This article may not exactly replicate the final version published in the APA journal. It is not the copy of record. The final published version can be obtained from the following:
Arens, A. K., Marsh, H. W., Pekrun, R., Lichtenfeld, S., Murayama, K., & vom Hofe, R. (2017). Math self-concept, grades, and achievement test scores: Long-term reciprocal effects across five waves and three achievement tracks. *Journal of Educational Psychology, 109*(5), 621-634.
doi:10.1037/edu0000163

Math Self-Concept, Grades, and Achievement Test Scores:

Long-Term Reciprocal Effects across Five Waves and Three Achievement Tracks

Date of submission: 19 September 2016

A. Katrin Arens
German Institute for International Educational Research, Germany
Herbert W. Marsh
Australian Catholic University, Australia; Oxford University, UK; King Saud University, Saudi Arabia
Reinhard Pekrun
University of Munich, Germany; Australian Catholic University, Australia
Stephanie Lichtenfeld
University of Munich, Germany
Kou Murayama
University of Reading, UK
Rudolf vom Hofe
University of Bielefeld, Germany

Author Note

This research was supported by four grants from the German Research Foundation (DFG) to R. Pekrun (PE 320/11-1, PE 320/11-2, PE 320/11-3, PE 320/11-4). We would like to thank the German Data Processing and Research Center (DPC) of the International Association for the Evaluation of Educational Achievement (IEA) for organizing the sampling and performing the assessments.

Correspondence concerning this article should be addressed to A. Katrin Arens, Center for Research on Individual Development and Adaptive Education of Children (IDeA), German Institute for International Educational Research, Schloßstraße 29, D-60486 Frankfurt am Main, Germany, Telephone: + 49 69 24708 138, Email: arens@dipf.de

Abstract

This study examines reciprocal effects between self-concept and achievement by considering a long time span covering grades 5 through 9. Extending previous research on the reciprocal effects model (REM), this study tests (1) the assumption of developmental equilibrium as time-invariant cross-lagged paths from self-concept to achievement and from achievement to self-concept, (2) the generalizability of reciprocal relations of self-concept when using school grades and standardized achievement test scores as achievement indicators, and (3) the invariance of findings across secondary school achievement tracks. Math self-concept, school grades in math, and math achievement test scores were measured once each school year with a representative sample of 3,425 German students. Students' gender, IQ, and socioeconomic status (SES) were controlled in all analyses. The findings supported the assumption of developmental equilibrium for reciprocal effects between self-concept and achievement across time. The pattern of results was found to be invariant across students attending different achievement tracks and could be replicated when using school grades and achievement test scores in separate and in combined models. The findings of this study thus underscore the generalizability and robustness of the REM.

Keywords: math self-concept; math achievement; reciprocal effects; school tracks

Educational Impact and Implications Statement

The present research shows that students' math self-concept (i.e., their self-perceptions of competence in math) influences their school grades and achievement test scores in this domain, and that grades and achievement test scores in math in turn influence students' math self-concept. Hence, math self-concept and math achievement are reciprocally related to each other, and these relations are found across a long period of time covering the first five years of secondary schooling (grades 5 to 9). Moreover, the reciprocal relations between math self-concept and math achievement are independent from the influence of students' gender, socioeconomic status, and IQ, and are found for students attending different achievement tracks of secondary school. These findings highlight the role of self-concept as an important predictor as well as an outcome of students' achievement. As such, educational interventions should adopt a dual approach by simultaneously fostering students' self-concept and achievement.

Self-Concept, Grades, and Achievement Test Scores in Mathematics: Long-Term Reciprocal Effects across Five Waves and Three Academic Tracks

Date of submission: 19 September 2016

Academic self-concept is defined as students' self-perceptions of competence in academic domains (Shavelson, Hubner, & Stanton, 1976). It has been a prominent construct in educational psychology over the last several decades as it has been found to share substantial relations to outcome variables including academic achievement (Marsh, 2007; Marsh & O'Mara, 2008b; Valentine, DuBois, & Cooper, 2004). In this context, many studies have supported reciprocal relations between academic self-concept and achievement involving important theoretical implications for self-concept formation and practical implications for the enhancement of both self-concept and achievement (for an overview see Marsh & Craven, 2006). However, several issues remain to be clarified. These include the assumption of developmental equilibrium, the interplay of school grades and standardized achievement test scores as two alternative achievement measures, and the generalizability of findings across school tracks. These issues are targeted in the present study.

Relations between Academic Self-concept and Achievement

When examining the link between academic self-concept and academic achievement, many studies have attested substantial cross-sectional relations (e.g., Arens, Yeung, Craven, & Hasselhorn, 2011; Marsh et al., 2013). Studies scrutinizing longitudinal relations have attracted even more attention because they help elucidate causality in the relation between self-concept and achievement. Thus, a critical question has been whether self-concept is an outcome of achievement or whether achievement is an outcome of self-concept. Calsyn and Kenny (1977) posited two models for the temporal relation between self-concept and achievement. While the skill development model suggests that achievement predicts self-concept, the self-enhancement model suggests that self-concept predicts achievement. Originally, the skill development and self-enhancement models were strictly contrasted but

recent research indicates that such a clear either-or stance is inappropriate because self-concept and achievement share mutually reinforcing relations. Therefore, in contemporary self-concept research, the reciprocal effects model (REM) prevails for depicting the relations between self-concept and achievement. Accordingly, self-concept is both an outcome of former and a predictor of subsequent achievement (e.g., Huang, 2011; Marsh & Craven, 2006; Möller, Retelsdorf, Köller, & Marsh, 2011; Niepel, Brunner, & Preckel, 2014).

Number of Waves and Developmental Equilibrium

Studies integrating two measurement waves can already serve to test the temporal ordering of relations between self-concept and achievement (Marsh, Byrne, & Yeung, 1999). However, the inclusion of three or more waves would allow for the examination and comparisons *among* skill development and self-enhancement effects over time. The assumption of developmental equilibrium would expect skill development and self-enhancement paths to be of similar size from one wave to the next (for studies integrating related assumptions see for example Marshall, Parker, Ciarrochi, & Heaven, 2014). Hence, in this case the effect from achievement (self-concept) to self-concept (achievement) would be of similar size across the different time lags, e.g., across waves 1 and 2 and waves 2 and 3.

Developmental equilibrium is not essential to providing evidence of reciprocal effects, but support for this assumption has a number of important advantages. For complex models resulting from the assessment of self-concept and achievement across many waves with many items used in each wave, the added parsimony provides more robust and precise estimates and facilitates the presentation and interpretation of results. Moreover, support for developmental equilibrium offers some protection against alternative interpretations of the results based on potential other variables not considered (Kenny, 1975). More importantly, if developmental equilibrium can be supported, self-concept (achievement) exerts a similar influence on later achievement (self-concept) at different time points. Hence, studies testing developmental equilibrium are best based on a large number of measurement waves which cover a long and

relevant period of time. For instance, it would be interesting to examine the robustness of skill development and self-enhancement effects across adolescence or secondary school years.

However, in a meta-analysis, Huang (2011) revealed that out of 32 studies examining longitudinal relations between self-concept and achievement, 19 studies relied on a two-wave design, eight studies had three measurement waves, two studies were respectively based on four and five measurement waves, and only one study covered six measurement waves. Thus, there seems to be a need for further studies that include more than two or three measurement waves. In line with these considerations, the present study covers five waves tracking German students' from the fifth to the ninth grade. Students' self-concept and achievement were collected once every school year. Therefore, the present study can replicate findings on the REM over an exceptionally long time interval and examine developmental equilibrium across students' fifth to ninth grade, the years of mandatory secondary schooling in Germany.

First-order and Higher-order Paths

The REM is commonly tested by cross-lagged panel models embedded in the framework of structural equation modeling (SEM; Curran & Bollen, 2001; Marsh et al., 1999). This modeling approach includes autoregressive or stability paths estimating the effect of one variable on the same variable across subsequent measurement waves, e.g., the relation between self-concept measured at the time 1 (t1) and self-concept measured at t2. In addition, cross-lagged paths represent the reciprocal relations of one variable on another variable between measurement waves (i.e., effects of self-concept at t1 on achievement at t2, and effects of achievement at t1 on self-concept at t2). In studies that cover more than two measurement waves, it is possible to include both first-order and higher-order paths. First-order paths depict the relation between two directly adjacent time points, i.e., the effect of self-concept at t1 on self-concept at t2 as an example of a first-order stability path, and the effect of self-concept at t1 on achievement at t2 as an example of a first-order cross-lagged path. First-order paths thus describe "lag 1" paths referring to the effect of one variable on the

same variable across two adjacent measurement waves. Higher-order paths describe the relations between constructs measured at more distal time points. Second order paths refer to "lag 2" paths among constructs measured at t1 and t3, third order paths refer to "lag 3" paths among constructs measured at t1 and t4 and so on. Thus, for instance, second-order stability addresses the relation between self-concept measured at t1 and self-concept measured at t3, and second-order cross-lagged paths depict the reciprocal relations between self-concept at t1 and achievement at t3.

Beyond the inherent inclusion of first-order paths, it might be worthwhile to consider higher-order paths in cross-lagged panel models for the reciprocal relation between self-concept and achievement as this allows for examining long term relations (Marsh & O'Mara, 2008a). Indeed, in a four-wave model for studying the relations between reading self-concept and reading achievement, besides first-order stability paths, Retelsdorf, Möller, and Köller (2014) found significant higher-order (i.e., second-order and third-order) stability estimates for reading self-concept and reading achievement. In addition to the corresponding first-order path, there was also evidence of significant higher-order cross-lagged paths for the relation between former reading achievement and later reading self-concept. Marsh, Gerlach, Trautwein, Lüdtke, and Brettschneider (2007) conducted a three-wave study examining the longitudinal relations between physical self-concept and physical achievement. The results revealed first-order and second-order stability estimates for both physical self-concept and physical performance. In addition, there was evidence of first-order and second-order cross-lagged paths between physical achievement and physical self-concept. Accordingly, the recommendations formulated by Marsh et al. (1999) for "ideal" studies on the REM include the advice to start with a full-forward model which incorporates the estimations of all paths. Researchers are then advised to compare this complete model with more parsimonious alternative models. Therefore, in this study, we use a full-forward model including second-

order, third-order, and fourth-order paths as a starting point to examine reciprocal relations between self-concept and achievement across five measurement waves.

Achievement Indicators: School Grades versus Test Scores

School grades and standardized achievement test scores are the two most commonly used indicators of students' achievement. School grades are very salient to students as they are directly communicated, easy to compare among classmates, and entail important implications for students' school careers. School grades do not only narrowly represent student achievement but also refer to other student characteristics such as students' effort or classroom behavior (Brookhart, 1993; McMillan, Myran, & Workman, 2002; Zimmermann, Schütte, Taskinen, & Köller, 2013). On the other hand, students are often unaware of their relative performance on standardized achievement tests. Therefore, students' self-concept has been found to be more strongly related to school grades than to standardized achievement test scores (Marsh et al., 2014; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005).

Nonetheless, school grades suffer from idiosyncrasies as teachers have been found to grade on a curve, allocating the best grades to the relatively best performing students within a classroom and the poorest grades to the relatively poorest performing students (Marsh et al., 2014). Hence, teachers use the classroom as a narrow frame of reference in their grading procedure. Accordingly, the same student with the same level of objective achievement can receive divergent grades depending on the average achievement and achievement standards in the individual student's class. For this reason, school grades are difficult to compare across classes, schools, and nations whereas standardized achievement tests are particularly designed for the purpose of such comparisons. Thus, school grades and standardized achievement test scores each have their advantages and disadvantages when they are used as achievement indicators, emphasizing their distinct yet complementary nature.

Therefore, it appears worthwhile to use both kinds of achievement indicators to examine reciprocal relations to self-concept. Nonetheless, the majority of studies supporting

reciprocal relations between self-concept and achievement have included school grades as achievement indicators (e.g., Marsh, 1990; Niepel et al., 2014; for an overview see Huang, 2011). Yet, there is also evidence that reciprocal relations between self-concept and achievement exist when using standardized achievement test scores as achievement indicators (Möller, Zimmermann, & Köller, 2014; Retelsdorf et al., 2014; Seaton, Parker, Marsh, Craven, &Yeung, 2014). However, most previous studies have considered school grades or achievement test scores separately while a more sophisticated approach would be to consider both achievement indicators simultaneously. The study of Marsh, Trautwein et al. (2005) provided evidence of the REM for math self-concept and math achievement when analyzing math grades and math test scores separately as well as when combining them into one model. As this study included only two measurement waves, there still seems to be a need for studies that combine both achievement indicators and consider multiple waves across a longer time interval to more adequately test the validity of the REM and the assumption of developmental equilibrium for both kinds of achievement indicators.

Generalizability across School Tracks

A number of studies have documented the applicability of the REM to both academic and non-academic domains such as reading (Retelsdorf et al., 2014), math (Marsh, Trautwein et al., 2005), and physical ability (Marsh et al., 2007), indicating its generalizability across different content domains. The REM has also been tested regarding its generalizability across different student characteristics such as age (Guay, Marsh, & Boivin, 2003) and gender (Marsh, Trautwein et al., 2005). Other studies further indicated the cross-cultural generalizability of the REM since reciprocal relations between self-concept and achievement have been found with Australian (Seaton et al., 2014), US-American (Marsh & O'Mara, 2008a), German (Marsh, Trautwein et al., 2005), Hong Kong (Marsh, Hau, & Kong, 2002), and French-Canadian (Guay et al., 2003) students.

To the best of our knowledge, research is lacking regarding the generalizability of the REM across students attending different school tracks. This is surprising because many educational systems implement at least some kind of tracking, particularly in secondary education (Chmielewski, Dumont, & Trautwein, 2013; LeTendre, Hofer, & Shimizu, 2003). Students attending different achievement tracks were found to differ in various respects. For instance, they have been found to reveal different levels of academic achievement, motivation (e.g., interest) and academic self-concept (e.g., Baumert, Watermann, & Schümer, 2003; Becker, Lüdtke, Trautwein, & Baumert, 2006; Hanushek & Wößmann, 2006; Köller & Baumert, 2001). This finding might be partly due to differences in the students' learning environments. Students attending high-achievement tracks were found to receive higher levels of instructional quality and to experience fewer disciplinary problems in the classroom, but also to get lower levels of individual learning support from the teacher compared to students in lower achievement track schools (Klieme & Rakoczy, 2003; Kunter et al., 2005). The present study aims to investigate and compare the REM among students attending different secondary school tracks as this provides an opportunity to test the generalizability of the REM across students experiencing different learning environments and educational opportunities.

Controlling for Covariates

The REM posits achievement to be a major determinant of self-concept and self-concept to be a major determinant of achievement. However, students' socioeconomic status (SES), IQ, and gender are also known to affect students' achievement as well as students' self-concept. Hence, studies aiming to establish reciprocal relations between self-concept and achievement would do well to consider these background variables.

Students from lower SES families demonstrate lower levels of achievement (Bradley & Corwyn, 2002; Sirin, 2005). High levels of student achievement have also often been linked to a high IQ (Frey & Detterman, 2004; Furnham & Monsen, 2009; Spinath, Spinath, & Plomin, 2008). Furthermore, student achievement is associated with gender. Specifically, girls

display higher achievement in verbal subjects (De Fraine, Van Damme, & Onghena, 2007; Van de gaer, Pustjens, Van Damme, & De Munter, 2006) including reading (Lietz, 2006; Mullis, Martin, Kennedy, & Foy, 2007). The findings related to math are less clear and seem to vary contingent upon the achievement indicator. Some studies have found higher scores for boys on standardized math achievement tests (Brunner, Krauss, & Kunter, 2008; Matteucci & Mignani, 2011; Van de gaer et al., 2008), but other studies have reported no or only small gender differences (Hyde, Fennema, & Lamon, 1990; Nowell & Hedges, 1998). When considering school grades in math, girls were found to obtain higher grades than boys (Marsh & Yeung, 1998), although other studies could not find any gender differences (Marsh, Trautwein et al., 2005). Regarding self-concepts, there is consistent evidence for gender differences which are in line with gender stereotypes. Hence, girls show higher levels of verbal self-concept whereas boys display higher levels of math self-concept (Fredricks & Eccles, 2002; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002; Skaalvik & Skaalvik, 2004).

The Present Study

Using a large representative sample of German secondary school students, the present study examines reciprocal relations between math self-concept and math achievement. As an innovative contribution to existing research, the study covers a long time span with five measurement waves enabling a proper investigation of the assumption of developmental equilibrium across German students' mandatory secondary school years. Moreover, this study examines the generalizability of reciprocal effects between math self-concept and math achievement across German students attending three different achievement tracks. The analytic approach starts with a complex full-forward model before turning to more parsimonious models in order to consider and test the adequacy of first-order and higher-order stability and cross-lagged paths. School grades and standardized achievement test scores are simultaneously considered in combined models to account for the two most widely used yet distinctive achievement indicators (Marsh, Trautwein et al., 2005; Marsh et al., 2014). Finally,

gender, IQ, and SES are considered as covariates to control for other important variables influencing students' self-concept and achievement.

Method

Sample

The data analyzed in this study originate from the *Project for the Analysis of Learning and Achievement in Mathematics* (PALMA; Frenzel, Goetz, Lüdtke, Pekrun, & Sutton, 2009; Frenzel, Pekrun, Dicke, & Goetz, 2012; Marsh et al., 2016; Murayama, Pekrun, Lichtenfeld, & vom Hofe, 2013; Murayama, Pekrun, Suzuki, Marsh, & Lichtenfeld, 2015; Pekrun, Lichtenfeld, Marsh, Murayama, & Goetz, in press). PALMA is a large-scale longitudinal study investigating the development of math achievement and its determinants (e.g., math-related motivation, classroom instruction, family variables) during secondary school in Germany. The study was conducted in the German federal state of Bavaria and covers six measurement waves spanning grade levels 5 to 10 with one measurement point each school year. Sampling and the assessments were conducted by the German Data Processing and Research Center (DPC) of the International Association for the Evaluation of Educational Achievement (IEA). The samples represented the typical student population in the German federal state of Bavaria in terms of secondary school achievement tracks and student characteristics such as gender, urban versus rural location, and SES. Participation rate at the school level was 100%. At the first measurement wave in grade 5, the sample comprised 2,070 students (49.6% female, 37.2% low-achievement track students, 27.1% middle-achievement track students, and 35.7% high-achievement track students). The students then had a mean age of 11.75 (SD = 0.68) which is the typical age for fifth grade students in Germany. A number of 42 schools participated in the study and two classes were randomly drawn within each school for final participation. For the subsequent data collections, the study did not only track the students who had already participated in the earlier assessments, but also included students who more recently entered classrooms participating in the PALMA

study and thus had not yet participated in the study (for more details on the sampling procedure, see Pekrun et al., 2007).

In the German federal state of Bavaria, beginning in grade five, students are allocated to either low-achievement (Hauptschule), middle-achievement (Realschule), or high-achievement (Gymnasium) tracks. This decision is mainly based upon students' achievement in the fourth grade of elementary school. Low-achievement track students commonly leave school after the ninth grade with a qualification allowing them to apply for an apprenticeship, middle-achievement track schools end after the tenth grade and students may begin vocational training, and high-achievement track students attend school until the thirteenth grade after which they may enter university. For reasons of including low-achievement track students, the present study focuses on the first five measurement waves covering students' grade levels 5 to 9. The final sample of the present study consists of N = 3,425 [N = 1,714 (50.0% girls), 1,710 (49.9%) boys, 1 (0.01%) indicated no gender] and included all students who participated in at least one of the five assessments. Among this final sample, n = 1,187 students attended the high-achievement track, n = 1,050 the middle-achievement track, and n = 1,188 the low-achievement track. Of the final sample, 38.7% participated in all five measurement waves (i.e., grades 5 to 9), and 9.0%, 18,9%, 15.1%, and 18.3% took part in four, three, two, or one of the assessments, respectively.

The students answered a questionnaire towards the end of each successive school year. All instruments were administered in the students' classrooms by trained external test administrators. Participation in the study was voluntary and parental consent was obtained for every student. Each survey was depersonalized to ensure participant confidentiality.

Measures

Math self-concept. Math self-concept was measured by the PALMA six-item math self-concept scale at each measurement wave. The items (i.e., "In math, I am a talented student"; "It is easy to understand things in math"; "I can solve math problems well"; "It is

easy for me to write math tests"; "It is easy for me to learn something in math"; "If the math teacher asks a question, I can answer it correctly most of the time") were answered using a 5-point Likert scale (1 = *not at all true* to 5 = *completely true*). The scale showed high reliability at each of the five measurement waves both when using the coefficient alpha reliability estimate (α) and when using the scale reliability estimate (ρ; also labelled as composite or instrument reliability) which was explicitly established within the framework of SEM (Raykov, 2009): t1: α = .876, ρ = .879; t2: α = .895, ρ = .896; t3: α = .893, ρ =.894; t4: α = .910, ρ =.910; t5: α = .920, ρ =.921.

Math achievement. Students' math achievement was measured both in terms of school grades and standardized achievement test scores. Math school grades were retrieved from school documents in terms of the report cards students received at the end of each school year, reflecting students' average math accomplishments throughout the school year. In Germany, school grades range from 1 to 6 with 1 depicting the highest and 6 the lowest achievement. For ease of interpretation, the grades were recoded prior to all analyses so that higher grades represent higher achievement.

The Regensburg Mathematical Achievement Test (vom Hofe, Pekrun, Kleine, & Götz, 2002; vom Hofe, Kleine, Blum, & Pekrun, 2005) was used to assess students' standardized math achievement test scores. This test was explicitly designed for the PALMA study serving to measure students' development of math competencies across secondary school years. Thus, different test versions are available for the different grade levels. The conception of this test has substantially and methodologically been linked to the concept underlying the math tests applied in the Programme for International Student Assessment (PISA). Following the construct of mathematical literacy, the test operationalizes math competences as mathematical modeling and problem solving in terms of students' abilities to convert real-world problems into mathematical models, to solve these problems in the context of mathematical models, and to transfer the solutions to reality. Based on this conceptualization, the test targets students'

modeling competencies and algorithmic competencies in arithmetic, algebra, and geometry. Methodologically, the test was constructed within the framework of item response theory (IRT; Wu, Adams, Wilson, & Haldane, 2007). At each measurement point, students worked on one of two different parallel test versions with 60-90 items each, the exact number of items varying across waves. The items were formulated either as multiple-choice or open-ended items. Pre-specified guidelines were given to two trained raters to score the open-ended items. The ratings showed a high level of inter-rater agreement supporting their objectivity [i.e., inter-rater disagreement for the test version A (the parallel version B) was 0.04% (0.13%), thus 0.085% on average]. Depending on the measurement wave, approximately 20 anchor items served to link the two parallel tests within each measurement point and the different tests across the five measurement points. Achievement test scores were scaled using one-parameter logistic IRT applying concurrent calibration (Rasch scaling; Wu et al., 2007) that has been found to have many advantages (e.g., model parsimony, parameter linearity) relative to alternative models (Liu, 2010; Wright, 1999) and that was also used in previous studies utilizing the PALMA math achievement test (Murayama et al., 2013). The reliability of item separation in IRT scaling was 0.99.

Covariates. Students' gender, IQ, and SES measured at t1 served as covariates. Students' IQ was measured using the German adaptation of Thorndike's Cognitive Abilities Test (Kognitiver Fähigkeitstest, KFT 4-12+R; Heller & Perleth, 2000). Reliability of the 25 item scale was $\alpha = .934$. Supporting validity, the IQ test scores were found to be substantially correlated with students' math achievement test scores (t1: $r = .577$, $p < .01$) and to discriminate between students of the different achievement tracks (t1: $F(2, 1987) = 327.778$, $p < .001$), with students attending the high-achievement track displaying the highest mean levels ($M = 110.03$, $SD = 10.36$), followed by the middle-achievement track students ($M = 104.535$, $SD = 9.47$). Students from the low-achievement track displayed the lowest IQ mean level ($M = 96.20$, $SD = 11.91$).

SES was assessed by parental report using the Erikson Goldthorpe Portocarero (EGP) social class scheme (Erikson, Goldthorpe, & Portocarero, 1979). The EGP consists of six ordered categories of parental occupational status wherein higher values represent higher SES.

Statistical Analyses

The analyses were conducted within the SEM framework with *Mplus* 7.5 (Muthén & Muthén, 1998-2015). All models were conducted using the robust maximum likelihood estimator (MLR) which is robust against non-normality of the observed variables (Hox, Maas, & Brinkhuis, 2010; Muthén & Muthén, 1998-2015). The *Mplus* option <type = complex> was used to accommodate the hierarchical nature of the study. Specifically, students were nested within the 42 participating schools and students attending the same school might be more similar to each other than students from different schools, resulting in non-independence of observations. Failure to attend to the hierarchical nature of the data could lead to biased standard errors – a miscalculation that is corrected by this *Mplus* model command (Muthén & Satorra, 1995).[1]

As inherent in any longitudinal study, the data set consisted of missing values on the measured variables which should be appropriately dealt with. In this study, missingness on variables mainly originates from the fact that students entered the study at later measurement waves without having completed the measures at earlier waves. The attrition rate could be kept rather low during the time period covered in the present study, i.e., up to grade level 9 after which the attrition rate is higher due to the low-achievement track students' leaving school. More concretely, among the total sample of the study across all measurement waves, 60.4%, 60.1%, 69.9%, 70.3%, and 73.6% participated in the first, second, third, fourth, and fifth measurement wave respectively. Within each wave, the number of missing values was low for the self-concept measures (t1: 0.68% to 1.74%; t2: 0.29% to 1.70%; t3: 0.50% to 1.67%; t4: 0.46% to 1.78%; t5: 0.63% to 1.31%), the math achievement test scores (0.00% to 0.28%), and school grades in math at t1 to t4 (0.00% to 2.39%). Missing values were handled

by the full information maximum likelihood estimator (FIML) implemented in *Mplus* by default (see Wang & Wang, 2012). FIML has been found to result in trustworthy, unbiased estimates for missing values (Enders, 2010; Graham, 2009) and represents an adequate means of managing missing data in longitudinal study designs (Jeličič, Phelps, & Lerner, 2009). However, *Mplus* excludes cases with missing data on any covariates if only defined as exogenous variables (Muthén & Muthén, 1998-2015). In order to use FIML for missing data on the continuous covariates (i.e., IQ and SES) as well, covariances among these covariates were estimated.

The amount of missing values was high (27.57%) regarding math school grades for the last measurement wave (t5) as the low-achievement track students left school after grade 9. Because of this high amount of missing data, we applied the technique of multiple imputation to handle missing data on students' math grades for t5 which were imputed using the math grades the students had obtained at the previous waves. Five sets of imputed data were created which were used for all analyses involving school grades and combined afterwards (Little & Rubin, 2002) while retaining the FIML approach for estimating missing data on the remaining variables (i.e., self-concept at all waves, test scores at all waves, and grades at t1 to t4).

In all models, one factor for math self-concept was assumed for each measurement wave defined by the six self-concept items answered by the students at the corresponding waves. Correlated uniquenesses for the same self-concept items over time were included in these models to account for the shared method variance due to the repeated use of the same items (Marsh & Hau, 1996). In addition, for each measurement wave, the models included two single-indicator achievement factors defined by students' school grades in math and their math achievement test scores, respectively.

The analyses started with a longitudinal confirmatory factor analyses (CFA) model assuming separate self-concept and achievement factors for each of the five waves. In this model, the self-concept and achievement factors were freely estimated across time with the

same set of items used to define the same number of factors at each measurement wave (configural invariance, Millsap, 2011). The analyses continued with a model of longitudinal measurement invariance by constraining the factor loadings to be of equal size across measurement waves (weak measurement invariance; Millsap, 2011). This model served to test whether the same constructs were measured at the different measurement waves (Widaman, Ferrer, & Conger, 2010).

In order to examine the generalizability of findings across students from different achievement tracks, students' attended school track (high-achievement, middle-achievement, or low-achievement track) was entered as a grouping variable in all models. To test whether the same constructs were measured in all groups of school tracks, we estimated a model assuming invariant factor loadings in the three groups of school tracks. We then stated an even more restrictive model by constraining the factor loadings to be simultaneously invariant across measurement waves and group to ensure that the same constructs were measured at each measurement wave in the three groups of students from different achievement tracks.

In order to test reciprocal effects between self-concept and achievement, we applied cross-lagged panel models and started with a full-forward model. The full-forward model included all possible (i.e., first-order and higher-order) paths for the stability and the cross-lagged relations among the constructs, and additionally assumed the disturbances of constructs to be correlated within each wave (Figure 1; Marsh et al., 1999). Based on this complex model, we evaluated more parsimonious models with fewer paths. In this context, we first assessed the need to include *both* first-order and higher-order paths by comparing the full-forward model with a model only including first-order stability and cross-lagged paths. Afterwards, we tested whether it is advantageous to incorporate first-order and higher-order paths for *both* the stability *and* cross-lagged paths. For this purpose, we estimated a model including first-order *and* higher-order stability paths but *only* first-order cross-lagged paths, and a model with first-order *and* higher-order cross-lagged paths but *only* first-order stability

paths. In the next step, the three covariates (gender, IQ, and SES) measured at t1 were included in the selected model to examine whether the findings were robust when controlling for these variables.

So far, the cross-lagged paths depicting the longitudinal effects between self-concept (achievement) and achievement (self-concept) were freely estimated across time. In order to test the assumption of developmental equilibrium, invariance constraints were imposed on these paths. The cross-lagged paths from one variable in one wave to another variable in a subsequent wave were assumed to be of equal size across all measurement points (e.g., achievement t1 → self-concept t2 = achievement t2 → self-concept t3 = achievement t3 → self-concept t4 = achievement t4 → self-concept t5). In a subsequent model, invariance constraints on the stability paths were included in terms that all stability estimates for one construct were restricted to be of the same size.

For presenting the relations between self-concept and achievement, we report the StdYX standardized coefficients provided by *Mplus* (Muthén & Muthén, 1998-2015), except for the effects of gender. The StdYX solution is based on the variances of both the continuous independent latent variable (X; e.g., math achievement) and the outcome (dependent) variable (Y; e.g., math self-concept) and interpreted as the mean change of Y in standard deviation units of Y for one standard deviation change in X. For gender as a binary variable, a proper standardized estimate results from standardizing the dependent variable Y only, which is provided by the StdY solution in Mplus and depicts the change in Y (e.g., math self-concept) in Y standard deviation units when X (i.e., gender) changes from zero to one.

To assess the fit of the models, we rely on a range of commonly applied descriptive goodness-of-fit indices (Marsh, Hau, & Grayson, 2005). We thus report the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). Values above .90 and .95

for the CFI and TLI represent acceptable and good fit, respectively (Hu & Bentler, 1999). With respect to the RMSEA, values near .05 imply "close fit", values near .08 indicate "fair fit", and values above .10 represent "poor fit" (Browne & Cudeck, 1993). SRMR values below .05 are interpreted as a good model fit (Diamantopoulos & Siguaw, 2000) but Hu and Bentler (1999) proposed a less strict cut-off criterion of .08.

The invariance models can be conceptualized as nested models which only differ from each other in the parameters which were set invariant across time and group. To evaluate invariance, we follow the guidelines proposed by Cheung and Rensvold (2002) and Chen (2007), according to which invariance should not be rejected if $\Delta CFI \leq -.01$ and $\Delta RMSEA \leq +.015$ for the more restrictive as compared with the less restrictive model. The cut-off values suggested for the evaluation of latent models including invariance models, should be considered as rough guidelines instead of golden rules. Researchers are rather advised to take all available information into account for an ultimate judgement of latent models, including parameter estimates, statistical conformity, and theoretical adequacy of the model besides the fit indices (Marsh, Hau, & Wen, 2004).

Results

The analyses reported herein are based on models including both school grades and standardized achievement test scores as achievement indicators, but the same series of models was also estimated using school grades and achievement test scores separately. The results from these models (i.e., using either school grades or achievement test scores) are reported in the Online Supplements (Models S1 to S7) and are fully consistent with the findings for the models combining school grades and achievement test scores.

The series of analyses started with a longitudinal measurement model assuming separate factors for math self-concept, math achievement test scores, and math grades at each measurement wave (Model 1 in Table 1). The fit of this model was excellent and largely maintained when imposing invariance of factor loadings across time (Model 2) indicating that

the same constructs were measured at each wave. Model 3 included students' achievement tracks as a grouping factor. The fit indices of this model still indicated good fit which was mostly retained when including invariant factor loadings across school tracks (Model 4). This finding indicates that the same constructs were measured in the three groups of students attending different achievement tracks. The fit indices also remained in the area of good model fit when assuming an even more restrictive model with invariant factor loadings both across measurement waves and school tracks (Model 5). This finding allowed for meaningful longitudinal analyses and comparisons across school tracks.

Model 6 is the full-forward model for describing the longitudinal relations between math self-concept, math achievement test scores, and math grades for students from different achievement tracks. This model incorporates all possible first-order and higher-order paths and only replaces the correlations among constructs by path coefficients. Therefore, it is statistically equivalent to Model 5 and results in the same fit. The path coefficients of Model 6 (Table S6 of the Online Supplements) suggest that multicollinearity might be at play since some coefficients for self-concept–achievement relations had implausible negative and small coefficients (Marsh, Dowson, Pietsch, & Walker, 2004). Hence, the full-forward model (Model 6) was compared to a less complex model (Model 7) which only included first-order stability and cross-lagged paths. The fit of Model 7 declined substantially compared to the full-forward Model 6 (ΔCFI = -.017; ΔRMSEA = +.008) suggesting that the inclusion of any higher-order paths seems to be warranted. In Model 8, we integrated first-order and higher-order stability paths, but only first-order cross-lagged paths. Here, the fit was highly similar to the fit resulting from the full-forward Model 6. However, when assuming first-order and higher-order cross-lagged paths along with first-order stability paths (Model 9), the model fit substantially declined compared to the full-forward Model 6 (ΔCFI = -.015; ΔRMSEA = +.008). The findings thus argue for the integration of higher-order stability paths, but not for the inclusion of higher-order cross-lagged paths leading us to retain Model 8.

The findings resulting from Model 8 were retained when adding the effects of the three covariates, i.e., gender, IQ, and SES, which were assumed to be related to students' math self-concept and math grades at t1 in Model 10. The covariates demonstrated significant effects on math self-concept, math grades, and math achievement test scores (Table S7 of the Online Supplements).

Model 11 then served to test the assumption of developmental equilibrium. For this purpose, the first-order cross-lagged paths between former self-concept and later achievement (both school grades and test scores) and the first-order cross-lagged paths between former achievement and later self-concept were respectively set to be equal across all time lags and groups of school tracks (Model 11). The various goodness-of-fit indices of Model 11 still indicated good model fit and did not demonstrate a substantial decline (ΔCFI = -.003; ΔRMSEA = +.001) relative to the precedent less restrictive Model 10 so that the assumption of developmental equilibrium invariant across school tracks could be supported.

Model 12 extends Model 11 in terms of invariance constraints on the stability coefficients. More concretely, Model 12 assumed the first-order and higher-order stability paths of all three constructs (i.e., math self-concept, math grades, and math achievement test scores) to be of equal size across measurement waves for the three school tracks. The fit of this restrictive model remains comparable to the fit of Model 11 arguing for its tenability.

Finally, Model 13 included the invariance of covariate paths meaning that the effects of the three covariates (gender, IQ, and SES) on math self-concept, math grades, and math achievement test scores at t1 are of similar size across groups of school tracks. The model fit remained stable supporting the appropriateness of this highly restrictive model. When considering the resulting standardized path coefficients of this final model (Table 2), it is obvious that math self-concept and math achievement (both school grades and test scores) were best predicted by their former levels given the substantial positive coefficients for the first-order stability paths. However, the significant higher-order stability estimates imply that

the stability of constructs does not only address consecutive waves but goes further. Second, the findings demonstrated reciprocal relations between math self-concept and math achievement both in terms of school grades in math and math achievement test scores. The cross-lagged paths leading from math self-concept to math achievement and those leading from math achievement to math self-concept were positive and significant across all measurement waves in the three groups of students irrespective of whether achievement was operationalized by school grades or achievement test scores. Considering the effects of the covariates, gender was found to have a significant effect on math self-concept with boys displaying higher levels. Moreover, students with a higher IQ demonstrated higher levels of math self-concept whereas students' SES was found to be unrelated to math self-concept. Boys and girls were found to obtain similar school grades in math, while students of higher SES and higher IQ were found to earn higher math grades. Regarding math achievement test scores, boys, students with higher IQ levels, and students of higher SES were found to demonstrate higher test scores. All these results were invariant across students from different school tracks indicating a high level of generalizability.[2]

Discussion

Even though the REM for self-concept–achievement relations has been extensively studied (Huang, 2011; Marsh & Craven, 2006), our study extends previous research and provides some of the strongest evidence for the REM so far. In essence, the present study revealed reciprocal relations between math self-concept and math achievement that were robust and generalizable in various ways.

The robustness and generalizability of the REM first becomes evident in terms of generalizability across time. The results supported the assumption of developmental equilibrium since both types of cross-lagged paths were found to be of similar sizes across the extensive time span of this study including five waves. Thus, within skill development effects and within self-enhancement effects, the effects seem to be invariant at least throughout the

years of German students' mandatory secondary schooling. Interventions should therefore pursue a dual approach targeting the enhancement of both students' self-concept and achievement (Craven, Marsh, & Burnett, 2003; O'Mara, Marsh, Craven, & Debus, 2006).

The generalizability of reciprocal effects between math self-concept and math achievement, including developmental equilibrium, was further supported by considering students from different achievement tracks of the German secondary school system. Previous studies demonstrated the generalizability of the REM across different student characteristics such as age (Guay et al., 2003), gender (Marsh, Trautwein et al., 2005), and culture (Chen, Yeh, Hwang, & Lin, 2013; Marsh et al., 2002). This study broadens this line of research by demonstrating generalizability of the REM across students attending different achievement tracks who might experience different learning environments (Klieme & Rakoczy, 2003; Kunter et al., 2005). Practically, this finding implies that the above mentioned dual intervention approach for enhancing students' self-concept and achievement is beneficial for a wide range of students.

The robustness of reciprocal effects between math self-concept and math achievement is further reflected by the fact that the resulting pattern of relations persists when including students' gender, SES, and IQ as covariates. Hence, the assumptions of the REM even remain in place when controlling for other major determinants of students' math self-concept and math achievement.

Finally, the generalizability and robustness of the REM as demonstrated in our study addresses achievement indicators. Given that school grades and standardized achievement test scores each have their advantages and disadvantages, research benefits from including both achievement indicators in empirical studies (Marsh et al., 2014). Previous studies have indicated that the REM holds when considering school grades and achievement test scores separately (e.g., Möller et al., 2014), but only one study so far has integrated school grades and achievement test scores in a combined model (Marsh, Trautwein et al., 2005). Given that

the latter study only included two measurement waves, the present study spanning five waves is a considerable enrichment. In fact, it supported reciprocal self-concept–achievement relations for both school grades and achievement test scores in math in combined models across five measurement waves, additionally demonstrating developmental equilibrium and invariance across school tracks.

Besides providing evidence of the strong robustness and generalizability of the REM, the present study contributes to methodological approaches to the REM. It illustrates the advantage of starting with a full-forward model in which all possible paths are estimated and which thus includes first-order and higher-order stability and cross-lagged paths. As exemplified in this study, such a complex model can serve as the starting point for deriving and empirically testing more parsimonious and less complex models. Accordingly, we could demonstrate that there was no additional benefit of including both first-order and higher-order cross-lagged paths, but the incorporation of both first-order and higher-order stability paths contributed to significantly better models. Substantively, this leads to the conclusion that self-concept and achievement are of high stability that lasts longer than across two immediately adjacent measurement waves (Marsh & O'Mara, 2008a).

A further methodological advice that can be derived from this study targets the need to include covariates which also relate to students' self-concept and achievement. Consistent with previous studies (Fredricks & Eccles, 2002; Jacobs et al., 2002; Watt, 2004), boys were found to display higher levels of math self-concept. Boys were also found to perform better on the math achievement test, but boys and girls obtained similar school grades in math. This finding corresponds to previous studies indicating that gender differences in math achievement might vary contingent upon the achievement indicator used, and that despite boys' consistent superior levels of math self-concept, gender differences in math achievement are less consistent (Hyde et al., 1990; Leahey & Guo, 2001; Lindberg, Hyde, Petersen, & Linn, 2010). Future research should also consider the situation and environmental

circumstances in which students' math achievement is assessed. For example, according to the stereotype threat paradigm (Nguyen & Ryan, 2008; Steele, 1997), females' math achievement might be lower when the stereotype that girls are poorer in math than boys is activated, as compared to test situations when this gender stereotype is not prevalent.

This study followed a traditional cross-lagged modeling approach to investigate reciprocal effects between self-concept and achievement which delivers easily interpretable results and facilitates comparability across numerous previous studies on the REM that also utilized this approach (e.g., Guay et al., 2003; Marsh et al., 2007; Marsh, Trautwein et al., 2005; Möller et al., 2011, 2014; Niepel et al., 2014; Seaton et al., 2014). However, the standard cross-lagged panel modeling approach has recently been criticized (Hamaker, Kuiper, & Grasman, 2015) mainly because of the lacking separation between the within-person level and the between-person level which would enable consideration of trait-like (i.e., stable) individual differences. Hence, it might be worthwhile to consider the application of proposed alternative models to the REM in the future. Alternative models should also be taken into account for the math achievement test. In this study, achievement test scores were scaled based on a one-parameter logistic IRT model, but alternative estimations including two-parameter models could be used in order to test the generalizability of findings. Indeed, there has been a long debate on the advantages of one-parameter relative to two-parameter IRT models (Bergan, 2013), and this debate might benefit from the application and comparison of both approaches to the same study and research question.

In light of the consistently demonstrated separation between math and verbal self-concepts (Möller, Pohlmann, Köller, & Marsh, 2009), further investigations are needed to generalize the present findings to the verbal domain. In this context, it might not only be worthwhile to study the math and verbal domains separately but to also investigate different domains simultaneously (Marsh et al., 2014). Furthermore, since only self-concept operationalized as students' perceptions of competence was considered, further variables for

students' self-perceptions should be addressed such as affect self-perceptions (Arens et al., 2011; Marsh et al., 2013). Finally, beyond achievement, it might be worthwhile to take a broader range of outcome variables such as goal orientations (Seaton et al., 2014), effort (Trautwein, Lüdtke, Schnyder, & Niggli, 2006) or emotions (Pekrun, 2006) into account.

Given that this study investigated students attending the secondary school years, further long-term studies should focus on preschool or elementary school years as the present findings cannot be generalized to younger students. It can be assumed that secondary school students have established a self-concept that is sufficiently stable to impact on later achievement (Wigfield & Karpathian, 1991). This might, however, not yet be the case in preschool and elementary school years when self-enhancement effects might predominate (Arens et al., 2016; Chapman & Tunmer, 1997; Chen et al., 2013; Helmke & van Aken, 1995). Studies covering a wide time frame would be worthwhile to gain insight into the onset of reciprocal relations and developmental equilibrium in these relations.

In sum, the present study provides relevant insights into research on reciprocal relations between self-concept and achievement. In essence, the assumption of developmental equilibrium could be supported, substantiating the robustness of relations from self-concept to achievement and from achievement to self-concept across a long time interval of five waves. The robustness of reciprocal effects was further substantiated by the generalizability of the findings across achievement indicators and school tracks even when controlling for important covariates. As such, although the REM has been well established and become an inherent characteristic of the self-concept construct (Marsh & Craven, 2006), the present study has pointed out remaining important questions on reciprocal self-concept–achievement relations, delivered answers to these questions, and once again underscored the generalizability and robustness of the REM at least for secondary school students.

Footnotes

[1] It was not possible to use students' classes as a clustering variable because the composition of students' classes changed across time.

[2] To check the robustness of the findings, the same series of analyses was conducted using sampling weights. The results are reported in the Online Supplements (Tables S8 to S15).The results are the same as those presented herein when not using sampling weights, documenting the robustness of the analysis.

References

Arens, A.K., Marsh, H.W., Craven, R.G., Yeung, A.S., Randhawa, E., & Hasselhorn, M. (2016). Math self-concept in preschool children: Structure, achievement relations, and generalizability across gender. *Early Childhood Research Quarterly, 36*, 391-403.

Arens, A.K., Yeung, A.S., Craven, R.G., & Hasselhorn, M. (2011). The twofold multidimensionality of academic self-concept: Domain specificity and separation between competence and affect components. *Journal of Educational Psychology, 103*, 970-981.

Baumert, J., Watermann, R., & Schümer, G. (2003). Disparitäten der Bildungsbeteiligung und des Kompetenzerwerbs. Ein institutionelles und individuelles Mediationsmodell [Disparity of educational participation and acquisition of skills. An institutional and individual model of mediation]. *Zeitschrift für Erziehungswissenschaft, 6*, 46-71.

Becker, M., Lüdtke, O., Trautwein, U., & Baumert, J. (2006). Leistungszuwachs in Mathematik: Evidenz für einen Schereneffekt im mehrgliedrigen Schulsystem [Achievement gains in mathematics: Evidence for differential trajectories in a tracked school system]? *Zeitschrift für Pädagogische Psychologie, 20*, 233-242.

Bergan, J.R. (2013). *Rasch versus Birnbaum: New arguments in an old debate*.Tucson, Arizona, USA. Assessment Technology, Incorporated.

Brookhart, S. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30*, 123-142.

Bradley, R.H., & Corwyn, R.F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*, 371-399.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Brunner, M., Krauss, S., & Kunter, M. (2008). Gender differences in mathematics: Does the story need to be rewritten? *Intelligence, 36,* 403-421.

Caslyn, R., & Kenny, D. (1977). Self-concept of ability and perceived evaluation by others: Cause or effect of academic achievement? *Journal of Educational Psychology, 69,* 136-145.

Chapman, J.W., & Tunmer, W.E. (1997). A longitudinal study of beginning reading achievement and reading self-concept. *British Journal of Educational Psychology, 67,* 279-291.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233-255.

Chmielewski, A.K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type: An international comparison of students' mathematics self-concept. *American Educational Research Journal, 50*, 925-957.

Chen, F.F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling, 14,* 464-504.

Chen, S-K., Yeh, Y-C., Hwang, F-M., & Lin, S.S.J. (2013). The relationship between academic self-concept and achievement: A multicohort–multioccasion study. *Learning and Individual Differences, 23,* 172-178.

Craven, R.G., Marsh, H.W., & Burnett, P. (2003). Cracking the self-concept enhancement conundrum. A call and blueprint for the next generation of self-concept enhancement research. In H.W. Marsh, R.G. Craven, & D.M. McInerney (Eds.), *International advances in self research: Speaking to the future* (pp. 91-126). Grennwich, CT: Information Age.

Curran, P.J., & Bollen, K.A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L.M. Collins, & A.G. Sayer (Eds.), *New methods for the*

analysis of change (pp. 105-136). Washington, DC: American Psychological Association.

De Fraine, B., Van Damme, J., & Onghena, P. (2007). A longitudinal analysis of gender differences in academic self–concept and language achievement: A multivariate multilevel latent growth approach. *Contemporary Educational Psychology, 32*, 132-150.

Diamantopoulos, A., & Siguaw, J.A. (2000). *Introducing LISREL*. London: Sage Publications.

Erikson, R., Goldthorpe, J.H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies. *Birtish Journal of Sociology, 30,* 415-441.

Enders, C.K. (2010). *Applied missing data analysis*. New York: Guilford.

Frenzel, A.C., Goetz, T., Lüdtke, O., Pekrun, R., & Sutton, R. (2009). Emotional transmission in the classroom: Exploring the relationship between teacher and student enjoyment. *Journal of Educational Psychology, 101,* 705-716.

Frenzel, A.C., Pekrun, R., Dicke, A.L., & Goetz, T. (2012). Beyond quantitative decline: Conceptual shifts in adolescents' development of interest in mathematics. *Developmental Psychology, 48,* 1069-1082.

Fredricks, J., & Eccles, J.S. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-domains. *Developmental Psychology, 38,* 519-533.

Frey, M.C., & Detterman, D.K. (2004). Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science, 15,* 373-378.

Fredricks, J.A., & Eccles, J.S. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology, 38,* 519-533.

Furnham, A., & Monsen, J. (2009). Personality traits and intelligence predict academic school grades. *Learning and Individual Differences, 19,* 28-33.

Graham, J.W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549-576.

Guay, F., Marsh, H.W., & Boivin, M. (2003). Academic self-concept and academic achievement: A developmental perspective on their causal ordering. *Journal of Educational Psychology, 95,* 124-136.

Hamaker, E.L., Kuiper, R.M., & Grasman, R.P.P.P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20,* 102-116.

Hanushek, E., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal, 116,* 63-76.

Heller, K.A., & Perleth. C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Beltz Test GmbH, Göttingen

Helmke, A., & van Aken, M.A.G. (1995). The causal ordering of academic achievement and self-concept of ability during elementary school: A longitudinal study. *Journal of Educational Psychology, 87,* 624-637.

Hox, J.J., Maas, C.J.M., & Brinkhuis, M.J.S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64,* 157-170.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology, 49,* 505-528.

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

Jacobs, J.E., Lanza, S., Osgood, W.D., Eccles, J.S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development, 73,* 509-527.

Jeličič, H., Phelps, E., & Lerner, R.M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology, 45,* 1195-1199.

Kenny, D.A. (1975). Cross-lagged panel correlation: A test for spuriousness. *Psychological Bulletin, 82,* 887-903.

Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In J. Baumert, C. Artelt, E. Klieme, J. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, & K.-J. Tillmann (Eds.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 334-359). Opladen: Leske + Budrich.

Köller, O., & Baumert, J. (2001). Leistungsgruppierungen in der Sekundarstufe I. Ihre Konsequenzen für die Mathematikleistung und das mathematische Selbstkonzept der Begabung [Ability grouping at secondary level 1: Consequences for mathematics achievement and the self-concept of mathematical ability]. *Zeitschrift für Pädagogische Psychologie, 15,* 99-110.

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W., Jordan, A., & Neubrand, M. (2005). Der Mathematikunterricht der PISA-Schülerinnen und -Schüler. Schulformunterschiede in der Unterrichtsqualität [Quality of mathematics instruction across school types: Findings from PISA 2003]. *Zeitschrift für Erziehungswissenschaft, 4,* 502-520.

Leahey, E., & Guo, G. (2001). Gender differences in mathematical trajectories. *Social Forces, 80,* 713-732.

LeTendre, G.K., Hofer, B.K., & Shimizu, H. (2003). What is tracking? Cultural expectations in the United States, Germany, and Japan. *American Educational Research Journal, 40,* 43-89.

Lietz, P. (2006). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation, 32*, 317-344.

Lindberg, S.M., Hyde, J.S., Petersen, J.L., & Linn, M.C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136,* 1123-1135.

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data*. New York: John Wiley.

Liu, X. (2010). *Using and developing measurement instruments in science education: A rasch modeling approach.* Charlotte, NC: Information Age Publishing.

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

Marsh, H.W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology*. Leicester, UK: British Psychological Society.

Marsh, H.W., Abduljabbar, A.S., Abu-Hilal, M., Morin, A.J.S., Abdelfattah, F., Leung, K.C., Xu, M.K., Nagengast, B., & Parker, P. (2013). Factor structure, discriminant and convergent validity of TIMSS math and science motivation measures: A comparison of USA and Saudi Arabia. *Journal of Educational Psychology, 105*, 108-128.

Marsh, H.W., Byrne, B.M., & Yeung, A.S. (1999). Causal ordering of academic self-concept and achievement: Reanalysis of a pioneering study and revised recommendations. *Educational Psychologist, 34*, 155-167.

Marsh, H.W., & Craven, R.G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective. Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1*, 133-163.

Marsh, H.W., Dowson, M., Pietsch, J., & Walker, R. (2004). Why multicollinearity matters: A reexamination of relations between self-efficacy, self-concept, and achievement. *Journal of Educational Psychology, 96,* 518-522.

Marsh, H.W., Gerlach, E., Trautwein, U., Lüdtke, O., & Brettschneider, W.-D. (2007). Longitudinal study of preadolescent sport self-concept and performance: Reciprocal effects and causal ordering. *Child Development, 78,* 1640-1656.

Marsh, H.W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64*, 364-390.

Marsh, H.W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares, & J. McArdle (Eds.), *Contemporary psychometrics. A Festschrift for Roderick P. McDonald.* Mahwah NJ: Erlbaum.

Marsh, H.W., Hau, K.-T., & Kong, K.W. (2002). Multilevel causal ordering or academic self-concept and achievement: Influence of language of instruction (English vs. Chinese) for Hong Kong students. *American Educational Research Journal, 39,* 727-763.

Marsh, H.W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to cutoff values for fit indexes and dangers in overgeneralizing Hu & Bentler's (1999). *Structural Equation Modeling, 11,* 320-341.

Marsh, H.W., Kuyper, H., Seaton, M., Parker, P.D., Morin, A.J.S., Möller, J., & Abduljabbar, A.S. (2014). Dimensional comparison theory: An extension of the internal/external frame of reference effect on academic self-concept formation. *Contemporary Educational Psychology,* 39, 326-341.

Marsh, H.W., & O'Mara, A. (2008a). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept. *Personality and Social Psychology Bulletin,* 34, 542-552.

Marsh, H.W. & O'Mara, A.J. (2008b). Self-concept is as multidisciplinary as it is multidimensional. In H.W. Marsh, R.G. Craven, & D.M. McInerney (Eds.), *Self-processes, learning, and enabling human potential. Dynamic new approaches* (pp. 87-115). Charlotte, NC: Information Age.

Marsh, H.W., Pekrun, R., Parker, P.D., Murayama, K., Guo, J., Dicke, T., & Lichtenfeld, S. (2016). Long-term positive effects of repeating a year in school: Six-year longitudinal study of self-beliefs, anxiety, social relations, school grades, and test scores. *Journal of Educational Psychology.* Advance online publication. doi:10.1037/edu0000144

Marsh, H.W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades and standardized test scores: Reciprocal effects model of causal ordering. *Child Development, 76,* 397-416.

Marsh, H.W., & Yeung, A.S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal, 35,* 705-738.

Marshall, S., Parker, P.D., Ciarrochi, J., & Heaven, P.C.L. (2014). Is self-esteem a cause or consequence of social support? A 4-year longitudinal study, *Child Development, 85* 1275-1291.

Matteucci, M., & Mignani, S. (2011). Gender differences in performance in mathematics at the end of lower secondary school in Italy. *Learning and Individual Differences, 21,* 543-548.

McMillan, J.H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research,* 95, 203-213.

Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Möller, J., Pohlmann, B., Köller, O., & Marsh, H.W. (2009). Meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79*, 1129-1167.

Möller, J., Retelsdorf, J., Köller, O., & Marsh, H.W. (2011). The reciprocal internal/external frame of reference model: An integration of models of relations between academic achievement and self-concept. *American Educational Research Journal, 48*, 1315-1346.

Möller, J., Zimmermann, F., & Köller, O. (2014). The reciprocal internal/external frame of reference model using grades and test scores. *British Journal of Educational Psychology, 84*, 591-611.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries.* Chestnut Hill, MA: Boston College.

Murayama, K., Pekrun, R., Lichtenfeld, S., & vom Hofe, R. (2013). Predicting long-term growth in students' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development, 84*, 1475-1490.

Murayama, K., Pekrun, R., Suzuki, M., Marsh, H.W., & Lichtenfeld, S. (2015). Don't aim too high for your kids: Parental over-aspiration undermines students' learning in mathematics. *Journal of Personality and Social Psychology.* Online first: doi: 10.1037/pspp0000079

Muthén, L.K., & Muthén, B.O. (1998-2015). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P.V. Marsden (Ed.), *Sociological Methodology* (pp. 267-316). Washington, DC: American Sociological Association.

Nguyen, H.D., & Ryan, A.M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93,* 1314-1334.

Niepel, C., Brunner, M., & Preckel, F. (2014). The longitudinal interplay of students' academic self-concepts and achievements within and across domains: Replicating and extending the reciprocal internal/external frame of reference model. *Journal of Educational Psychology, 106,* 1170-1191.

Nowell, A., & Hedges, L.V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles, 39,* 21-43.

O'Mara, A.J., Marsh, H.W., Craven, R.G., & Debus, R.L. (2006). Do self-concept interventions make a difference? A synergistic blend of construct validation and meta-analysis. *Educational Psychologist, 41,* 181-206.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18,* 315-341.

Pekrun, R., vom Hofe, R., Blum, W., Frenzel, A.C., Goetz, T., & Wartha, S. (2007). Development of mathematical competencies in adolescence: The PALMA longitudinal study. In M. Prenzel (Ed.), *Studies on the educational quality of schools* (pp. 17-37). Münster, Germany: Waxmann.

Pekrun, R., Lichtenfeld, S., Marsh, H.W., Murayama, K., & Goetz, T. (in press). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child Development.*

Raykov, T. (2009). Evaluation of scale reliability for unidimensional measures using latent variable modeling. *Measurement and Evaluation in Counseling and Development, 42,* 223-232.

Retelsdorf, J., Köller, O., & Möller, J. (2014). Reading achievement and reading self-concept. Testing the reciprocal effects model. *Learning and Instruction, 29*, 21-30.

Seaton, M., Parker, P., Marsh, H.W., Craven, R.G., & Yeung, A.S. (2014). The reciprocal relations between self-concept, motivation and achievement: Juxtaposing academic self-concept and achievement goal orientations for mathematics success. *Educational Psychology, 34*, 49-72.

Shavelson, R.J., Hubner, J.J. & Stanton, G.C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46*, 407-441.

Sirin, S.R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*, 417-453.

Skaalvik, S., & Skaalvik, E.M. (2004). Gender differences in math and verbal self-concept, performance expectations, and motivation. *Sex Roles, 50*, 241-252.

Spinath, F.M., Spinath, B., & Plomin, R. (2008). The nature and nurture of intelligence and motivation in the origins of sex differences in elementary school achievement. *European Journal of Personality, 22*, 211-229.

Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology, 98*, 438-456.

Steele, C.M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613-629.

Valentine, J.C., DuBois, D.L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39*, 111-131.

Van de gaer, E., Pustjens, H., Van Damme, J., & De Munter, A. (2008). Mathematics participation and mathematics achievement across secondary school: The role of gender. *Sex Roles, 59*, 568-585.

Widaman, K.F., Ferrer, E., & Conger, R.D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4,* 10-18.

Wu, M.L., Adams, R.J., Wilson, M.R., & Haldane, S.A. (2007). *ACERConQuest Version 2: Generalised item response modelling software.* Camberwell: Australian Council for Educational Research.

Vom Hofe, R., Pekrun, R., Kleine, M., & Götz, T. (2002). Projekt zur Analyse der Leistungsentwicklung in Mathematik (PALMA): Konstruktion des Regensburger Mathematikleistungstests für 5.-10. Klassen [Project for the Analysis of Learning and Achievement in Mathematics (PALMA): Development of the Regensburg Mathematics Achievement Test for grades 5 to 10]. *Zeitschrift für Pädagogik, 45,* 83-100.

Vom Hofe, R., Kleine, M., Blum, W., & Pekrun, R. (2005). On the role of "Grundvorstellungen" for the development of mathematical literacy. First results of the longitudinal study PALMA. *Mediterranean Journal for Research in Mathematics Education, 4,* 67-84.

Wang , J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Chichester, UK: John Wiley & Sons.

Watt, H.M.G. (2004). Development of adolescents' self-perceptions, values, and task perceptions according to gender and domain in 7th- through 11th-grade Australian students. *Child Development, 75,* 1556-1574.

Wigfield, A., & Karpathian, M. (1991). Who am I and what can I do? Children's self-concepts and motivation in achievement solutions. *Educational Psychologist, 26,* 223-261.

Wright, B.D. (1999). Fundamental measurement for psychology. In S.E. Embretson, & S.I. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know*. Hillsdale, NJ: Erbaum.

Zimmermann, F., Schütte, K., Taskinen, P., & Köller, O. (2013). Reciprocal effects between adolescent externalizing problems and measures of achievement. *Journal of Educational Psychology, 105*, 747-761.

Table 1

Goodness-of-fit Indices of the Models including School Grades and Test Scores as Achievement Indicators

	χ^2	df	CFI	TLI	RMSEA	SRMR	
1	1121.292	585	.990	.987	.016	.023	CFA longitudinal measurement model
2	1206.239	605	.989	.986	.017	.027	CFA longitudinal measurement model; invariance of factor loadings across time
3	2634.556	1755	.984	.978	.021	.029	Multi-group CFA longitudinal measurement model
4	2695.456	1805	.984	.979	.021	.030	Multi-group CFA longitudinal measurement model; invariance of factor loadings across school tracks
5	2780.448	1824	.982	.977	.021	.034	Multi-group CFA longitudinal measurement model; invariance of factor loadings across school tracks and time
6	2780.446	1824	.982	.977	.021	.034	Full-forward cross-lagged panel model; all paths freely estimated across school tracks and time
7	3857.313	1986	.965	.959	.029	.058	Cross-lagged panel model; only first-order stability and cross-lagged paths; all paths freely estimated across school tracks and time
8	2951.810	1932	.981	.977	.022	.036	Cross-lagged panel model; first-order and higher-order stability paths, but only first-order cross-lagged paths; all paths freely estimated across school tracks and time
9	3641.541	1878	.967	.959	.029	.052	Cross-lagged panel model; first-order and higher-order cross-lagged paths, but only first-order stability paths; all paths freely estimated across school tracks and time
10	3540.821	2265	.977	.972	.022	.037	Cross-lagged panel model; first-order and higher-order stability paths, but only first-order cross-lagged paths; inclusion of control variables; all paths freely estimated across school tracks and time
11	3769.556	2331	.974	.970	.023	.042	Cross-lagged panel model; first-order and higher-order stability paths, but only first-order cross-lagged paths; inclusion of control variables; invariance of cross-lagged paths across school tracks and time (developmental equilibrium)
12	3996.579	2409	.971	.968	.024	.052	Cross-lagged panel model; first-order and higher-order stability paths, but only first-order cross-lagged paths; inclusion of control variables; invariance of cross-lagged paths and (first-order and higher-order) stability paths across school tracks and time
13	4031.404	2427	.971	.968	.024	.052	Cross-lagged panel model; first-order and higher-order stability paths, but only first-order cross-lagged paths; inclusion of control variables; invariance of cross-lagged paths, (first-order and higher-order) stability, and covariates paths across school tracks and time

Note. All models are estimated with the Robust Maximum Likelihood (MLR) estimator; all χ^2 are significant ($p < .05$).
CFA = confirmatory factor analyses; CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

Table 2
Standardized Paths Coefficients of Model 13

	High-achievement track	Middle-achievement track	Low-achievement track	High-achievement track	Middle-achievement track	Low-achievement track	High-achievement track	Middle-achievement track	Low-achievement track
Stability									
		Math self-concept			**Math grades**			**Math test scores**	
t1-t2	.506*	.537*	.538*	.466*	.460*	.473*	.525*	.499*	.492*
t1-t3	.140*	.147*	.150*	.129*	.132*	.132*	.201*	.201*	.196*
t1-t4	.067*	.071*	.069*	.083*	.082*	.086*	.105*	.103*	.097*
t1-t5	.035	.036	.037	.022	.021	.022	.044*	.045	.041*
t2-t3	.522*	.518*	.525*	.416*	.430*	.417*	.445*	.468*	.462*
t2-t4	.143*	.143*	.139*	.122*	.123*	.124*	.182*	.187*	.179*
t2-t5	.068*	.067*	.068*	.080*	.078*	.080*	.089*	.095*	.088*
t3-t4	.518*	.522*	.500*	.439*	.428*	.447*	.475*	.466*	.451*
t3-t5	.143*	.141*	.140*	.131*	.124*	.131*	.182*	.184*	.173*
t4-t5	.520*	.509*	.530*	.448*	.434*	.441*	.446*	.460*	.446*
Cross-lagged paths									
	Math grades → math self-concept			**Math self-concept → math grades**			**Math self-concept → math test scores**		
t1-t2	.088*	.090*	.093*	.047*	.049*	.049*	.056*	.054*	.056*
t2-t3	.083*	.085*	.086*	.046*	.046*	.045*	.056*	.054*	.056*
t3-t4	.087*	.087*	.087*	.046*	.046*	.046*	.059*	.056*	.055*
t4-t5	.086*	.087*	.087*	.048*	.045*	.048*	.058*	.057*	.056*
	Math test-scores → math self-concept			**Math test scores → math grades**			**Math grades → math test scores**		
t1-t2	.089*	.093*	.088*	.164*	.166*	.157*	.132*	.123*	.132*
t2-t3	.077*	.084*	.082*	.134*	.148*	.137*	.122*	.120*	.124*
t3-t4	.079*	.082*	.077*	.137*	.140*	.138*	.134*	.126*	.129*
t4-t5	.075*	.078*	.079*	.134*	.135*	.139*	.131*	.131*	.125*
Covariates									
	Effects on math self-concept (t1)			**Effects on math grades (t1)**			**Effects on math test scores (t1)**		
Gender	.617*	.599*	.580*	.057	.058	.054	.367*	.362*	.370*
IQ	.207*	.190*	.200*	.350*	.335*	.341*	.421*	.392*	.435*
SES	.016	.017	.016	.087*	.091*	.083*	.068*	.070*	.069*

Note. Gender is coded 0=female, 1=male. Coefficients based on StdYX standardization within *Mplus* (i.e., standardization of independent and dependent variables) are provided for all effects except effects involving gender. For effects involving gender, coefficients based on StdY standardization are provided.
* $p < .05$.

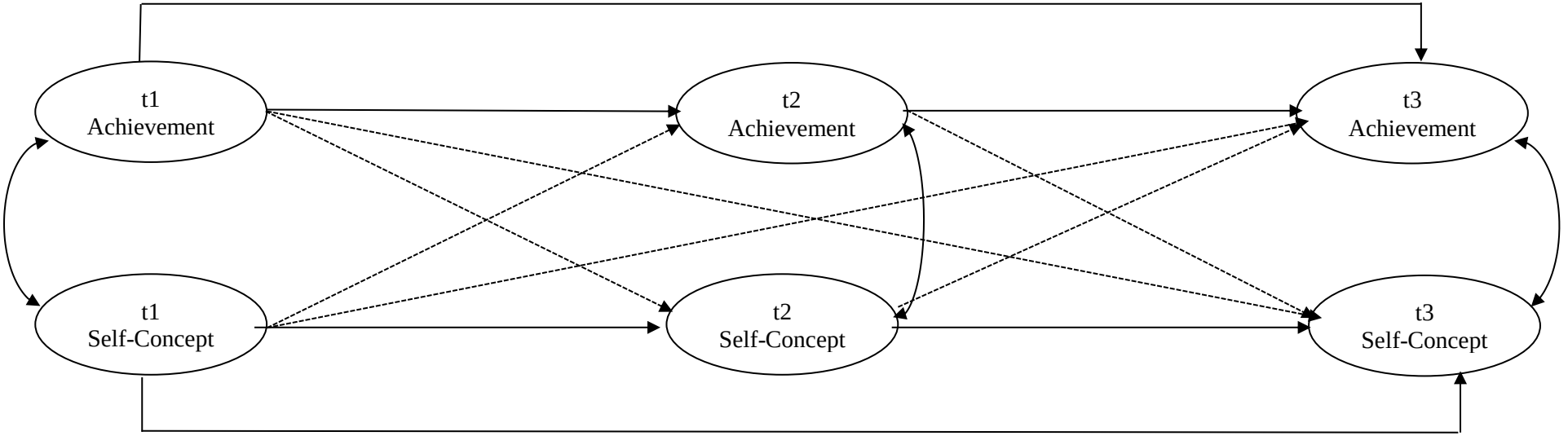

Figure 1. Prototype cross-lagged effects model for reciprocal relations between self-concept and achievement. For simplification, only three measurement waves are presented. Ovals represent latent constructs (self-concept and achievement factors); straight dashed arrows represent first-order and higher-order (here: second-order) cross-lagged effects paths; straight solid arrows represent first-order and higher-order (here: second-order) stability paths; curved arrows represent covariances between factors.