European Journal of Psychology of Education
1986, Vol. I, N.º 3, 81-91

The Role of Classroom Differences in Achievement Changes

Wolfgang Schneider
Andreas Helmke
Max-Planck-Institute for Psychological Research, Germany

A combined criterion involving the regression slopes of pretest-posttest achievement scores and achievement gain scores was used to classify similar types of classrooms. Mathematics achievement differences among 632 fifth graders were analysed in a longitudinal design and explained in a structural equation framework provided by LISREL, separately for four types of classrooms. The results replicated the findings of an earlier study (Schneider & Treiber, 1984) in that the local nature of achievement models could be demonstrated. That is, the structural components of the causal models could not be generalized across the four groups of classrooms. The inclusion of a second grouping criterion (i. e., achievement gain) proved useful in that a better model fit was always obtained for classrooms with high achievement gains. As a global model test ignoring group and classroom membership did mask the differential validity of the achievement model in the various subgroups, the need for multilevel approaches was emphasized.

One of the fundamental issues in educational psychology has been the identification of student aptitudes and classroom instructional characteristics that determine changes in student scholastic achievement. More recent theoretical models of school learning emphasized the fact that classroom learning is a multiplicative function of pupil background, intrinsic motivation, teacher activities, and quantity of schooling (for a review see Haertel, Walberg & Weinstein, 1983). Although these models of school learning are explicit enough to be empirically tested by experimental or correlational methods, only a few recent studies have tried to explore the causal/ /structural dependencies among individual aptitude, instructional treatment, and resulting achievement changes by using more complex causal modeling techniques (cf. Helmke, Schneider & Weinert, 1986; Parkerson, Lomax, Schiller & Walberg, 1984; Schneider & Treiber, 1984; Schneider & Helmke, 1986). Taken together,

The study was financially supported by the German National Science Foundation (Deutsche Forschungsgemeinschaft). We would like to thank Raymond Perry from the University of Manitoba in Winnipeg, Canada, for his helpful comments on an earlier version of the manuscript.

the empirical studies have unequivocally demonstrated the central role of student's cognitive entry characteristics (i. e., general aptitude and prior knowledge) for subsequent learning outcomes. However, the role of instructional characteristics in predicting achievement changes has remained unclear, which is mainly due to the fact that the theoretical constructs and measures used to represent instructional quality were not directly comparable in the different studies. The same is true for the chosen level of analysis. That is, whereas some of the studies used global model tests solely based on students or classrooms as the unit of analysis (e. g., Parkerson et al., 1984), others preferred a multilevel approach (Schneider & Helmke, 1986).

In the present study, an attempt was made to further clarify the relative impact of student entry abilities and classroom instructional characteristics on measures of achievement. In particular, the major goal was to investigate whether the findings of a recent study by Schneider and Treiber (1984) could be validated and extended to an independent sample. Their procedure for testing models of school learning was chosen for two reasons. First, it followed a more elaborate methodological approach by combining various components not considered simultaneously in most other empirical studies into the problem. That is, it used a latent variable causal modeling approach (LISREL) to compare the impact of aptitude and instruction on students' mathematics achievement. A longitudinal component was included that described achievement changes in a multiwave growth model involving four measurement points. Second, the problem that education is a multilevel enterprise in which students are nested within classes and classes are nested within schools was also addressed. An explicit attempt was made to determine whether the same achievement model could be applied to classrooms with differing instructional «histories». Within-classroom slope differences in their pre- and posttest achievement regression were taken as a grouping criterion. According to several researchers (see Burstein, 1980, for a review), differences in slopes across groups can reflect substantive educational effects. Classrooms with greater opportunities to learn may have allowed certain students to take fuller advantage of their abilities, and may also have caused less able students to fall behind their classmates. Thus a steep slope will result when the posttest scores of these students are regressed on to their pretest achievement scores. In contrast, the slope of achievement should be flatter in classrooms where the teachers are using a compensatory or remedial approach. Here, entry characteristics would then predict posttest achievement only minimally. Hence, slope heterogeneity across classrooms may be regarded as a function of instructional treatment differences between them.

In the Schneider and Treiber study, two extreme groups of classrooms with widely differing slopes were preselected on the basis of a regression analysis of pre- and posttest achievement to determine whether different achievement models would hold for the two subgroups. As a main result, the assumption of «local» applicability of achievement models (cf. Snow, 1977) was confirmed. That is, the causal model (LISREL) fit the data only for the subgroup with steep regression slopes (so called «High-Slope» classes), thus indicating that the proposed educational achievement model could be confirmed for classrooms where a «meritocratic» approach (Rachman-Moore & Wolfe, 1984) was preferred to compensatory efforts.

The second reason for considering the Schneider and Treiber procedure in the present study was the assumption that some of its inherent limitations may also restrict the validity of its findings. For example, one problem of that study was that only two (group-level) indicators were available to assess instructional processes. Moreover, math achievement was represented by single test scores. Further, it could be criticized that structural modeling procedures were restricted to the analysis of extreme groups, and that no information was given about a general test of the proposed model of school achievement based on the total sample. Finally, the use of regression slopes as the sole grouping criterion appears questionable because it does not provide us with information concerning achievement gains in the different groups.

To overcome these restrictions in the present study, a combined criterion involving the regression slope *and* achievement gain were simultaneously used to classify similar types of classrooms. Results of non-hierarchical clustering algorithms were compared with median-split procedures to obtain a complete classification of all classrooms involved in the study rather than considering extreme groups. A general model test was included to assess the global validity of the proposed achievement model. Finally, multiple indicators of math achievement and several individual-based indicators of instructional quality were available for analysis. Taken together, all these extensions of the Schneider and Treiber approach were regarded as essential for testing the validity of their basic findings.

Method

Subjects

The subjects were 632 students from 34 fifth-grade classrooms in German primary schools. All students were selected from rural and urban schools in the Munich area. The 34 teachers participating in the study were the regular classroom and math teachers for these classes.

Instruments

A general aptitude test included three subtests, namely spatial ability, inferential reasoning, and verbal ability («Kognitiver Fähigkeitstest» KFT by Heller, Gaedeke & Weinläder, 1976). In sum, the test consisted of 70 items and can be considered as sufficiently reliable (Cronbach's alpha = .91).

A questionnaire was used to assess students' perception of instruction. From this measure, the following aspects of instructional quality and classroom management were selected:

(1) Clarity, e.g. «The teacher explains the material in a way that is easy for me to understand»;

(2) Appropriateness, comprising components of both individualization and remedial help, e.g. «The teacher gives me extra help with work I find difficult» or «Does the teacher give you tasks that are too difficult for you?»;

(3) Task orientation, e.g. «The teacher sticks to classwork and doesn't get sidetracked», and

(4) Management, e.g. «The teacher knows what's going on in the classroom».

Given the fact that only short subscales varying between five and nine items were used, the reliability coefficients (Cronbach's alpha) ranging between .64 and .75 appeared sufficient.

The two math tests tapped the content covered during the 5th graders' math instruction, but dealt with different topics. One test (29 items) focused on arithmetic skills, and the other test (23 items) consisted mainly of word problems and thus required less algorithmic skills, but more comprehension and application abilities. Both tests were given in a free answer format, allowing for an analysis of the individual's errors as well as avoiding the «correcting-for-guessing»-problems related to multiple-choice tests.

Procedure

The study started at the beginning of the school year (September 1983) and lasted for approximately 2 years, including five measurement points. Only the three first measurement points are considered in the present analysis (cf. Helmke, Schneider & Weinert (1986), Helmke (1986), or Weinert & Helmke (1984) for an overview of the complete study). All subjects were tested in groups within their classrooms. Pretest sessions were given immediately after beginning of the academic year to make sure that teacher effects could be neglected (all teachers were new to the classes).

Administering the questionnaires assessing, among others, students' perception of instruction took two subsequent hours. Subjects were told that their own opinions were important (and not those of their neighbours), and that there were no right or wrong answers. It was emphasized that neither parents nor teachers could find out the answers.

Completing the math tests took another two consecutive hours, one hour for each of the two tests. The sequence of test presentation was counterbalanced within classrooms to prevent cheating: seat neighbours always got two different versions of tests, that is, either the test focusing on arithmetic problems or the test dealing with word problems.

The same procedure was repeated at the intermediate measurement point (December 1983) as well as at the third measurement point (April 1984).

Steps of Statistical Analysis

A sequential strategy was used to analyze the data. First, group-level analyses were carried out to identify subgroups or types of similar classrooms. Within-classroom regression slopes from math posttest on pretest and achievement gain scores, aggregated on classroom level, served as input variables for the SAS «Fastclus» clustering procedure. In addition to cluster analysis, two simple descriptive statistics, the median of slope coefficients and achievement gain scores, were computed to classify the classrooms into the following four groups: High-Slope/High-Gain, High--Slope/Low-Gain, Low-Slope/High-Gain, and Low-Slope/Low-Gain.

The next step was to estimate the proposed achievement model by using LISREL VI (Jöreskog & Sörbom, 1984). Simultaneous group comparisons were done to test the hypothesis that the same achievement model would hold for all groups. In case of rejection of the hypothesis, separate model tests were planned for the different types of classrooms. Finally, the last step of the analysis involved the general test of the model, that is, a model test based on the total sample of subjects with group membership ignored.

Results

Preliminary analyses with the non-hierarchical clustering algorithm did not lead to interpretable results. That is, it was not possible to specify a cluster solution that was superior to theoretically possible alternatives, i.e., solutions with differing numbers of clusters. Moreover, inspection of cluster means did not prove helpful in classifying groups that significantly differed with regard to average slopes or achievement gains. Finally, the sample size of the different clusters differed remarkably, and there were severe outlier problems. Thus, the simple, descriptive approach leading to the four groups described above was preferred instead.

Descriptive results

Table 1 depicts some descriptive data on the four groups of classrooms.

Table 1: *Descriptive Characteristics of the 4 Groups of Classrooms (z-standardized scores)*

	High Slope		Low Slope	
	High Gain	Low Gain	High Gain	Low Gain
N =	228	154	131	124
Observed indicators of instructional quality				
Use of instructional time	.34	.22	—.26	—.44
Effective classroom management	.11	.36	.27	—.65
Frequency of cues facilitating comprehension	.47	—.19	—.20	—.28
Clarity of instruction	.17	—.09	.19	—.32
Consolidation of learned material by exercises (seatwork)	.50	—.30	—.23	—.18
Amount of homework	—.27	—.84	.60	.61
Growth (improvement or worsening) of affective variables				
Positive attitude toward mathematics	—.44	.10	.26	.27
Positive attitude toward school	—.39	—.00	.03	.46
Self concept of aptitude	—.23	—.05	—.17	.41

Note. N = 637 students from thirty-four 5th-grade classrooms.

Tableau 1: *Caractéristiques descriptives des quatre groupes de classes*

Perhaps the most remarkable result concerns the pattern for the «High Gain/ /High Slope» group. Here, almost all indicators of instructional quality and effective classroom management (Helmke et al., 1986) were very positive compared with the other groups. Of particular interest was the outstanding role of facilitating cues and of consolidation of learned material by individualized, subject-related support during seat work (cf. Weinert & Helmke, in press).

On the other hand, it is also apparent that the strong achievement gain of this group had corresponding penalties since there were significant negative side effects with regard to affective learning outcomes. The High-Slope/High-Gain group is by far the most unfavorable one in that both positive attitudes toward mathematics and school, as well as student's self concept of aptitude, deteriorated significantly.

Results of Causal Modeling

From a theoretical perspective, the most interesting issue was whether the same proposed achievement model could hold for the four subgroups of classrooms, or whether different groups require different theoretical models. The LISREL model can be used to analyze data from several groups simultaneously. As Bentler (1980) pointed out, different types of model tests are available for multi-sample analyses. For example, a 'tight' model test would require all parameters of the model to have identical estimates in all groups. In a 'moderate' model test, only some critical theoretical parameters, like factor loadings, should be held constant for all groups, whereas only the replication of the same pattern of model parameters is necessary in a 'loose' model test.

When the 'loose' model test was first conducted for the complete achievement model including all three measurement points, the major finding was that estimates were considerably biased for all groups because of high multicollinearity among achievement measures. To overcome this problem, only the first (pretest) and last (posttest) measurement point were included in data analysis. It would seem that the time interval between pretest and the intermediate test (about four months) was too short to yield significant achievement changes.

Results for the 'loose' model test showed that the proposed achievement model did not hold for all four groups of classrooms ($\chi^2 = 213.01$, $df = 156$, $p < .002$). Not surprisingly, the data fit was even worse when a 'moderate' model test requiring identical measurement models for all groups was done ($\chi^2 = 282.34$, $df = 213$, $p < .001$). As a consequence, separate analyses were carried out for the four types of classrooms, starting with the full school achievement model and all possible relationships among exogeneous and endogeneous constructs specified in the model.

The results showed that, for all groups, the path between the aptitude factor and math posttest achievement, as well as the path between the instruction factor and math posttest achievemment, could be omitted without any loss of information. However, the resulting LISREL solutions for the reduced model still did not fit the data, regardless of group. Inspection of the first order derivatives revealed that measurement errors between corresponding indicators of math pre - and posttest achievement should be allowed to intercorrelate for all groups. Indeed, this minor model modification lead to a considerable drop in χ^2. The decreases in χ^2 for the High-Slope/High-Gain, High-Slope/Low-Gain, Low-Slope/High-Gain and Low-Slope/Low-Gain groups were 42.17, 11.68, 18.17, and 14.31, respectively. All these changes were statistically significant ($p < .01$).

The resulting achievement models are depicted in Figures 1 to 4. Although these models show a similar structure, it should be noted that it was only for the two High-Slope subgroups that the achievement model actually fit the data. The best data fit was obtained for the High-Slope/High-Gain subgroup, as is also indicated by the goodness-of-fit index (GFI) and the root mean square residual (RMS). In contrast, the model obviously did not fit the data of Low-Slope/Low-Gain group.

Interestingly, the model test based on the total number of subjects (N = 632) lead to a different result. The χ^2 goodness-of-fit statistic seems to indicate lack of fit ($\chi^2 = 86.22$, $df = 39$, $p < .001$), but the GFI and the RMS indices are more reliable in this case because they are not affected by large sample size. Both measures indicated excellent data fit, and thus suggest that the achievement model was empirically confirmed for all subjects.

Figure 1: *Resulting LISREL model for the «High Slope» classrooms with high achievement gains*

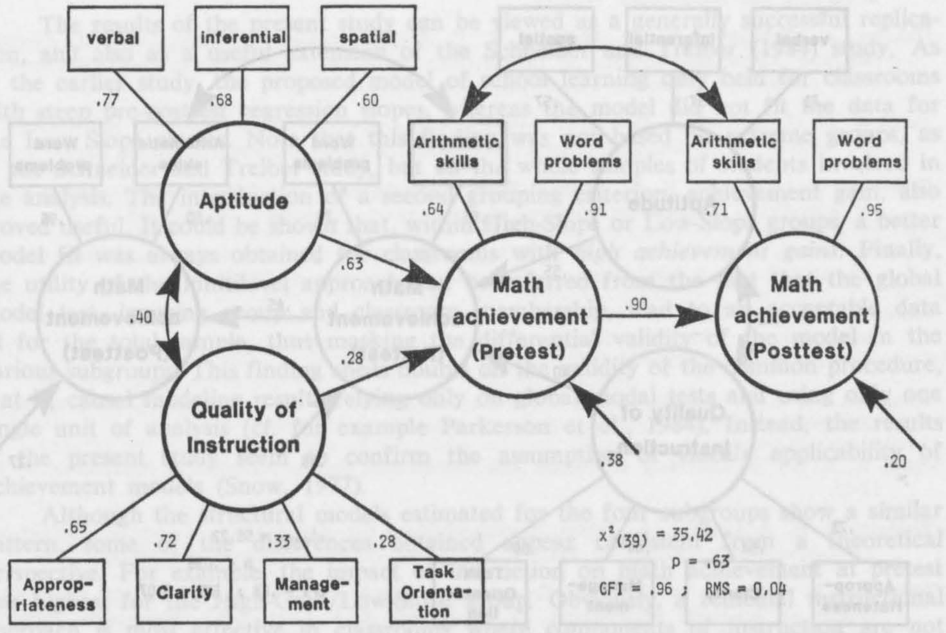

Figure 1: *Résultats obtenus selon le modèle LISREL pour les classes «High Slope» à fort progrès*

Figure 2: *Resulting LISREL model for «High Slope» classrooms with low achievement gains*

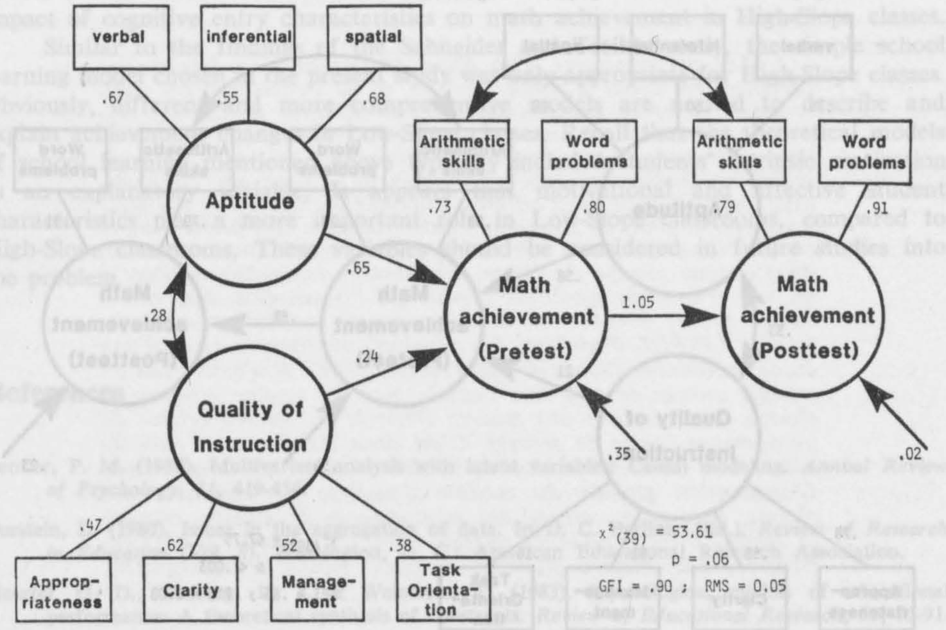

Figure 2: *Résultats obtenus selon le modèle LISREL pour les classes «High Slope» à faible progrès*

Figure 3: *Resulting LISREL model for «Low Slope» classrooms with high achievement gains*

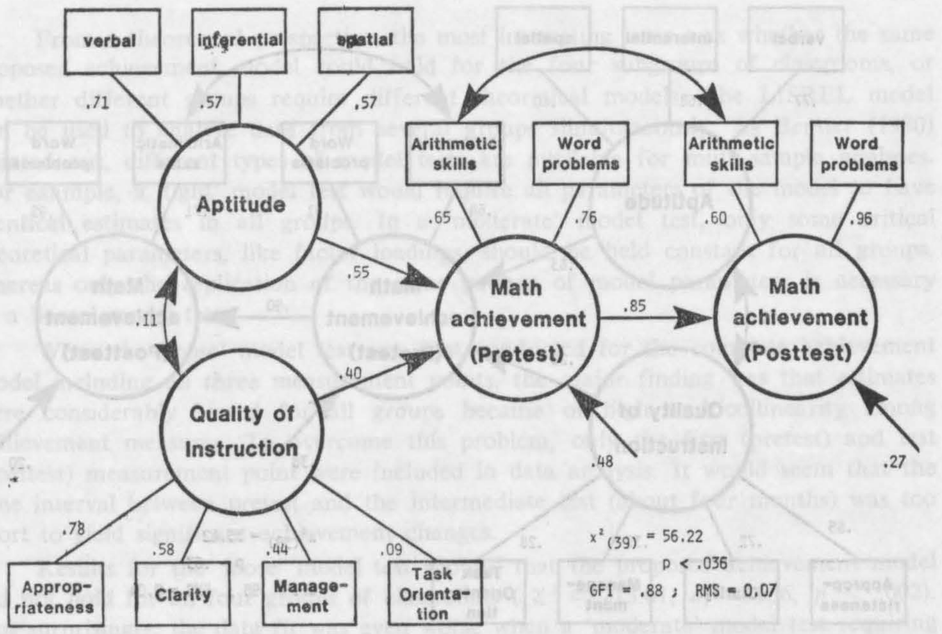

Figure 3: *Résultats obtenus selon le modèle LISREL pour les classes «Low Slope» à fort progrès*

Figure 4: *Resulting LISREL model for «Low Slope» classrooms with low achievement gains*

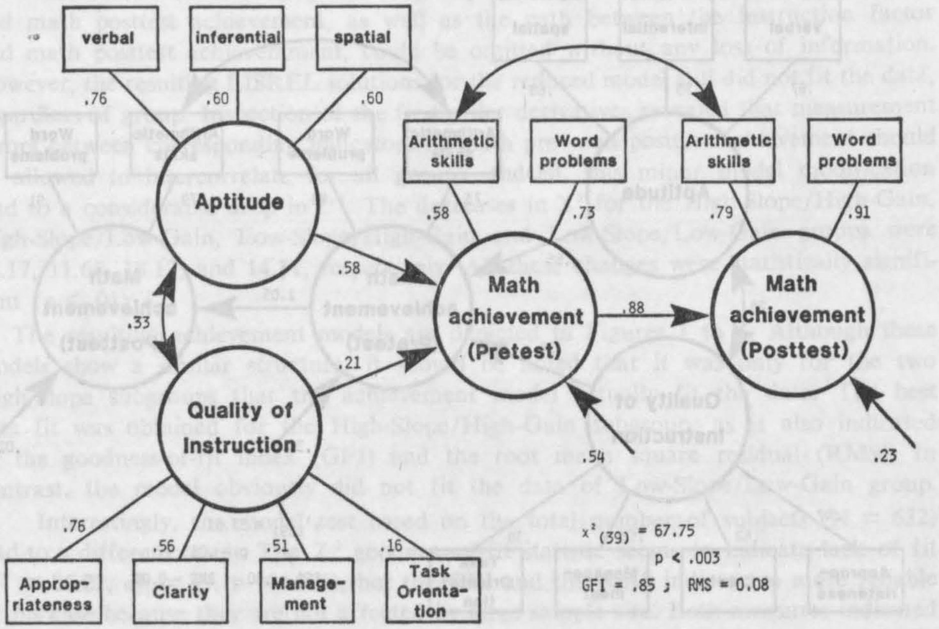

Figure 4: *Résultats obtenus selon le modèle LISREL pour les classes «Low Slope» à faible progrès*

Discussion

The results of the present study can be viewed as a generally successful replication, and also as a useful extension of the Schneider and Treiber (1984) study. As in the earlier study, the proposed model of school learning only held for classrooms with steep pre-posttest regression slopes, whereas the model did not fit the data for the Low Slope classes. Note that this finding was not based on extreme groups, as in the Schneider and Treiber study, but on the whole samples of students involved in the analysis. The introduction of a second grouping criterion, achievement gain, also proved useful. It could be shown that, within High-Slope or Low-Slope groups, a better model fit was always obtained for classrooms with *high achievement gains.* Finally, the utility of the multilevel approach can be inferred from the fact that the global model test, ignoring group and classroom membership, lead to an acceptable data fit for the total sample, thus masking the differential validity of the model in the various subgroups. This finding sheds doubts on the validity of the common procedure, that is, causal modeling results relying only on global modal tests and using only one single unit of analysis (cf. for example Parkerson et al., 1984). Instead, the results of the present study seem to confirm the assumption of «local» applicability of achievement models (Snow, 1977).

Although the structural models estimated for the four subgroups show a similar pattern, some of the differences obtained appear consistent from a theoretical perspective. For example, the impact of instruction on math achievement at pretest was highest for the High-Gain/Low-Slope group. Obviously, a remedial instructional approach is most effective in classrooms where components of instruction are not severely confounded with cognitive entry characteristics which may contribute independently to math achievement. In these classrooms, the relative impact of instruction is almost comparable to that of aptitude. On the other hand, it makes sense that pretest math achievement can be better explained for High-Slope than for Low-Slope classes (cf. the disturbance terms in Figures 1 to 4), given the relatively pronounced impact of cognitive entry characteristics on math achievement in High-Slope classes.

Similar to the findings of the Schneider and Treiber study, the simple school learning model chosen in the present study was only appropriate for High-Slope classes. Obviously, different and more comprehensive models are needed to describe and explain achievement changes in Low-Slope classes. Recall that the theoretical models of school learning mentioned above typically included students' intrinsic motivation as an explanatory variable. It appears that motivational and affective student characteristics play a more important role in Low-Slope classrooms, compared to High-Slope classrooms. These variables should be considered in future studies into the problem.

References

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31,* 419-456.

Burstein, L. (1980). Issues in the aggregation of data. In D. C. Berliner (Ed.), *Review of Research in Education* (Vol. 8). Washington, D. C.: American Educational Research Association.

Haertel, G. D., Walberg, H. J. & Weinstein, T. (1983). Psychological models of educational performance: A theoretical synthesis of constructs. *Review of Educational Research, 53,* 75-91.

Heller, M. S., Gaedike, A. & Weinläder, H. (1976). *Kognitiver Fähigkeitstest für 4. bis 13 klassen (KFT 4-13).* Weinheim: Beltz.

Helmke, A. (1986). Student attention during instruction and achievement. In: S. E. Newstead S. H. Irvine & P. D. Dann (Eds.), *Human assessment: cognition and motivation*, (pp. 273-286) Dordrecht/The Netherlands: Nijhoff.

Helmke, A., Schneider, W. & Weinert, F. E. (1986). Quality of instruction and classroom learning outcomes — Results of the German contribution to the Classroom Environment Study of the IEA. *Teaching and Teacher Education*, *2*, 1-18.

Jöreskog, K. G. & Sörbom, D. (1984). LISREL VI. *Analysis of linear structural relationships by the method of maximum likelihood*. Mooresville, Indiana: Scientific Software. Inc.

Parkerson, J. A., Lomax, R. G., Schiller, D. P. & Walberg, H. J. (1984). Exploring causal models of educational achievement. *Journal of Educational Psychology*, *76*, 638-646.

Rachman-Moore, D. & Wolfe, R. G. (1984). Robust analysis of a nonlinear model for multilevel educational survey data. *Journal of Educational Statistics*, *9*, 277-293.

Schneider, W. & Helmke, A. (1986). Mehrebenenanalytische Ansätze zur Erklärung von Schulleistungen. In: M. von Saldern (Ed.), *Mehrebenenanalyse — Beiträge zu einem jungen Forschungsansatz*. Bern: Huber.

Schneider, W. & Treiber, B. (1984). Classroom differences in the determination of achievement changes. *American Educational Research Journal*, *21*, 195-211.

Snow, R. E. (1977). Individual differences and instructional theory. *Educational Researcher*, *6* (10), 11-15.

Weinert, F. E. & Helmke, A. (1984). *Zwischenbericht für das Projekt «Unterrichtsqualität und Leistungszuwachs»*. München: Max-Planck-Institut für psychologische Forschung.

Weinert, F. E. & Helmke, A. (in press). Compensatory effects of student self-concept and instructional quality on academic achievement. In: F. Halisch & J. Kuhl (Eds.), *Motivation, intention, volition*. Berlin: Springer.

Influence de la classe sur les performances d'un élève

Les performances d'un élève dépendent à la fois de ses capacités en début d'année scolaire et de l'enseignement qu'il reçoit. L'objet de cette étude était de mesurer l'influence respective de ces deux facteurs sur l'apprentissage des mathématiques dans une population de 632 élèves de 5ème année d'école primaire. Quatre groupes homogènes de classes ont été constitués sur la base de deux critères: analyses de régression pré-test/ /post-test et gains de scores en mathématiques. Les performances en mathématiques ont fait l'objet d'un recueil longitudinal et, dans chaque groupe de classe, les données (capacités initiales à un test d'aptitude générale, qualité de l'instruction d'après les élèves, pré et post-test en mathématiques) ont été soumises au modèle LISREL d'analyse. Les résultats confirment ceux d'une étude précédente (Schneider & Treiber, 1984). Il n'existe pas un modèle unique de structure causale pour les quatre groupes de classes. Les résultats ont montré l'intérêt du second critère de groupement (gains de scores). C'est dans les classes à gain de score élevé que l'adéquation du modèle s'est avérée la meilleure. L'application globale du modèle d'analyse à l'ensemble des classes, sans tenir compte des groupements de classes, masque la validité différentielle du modèle selon les groupes. D'où l'importance d'analyses à plusieurs niveaux: global et différentiels.

Key words: Models of school learning, Causal modeling (LISREL), Motivation and achievement, Multilevel analysis.

Received: October 31, 1985
Revision received: March 25, 1986

Wolfgang Schneider. Max-Planck-Institut für psychologische Forschung Leopoldstr. 24, D-8000 München 40, FRG.

Current theme of research:

Prediction of school achievement, cognitive development in children

Most relevant publications in the field of Educational Psychology:

Schneider, W. (1980). *Bedingungsanalysen des Rechtschreibens* (determinants of spelling skills). Bern, Switzerland: Huber,

Schneider, W. (1979). Educational Psychology. *The German Journal of Psychology, 3,* 236-266 (with F. E. Weinert and B. Treiber).

Schneider, W. (1984). Classroom differences in the determination of achievement changes. *American Educational Research Journal, 21,* 195-298 (with B. Treiber).

Schneider, W. (1985). Exploratorische Analysen zu Komponenten des Schulerfolgs (exploratory analyses on the predictatibility of academic success). *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 17,* 325-340 (with K. Bös).

Andreas Helmke. Max-Planck-Institut für psychologische Forschung, Leopoldstr. 24, D-8000 München 40, FRG.

Current theme of research:

Motivation and academic achievement; teacher effectiveness, test anxiety.

Most relevant publications in the field of Educational Psychology:

Helmke, A. (1986). Student attention during instruction and achievement. In S. E. Newstead, S. H. Irvine & P. D. Dann (Eds.). *Human assessment: cognition and motivation.* Dordrecht/ /The Netherlands: Nijhoff.

Helmke, A., Schneider, W. & Weinert, F. E. (1986). Quality of instruction and classroom learning outcomes. Results of the German contribution to the Classroom Environment Study of the IEA. *Teaching and Teacher Education, 2,* 1-18.

Helmke, A. & Schrader, F. W. (1986). *Interactive effects of teacher diagnostic competence and instructional quality on student academic achievement.* Manuscript submitted for publication.

Weinert, F. E. & Helmke, A. (in press). Interactive effects of student motivation and instructional quality on academic achievement. In F. Halisch & J. Kuhl (Eds.), *Motivation, intention, volition.* Berlin: Springer.