be primarily concerned with output, and therefore should not dwell upon views of organizational effectiveness in which input and process characteristics are also seen as effectiveness criteria (cf. Cameron & Whetten,

‡Order of authors was determined alphabetically.

1983). From these points of view our basic perspective on school effectiveness is that of *productivity* (see also Scheerens & Stoel, 1988).

Achievement or Attainment Measures

The predominant criterion in school effectiveness studies from various disciplinary origins is achievement. Attainment measures depend on formal levels in the school careers of pupils. Roughly speaking educational attainment scores express the level that individuals or groups of pupils have reached after a certain number of years of schooling. Examples of discrete attainment levels are the end of the primary school period and the end of the secondary school period. Particularly when an educational system consists of many school types, to which societal value is attributed in various degrees, attainment scales can become quite differentiated (see e.g., Bosker & van der Velden, 1989).

When discussing the option of either choosing attainment measures or achievement measures, various underlying dimensions for this choice can be discerned. First of all the choice may depend on different connotations of effectiveness (e.g., maximization of output vs. enhancing quality). Secondly, preferences concerning band width vs. specifity of output measures could determine the choice, i.e., the question whether an overall output measure or a more narrowly defined performance indicator is to be preferred. Thirdly, the question of the predominance of a more practical vs. a scientific interest in establishing effectiveness may lie at the background of this choice.

Attainment measures are close to the economic notion of effectiveness as maximization of output, where output is measured as the amount of products that results from a particular production process. In education pupils that pass their exams can be seen as the products of the process of schooling. Achievement, on the other hand, fits better in an interpretation of effectiveness in terms of quality. Achievement tests as effectiveness criteria capitalize on more fine-grained quality differences of the units of output.

Attainment measures are cruder output measures than achievement tests, but at the same time they usually imply a broader coverage of the whole spectrum of educational objectives. The passing of a final examination (attainment indicator) depends on achievement in many subjects, whereas achievement tests in school effectiveness studies are often limited to arithmetic and language tests.

School effectiveness is both a subject of scientific inquiry and an applied field of interest in educational policy and management. When issues of consumer demands, monitoring of schools and accountability are at stake one cannot do without attainment indicators. In the case of inquiry into determinants of school effectiveness, i.e., input–output, or input-process–output studies, researchers will want output indicators that differentiate more strongly between qualities of the units of output and prefer achievement tests.

In summary, it is our contention that attainment measures are called for when purely economic and applied perspectives of effectiveness predominate or in case one wishes to explore, in the tradition of sociology of education, the contribution of schools to a person's status attainment. Achievement measures are more likely to be chosen when quality of education is at stake and when a more psychological interest in cognitive development (or an educational interest in schools as organizations) has the upper hand.

Finally, it should be mentioned that the choice between attainment or achievement can be avoided in two ways: (a) by using both, and (b) in the case of a decision oriented use of

achievement tests (as when performance standards in the form of cutting scores on tests determine further career options).

Intermediate and 'Ultimate' Effectiveness Measures

Are attainment or achievement measures obtained at the end of a particular period of schooling to be considered as the ultimate productivity measures or would only more long term civil effects of schooling, such as employment or job level reached by graduates, qualify as such? Or, moving into the other direction on the scale of ultimateness of effectiveness measures, could we use intermediate effects like attendance and drop-out rates as substitute effectiveness criteria? (e.g., Rutter, Maughan, Mortimore, & Ouston, 1979).

Searching for ultimate school effects is like looking for the holy grail, since one can go on and on in stating even more ultimate effects. The most likely, be it arbitrary, points in the school careers of pupils to measure school effects are indeed when a particular period of schooling is terminated and transition to a higher school type or into the labour market takes place. Post-school effect measures could be seen as important in macro-level applications of educational indicators for purposes of monitoring national school-systems. Also, post-school effect-measures could be seen as important criteria to gain insight into the predictive validity of effect measures at the end of the period of schooling. Attendance and drop-out rates are better treated as process-measures in school-effectiveness studies, because they generally function as means rather than as desired ends of schooling.

General vs. Curriculum Specific Achievement Tests

When the decision is taken to use achievement rather than attainment output data, there is a further option in the choice of the type of test. Madaus, Kellaghan, Rakow, and King (1979) have provided arguments in favour of curriculum-specific tests and against the use of general achievement tests (like the Scholastic Aptitude Test). One of their arguments is that larger school (or class) effects are demonstrated when curriculum specific tests (exams in their case) are used. Before offering a few lines of thinking in determining the choice of output measure along this particular dimension, it should be remarked that general vs. curriculum specific achievement measures should be seen as a continuum with many discrete scale-points rather than a dichotomous choice between two extremes.

Varying from curriculum specific to general aptitude measures one could discern the following types of measures:

— trained test items;
— content specific measures;
— Rasch-scales of narrow content areas;
— subject specific tests;
— general scholastic aptitude tests;
— intelligence tests.

A general guideline to choose from these alternatives would be to use the more specific measures up to the degree that the application purpose is closer to the micro situation of classroom-instruction. A line of thinking which perhaps offers a more fundamental

solution to this problem of choice, within the content of school effectiveness research, would be to choose the type of outcome measure that has the greatest predictive validity with respect to the more ultimate educational effects. To give an example: when measuring achievement at the end of a specific stream of vocational education, we might prefer content-specific measures, assuming a close connection between the curriculum and skills that are required in the job-situation. It should be noted here, however, that this latter kind of criterion choice further depends on the theory one holds about the relationship between education and the labour market. Departing from a credentials or screening theory, certification itself would be the best criterion for school effectiveness, whereas achievement is more connected to the human capital philosophy.

Controlling for Confounding of Measures of Effect

So far it was assumed that the measures on attainment or achievement scales can indeed be interpreted unequivocally as criteria of school effectiveness. This is not so. One source of confounding of these output measures that is well-known is the initial or even innate ability of pupils. In school effectiveness studies it should be attempted to separate the contribution of innate characteristics of pupils from the net-effect of school characteristics. Although this source of confounding the interpretation of output measures is well-known, critical reviewers of effective schools research have noted that in many cases control for innate characteristics of pupils is inadequate (Purkey & Smith, 1983).

A second source of possibly biased interpretations of school effects are selective policy-measures at the school level. Examples are: lenient vs. a strict policy in letting pupils pass from one grade to the next, a more or less conservative policy in allowing students in secondary education to go in for their final examination, a more or less reluctant attitude when it is to be decided to send pupils to special schools (primary education) or to lower categories of secondary education (when a country has a differentiated secondary education system). This type of selectivity bias is usually accounted for when economic measures of school output are used, for instance, number of graduates divided by the total number of pupils in the cohort that entered the school as many years back as the normal duration of the school-period. Usually in individual level measures of attainment selectivity will also be accounted for, because drop-outs will still receive a score on the attainment scale. When individual level achievement data are used as the effect-criterion this type of selectivity bias is usually neglected, however. The consequence of this practice is that the corresponding estimation of school effects might well be confounded and corresponding policy- or managerial decisions consequently unfair. The general solution to this problem would be to obtain some kind of measure of the selectivity policy of schools and use this information as an additional independent variable (defined at the school level) to separate its effect from the other independent variables that stand for the more genuine determinants of school effectiveness. Statistical techniques to model attrition bias may help in solving the problem of separating selectivity effects from the influence of other independent variables (see e.g., Hausman & Wise, 1979).

Effect Size

First and foremost it should be realized that what we call school effects are comparative

rather than absolute measures; we do not compare the effects of schooling to no schooling at all but we compare variations in schooling, since we are bound to what actually happens in educational practice.

Comparative school effects are usually expressed in two ways: as an overall effect of schools in terms of the between school variance relative to the total variance in pupil achievement and in terms of specific effects of particular school characteristics, either expressed as a proportion of the between school variance accounted for or as the proportion of variance in individual achievement that is accounted for by the specific school characteristic.

Effect Size as a Function of the Choice of Dependent and Independent Variables and the Unit of Analysis

Effect size depends, up to a degree, on the choice of the *dependent* variable. As is to be expected specific characteristics of instructional processes (particularly content covered) will show up more clearly when curriculum specific tests rather than general scholastic aptitude tests are used as the dependent variable. This phenomenon was empirically demonstrated by Madaus *et al.* (1979). Effect sizes are also relatively higher when subject matter areas are tested that depend more exclusively on schooling and instruction; so, effects in arithmetic and mathematics are usually somewhat higher than effects in a subject like (native) language, which is also learned at home.

As was already implied in the above, *independent* variables that are closer to the output measures — like content covered and time on task will most likely explain more variance than school variables (like leadership-styles and variations in organizational structure) that are further removed from the instructional process. As Rutter (1983) points out the strength of association between school characteristics and achievement will be depressed because of the smaller variance in the former characteristics. If we could manage to measure curricular characteristics like opportunity to learn at the level of individual pupils we might even expect larger effect sizes.

Empirical Estimates of the Size of School Effects

A method to gain insight into the strengths of relationships between school measures and output data would be to average effect sizes found in a number of studies. A modest effort to do so was made by looking at 12 Dutch effectiveness studies.

We examined the average between school variance. Only in 6 studies (out of 12) was the between school variance computed. The average percentage of the total variance accounted for by the factor school is 12. This estimate is in line with the findings in major Anglo-Saxon studies, like, for instance, Coleman, 1967 (ca. 9%) and Mortimore, Sammons, Stoll, Lewis, & Ecob (1988) (11%). It should be noted that effects on various types of dependent variables were all thrown in one hat, which is of course a bit of a rough procedure. Because of the different statistics that were used, we did not attempt to synthesize effects of individual school variables.

This rudimentary attempt at research synthesis leaves us with the impression that — unless perhaps one could get access to the primary data of studies — it is at present very

difficult to quantitatively summarize the results of school effectiveness studies. Research syntheses would strongly benefit from a complete and standardized manner of reporting school effects. In the subsequent section we will turn to the issue of alternative ways of expressing school effects.

Towards More Insightful Ways of Expressing School Effects

The most common way of expressing school effects is to report the between school variance relative to the total variance. Next, in subsequent analyses, one could establish the contribution of specific school variables to the between school variance (see e.g., Madaus *et al.*, 1979).

Yet, expressing effects in terms of proportions of variance accounted for still leaves many questions open about the meaning of effects.

A further step towards better interpretable effect standards would be to try and express effects in terms of intervals on the scale of the output variable. In experimental or quasi-experimental studies the difference between means of experimental conditions are useful points of reference for such intervals (cf. Cahan, Davis, & Cohen, 1987), for instance, by expressing the interval between the experimental and control group mean in terms of the number/fraction of standard deviations. One could thus adapt — be it arbitrary — conventions like: an effect of 0.3 standard deviations or greater is indicative of a meaningful difference. Since school effectiveness studies are nonexperimental, schools could only be grouped on an *ad hoc* basis in, for instance, the highest scoring 20%, the middle 60 and the lowest scoring 20%. Though this procedure would inevitably imply exploiting chance, it might still be adopted to make the results of school effectiveness studies amenable to the interpretation of effect sizes according to established conventions. Purkey and Smith (1983, p. 428) conclude in their review of the school effectiveness literature that, when comparing the bottom 20% of low-scoring schools with the top 20% of high-scoring schools, the difference in average achievement for sixth grade pupils is about two thirds of a standard deviation.

The most insightful way of expressing school effects would be to combine the contrasting of successful and unsuccessful schools with the attachment of some kind of societal value to score-levels on the output variable. Some examples are: to indicate the increased chance of pupils who have visited an elementary school with a certain favorable characteristic (e.g., emphasis on cognitive objectives) of being referred to a higher type of secondary education, to express attainment measures in monetary values (Stoel, 1984), and to express differences in successful and unsuccessful schools in terms of IQ-points, or other scores that are geared to age norms — so that effects can be expressed in average gain in, for instance, reading age (Rutter, 1983; Mortimore *et al.*, 1988). These applications show that though school effects might be moderate or low in terms of percentage of variance accounted for, they may still have quite significant societal consequences. For instance, Rosenthal and Rubin (1982), investigated effect sizes for cases where the criterion variable is measured as the success or failure rate. They show that only 10% of variance accounted for, implies a point-biserial correlation of the dichotomous treatment variable with the criterion variable of .32 and a difference in success rate from 34% to 66%. In this case the success rate might express the reduction in illness rate, or, in the context of education, the percentage of exam passes. And, to give a final example, the effect size of 2/3 of a standard

deviation reported by Purkey and Smith (1983) would mean roughly one full grade level of achievement, in other words, the average pupil in the 20% highest scoring school would be roughly 1 school year ahead of the average pupil in the 20% lowest scoring school.

We are painfully aware of the fact that we have not given numerical answers to the question 'how large should a school effect be to be called significant?' We will make so bold as to give an impression: when 15% of the variance in the individual level output variable is explained by the factor school — after relevant background characteristics of pupils have been controlled — school characteristics that explain 2 or 3% of the individual level variance are probably meaningful.

Stability of Effects

Introduction

Stability is a vital issue in school effectiveness research since it is a necessary condition for further theory development. The assumption of consistency in achievement and/or attainment is rather crucial in more than one way. Since we might expect that organizational characteristics of schools are more or less stable over time, we must know if the rank order of schools on output remains the same no matter when we measure the effect. If schools affect achievement in another way, each year, and organizational features remain more or less the same, the resulting overall correlation between the school characteristics and the output index must be near zero. Another aspect of stability over time is the possible existence of grade-specific effects. If school effects vary across grades this would mean — especially in primary education — that these effects are in fact teacher effects.

Finally, models of school effectiveness usually assume that they are invariant as far as the operationalization of school success is concerned. The question of whether or not process–output relationships are consistent across effectiveness criteria can be treated as another specific instance of stability, however.

In the next section some of the school effectiveness research in this area is reviewed in order to see what empirical evidence is available to check these various types of consistency or stability of school effects.

Empirical Evidence

Stability Over Time

The most fundamental requirement for a school effectiveness theory is the psychometric test-retest conceptualization of reliability: is the correlation of two independent measures with the same instrument of a latent trait high enough, i.e., is a school effective irrespective of the year in which effectiveness is measured. Research on this topic indicates stable school effects across years. Rutter *et al.* (1979) were the first to present results in this area. Their research shows that examination results and delinquency-rates are stable across years (associations — rank-correlations — of near .80). Yet, on the basis of Rutter's findings one cannot be sure whether stability was just a function of the balance of intake or could indeed be attributed to certain organizational or curricular school characteristics.

The correlation over years of properly assessed (secondary) school effects, that is to say assessed in a multilevel framework, amounts to approximately .80 res. .60 in the U.K. (Willms, 1987; Goldstein, 1987) and to approximately .75 — when analyzing separate school types (Roeleveld & de Jong, 1989) — to .96 overall in the Netherlands (Bosker, Guldemond, Hofman, & Hofman, 1988). Research in elementary schools (grade one to four) in the U.S.A. gives results ranging from .34 to .66 (Mandeville, 1987), though these figures might be somewhat deflated because of the inadequacy of the statistical models used. Mandeville does not separate sample variance from true parameter variance. In not doing so, the stability is underestimated since measurement noise, that sometimes amounts to near 80% of the observed between year variance, confounds the effects (see Willms & Raudenbush, 1988).

From a theoretical point of view stability across grades is a more interesting question. Only little is known about this topic. Mandeville and Anderson report correlations across grades near .10 (grade 1 to 4 in elementary schools) when using mathematics and language effects. Their explanation for this inconsistency of school effects across grades is plausible. In using curriculum specific tests, some variation may occur as a function of the degree to which the specific subject is taught to the pupils. In an analysis of a large Dutch sample of primary schools Bosker (1989) found an intra-school-correlation between grades for arithmetic of .50 and for language of .47. Bosker *et al.* (1988) found rank-correlations ranging from .40 to .80 between grades in Dutch secondary education, but these figures may be somewhat inflated because of the dependency in the observations (the same pupils are tested in these grades, and the criterion variable is cumulative over years). Rutter reports rank-correlations across grades for pupil misbehavior ranging from .23 to .65. His figures for school attendance are more encouraging (near .80).

Stability Across Classes

The results presented so far suggest that teacher effectiveness may be a more probable cause of differences between schools than characteristics defined at the school level. This conclusion may be corroborated when the differences between classes within grades are examined.

Data reported by Ecob for the U.K. (in Mortimore *et al.*, 1988, p.130. columns 1 and 2) only partly contradict this proposition.

Table 4.1
Stability Across Classes Within Grades

	Grade 2	Grade 3
Reading	.63	.93
Mathematics	.46	.65

Stability Across Effect Criteria

Mandeville and Anderson (1986) as well as Mandeville (1987) investigated the correlation between primary school effect indices for different subject areas (mathematics and reading). The results indicate great stability (near .70), but again the same pupils are

tested so this figure too might be somewhat inflated. The same objections could be raised to the results on the Brandsma and Knuver data for elementary schools in the Netherlands. Here a correlation of .72 was found (Bosker, 1989). Cuttance (1987) reports correlations for secondary schools in Scotland of .47 and .74 for English and arithmetic respectively with an overall achievement indicator.

Stability across subject area is a specific topic of the general problem of stability across effect criteria.

From Rutter *et al.*'s study it seems that the choice of the right criterion is rather an academic problem, since the author computed associations between three effect criteria, and came up with correlations near .70. Research in Holland by Bosker *et al.* (1988), however, shows that the choice of the effect criterion is meaningful indeed, since two attainment variables (one measuring efficiency and the other educational perspectives) correlated only .35. For primary schools the correlation between attainment and achievement school effects, corrected for intake differences, turned out to be dramatically low (near .03) (Bosker, 1989).

Implications

First the results on the stability of school effects are summarized.

Table 4.2
Range of Stability Estimates for School Effects

	Primary	Secondary
Across years	.35–.65	.70–.95
Across grades	.10–.65	.25–.90
Across classes	.45–1.00	—
Across subjects	.70–.75	.45–.75
Across criteria	.00–.05	.35–.70

The presented figures in the table are mostly correlations. So even when in some cases the correlations are .60 or more, one should realize that an important part of the variance is not accounted for. The theoretical implications are quite interesting. With respect to the inconsistency of school effects across criteria, we now know that the questions that were raised about the choice of criteria are significant indeed. As for the grade specific effect we might refer to the Mandeville and Anderson hypothesis of school and grade specific curricula and also to the *teacher* effectiveness hypotheses (i.e., that not the school but individual teachers are to be seen as the primary locus of effectiveness). Considering this type of instability it could be argued that school effectiveness had better be assessed at the end of a particular period of schooling, since in this case it at least incorporates the cumulative — and not the individual — effects of teachers.

Contingency theory points to another cause of instability of effects, since organizations, being adaptive to external circumstances, might be instable themselves.

Finally Goldstein (1987) gives an interesting explanation of the variance not accounted for by schools. He found that in assessing school and year effects, without adjusting for

intake differences, the between year component decreases to 5% of the school and year variance. Besides, only the intake differences seem to vary across years and across schools, which leads him to the conclusion "that the schools may tend to compensate for yearly intake differences in achievement in order to produce only small year-to-year differences in the overall examination results" (pp. 59,60).

Conclusions

The general conclusion is that the choice of an effect criterion strongly depends on the effectiveness perspective and the theory one wishes to corroborate. It seems necessary to specify more criterion specific effectiveness theories as a necessary prerequisite to overall effectiveness models (i.e., a stable set of predictor variables for various criteria).

The most important observation concerning the effect size is, that even small effects may be relevant, if only because the effects should be multiplied by the number of pupils benefiting from outstanding schools. The best way to make effects more insightful might be to translate them to their societal impetus by means of, for instance, cutting scores.

With respect to the stability issue the main conclusion is that school effects do exist even though they may vary across grades, classes, time and criteria. A final remark is, that school effectiveness theories have led to a shift in focus from pupils to schools as the central unit of interest. Despite this shift research designs concentrate on the pupil. We feel that more refined designs are needed at the school level together with the sophisticated designs we usually employ when investigating variation between pupils. Next to the longitudinal designs for pupils, a repeated measurement design at the school level would surely help us to better locate the sources of variations in outcomes. Apart from more refined research designs school effectiveness research is badly in need of further theory development.

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *The Journal of the Royal Statistical Society, Series A (General)*, **149**, Part 1, 1–43.

Anderson, L. W., & Mandeville, G. K. (1986). *Towards a solution of problems inherent in the identification of effective schools.* AERA-paper. San Francisco.

Bosker, R. J. (1989). *Theory-development in school effectiveness research: in search for stability of effects.* Paper presented at the multi-level conference, Nijmegen.

Bosker, R. J., Guldemond, H., Hofman, R. H., & Hofman, W. H. A. (1988). *Kwaliteit in het voortgezet onderwijs.* Groningen: RION.

Bosker, R. J., & Velden, R. K. W., van der (1989). The effects of schools on the educational career of disadvantaged pupils. In B. P. M. Creemers & D. Reynolds (Eds.), *The proceedings of the 2nd International Congress for Schooleffectiveness*, Rotterdam.

Brandsma, H. P., & Knuver, J. W. M. (1988). Organisatorische verschillen tussen basisscholen en hun effect op leerlingprestaties. *Tijdschrift voor Onderwijsresearch*, **13** (14), 201–212.

Cahan, S., Davis, D., & Cohen, N. (1987). The definitional interpretation of effects in decision oriented evaluation studies. *International Journal of Educational Research*, **11**, 91–104.

Cameron, K. S., & Whetten, D. A. (1983). *Organizational effectiveness: A comparison of multiple models.* New York: Academic Press.

Cuttance, P. (1987). *Modelling variation in the effectiveness of schooling.* Edinburgh: CES.

Doddema-Winsemius, H., & Hofstee, W. K. B. (1987). Enkele controversiële onderwijsdoelstellingen in de context van evaluatie. *Pedagogische Studiën*, **64** (5), 192–201.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Charles Griffin.

Gray, J., Jesson, D., & Jones, B. (1986). The research for a fairer way of comparing schools examination results. *Research papers in Education*, **1** (2).

Hausman, J. A., & Wise, D. A. (1979). Attrition bias in experimental and panel data: the Gary Income Maintenance Experiment. *Econometrica*, **47**, 455–473.

Knuver, A. (1987). *Schoolkenmerken en leerlingfunctioneren; een replicatie-onderzoek*. Groningen: Rijksuniversiteit.

Linden, W. J., van der (1987). *Het zwalkend niveau van ons onderwijs*. Diës-rede, Universiteit Twente.

Mandeville, G. K. (1987). *The stability of school effectiveness indices across years*. NCME-paper. Washington.

Mandeville, G. K., & Anderson, L. W. (1986). *A study of the stability of school effectiveness measures across grades and subject areas*. AERA-paper. San Francisco.

Madaus, G. F., Kellaghan, T., Rakow, E. A., & King, D. (1979). The sensitivity of measures of school effectiveness. *Harvard Educational Review*, **49**, 207–230.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *The junior school project; technical appendices*. London: ILEA, Research and Statistics Branch.

Purkey, S. C., & Smith, M. S. (1983). Effective schools: a review. *The Elementary School Journal*, **83**, 427–452.

Raudenbush, S. W. (1988). *The analysis of longitudinal multilevel data*. Paper for the seminar on policy applications of multilevel analysis. Washington.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, **59**, 1–17.

Roeleveld, J., & Jong, U., de (1989). Evaluating effectiveness of secondary schools in the Netherlands: models and stability. In B.P.M. Creemers & D. Reynolds (Eds.), *The proceedings of the 2nd International Congress for Schooleffectiveness*, Rotterdam.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, **74**, 166–169.

Rowan, B., Bossart, S. T., & Dwyer, D. C. (1983). Research on effective schools. A cautionary note. *Educational Researcher*, April, 24–31.

Rutter, M. (1983). School effects on pupil progress. Research findings and policy implications. *Child Development*, **54**, 1–29.

Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. (1979). *Fifteen thousand hours. Secondary schools and their effects on children*. Somerset: Open Books.

Scheerens, J., & Stoel, W. G. R. (1988). *Theory development on school effectiveness*. AERA-paper. New Orleans.

Stoel, W. G. R. (1984). Vergroting klassen tast kwaliteit voortgezet onderwijs aan. *Didaktief*, januari.

Stoel, W. G. R. (1986). *Schoolkenmerken en het gedrag van leerlingen en docenten in het voortgezet onderwijs*. Groningen: RION (interne publicatie).

Willms, J. D. (1987). Differences between Scottish education authorities in their examination attainment. *Oxford Review of Education*, **13** (2), 211–232.

Willms, J. D. (1988). *Estimating the stability of school effects with a longitudinal, heirarchical linear model*. AERA-paper. New Orleans.