Journal of Educational Psychology
1995, Vol. 87, No. 4, 624–637

0022-0663/95/$3.00

The Causal Ordering of Academic Achievement and Self-Concept of Ability During Elementary School: A Longitudinal Study

Andreas Helmke
University of Landau

Marcel A. G. van Aken
University of Nijmegen

This article addresses the question of the causal ordering of self-concept of ability and academic achievement during elementary school. The questions were (a) Do self-concept and achievement influence each other? and (b) Does it make a difference whether achievement is assessed by marks or by tests? The sample consisted of 697 students from 54 German elementary school classes. The design of the study allowed 3 measurement waves from Grade 2 to Grade 4. Mathematics achievement was measured both by marks and by tests. The results of the structural equation modeling analyses show that it makes a difference whether achievement is (as usually) measured with only one indicator (either mark or test performance), or if both indicators are integrated in the model. The latter model clearly supports the skill development model: In elementary school, prior self-concept does not significantly contribute to the prediction of subsequent achievement.

The direction of causality between academic achievement and academic self-concept has been the subject of considerable interest and speculation in educational psychology (for an overview, see Byrne, 1984; Helmke, 1992; Marsh, 1990a). On the one hand, there can be no doubt that academic self-concept is formed at least in part by prior achievement. Achievement-related successes and failures influence self-concept through various means, in particular through the evaluation of significant others (e.g., teacher and parents). This consideration underlies the skill-development approach, which maintains that self-concept is primarily the result of past achievement rather than a cause for subsequent achievement. The opposite position is held by the self-enhancement model, which claims that academic achievement depends not only on prior achievement but that prior self-concept also contributes significantly to the prediction of subsequent achievement (cf. Calsyn & Kenny, 1977). For example, a high self-concept of ability may be a favorable precondition for the initiation and persistence of effort in learning and achievement situations (Helmke, 1989, 1991, 1992). Also, students with a low self-concept might avoid critical learning situations that could threaten their self-concept and thus might show less effort in school. Furthermore, as suggested by self-worth theory (Covington, 1984), students with low success expectations are prone to develop failure-avoiding tactics (e.g., avoidance behavior, procrastination, or intrapsychic defensive processes). These activities may yield temporary relief, but in the long run they are mostly counterproductive and impair academic achievement.

The controversy between the self-enhancement and skill-development positions (i.e., whether prior self-concept has an impact on the prediction of subsequent achievement that is independent of the impact of prior achievement) is not only theoretically interesting but also of considerable practical importance. Self-enhancement programs, for example (see Scheirer & Kraut, 1979), rely on the assumption that an improvement in self-concept will lead to a gain in academic achievement.

Given the theoretical and practical significance of the problem, surprisingly little sound research concerning the causal predominance of self-concept or achievement exists. Rather than giving a report of the literature (for a comprehensive review, see Byrne, 1986; Marsh, 1990a), we restrict our short review to those few longitudinal studies investigating the issue of the causal predominance of self-concept of ability and academic achievement that satisfy the three central prerequisites noted in Byrnes's (1984) review, namely (a) establishment of a statistical relationship, (b) establishment of a clear time precedence, and (c) testing of a causal model. The collection of these studies has been facilitated by the prior work of Marsh (1990b), who reported six studies fulfilling the criteria just mentioned. We extend the list by adding some more recent studies (see Table 1). For methodological reasons, we have not considered longitudinal studies that used the statistical method of cross-lagged panel correlation (e.g., Calsyn & Kenny, 1977; Chapman, Cullen, Boersma, & Maguire, 1981; Marsh, 1987; Pottebaum, Keith, & Ehly, 1986), because this method has been strongly criticized by statisticians (Rogosa, 1980). Traditional path analyses (e.g., Bachman &

Andreas Helmke, Department of Psychology, University of Landau, Landau, Germany; Marcel A. G. van Aken, Department of Developmental Psychology, University of Nijmegen, Nijmegen, The Netherlands.

Correspondence concerning this article should be addressed to Andreas Helmke, Department of Psychology, University of Landau, Im Fort 7, D-76829, Landau, Germany. Electronic mail may be sent via Internet to helmke@uni-landau.de.

Table 1
An Overview of Longitudinal Studies Using Structural Equation Causal Modeling on the Causal Predominance of Academic Self-Concept and Academic Achievement

Study	Achievement measure	SCA measure	Waves	Grades	Time interval	Sample size	SEM-method	Causal predominance
Byrne (1986)	Marks and tests (reading)	2 scales	2	9–12	6 months	929	LISREL	No cross-lagged effect
Helmke (1992)	Marks and tests (mathematics)	4 scales	4	5–6	½–1 year	341	PLS[a]	Reciprocal
Jerusalem (1983)	Marks (mathematics)	Scale	5	5–6	4–6 months	510	LISREL	Reciprocal
Marsh[c] (1990b)	GPA (self-report) (3 subjects)	3 items	4	10–13	1 year	1,456	LISREL	Self-concept
Newman[d] (1984)	Test (mathematics)	1 item	3	2–10	3–5 years	185	LISREL	Achievement
Pekrun (1987)	GPA (3 subjects)	Scales	4	4–9	1–2 years	365	LISREL[b]	Reciprocal
Shavelson & Bolus (1982)	GPA (3 subjects)	Scales	2	7–8	4 months	99	LISREL	Self-concept
Skaalvik & Hagtvet (1990)	Teacher ratings (3 subjects)	Scale	2	3–4 6–7	18 months	271 364	LISREL	Reciprocal

Note. SCA = self-concept of ability; SEM = structural equation modeling; GPA = grade point average.
[a] Partial least squares method. [b] Without latent variables. [c] Reanalysis of the Youth in Transition Study. [d] However, cf. Marsh's (1988) reanalysis of the Newman data led to a different conclusion.

O'Malley, 1977; Marsh, 1987; Pugh, 1976) and the comparison of simple correlations (e.g., Bridgeman & Shipman, 1978) have been criticized for methodological reasons as well. There is now broad consensus that structural equation modeling (SEM) is the preferable method for a sound analysis of longitudinal panel data. Furthermore, we have not mentioned studies that did not consider both self-concept and achievement at more than one measurement point (e.g., Felson, 1984; Maruyama, Rubin, & Kingsbury, 1981; Skaalvik & Rankin, 1990).

What can be concluded on the basis of the causal modeling studies reported in Table 1? Obviously, the pattern of results with regard to the causal predominance of self-concept or achievement is quite heterogeneous, as an inspection of the last column in Table 1 reveals. There are studies supporting either the skill development or the self-enhancement approach, but there are also several studies with reciprocal effects and even one with no cross-lagged effects at all. What could be the reason for this heterogeneous pattern of results?

First, considerable differences concerning the design and the sample of the various studies exist. For example, the number of measurement points varies from two (e.g., follow-up studies) to four or five (longitudinal studies); grade levels vary from 2nd grade to 1 year after 12th grade; and time intervals range from 4 months to 5 years! Second, the operationalization of academic self-concept varies from assessment by means of a single item to comprehensive scales. Third, the studies differ with respect to the domain under consideration. Half of the studies focused on one subject, such as mathematics or reading, whereas other studies investigated several domains, mostly relying on composite measures (i.e., aggregating both self-concept and academic achievement across various domains). However, some studies (e.g., Faber, 1992; Marsh, 1986, 1990c, 1992; Shavelson & Bolus, 1982; Pekrun, 1987) have emphasized that there are considerable differences in the pattern of causal dominance if this issue is analyzed separately for different domains. Finally, the database for academic achievement is represented by the annual teacher ratings (marks), or objective achievement tests, or both.

In our view, the last aspect—the type of criterion chosen as an indicator of achievement—is of particular theoretical interest and represents a main focus of this study. The predominant use of teacher ratings as criteria for academic achievement may lead to an inadequate view with regard to the issue of the causal predominance of self-concept versus academic achievement. There has been a long debate as to whether and to what degree teacher ratings are at all reliable judgments (see the critique by Hansford & Hattie, 1982, and Marsh's, 1990b, overview) and whether they actually reflect student achievement. First, it is a well-known fact that marks often serve pedagogical functions too (e.g., disciplining students or encouraging and reinforcing student effort). Second, some studies (e.g., Schrader & Helmke, 1990) have found that students' motivational characteristics (including self-concept) may play a significant role for teachers' judgments of achievement, which is independent of student's actual achievement.

Third, there are substantial differences between marks and test performance as measures of academic achievement. Marks are usually communicated to the students and represent the main database for students' social comparison processes within their classroom. This is presumably not (or at least seldomly) the case for test performance, because

students in most of the studies that have used achievement tests did not experience their test results or their relative standing. Furthermore, getting good marks is an important and positively evaluated goal for most of the students, whereas test performance may or may not be of personal significance.

Finally, marks are more predictable and controllable than test results, because deficiencies (e.g., in certain domains of mathematics or in problems in test-like written exercises) can be partially compensated by means of additional effort and persistence. These systematic differences between marks and test performance are reflected by the size of the correlations between these two achievement criteria, which generally range from .40 to .60.

An additional question is whether there are gender differences concerning the pattern of relations between self-concept and achievement in general and to the causal predominance in particular. Contrary to widespread beliefs, recent reviews and meta-analyses (Halpern, 1992; Hyde, Fennema, & Lamon, 1990) as well as cross-cultural studies (e.g., Lummis & Stevenson, 1990) have shown that the effect sizes of differences in achievement are very small, particularly for younger children, whereas substantial gender differences concerning achievement-related motives, attitudes, and self-concepts have been found as early as in elementary school (cf. Eccles, Wigfield, Harold, & Blumenfeld, 1993; Helmke, in press; Multon, Brown, & Lent, 1991; Stipek & Gralinski, 1991). Whereas this concerns the level of the constructs under consideration, much less is known about gender differences in the pattern of relations between self-concept and achievement. Although this issue is not the focus of this article, we will examine gender differences in the patterns of relations between indicators of mathematics achievement (marks and test performance) and self-concept of mathematics aptitude.

In sum, this study attempts to fill a gap in the literature on the interrelation of self-concept and academic achievement with respect to two deficiencies: First, there is as yet little empirical evidence covering the elementary school period. Second, attempts to systematically compare the differences in the causal pattern arising from marks versus test performance as measures of achievement are lacking.

The study described here is part of the 4-year longitudinal project SCHOLASTIC (School Learning and the Socialization of Talents, Interests, and Competencies). The project was launched in 1988 to investigate the development of children's academic achievement and achievement-related motives and beliefs during elementary school, dependent on student entry characteristics, classroom context, and instructional quality (Helmke, in press; Weinert & Helmke, 1995a, 1995b, in press). The data in the present paper concern mathematics as a domain of academic competence, because the instruments were most comprehensive for this area. In a multiwave, multipanel design, data are analyzed for mathematics achievement in the form of mathematics test performance and math marks, as well as self-concept of academic ability in mathematics.

Method

Sample

The sample consisted of 1,023 students from 54 elementary school classes from urban and rural regions in and around Munich, Germany. Because data provided by teachers were sometimes missing on a nonrandom basis (i.e., for an entire school class), data reported in this article stem from those 697 students (358 boys and 339 girls) for whom there was complete data for all variables.

Instruments

Mathematics tests. Math test performance was measured with a math test that comprised several subtests (Stern, 1989, 1993). To form two indicators of mathematical competence for the causal modeling procedure, several subtests were combined to form two scales: (a) basic arithmetic skills (speed tests of addition, subtraction, multiplication, etc.) and (b) tasks requiring solving word problems and application and transfer of mathematical knowledge (such as reversing basic arithmetic procedures). Both kinds of tasks—arithmetic skills and word problems—are part of the German curriculum in elementary school mathematics. The content and difficulty of the tasks varied from second to fourth grade according to the curriculum of the respective class. Depending on the measurement wave, the number of items for arithmetic skills ranged from 40 to 55 and for word problems from 15 to 20.

Self-concept of ability in mathematics. This was defined as students' self-evaluation in various domains of math competence. Three indicators were formed at each wave, namely the self-evaluation of (a) paper-and-pencil arithmetic tasks and mental arithmetic skills, (b) the ability to solve math word problems, and (c) an overall self-evaluation of competence in mathematics in general. All judgments required explicit social comparison processes: Students were asked to rate their relative standing in the respective domain as compared with their classmates.

Marks. We used the official marks in the report cards,[1] which are provided by teachers at the end of each school year.

Procedure

The mathematics tests and the student questionnaires (containing, among others, the self-concept instruments) were administered annually to the intact classroom groups. The math tests and the questionnaire each took one class period (about 45 min) at each of the measurement points. Self-concept questionnaires and mathematics tests were administered shortly before marks were given. The three measurement points (waves) for the assessment of the

[1] Regarding the use of marks provided by the teacher in class, one might choose between either standardizing the marks per class or not. Standardizing by class would mean that the child's score reflects the actual *rank order* the child has in the class. This reflects important information for the child in determining his or her self-concept of ability. On the other hand, in the German school system the marks also have an *absolute value.* The child is aware of this absolute value, and this value may determine the reactions of parents and friends to the marks received by the child, so it might also be important for determining the child's self-concept of ability. Repeating all the LISREL analyses in the article using marks standardized by class led to a somewhat lower fit of the models but not to any differences in pattern or magnitude of the regression coefficients in the models.

Table 2
Intercorrelations of the Manifest Variables

Variable/test	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Marks																				
1. Mark2	4.59	0.95	—																	
2. Mark3	4.31	1.00	.577	—																
3. Mark4	4.21	1.02	.552	.754	—															
Self-concept																				
4. Self2g	0.84	1.00	.333	.306	.303	—														
5. Self2s	0.94	0.89	.195	.213	.209	.598	—													
6. Self2m	0.78	1.16	.210	.218	.197	.511	.535	—												
7. Self3g	0.53	0.88	.370	.402	.391	.405	.328	.338	—											
8. Self3s	0.43	0.71	.272	.306	.269	.398	.374	.395	.684	—										
9. Self3m	0.30	0.83	.304	.328	.323	.389	.339	.413	.601	.645	—									
10. Self4g	0.45	0.83	.334	.453	.508	.362	.269	.329	.407	.358	.379	—								
11. Self4s	0.38	0.68	.353	.389	.430	.332	.311	.331	.447	.383	.428	.658	—							
12. Self4m	0.36	0.88	.327	.425	.403	.354	.334	.373	.456	.421	.502	.636	.660	—						
Mathematics																				
13. Test2a	27.51	9.52	.338	.441	.437	.246	.183	.185	.279	.253	.276	.286	.287	.288	—					
14. Test2k	14.25	5.76	.470	.562	.570	.341	.283	.263	.363	.268	.312	.355	.325	.339	.661	—				
15. Test3a	15.93	5.86	.356	.431	.399	.231	.144	.224	.351	.241	.308	.309	.366	.327	.544	.446	—			
16. Test3k	5.48	2.33	.462	.548	.508	.309	.238	.279	.380	.286	.295	.276	.297	.312	.464	.549	.462	—		
17. Test4a	14.41	3.26	.244	.359	.417	.164	.106	.148	.319	.220	.268	.305	.374	.281	.319	.302	.544	.320	—	
18. Test4k	7.68	4.24	.454	.575	.600	.297	.247	.238	.405	.302	.335	.417	.422	.421	.466	.577	.523	.507	.490	—

Note. The range of marks was from 6 (high) to 1 (low); the self-concept subscales ranged from −2 (low) to +2 (high). The mathematics subtests reflect the number of correct responses, varying from scale to scale (depending on the number of items). Marks 2–4 represent marks at the end of Grades 2, 3, and 4. Selfs 2–4 g, s, and m represent self-concept of mathematical ability concerning a general overall self-evaluation (g), arithmetic skills (s), and mathematical word problems (m) at Grades 2–4. Tests 2–4 a and k represent mathematical test performance in the subdomains arithmetic skills (a) and application and transfer of mathematical knowledge (k) at Grades 2–4.

relevant data are labeled according to the respective grades (2, 3, or 4).

Statistical Analyses

Structural equation modeling (LISREL 7; Jöreskog & Sörbom, 1989) was used to test our hypotheses about the causal predominance of math achievement or math-related academic self-concept and about the appropriateness of using tests and marks as indicators of one single latent variable. Three classes of models were examined:

Model 1. In the first step, a model with math tests and marks as indicators of one and the same latent variable (achievement) was tested. This model assumed an initial correlation between the latent variables achievement and self-concept at Wave 2, and cross-lagged effects of achievement at Time t on self-concept at Time $t + 1$ (representing the skill-development model), as well as cross-lagged effects of self-concept at Time t on achievement at Time $t + 1$ (representing the self-enhancement model).

Model 2. Second, two separate models were analyzed: one concerning the relations between marks and self-concept (*Model 2a*) and one concerning the relations between math tests and

Table 3
Goodness-of-Fit Indices for Alternative Models

Model	χ^2	*df*	χ^2/df	GFI	BBI	TLI	Description
1	351.61	108	3.26	.94	.94	.94	Model with one latent variable (achievement)
1-auto	493.97	112	4.23	.93	.90	.92	Model 1 with autoregressive relations only
2a	158.85	39	4.07	.96	.96	.95	Model with one latent variable (marks)
2a-auto	300.07	43	6.98	.94	.88	.90	Model 2a with autoregressive relations only
2b	167.68	66	2.54	.97	.94	.96	Model with one latent variable (tests)
2b-auto	279.49	70	3.99	.95	.91	.93	Model 2b with autoregressive relations only
3a	522.59	106	4.93	.92	.92	.90	Model with cross-lagged effects between Mark2 on Test3, Mark3 on Test4, Test2 on Mark3, and Test3 on Mark4
3-auto	690.04	114	6.05	.90	.85	.87	Model 1 with autoregressive effects only
3b	443.65	104	4.27	.93	.93	.92	Model 3, with additional within-grade correlations between Marks and Tests for Waves 3 and 4
3c	310.43	102	3.04	.95	.95	.95	Model 3, with cross-lagged effects from Mark2 on Test3, Mark3 on Test4, Test2 on Mark3, and Test3 on Mark4.

Note. GFI = goodness-of-fit index; BBI = Bentler–Bonett index; TLI = Tucker–Lewis index.

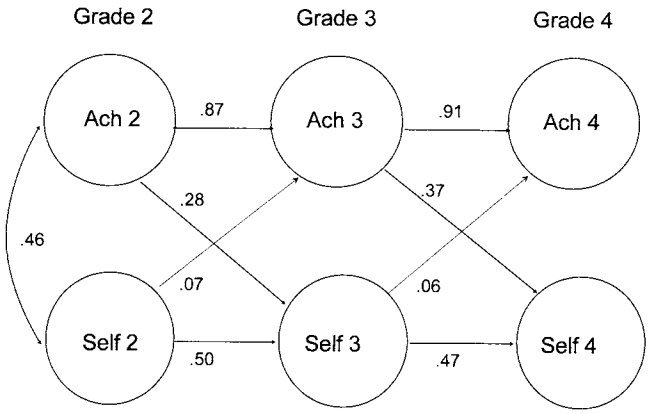

Figure 1. LISREL model with math achievement aggregated across math test performance and marks. Ach = achievement in mathematics, aggregated across mark and grade; Self = self-concept of ability in mathematics.

self-concept (*Model 2b*). In both models, similar initial correlations and cross-lagged effects as in the first model are assumed.

Model 3. Third, all variables were again included in one model, but now with the latent variable achievement at each wave split into two separate latent variables, namely marks (indicated by children's marks for mathematics at that measurement point) and tests (indicated by both math tests at that measurement point). Again, initial correlations between marks, self-concept, and tests were assumed, as were similar cross-lagged effects as in the first and second models.

Because we were interested in the relations between the latent variables rather than in the measurement per se, in contrast to some other studies (see e.g., Skaalvik & Hagtvet, 1990), no factorial invariance between the measurement waves was assumed. That is, the magnitude of the loadings of the observed variables on the latent variables was allowed to vary between measurement waves. Loadings for the respective models are presented in Tables 4 to 6.

In longitudinal research, the residuals of a manifest variable used in subsequent measurement waves are often correlated, indicating correlated measurement errors. Allowing these correlated errors to be estimated generally leads to a better fit of the model and to more accurate estimates of the stability of the latent variables. Therefore, all models were estimated assuming correlated measurement errors (or correlated uniqueness) between all identical indicators for self-concept, marks, and mathematics tests across the three waves (e.g., between the speed tests in Grade 1 and Grade 2, in Grade 1 and Grade 3, and in Grade 2 and Grade 3). The use of single indicators for marks in Model 2a and Model 3 led to problems in the estimation of correlated measurement errors, as will be discussed in the next paragraph. Therefore, for the indicators of marks, a small amount of correlated measurement errors will be fixed in the models. Again, between all identical indicators for self-concept and tests across the three waves, correlated measurement errors will be estimated.

As indicators of the latent variable marks at each wave, only one indicator was available, namely the marks the pupil received in that period. The use of one single indicator for a latent variable has been criticized elsewhere for methodological reasons (although it should be noted that in German schools only one official mark per year and domain actually exists). For example, Marsh (1988, 1990b) pointed out that two problems may arise. First, relations involving these latent constructs cannot be corrected for unreliability of the indicators. Second, the existence of correlated measurement errors for these indicators is not testable and controllable. Therefore, initial analyses with marks as a single indicator (i.e., in Model 2a and in Model 3) were conducted in which the assumed reliability of the construct was .90 (cf. Jöreskog & Sörbom, 1989, p. 153; Marsh, 1990b), and the correlated residuals between the different grade estimates were set at 12.5% of the residual variance. In subsequent sensitivity analyses of Models 2a and 3a, the implications of these a priori assumptions were explored.

For the evaluation of our models in terms of their goodness-of-fit, we will report the χ^2/df ratio and the indices proposed by Bentler and Bonett (1980; BBI) and Tucker and Lewis (1973; TLI). It is now widely known that the chi-square likelihood ratio test is sensitive to sample size and that in large samples, the chi-square test will almost always reach significance, indicating a difference between the observed and the predicted covariance matrices. As an alternative, the χ^2/df ratio has been proposed. The acceptable numerical values for this ratio that have been proposed by different authors cover a range from 1 to 5 (Byrne, 1989). In an evaluation of the most widely used indices of goodness-of-fit, Marsh, Balla, and McDonald (1988) concluded that the TLI was relatively unbiased, independent of sample size, and penalized model complexity. Values of greater than .9 are regarded as indicators of an acceptable fit of the model. Both BBI and TLI are computed by comparing the target model with a null model, assuming complete independence of all observed measures, and can be roughly seen as indicating the proportion of covariance explained by the model. No serious problems during estimation

Table 4
Loadings of Manifest on Latent Variables for Model 1 (Figure 1)

Variable	Achievement 2	Achievement 3	Achievement 4	Self 2	Self 3	Self 4	Error/ uniqueness
Marks							
Mark2	58						67
Mark3	00	73					46
Mark4	00	00	74				45
Self-concept							
Self2g	00	00	00	78			39
Self2s	00	00	00	75			43
Self2m	00	00	00	68			54
Self3g	00	00	00	00	82		34
Self3s	00	00	00	00	83		32
Self3m	00	00	00	00	76		42
Self4g	00	00	00	00	00	79	38
Self4s	00	00	00	00	00	82	33
Self4m	00	00	00	00	00	80	35
Mathematics							
Test2a	72						48
Test2k	86						27
Test3a	00	63					60
Test3k	00	72					48
Test4a	00	00	54				71
Test4k	00	00	83				32

Note. Decimal points have been omitted. These values are for a completely standardized solution. Marks 2–4 represent marks at the end of Grades 2, 3, and 4. Selfs 2–4 g, s, and m represent self-concept of mathematical ability concerning a general overall self-evaluation (g), arithmetic skills (s), and mathematical word problems (m) at Grades 2–4. Tests 2–4 a and k represent mathematical test performance in the subdomains arithmetic skills (a) and application and transfer of mathematical knowledge (k) at Grades 2–4.

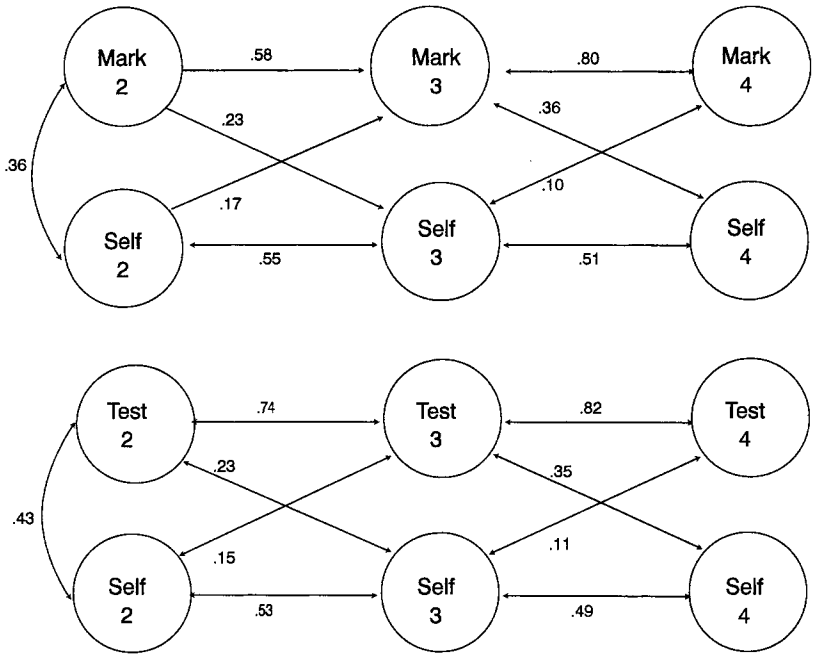

Figure 2. Two LISREL models, displayed separately for marks (upper part) and tests (lower part) as indicators of mathematics achievement. Mark = mark at the end of each school year; Self = self-concept of ability in mathematics; Test = mathematics test.

were encountered, all models converged within 40 iterations, all variances were positive, and no standardized coefficients exceeded 1.00.

Results

The intercorrelations between the 18 manifest variables as well as the means and standard deviations of the raw variables (which were later z transformed with $M = 0$ and $SD = 1$ because of different ranges of the various subscores) are presented in Table 2.

The various goodness-of-fit indices for the alternative models are presented in Table 3.[2] For our first model (Model 1), assuming one latent variable, achievement, an acceptable fit was found. The significant ($p < .05$) relations between the latent variables in this model are presented in Figure 1. Loadings of the manifest variables on their designated latent variables and their error–uniqueness for this model are presented in Table 4.

Because this is the model in which the possibility of correlated measurement errors is largest (because all latent variables were measured with more than one indicator), we will describe these correlated measurement errors here, providing their completely standardized values. For self-concept of arithmetic skills, significant correlated measurement errors were found between Grades 2 and 4 (.05); for self-concept of problem solving, no correlated measurement errors were found; and for overall self-concept of mathematical competence, significant correlated measurement errors were found between Grades 2 and 3 (.06) and between Grades 3 and 4 (.08). For basic arithmetic skills, significant correlated measurement errors were found between Grades 2 and 3 (.15) and between Grades 3 and 4 (.24). For word problems, significant correlated measurement errors were found between Grades 3 and 4 (−.05). For marks, significant measurement errors were found between Grades 2 and 3 (.17), between Grades 3 and 4 (.23), and between Grades 2 and 4 (.15). The importance in assuming these measurement errors was also illustrated by the fact that a model without these errors showed a poorer fit: χ^2 (126, N = 697) = 676.05, $\chi^2/df = 5.37$, GFI = .89, BBI = .89, TLI = .89; significance of the difference with Model 1, χ^2 (18, N = 697) = 324.44, $p < .001$. However, the robustness of the solution in Figure 1 is illustrated by the fact that only small differences in the path coefficients occurred (mean difference = .035, maximum difference = .08; no changes in pattern of significance of path coefficients).[3]

The two models in the second step, separately modeling

[2] We tested all models against their autoregressive counterparts, in which only initial correlations but no cross-lagged effects were assumed. In all cases, the addition of cross-lagged effects led to significant ($p < .001$) improvements in the fit of the models. This indicates that achievement in mathematics and math-related self-concept are related not only in the beginning of our measurement period but also have an additional influence on each other over time.

[3] Note that although the effects of correlated measurement errors in Model 1 are only discussed here, correlated measurement errors were assumed in all estimated models throughout the study. However, because the effects of Models 2a, 2b, 3a, 3b, and 3c were largely similar to the effects in Model 1, they will not be discussed.

Table 5
Loadings of Manifest on Latent Variables for Model 2a and 2b (Figure 2)

Variable	Grade			Self			Tests			Error/ uniqueness
	2	3	4	2	3	4	2	3	4	
Marks										
Mark2	95									10
Mark3	00	95								10
Mark4	00	00	95							10
Self-concept										
Self2g	00	00	00	79/78						38/39
Self2s	00	00	00	75/75						44/43
Self2m	00	00	00	68/68						54/53
Self3g	00	00	00	00	81/81					34/34
Self3s	00	00	00	00	83/83					32/31
Self3m	00	00	00	00	76/76					42/42
Self4g	00	00	00	00	00	79/79				37/38
Self4s	00	00	00	00	00	81/82				34/33
Self4m	00	00	00	00	00	80/80				36/35
Mathematics										
Test2a	00	00	00	00	00	00	76			42
Test2k	00	00	00	00	00	00	87			25
Test3a	00	00	00	00	00	00	00	66		56
Test3k	00	00	00	00	00	00	00	71		49
Test4a	00	00	00	00	00	00	00	00	55	70
Test4k	00	00	00	00	00	00	00	00	89	21

Note. Decimal points have been omitted. Values are for a completely standardized solution. Loadings are the result of the analyses of separate models. For self-concept, first loadings are from Model 2a, second loadings are from Model 2b. Marks 2–4 represent marks at the end of Grades 2, 3, and 4. Selfs 2–4 g, s, and m represent self-concept of mathematical ability concerning a general overall self-evaluation (g), arithmetic skills (s), and mathematical word problems (m) at Grades 2–4. Tests 2–4 a and k represent mathematical test performance in the subdomains arithmetic skills (a) and application and transfer of mathematical knowledge (k) at Grades 2–4.

the relation between either marks and self-concept (Model 2a) or tests and self-concept (Model 2b), both show an acceptable fit. The significant ($p < .05$) relations between the latent variables in both models are presented in Figure 2. Loadings of the manifest variables on their designated latent variables and the error–uniqueness for these models are presented in Table 5.

As discussed earlier, the use of a single indicator can be seen as a weakness of the models presented here. In a sensitivity analysis of Model 2a (where an a priori reliability estimate of .90 and a correlated residual of 12.5% of the uniqueness were assumed), we varied both the reliability estimates for the single indicator and the amount of correlated residuals. Table 6 presents a comparison of the results from Model 2a in Figure 2, with estimates based on a reliability of 1.0, .95, .85, and .80, and a correlation between the residuals of 0%, 12.5%, 25%, and 50% of the corresponding residual variances (see Marsh, 1990b, for a description of this procedure). Table 6 shows that the stability of the latent variable "grade" varied inversely with the assumed reliability of the single indicators. The assuming of correlated residuals compensated marginally for this effect. Important for the interpretation of our models involving the single-indicator construct marks is that the cross-lagged paths varied only marginally with varying reliability of the indicators and varying correlated residuals.

For our third step, also assuming that marks and tests should be regarded as two separate latent variables, but now including them in one model, a model with a marginally acceptable fit was found (see Model 3a in Table 3). The significant ($p < .05$) relations between the latent variables in this model are presented in Figure 3. Loadings of the manifest variables on their designated variables and their uniqueness–error for this model are presented in Table 7. Again, the use of a single indicator was tested in a sensitivity analysis of Model 3a according to the same procedure as for Model 2a.

The results of this analysis are presented in Table 8. As established for Model 2a, for Model 3a we found that although the stability of the variable marks varied inversely with the assumed reliability of the single indicators, the cross-lagged paths involving this variable varied only marginally, with varying reliability of the indicators and varying correlated residuals. Inspection of the modification indices (Jöreskog & Sörbom, 1989, p. 45) indicated that the fit of the model could be improved by allowing for additional covariance between the residuals of the latent variables tests and marks in Measurement Waves 3 and 4. Remember that the covariance between tests and marks in Measurement Wave 2 is already estimated in the model. Therefore, these additional covariances between the two residuals cannot by attributed to correlations assumed in an earlier wave or to the effects of the latent variable self-concept. Two alternative models were formulated to account for these additional covariances. First, Model 3b, where correlations between the residuals of the latent variables tests and marks within

Table 6
Sensitivity Analysis: Path Coefficients for Alternative Versions of Model 2a With Different Reliabilities and Correlated Residuals for the Single-Indicator Construct "Marks"

Correlated uniqueness	From:[a] / To:[b]	Mark2			Mark3		Self2		Self3	
		Self2	Mark3	Self3	Self4	Mark4	Mark3	Self3	Mark4	Self4
				Uniqueness = 0, reliability = 1.00						
.00		.34	.52	.21	.31	.71	.18	.57	.13	.54
				Uniqueness = .05, reliability = .95						
.00		.35	.56	.22	.33	.76	.17	.55	.12	.53
.00625		.35	.55	.22	.33	.75	.18	.56	.12	.53
.01250		.35	.54	.22	.33	.74	.18	.56	.12	.53
.02500		.35	.52	.22	.33	.72	.18	.56	.13	.53
				Uniqueness = .10, reliability = .90						
.00		.36	.60	.23	.36	.81	.16	.55	.09	.51
.01250		.36	.58	.23	.36	.80	.17	.55	.10	.51
.02500		.36	.56	.23	.35	.77	.17	.55	.11	.52
.05000		.36	.53	.22	.34	.74	.19	.55	.13	.52
				Uniqueness = .15, reliability = .85						
.00		.37	.64	.24	.38	.87	.15	.54	.07	.50
.01875		.37	.61	.24	.38	.84	.16	.54	.08	.50
.03750		.37	.59	.24	.38	.82	.17	.54	.09	.50
.07500		.38	.53	.23	.36	.76	.19	.55	.12	.51
				Uniqueness = .20, reliability = .80						
.00		.38	.68	.26	.39	.93	.14	.53	.04	.48
.02500		.38	.65	.25	.40	.90	.15	.53	.06	.48
.05000		.38	.61	.25	.40	.86	.17	.53	.08	.48
.10000		.39	.54	.24	.39	.79	.20	.54	.11	.49

Note. Correlation between uniqueness is set at 0%, 12.5%, 25%, or 50% of the uniqueness. Within a given model, the uniqueness for the three mark variables and the three uniqueness correlations among the three grade variables were assumed to be equal.
[a] The headings in this row that follow indicate variables from which path coefficients originate. [b] The headings in this row that follow indicate resultant variables of path coefficients.

Wave 3 and within Wave 4 were estimated. Second, Model 3c, where additional cross-lagged paths were estimated from marks in Wave 2 on tests in Wave 3, from marks in Wave 3 on tests in Wave 4, from tests in Wave 2 on marks in Wave 3, and from tests in Wave 3 on tests in Wave 4. Table 3 shows that both alternative models showed im-

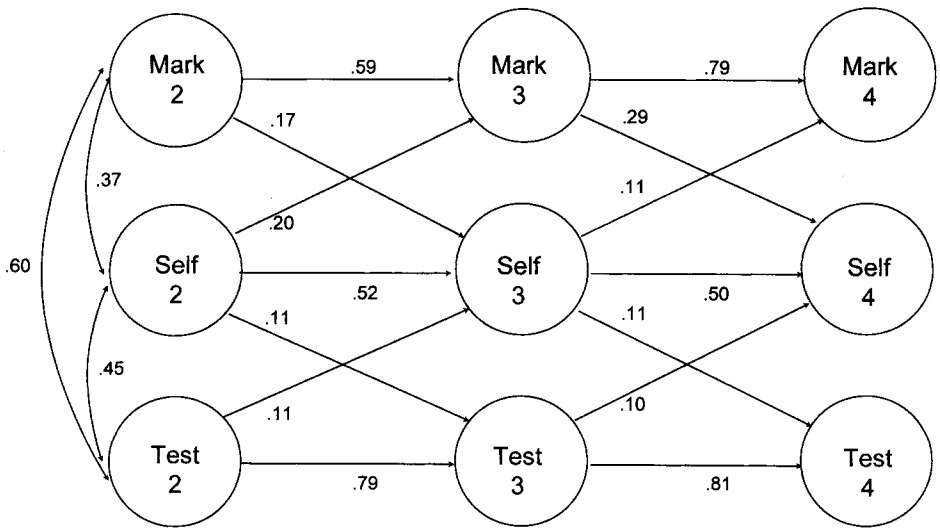

Figure 3. LISREL model with marks and test performance as separate indicators of mathematics achievement (Model 3a). Mark = mark at the end of each school year; Self = self-concept of ability in mathematics; Test = mathematics test.

Table 7
Loadings of Manifest on Latent Variables for Model 3a (Figure 3) and Model 3c (Figure 4)

Grade	Self			Tests			Marks			Error/ uniqueness
	2	3	4	2	3	4	2	3	4	
Marks										
Mark2	95/95									10/10
Mark3	00	95/95								10/10
Mark4	00	00	95/95							10/10
Self-concept										
Self2g	00	00	00	78/79						39/38
Self2s	00	00	00	74/75						45/44
Self2m	00	00	00	68/68						54/54
Self3g	00	00	00	00	82/82					33/33
Self3s	00	00	00	00	82/82					33/32
Self3m	00	00	00	00	76/76					42/42
Self4g	00	00	00	00	00	79/79				37/37
Self4s	00	00	00	00	00	81/81				34/34
Self4m	00	00	00	00	00	80/80				36/36
Mathematics										
Test2a	00	00	00	00	00	00	71/73			50/47
Test2k	00	00	00	00	00	00	91/90			17/19
Test3a	00	00	00	00	00	00	00	61/63		63/61
Test3k	00	00	00	00	00	00	00	81/78		35/40
Test4a	00	00	00	00	00	00	00	00	50/52	76/73
Test4k	00	00	00	00	00	00	00	00	96/90	08/19

Note. Decimal points have been omitted. Values are for a completely standardized solution. Loadings are the result of the analyses of separate models. First loadings are from Model 3a, second loadings are from Model 3c. Marks 2–4 represent marks at the end of Grades 2, 3, and 4. Selfs 2–4 g, s, and m represent self-concept of mathematical ability concerning a general overall self-evaluation (g), arithmetic skills (s), and mathematical word problems (m) at Grades 2–4. Tests 2–4 a and k represent mathematical test performance in the subdomains arithmetic skills (a) and application and transfer of mathematical knowledge (k) at Grades 2–4.

provements of fit as compared with Model 3a. Table 9 shows a comparison of the path coefficients of three versions of Model 3.

Table 9 shows that when within-wave correlations were assumed (Model 3b as compared with Model 3a), only marginal differences between the path coefficients occurred. The only meaningful difference occurred in the path from self-concept in Grade 2 to tests in Grade 3. This effect was slightly smaller (.11 vs. .17) in the version of the model assuming within-wave correlations between the residuals. The effect from mathematics tests in Grade 3 to self-concept in Grade 4 increased from .08 to .10 and passed the level of significance. However, when cross-lagged effects were assumed (Model 3c as compared with Model 3a), the model changed drastically. As was expected, the stability of marks and tests decreased somewhat, the effects from achievement (both marks and tests) to self-concept remained fairly the same, but the effects from self-concept to achievement almost completely disappeared. To illustrate the impact of these changes, the significant relations ($p < .05$) between the latent variables in Model 3c are presented in Figure 4. Loadings of the manifest variables on their designated variables and the uniqueness–error terms for this model are presented in Table 7.

In comparing the substantial results of the various types of models, we see that their implications for the question of causal predominance of self-concept versus achievement differed.

Assuming that marks and tests can be regarded as one latent variable, achievement, there was clear support for a skill-development model: The effect of the achievement on later self-concept was found repeatedly over both waves, whereas prior self-concept did not significantly contribute to the prediction of subsequent achievement. When the latent variable achievement was split into marks and tests, the picture was somewhat different.

In the final model, with latent variables comprising marks and tests (as separate constructs but within the same model), for marks a complete reciprocal model was found between Grades 2 and 3, but between Grades 3 and 4 the effect of marks on self-concept was somewhat stronger than self-concept on marks. For mathematics tests, the effects were reciprocal, both between Grades 2 and 3 and between Grades 3 and 4, although the effects were smaller than for marks. However, when effects of marks on tests and vice versa were allowed, the effects of self-concept on achievement completely disappeared, suggesting a pure skill-development model. Apparently, the small effects of self-concept on achievement can be understood in terms of relations between marks and test achievements, possibly caused by third variables underlying both.

A comparison of the separate models for marks and tests

Table 8
Sensitivity Analysis: Path Coefficients for Alternative Versions of Model 3a With Different Reliabilities and Correlated Residuals for the Single-Indicator Construct "Marks"

Correlated uniqueness	From:[a] / To:[b]	Mark2		Mark3		Self2			Self3			Test2		Test3	
		M3	S3	M4	S4	M3	S3	T3	M4	S4	T4	S3	T3	S4	T4
Uniqueness = 0, reliability = 1.00															
.00		.50	.15	.70	.23	.22	.53	.12	.14	.52	.11	.13	.78	.14	.80
Uniqueness = .05, reliability = .95															
.00		.55	.16	.75	.26	.21	.52	.12	.13	.51	.11	.12	.79	.12	.81
.00625		.54	.16	.74	.26	.21	.52	.12	.13	.51	.11	.12	.79	.12	.81
.01250		.53	.16	.73	.25	.22	.53	.12	.13	.51	.11	.12	.79	.12	.80
.02500		.51	.16	.72	.24	.22	.53	.12	.14	.51	.11	.13	.78	.13	.80
Uniqueness = .10, reliability = .90															
.00		.61	.18	.81	.29	.19	.52	.11	.10	.50	.11	.11	.80	.09	.81
.01250		.59	.17	.79	.29	.20	.52	.11	.11	.50	.11	.11	.79	.10	.81
.02500		.57	.17	.77	.28	.20	.52	.11	.12	.50	.11	.11	.79	.11	.81
.05000		.53	.17	.74	.26	.22	.52	.12	.13	.50	.11	.12	.78	.12	.80
Uniqueness = .15, reliability = .85															
.00		.67	.20	.87	.32	.17	.52	.10	.07	.49	.10	.09	.81	.07	.81
.01875		.64	.19	.84	.32	.18	.52	.10	.09	.49	.10	.09	.80	.08	.81
.03750		.61	.19	.82	.31	.19	.52	.11	.10	.49	.11	.10	.80	.08	.81
.07500		.54	.18	.76	.29	.21	.52	.12	.12	.50	.11	.11	.78	.11	.80
Uniqueness = .20, reliability = .80															
.00		.74	.23	.93	.35	.13	.51	.09	.04	.48	.10	.06	.82	.05	.81
.02500		.70	.22	.90	.35	.15	.51	.09	.06	.48	.10	.07	.81	.06	.81
.05000		.66	.21	.86	.34	.17	.51	.10	.08	.48	.10	.08	.80	.06	.81
.10000		.57	.19	.79	.31	.21	.51	.12	.11	.49	.11	.10	.79	.09	.80

Note. Correlation between uniqueness is set at 0%, 12.5%, 25%, or 50% of the uniqueness. Within a given model, the uniqueness for the three mark variables and the three uniqueness correlations among the three mark variables were assumed to be equal. M = mark; S = self; T = test.

(Models 2a and 2b) and the model with marks and tests as separate latent variables leads to somewhat different conclusions with regard to the effect of math tests on self-concept. The use of either marks or tests may lead to artificially high regression coefficients. The effects of marks and tests in separate models seem to be partly caused by common factors. Therefore, the inclusion of both in one model leads to more correct estimations of these effects, resulting particularly in lower estimates for the effect of math tests on self-concept.

We tested for gender differences in our results by including gender as a variable in all our structural models and reestimating them, assuming effects of gender on all latent variables (cf. Marsh et al., 1985; Skaalvik, 1990). The fit of these models was only slightly higher than that of the respective models without gender (the largest difference in χ^2/df ratio was .18 for Model 3a). Furthermore, a general pattern was found in which gender had an effect on self-concept, indicating that boys had a higher self-concept than girls (the path coefficients from gender to self-concept in Grade 2 were .33 in all models) and, to a lesser extent, an effect on achievement (for both test and marks), indicating that boys also had a slightly higher level of mathematics achievement than girls (path coefficients from gender to achievement ranged from .17 to .19). More important for this investigation, however, was the fact that the pattern of path coefficients presented in Figures 1–4 did not change when gender was included as a variable in the analyses. (The maximum change in path coefficients was .06, found in all models as a reduction of the initial [Grade 2] correlation between achievement and self-concept.) This suggests that gender effects, although they influence the level of achievement and self-concept, do not change the pattern of causal predominance found for the total sample.

Discussion

A main result of our study on the causal ordering of self-concept and academic achievement in elementary school is certainly the difference in the patterns of "cross-lagged effects" (effects of self-concept at Time 1 on achievement at Time 2 and vice versa) between simple models using only one indicator of achievement and more complex models that use both indicators of scholastic achievement. The former models, which use either test performance or marks (such as the majority of relevant studies, see Table 1), yield a reciprocal model, implying that self-concept in elementary school serves both as cause and as effect. However, there is an increasing dominance of paths leading from achievement to self-concept, as compared with the paths from self-concept to achievement, supporting the skill development model more than the self-enhancement model. Surprisingly, the pattern of effects is

Table 9
Comparison of Three Versions of Model 3

Variable	Model 3a[a]	Model 3b[b]	Model 3c[c]
Mark2			
with Self2	.37	.37	.36
with Test2	.60	.60	.54
with Mark3	.59	.54	.38
with Self3	.17	.17	.16
with Test3	—	—	.21
Mark3			
with Test3	—	.20	—
with Mark4	.79	.78	.66
with Self4	.29	.29	.27
with Test4	—	—	.21
Mark4			
with Test4	—	.11	—
Self2			
with Test2	.45	.45	.43
with Mark3	.20	.22	.03*
with Self3	.52	.53	.50
with Test3	.11	.17	.09
Self3			
with Mark4	.11	.11	.04*
with Self4	.50	.51	.48
with Test4	.11	.12	.08*
Test2			
with Self3	.11	.10	.15
with Test3	.79	.72	.69
with Mark3	—	—	.46
Test3			
with Self4	.10	.08*	.12
with Test4	.81	.82	.68
with Mark4	—	—	.24

Note. Mark = mark at the end of each year; Self = self-concept of ability in mathematics; Test = mathematics test performance. [a] Model 3a without relations between tests and marks in Grades 3 and 4. [b] Model 3b with correlations between tests and marks within Grades 3 and 4. [c] Model 3c with cross-lagged effects from marks in Grade T on tests in Grade T + 1 and vice versa.
* Not significant.

very similar for marks and test scores, indicating that using grades or test performance as an indicator of academic achievement makes no difference for the question of causal ordering. The inclusion of both achievement indicators as manifest variables of one construct (as done by Byrne, 1986) supports the skill development approach even more clearly: Later achievement depends almost completely only on prior achievement and not on prior self-concept.

However, the last, most complex model, which includes marks and tests as independent variables (as recommended by Marsh, 1990a), presented in Figures 3 and 4, tells a somewhat different, more subtle story. First, it underlines the imbalance of the pattern of cross-lagged effects in the relation between self-concept and achievement: In Figure 4, there is not a single significant path leading from self-concept to either later test performance or marks. Thus, these results of our study are not in accordance with the majority of existing studies dealing with this topic, which have—for older students—predominantly underlined the reciprocal character of the relationship between self-concept and achievement.

Possibly, the motivational properties of self-concept are not yet fully developed in elementary school. In other words, an efficient way to improve self-concept of ability in elementary school children is to improve their achievement competence. This result would have clearly been masked if only one indicator of achievement (either marks or test performance) had been used in our study. Furthermore, this model illustrates the relative impact of tests versus marks on subsequent self-concept. Whereas from second to third grade both influences are equally strong (.16 and .17), there is a tendency toward a stronger impact of marks on self-concept (.28) from third to fourth grade, as compared with the path from tests to self-concept (.12). Although the result is less clear than we had expected, it underlines the psychologically important differences between tests and marks. The latter are subjectively relevant for students because the mental anticipation of marks may represent a strong incentive for initiating learning activities, effort and persistence (Helmke, 1989). Also, whereas (in the German school system) marks are usually communicated to the students (often in public, which facilitates social comparison processes), test results are usually not reported to the students. Also, because the students did not know about the math tests in advance, preparation was not possible. Thus, the result of an achievement test probably reflects the actual competence of the student in that domain to a greater extent, but it is subjectively less important because success and failure on these tests have no crucial consequences. Finally, our complex model stresses the dynamics of the interrelationship between both indicators of achievement—grades and test performance—and its changes over time.

From the perspective of developmental psychology (cf. Harter, 1983), our results with regard to the determinants of students' self-concept of ability support the conception of two distinct sources of self-evaluation in children, namely (a) competence, in particular, speed and quality of performance, and (b) evaluation by significant others (here, teachers) and the notion that both of these sources (a and b) appear to operate (at least, to some degree) independently from each other. That is, not only marks (which reflect to a major extent teachers' evaluation of students' cumulative achievement) but also actual competence (or "mastery," indicated by test performance) proved to influence self-concept development. The experience of solving the test tasks (or failing to solve them) may affect students' self-evaluation even though they were not informed about their individual results. However, one cannot rule out the possibility that some informal within-classroom social comparison processes (e.g., immediately after the completion of the test) may have taken place.

Although the addition of cross-lagged paths to the model always led to a significant improvement of the fit when compared with "autoregressive models" (models comprising only effects between identical variables assessed at different times, i.e., no "cross-lagged" paths from one variable at Time 1 to another variable at Time 2, etc.), one could still ask why the cross-lagged effects are so low and why the autoregressive effects are so high. It should be noted that in our study the stability coefficients (i.e., the correlations of self-concept and achievement over time) were based on

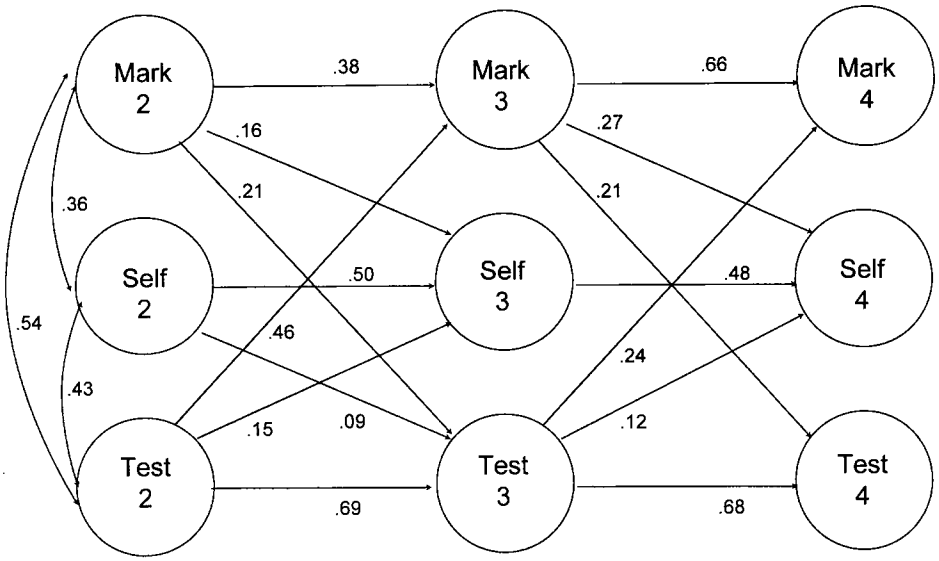

Figure 4. LISREL model with marks and test performance as separate indicators of mathematics achievement, allowing for cross-lagged effects from test performance to marks and vice versa (Model 3c). Mark = mark at the end of each year; Self = self-concept of ability in mathematics; Test = mathematics test.

measures that were constructed to be as similar as possible, because one goal of the study had been to analyze real gains. Thus, we probably have a maximum of stability and a minimum of cross-lagged effects in our study. This leads to a well-known general problem of longitudinal studies that attempt to model the dynamic of the interplay between various constructs: The more similar (in the extreme case, identical) the operationalizations for the same variable over time, the higher (other things being equal) their autocorrelations—and the smaller the size of possible cross-lagged effects. That is, because our math tests across proximal waves were characterized by a considerable amount of content overlap, a maximizing of temporal stability and a minimizing of cross-lagged effects were the consequences.

What is the optimal time interval for studies investigating the causal interrelationship between self-concept and achievement? On the one hand, it is obvious that very large time intervals (e.g., several years) may, depending on the domain and the age level under consideration, mask important developmental phenomena. On the other hand, if one is interested in changes of self-concept and achievement level, conceptualized as relatively stable personality characteristics, daily or weekly fluctuations are certainly less interesting. In our opinion, the smallest natural time lag that makes theoretical sense for the issue of causal ordering of self-concept and achievement is the time interval between marks, which ranges, depending on the school system, from a couple of months to a semester or even a whole year. Nevertheless, it is possible (and even probable) that in special cases and for some persons certain critical events (e.g., a specific rating or mark, or a comment made by a personally very important person, or another important experience) may lead to abrupt and perhaps even lasting changes in the level of self-concept. The same is true for sudden changes in the level of academic achievement, for example, shifts of competence caused by the emergence of qualitatively different stages of information processing or by sudden insights. To the degree that changes in self-concept or achievement occur discretely rather than in a continuous manner, designs providing fixed intervals (such as years) necessarily lead to an increase in the error variance.

Conclusion

The scope of this study has been to fill a gap with regard to the issue of the causal ordering of self-concept and achievement in elementary school, and especially to differentiate the global construct achievement into two components with different psychological meaning, specifically marks and test performance. As our results have shown, previous research on the causal ordering of self-concept and achievement based on either of the two aspects of achievement, or on a mixture of both, tells only part of the story. The most complex—and probably also the most realistic—model clearly supports the skill-development model. It indicates that during elementary school self-concept is mainly a consequence of cumulative achievement-related success and failure and that it does not have a significant impact on later achievement, neither on marks nor on test performance.

Nevertheless, many related questions concerning the causal ordering issue could not (or were not intended to) be answered in this research, and a couple of new questions may have arisen. For example, the results must leave unex-

plained an important paradox: Models that include only grades or test scores indicate effects of self-concept on subsequent achievement, whereas the model including both grades and test scores does not. At present, we do not know which mechanisms are responsible for this effect. To shed more light on this phenomenon, in-depth studies focusing on the dynamics of the mutual relationship between various components of academic achievement—including marks and test performance, as well as other indicators—and possible mediational processes between those indicators appear necessary.

Furthermore, although the complex model (including tests and marks as different constructs) used in our study and recommended for future investigations on this issue may appear complicated enough, one could go further and ask for group specificity, domain specificity, and context specificity of the results (Helmke & Weinert, in press-a, in press-b). In other words, where and to which degree is it possible to develop global and universal models, and where is it necessary to build local and domain-specific models (for a discussion of this point see Snow & Swanson, 1992)? For example, are there differences with regard to the domain or subject matter (cf. Faber, 1992; Marsh, 1990c, 1992), or to the classroom context (style of instruction and feedback, classroom composition and social climate, see Helmke, 1992; Rosenholtz & Simpson, 1984)? And finally, What are the crucial mediating psychological processes accounting for the causal effects of self-concept on achievement (and vice versa), and do they change with age? This list could easily be expanded. Much more theoretical and empirical work, preferably based on longitudinal and quasiexperimental field studies, must be done to gain a deeper understanding of the interplay between self-concept and achievement and its dynamics.

References

Bachman, J. G., & O'Malley, P. M. (1977). Self-esteem in young men: A longitudinal analysis of the impact of educational and occupational attainment. *Journal of Personality and Social Psychology, 35,* 365–380.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bridgeman, B., & Shipman, V. C. (1978). Pre-school measures of self-esteem and achievement motivation as predictors of third-grade achievement. *Journal of Educational Psychology, 70,* 17–28.

Byrne, B. M. (1984). The general/academic self-concept nomological network: A review of construct validation research. *Review of Educational Research, 54,* 427–456.

Byrne, B. M. (1986). Self-concept/academic achievement relations: An investigation of dimensionality, stability, and causality. *Canadian Journal of Behavioral Science, 18,* 173–186.

Byrne, B. M. (1989). *A primer of LISREL. Basic applications and programming for confirmatory factor analytic models.* New York: Springer.

Calsyn, R. J., & Kenny, D. A. (1977). Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievements? *Journal of Educational Psychology, 69,* 136–145.

Chapman, J. W., Cullen, J. L., Boersma, F. J., & Maguire, T. O. (1981). Affective variables and school achievement: A study of possible causal influences. *Canadian Journal of Behavioural Sciences, 13,* 181–192.

Chapman, J. W., Lambourne, R., & Silva, P. A. (1990). Some antecedents of academic self-concept: A longitudinal study. *British Journal of Educational Psychology, 60,* 142–152.

Covington, M. V. (1984). The motive for self-worth. In R. Ames, & C. Ames (Eds.), *Research on motivation in education: Student motivation* (Vol. 1, pp. 78–113). Orlando, FL: Academic Press.

Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development, 64,* 830–847.

Faber, G. (1992). Bereichsspezifische Beziehungen zwischen leistungsthematischen Schülerselbstkonzepten und Schulleistungen [Domain-specific relations between academic self-concepts and academic achievements]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 24,* 66–82.

Felson, R. B. (1984). The effect of self-appraisals of ability on academic performance. *Journal of Personality and Social Psychology, 47,* 944–952.

Halpern, D. F. (1992). *Sex differences in cognitive abilities.* Hillsdale, NJ: Erlbaum.

Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/performance measures. *Review of Educational Research, 52,* 123–142.

Harter, S. (1983). Developmental perspectives on the self-system. In P. H. Mussen (Ed.), *Handbook of child psychology: Vol. 4. Socialization, personality, and social development* (pp. 275–385; 4th ed.). New York: Wiley.

Helmke, A. (1989). Incentive value of success and failure in school: Developmental trends and impact on academic achievement. In F. Halisch & J. H. L. van den Bercken (Eds.), *International perspectives on achievement and task motivation* (pp. 225–237). Lisse: Swets & Zeitlinger.

Helmke, A. (1991). Entwicklung des Fähigkeitsselbstbildes vom Kindergarten bis zur dritten Klasse. In R. Pekrun & H. Fend (Eds.), *Schule und Persönlichkeitsentwicklung. Ein Resumee der Längsschnittforschung* [Development of self-concept of ability from kindergarten till Grade 3] (pp. 83–99). Stuttgart: Enke.

Helmke, A. (1992). *Selbstvertrauen und schulische Leistungen* [Self-confidence and scholastic achievement]. Göttingen: Hogrefe.

Helmke, A. (in press). From optimism to realism? Development of children's academic self-concept from kindergarten to Grade 6. In F. E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich Longitudinal Study.* Cambridge, England: Cambridge University Press.

Helmke, A., & Weinert, F. E. (in press-a). Bedingungsfaktoren schulischer Leistungen [Determinants of scholastic achievement]. In F. E. Weinert (Ed.), *Enzyklopädie der Psychologie, Band 3: Psychologie der Schule und des Unterrichts.* Göttingen, Germany: Hogrefe.

Helmke, A., & Weinert, F. E. (in press-b). Schooling and the development of achievement differences. In F. E. Weinert, & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich Longitudinal Study.* Cambridge, England: Cambridge University Press.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107,* 139–155.

Jerusalem, M. (1983). *Selbstbezogene Kognitionen in schulischen Bezugsgruppen. Eine Längsschnittstudie* [Self-related cogni-

tions in different reference groups: A longitudinal study]. Dissertation, Freie Universität, Berlin.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7—A guide to the program and applications* (2nd ed.). Chicago: SPSS, Inc.

Lummis, M., & Stevenson, H. W. (1990). Gender differences in beliefs and achievement: A cross-cultural study. *Developmental Psychology, 26,* 254–263.

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23,* 129–149.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79,* 280–295.

Marsh, H. W. (1988). Causal effects of academic self-concept on academic achievement: A reanalysis of Newman (1981). *Journal of Experimental Education, 56,* 100–104.

Marsh, H. W. (1990a). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review, 2,* 77–172.

Marsh, H. W. (1990b). Causal ordering of academic self-concept and academic achievement: A multiwave, longitudinal panel analysis. *Journal of Educational Psychology, 82,* 646–656.

Marsh, H. W. (1990c). Influences of internal and external frames of reference on the formation of math and English self-concepts. *Journal of Educational Psychology, 82,* 107–116.

Marsh, H. W. (1992). Content specificity of relations between academic achievement and academic self-concept. *Journal of Educational Psychology, 84,* 35–42.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analyses: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Marsh, H. W., Smith, I. D., & Barnes, J. (1985). Multidimensional self-concepts: Relations with sex and academic achievement. *Journal of Educational Psychology, 77,* 581–596.

Maruyama, G., Rubin, R. A., & Kingsbury, G. G. (1981). Self-esteem and educational achievement: Independent constructs with a common sense? *Journal of Personality and Social Psychology, 40,* 962–975.

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology, 38,* 30–38.

Newman, R. S. (1984). Children's achievement and self-evaluations in mathematics: A longitudinal study. *Journal of Educational Psychology, 76,* 857–873.

Pekrun, R. (1987). Die Entwicklung Leistungsbezogener Identität bei Schülern [Development of pupils' achievement-related identity]. In H. P. Frey & K. Haußer (Eds.), *Identität, Entwicklungen psychologischer und soziologischer Forschung* (pp. 43–57). Stuttgart: Enke.

Pottebaum, S. M., Keith, T. Z., & Ehly, S. W. (1986). Is there a causal relation between self-concept and academic achievement? *Journal of Educational Research, 79,* 140–144.

Pugh, M. D. (1976). Statistical assumptions and social reality: A critical analysis of achievement models. *Sociology of Education, 49,* 34–40.

Rogosa, D. (1980). A critique of cross-lagged correlation. *Psychological Bulletin, 88,* 245–258.

Rosenholtz, S. J., & Simpson, C. (1984). The formation of ability conceptions: Developmental trend or social construction? *Review of Educational Research, 54,* 31–63.

Scheirer, M. A., & Kraut, R. E. (1979). Increasing educational achievement via self-concept change. *Review of Educational Research, 49,* 131–150.

Schrader, F. W., & Helmke, A. (1990). Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? Eine Untersuchung zu Determinanten diagnostischer Lehrerurteile [Are teachers' grades influenced by non-achievement-related considerations? An analysis of the determinants of teachers' diagnostic competence]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 22,* 312–324.

Shavelson, R. J., & Bolus, R. (1982). Self-concept: The interplay of theory and methods. *Journal of Educational Psychology, 74,* 3–17.

Skaalvik, E. M. (1990). Gender differences in general academic self-esteem and in success expectations on defined academic problems. *Journal of Educational Psychology, 82,* 593–598.

Skaalvik, E. M., & Hagtvet, K. A. (1990). Academic achievement and self-concept: An analysis of causal predominance in a developmental perspective. *Journal of Personality and Social Psychology, 58,* 292–307.

Skaalvik, E. M., & Rankin, R. J. (1990). Math, verbal, and general academic self-concept: The internal/external frame of reference model and gender differences in self-concept structure. *Journal of Educational Psychology, 82,* 546–554.

Snow, R. E., & Swanson, J. (1992). Instructional psychology: Aptitude, adaptation, and assessment. *Annual Review of Psychology, 43,* 583–626.

Stern, E. (1989). Mathematiktest. In Max-Planck Institute for Psychological Research (Ed.), *Tätigkeitsbericht 1987–1989.* [Activity report 1987–1989] (pp. 36–38). München, Germany: Max-Planck-Institut für psychologische Forschung.

Stern, E. (1993). What makes certain arithmetic word problems involving the comparison of sets so difficult for children? *Journal of Educational Psychology, 85,* 7–23.

Stipek, D., & Gralinski, J. H. (1991). Gender differences in children's achievement-related beliefs and emotional responses to success and failure in mathematics. *Journal of Educational Psychology, 83,* 361–371.

Tucker, L. R., & Lewis, C. A. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Weinert, F. E., & Helmke, A. (1995a). Inter-classroom differences in instructional quality and interindividual differences in cognitive development. *Educational Psychologist, 30,* 15–20.

Weinert, F. E., & Helmke, A. (1995b). Learning from wise mother nature or big brother instruction: The wrong alternative for cognitive development. *Educational Psychologist.*

Weinert, F. E., & Helmke, A. (in press). The neglected role of individual differences in theoretical models of cognitive development. *Learning and Instruction.*

Received October 20, 1992
Revision received May 23, 1995
Accepted May 24, 1995 ■