Running Head: BREAKING THE DOUBLE-EDGED SWORD

Breaking the Double-Edged Sword of Effort/Trying Hard: Developmental Equilibrium and Longitudinal Relations Among Effort, Achievement, and Academic Self-Concept

Herbert W. Marsh, Australian Catholic University and King Saud University, Saudi Arabia

Reinhard Pekrun, University of Munich, Germany and Australian Catholic University

Stephanie Lichtenfeld, University of Munich, Germany

Jiesi Guo, Australian Catholic University

A. Katrin Arens, German Institute for International Educational Research

Kou Murayama, University of Reading, UK

5 June, 2015

Revised: 23 November, 2015

Revised: 28 March, 2016

Acknowledgements

This paper was supported in part by a grant from the Australian Research Council to H. Marsh (DP130102713), and 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). Requests for further information about this investigation should be sent to Professor Herbert W. Marsh, Institute for Positive Psychology and Education, Australian Catholic University. E-mail: *herb.marsh@acu.edu.au*

Abstract

Ever since the classic research of Nicholls (1976) and others, effort has been recognized as a double-edged sword: whilst it might enhance achievement, it undermines academic self-concept (ASC). However, there has not been a thorough evaluation of the longitudinal reciprocal effects of effort, ASC and achievement, in the context of modern self-concept theory and statistical methodology. Nor have there been developmental equilibrium tests of whether these effects are consistent across the potentially volatile early-to-middle adolescence. Hence, focusing on mathematics, we evaluate reciprocal effects models over the first four years of secondary school, relating effort, achievement (test scores and school grades), ASC, and ASCxEffort interactions for a representative sample of 3,421 German students (Mn age = 11.75 years at Wave 1). ASC, effort and achievement were positively correlated at each wave, and there was a clear pattern of positive reciprocal positive effects among ASC, test scores and school grades—each contributing to the other, after controlling for the prior effects of all others. There was an asymmetrical pattern of effects for effort that is consistent with the double-edged sword premise: prior school grades had positive effects on subsequent effort, but prior effort had non-significant or negative effects on subsequent grades and ASC. However, on the basis of a synergistic application of new theory and methodology, we predicted and found a significant ASC-by-effort interaction, such that prior effort had more positive effects on subsequent ASC and school grades when prior ASC was high—thus providing a key to breaking the double-edged sword.

Keywords: academic effort, academic self-concept, double-edged sword, reciprocal effects models, developmental equilibrium; multigroup longitudinal invariance

Breaking the Double-Edged Sword of Effort/Trying Hard: Developmental Equilibrium and Longitudinal Relations Among Effort, Achievement, and Academic Self-Concept

Our focus is on a longitudinal reciprocal effects model (REM) of the temporal ordering of academic effort, achievement, and academic self-concept (ASC) over the potentially volatile early-to-middle adolescent period. In pursuit of this aim, we introduce a developmental equilibrium hypothesis that posits the consistency of related effects over early-to-middle adolescence, and test it on the basis of longitudinal data. More specifically, in relation to a priori predictions, we formally test developmental equilibrium as the invariance of effects across four waves of data, providing a formal test of the hypothesis that the self-system attains a developmental balance. For the present purposes we focus specifically on the school subject domain of mathematics, which was the basis of the secondary database that we used, noting that the math domain is relevant as an important school subject, and also because of the related psychological processes we are studying, in that effort is likely to be needed to master the skills to be learned.

Self-concept, one of the oldest constructs in psychology, is an important psychological construct that is fundamental to psychological well-being and that facilitates the attainment of personal goals such as educational attainment (Marsh, 2007b). Particularly in developmental and education settings, a positive ASC is both a highly desirable goal and a means of facilitating subsequent academic accomplishments, academic achievement, subject choice and coursework selection, interest, positive emotions, academic persistence, and long-term educational attainment (e.g., Chen, Yeh, Hwang, & Lin, 2013; Goetz, Frenzel, Hall, & Pekrun, 2008; Guay, Larose, & Boivin, 2004; Marsh, 2007b; Marsh & Craven, 2006; Pekrun, 2006; Pinxten, De Fraine, Van Damme, & D'Haenens, 2010). For example, Marsh and Yeung (1997; see also Parker et al., 2014) found that ASCs predicted future coursework selection better than did academic achievement. Moreover, Marsh and O'Mara (2008, 2010) found that ASCs predicted long-term educational attainment five years after high school graduation, better than did school grades, IQ, standardized tests, or socio-economic status.

In the present investigation we briefly review the extensive literature and the well-established support for the reciprocal effects model (REM) of relations between achievement and ASC. We then explore alternative theoretical perspectives on effort as a potential mediator of the reciprocal effects relating ASC and achievement. More specifically, we focus on the roles that effort, trying hard, and attributions for

success and failure have on the reciprocal relations between achievement and ASC, and integrate important early research on this topic by Nicholls (1976, 1978, 1979, 1984), Covington and Omelick (1979, 1984), Marsh (1984), and Dweck (2006; Dweck & Legget, 1988; Mueller & Dweck, 1998).

Reciprocal Effects Models (REMs) of Relations Between Achievement and ASC: What is the Role of Effort?

REMs of ASC & Achievement

Although a growing body of research (e.g., Huang, 2011; Marsh, 2007b; Marsh, Kuyper et al., 2014; Marsh, Lüdtke, et al., 2015; Marsh & Martin, 2011) shows that ASC and academic achievement are substantially correlated, this does not address the critical question of the temporal ordering of these two constructs. Traditional approaches to this issue (Calsyn & Kenny, 1977; also see Marsh, 1990) have taken an "either-or" approach—either prior achievement leads to subsequent ASC (a skill development model) or prior ASC leads to subsequent achievement (a self-enhancement model). However, integrating theoretical and statistical perspectives, Marsh (1990) argued for a dynamic reciprocal effects model (REM) that incorporates both the skill development and self-enhancement models, such that both ASC and achievement are causes and also effects of each other. Following from this classic demonstration of the REM (Marsh, 1990), there have been increasingly sophisticated developments in the statistical methodology measures used to test the REM, and substantial support has been garnered for the generalizability of the findings over age, nationality, different self-concept instruments, and different ways of measuring achievement (Marsh & Craven, 2006; Guay, Marsh & Boivin, 2003; also see meta-analyses by Valentine, DuBois, & Cooper, 2004; Huang, 2011).

What is the Role of Effort in REMs of ASC and Achievement?

How does academic effort fit into the REM of relations between academic achievement and ASC over time? Students who have higher ASCs tend to be willing to invest more effort into their academic work (Levpušček, Zupančič & Sočan, 2013). Similarly, Trautwein, Lüdtke, Marsh et al. (2009) argue that a different theoretical perspectives on competence beliefs (e.g., self-concept, self-efficacy, expectancy beliefs) all suggest that "people who are confident of their competence in a specific field are more likely to invest effort, to persist, and to succeed than are people with less belief in their competence" (p. 1116). We also note that their juxtaposition addresses both cognitive (self-concept) and behavioral (effort) components of motivation (see Martin, 2010) that are posited to facilitate learning. Hence, increased effort could reasonably

be one of the most important processes through which ASC leads to enhanced achievement (e.g., Marsh & Craven, 2006)—particularly with school grades, which have potential to be more responsive to effort than are standardized test scores. Thus, Trautwein, Lüdtke, Roberts, et al. (2009) reported that the positive effects of ASC on subsequent school grades are mediated in part by conscientious and persistent homework effort, even after controlling for prior achievement. This would suggest that ASC should lead to increased effort, which in turn would lead to better achievement. However, this was not actually tested with an REM, as achievement was the only variable measured in more than one wave. Furthermore, other work by this research group (e.g., Trautwein, Schnyder, et al., 2009) demonstrates that relations over multiple waves of data vary as a function of how homework is measured. Thus, an REM with perceived time on homework tasks (in minutes), homework effort (a multi-item scale, e.g., I always finish my homework) and school grades, showed that, whilst the effect of prior homework effort on subsequent school grades was positive, the effect of prior homework time was negative. Trautwein, Schnyder et al. suggested that the negative effect of time was plausible, in that more able students require less time to complete a particular homework task; the negative effect might reflect inefficient, unmotivated work styles. However, this result might also reflect the focus of their study on specific homework tasks assigned by the teacher, rather than on more general study time and out-of-school learning activities. Hence, the effect of homework effort could be positive or negative, depending on the nature of the construct, and might be idiosyncratic to a focus on homework, as opposed to more general expenditure of effort on other study and out-of-school learning activities.

Pinxten, Marsh et al. (2014) asked how a more general measure of "perceived math effort expenditure" (e.g., I put a lot of effort into mathematics; I work hard for mathematics) is related to ASC and achievement. They noted a number of cross-sectional studies showing that ASC and effort are positively related, but emphasized the importance of longitudinal REMs that go beyond mere correlational/cross-sectional studies, lamenting that "empirical research focusing on antecedents and consequences of academic effort expenditure as an educational outcome on its own is rather scarce" (p. 5). In their longitudinal REM study spanning five years (Years 3–7), they found that prior achievement and especially ASC, had negative effects on subsequent effort after controlling for prior effort, whereas effects of prior effort tended to have negative effects on subsequent achievement and ASC. Although ASC and achievement were reciprocally related, achievement had positive effects on subsequent ASC but negative effects on subsequent effort.

Effort as a double-edged sword. In addressing the complex relation between effort and ASC, Nicholls (1976, 1978, 1979, 1984, 1989), Covington and Omelick (1979, 1984) and others have suggested that increased academic effort is a double-edged sword. Increased effort and trying hard are seen as leading to academic success, and are valued by teachers, parents, and students themselves. However, effort and trying hard potentially undermine ASC, particularly when followed by failure. Covington's self-worth theory (1984, 1992, 1998) argues that students strive to maintain a sense of self-worth in academic settings by protecting their ASC. Therefore, particularly if failure is a likely outcome—a possibility that might be signalled by low ASCs—students might not try hard, because trying hard and failing would further undermine their subsequent ASCs.

In classic developmental research, Nicholls (1978) showed film clips to young children and then interviewed them about the films. On the basis of their responses, he ascertained the age/developmental level at which children could distinguish between ability and effort, and fully differentiate them as separate causes of outcomes. Nicholls (1976, 1979, 1984) posited that ASC is positively related to attributing academic success to ability (success/ability attributions), but negatively correlated with success/effort attributions. Likewise, he reasoned that ASC should correlate negatively with failure/ability attributions but positively with failure/effort attributions. However, Marsh (1984; Marsh, Cairns et al., 1984) questioned the logical basis of this claim in relation to attributions, suggesting that it might reflect a methodological detail that was idiosyncratic to Nicholls' research. Specifically, Nicholls used a force-choice format, which forced ability and effort attributions to be negatively correlated. Even in his interview study, which was not based on a forced-choice measurement instrument (Nicholls, 1978), ability and effort were experimentally manipulated in the film clips so as to be uncorrelated. Marsh, Cairns et al. emphasised the contrasting conclusion reached by Covington and Omelich (1979, 1984): that students will choose ability over effort, but would prefer to have both. Consistently with this alternative perspective, using separate Likert scales to measure attributions of success to ability and effort, Marsh, Cairns et al. showed that these attributions were positively correlated, and that both were positively correlated to ASC, even though correlations with ability attributions were more positive. Following failure, ability and effort attributions were both negatively correlated with self-concept, although correlations with ability attributions were more negative.

In related research, Dweck's (2006; Dweck & Legget, 1988; Muller & Dweck, 1988) implicit theory of intelligence posits a continuum of beliefs about the nature of intelligence that varies from believing that

intelligence is a fixed trait (fixed mindset or entity belief of intelligence) to believing intelligence is a function of effort, hard work, and persistence (a growth mindset or incremental belief of intelligence). Theories of intelligence are logically related to attributions of the causes of success and, in particular, failure (Diener & Dweck, 1978). Children with a fixed mindset are more likely to attribute outcome to ability, whilst those with a growth mindset are more likely to attribute outcomes to effort. Having a fixed mindset is particularly worrisome, following academic failure, as it implies that students lack ability and that there is nothing that can be done about it; this should have a negative effect on ASC. In contrast, students with a growth mindset are more likely to attribute failure to a lack of effort, a problem that can be remedied by trying harder, so that failure should have less negative effects on ASC. Thus, a fixed mindset "puts self-esteem and self-development in conflict with each other" (Dweck, 1991, p. 209), in that students are more concerned with protecting or augmenting their self-worth than with task mastery; this can lead to self-protective and effort-reductive strategies such as self-handicapping (Martin, 2010; Martin, Marsh & Debus, 2001). Paradoxically, noncontingent positive praise in relation to ability following success can have negative effects through reinforcing a fixed mindset (Mueller & Dweck, 1988). Mueller and Dweck reported that following failure, children praised for ability showed less persistence, less enjoyment, more low-ability attributions, and worse task performance than did children praised for effort. Hence, Craven, Marsh and Debus (1991) combined attributional retraining and effective praise strategies to enhance ASC. More generally, the O'Mara, Marsh, Craven and Debus (2006) meta-analysis of self-concept intervention studies found that interventions administering noncontingent praise were substantially less effective than those utilizing attributional feedback, goal feedback, and contingent praise. They noted that these results were consistent with findings by Mueller and Dweck (1998) that not all forms of praise and feedback foster a positive self-worth, typically citing person or trait-related praise (noncontingent praise) as being a cause of helplessness, with the potential to undermine self-worth.

ASC-by-Effort interactions. Following from Nicholls' earlier work, Jagacinski and Nicholls (1990) subsequently argued that effort reduction strategies only make sense if students have low ASCs and expect to fail. They posited that for students with high ASCs, the best way to demonstrate academic competence is through doing well in school, and that this should be facilitated by trying hard. Although this apparently important theoretical prediction was not specifically pursued by Jagacinski and Nicholls, their premise suggests that the effect of prior effort on subsequent ASC should depend on (i.e., interact with) prior levels

of ASC. Hence, as argued by Jagacinski and Nicholls, for students with high ASCs who expect to succeed, the best way to ensure success is by trying hard, so that more effort will lead not only to better achievement but also to higher ASCs. For these students, trying hard is not a threat to their ASCs, because ASCs are already high and they expect to succeed. By contrast, for students with low ASCs who expect to fail, trying hard and still failing will further undermine their already low ASCs. By not trying hard (i.e., so-called effort reduction strategies of self-handicapping), these students can shelter their already fragile ASC by attributing their expected failure to lack of effort rather than lack of ability. Hence, for these students, the effects of effort on ASC should be negative. Combining these ideas, the interaction between prior ASC and prior effort should be positively related to subsequent ASC. In the present investigation, we extend the typical REM of relations between ASC and achievement by adding measures of academic effort and latent interactions between effort and ASC, to test this proposition rigorously. Specifically, we examine the effects of prior effort, prior ASC, and their interaction on subsequent achievement, effort and ASC, in a fully latent REM.

Achievement: Test scores vs. school grades. In ASC research there is ongoing concern about the relative merits of assessing achievement on the basis of school grades (i.e., teacher marks typically provided as feedback to students and parents on periodic report cards) and standardized achievement tests. Marsh (2007b; Marsh & Martin, 2011) provided a theoretical rationale as to why school grades, compared to standardized test scores, provide a more salient, local source of feedback to students about their accomplishments, and why they tend to be more correlated with ASCs. However, teachers typically grade-on-a-curve, such that the best and worst students in each class tend to get the highest and lowest grades respectively, independently of the average ability levels of students within each class. This undermines the comparability of grade scores across classrooms and schools. Furthermore, grades tend to be somewhat idiosyncratic to particular subjects and to individual teachers. As emphasized by Marsh (1987; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005), with low-stakes tests for which students have no opportunity to study and no incentive to do so, characteristics such as study habits, effort, and persistence are unlikely to have much effect on test performance. In contrast, with high-stakes tests where students are highly motivated to perform well, know the content of the examination, and are able to prepare for it, characteristics such as study, effort, and persistence are likely to have more impact—particularly if such characteristics are part of the marking process (e.g., students are penalized for sloppy work habits or for not completing assignments on time, but rewarded for conscientious effort). Hence, increasingly, ASC researchers (e.g., Marsh, Kuyper,

et al., 2014) have argued that test scores and school grades should ideally be juxtaposed as separate constructs within the same study. For similar reasons, academic effort should be more highly correlated with school grades than are standardized test scores, particularly as effort, trying hard, and conscientiousness might actually be components of grades assigned by teachers (e.g., McMillan, Myran & Workman, 2002).

Developmental equilibrium

Equilibrium is reached when a system achieves a state of balance between the potentially counter-balancing effects of opposing forces. The application of equilibrium, and related terms, has a long history in psychological theorizing more generally, but particularly in developmental psychology. Thus, for example, Piaget (e.g., Piaget & Cook, 1952) argued that the psychological system aims to achieve a steady state of equilibrium that allows children to accommodate new experiences using existing schemas, whereas disequilibrium forces children to change their cognitive structures to regain equilibrium. Here we evaluate support for developmental equilibrium—whether the self-system is in a state of balance in relation to support for a priori predictions during this potentially volatile period. We address developmental equilibrium through tests of the consistency of relations among critical variables over early-to-middle adolescence that have important theoretical, developmental, and substantive implications for how relations among variables and processes vary as a function of age or year in school (see Eccles, 2009; Marsh, 2007b; Marsh & O'Mara, 2008; Marsh, Seaton, et al., 2008; Murayuma, et al., 2013)—the relative size paths in the REM.

Here we use the term developmental equilibrium in a more statistical sense, based on formal models of the invariance of effects across multiple waves. Thus, for example, tests of the REM model of relations between achievement, ASC and other constructs, typically are based on two measurement waves to test the temporal ordering, but at least three waves—and preferably more—are required to test developmental equilibrium assumptions that the effects of one variable on another across any two waves are consistent over multiple waves. Statistical models of developmental equilibrium (invariance of effects over multiple waves) test whether the pattern of reciprocal effects from one wave to the next is consistent across multiple waves—that the state of development is in balance over the period under consideration. Thus, for example, Marshall et al. (2014) showed that a system of reciprocal effects between self-concept and social support had attained equilibrium by junior high school. Furthermore, support for such tests of developmental equilibrium also facilitates interpretation of the results, provides a more parsimonious model, and results in statistically stronger tests of a priori predictions (also see Little et al., 2007, for more general discussion of stationarity

assumptions in cross-lag panel studies).

The Present Investigation

The present investigation attempts to disentangle the temporal ordering in the development of ASC, effort, school grade and test scores collected over four years from early secondary school students (N = 3,421), and to test new theoretical predictions about how to break the double-edged sword of effort in relation to ASC. The student sample was drawn from the three major school tracks in the German secondary school system: low-ability, medium-ability, and high-ability school tracks (Hauptschule, Realschule, and Gymnasium). The K-12 education system of education varies across the German states; however, one common principle is that there is between-schools tracking based on students' achievement after primary school. In Bavaria, primary school ends at grade 4. Subsequently, students attend secondary schools of one of the three tracks: the low-achievement "Hauptschule" track, the medium-achievement "Realschule" track, and the high-achievement "Gymnasium" track. The decision about placement in one of the tracks is based on students' achievement in the final grade of elementary school. The low-, medium-, and high-achievement track students are obliged to attend school until grades 9, 10, and 12, respectively. Thus, the low- and medium-track students are expected to enter vocational training, and the high-track students are expected to enter university after graduating from secondary school. There are no common achievement tests used in German schools; consequently, it is not possible to compare GPAs across tracks.

Methodologically, the present study uses fully latent REMs based on psychometrically strong measures, within-group tests of invariance over time, between-group tests of invariance across the three school-type tracks, developmental equilibrium tests of the invariance of effects across four waves of data (see Figure 1), and latent ASCxeffort interactions at each wave of the study.

Research Hypotheses

1. **Correlations.** Consistently with previous research, we predict ASC, effort, test scores and school grades to be positively correlated with each other within each of the four waves of data.
2. **Path Coefficients**. REMs are used to test the causal ordering of ASC, effort, and achievement in our REM (see Figure 1). Initially, separate SEMs are conducted for school grades and test scores, and then both indicators of achievement are considered in the same model.
    a. **ASC and Achievement**. Consistently with a growing body of research, we predict that ASC and achievement are reciprocally related, and that at least the pattern of results will be consistent for

models with only test scores, only school grades, and both school grades and test scores (e.g., Marsh & Craven, 2006). However, the unique effects of prior ASC on both indicators of achievement, and particularly the effects of both indicators of achievement on subsequent ASC, are expected to be somewhat smaller when both indicators of achievement are considered in the same model.

b. **Effort and Achievement.** We predict that prior effort has a positive effect on subsequent achievement, particularly with school grades, which are likely to be more influenced by effort and by trying hard than are standardized test scores. Logically, prior achievement might be expected to have a positive effect on subsequent effort (i.e. the effort-achievement relation is reciprocal), but we note that there is apparently little research into this issue. Hence, we leave this as a research question.

c. **Effort and ASC.** The theoretical basis for paths relating effort and ASC is less clear-cut, but we offer some tentative suggestions based primarily on theoretical perspectives reviewed earlier, noting that empirical results are not always consistent with each other or with theoretical predictions. The logic of the double-edged sword suggests that prior effort will have a negative effect on ASC, after controlling for prior ASC and achievement—that having to try harder to achieve is perceived by students to be indicative of lack of ability and will lead to a diminished ASC. However, expectancy-value theory (e.g., Eccles, 2009; Pekrun, 1993, 2006; Guo, et al., 2015) provides a clear theoretical prediction that ASC, as an indicator of expectancy, should have a positive effect on subsequent effort (but also see empirical results by Pinxten et al., 2014, which are inconsistent with this prediction). Bringing together these two disparate lines of theory suggests a paradoxical pattern of asymmetrical reciprocal effects involving effort (negative effort➔ASC, but positive ASC➔effort): this is apparently a new contribution.

d. **ASC-by-Effort interactions**. Also, to be consistent with the logic of the double-edged sword, and with suggestions by Jagacinski and Nicholls (1990), the effect of prior effort on subsequent ASC and achievement should be more positive (or less negative) for students with higher prior ASCs. High-ASC students are likely to achieve well—which implies that they can maintain or enhance their ASC by further boosting their achievement through the investment of effort. For these students, trying hard would be perceived to be less indicative of a lack of ability, because

such an interpretation would contradict their ASC. In contrast, low-ASC students may be motivated to protect further threats to their ASC by not investing effort, which would make it possible to attribute failure to lack of effort rather than to lack of ability. However, this comes at the potential cost of reduced achievement. For low-ASC students, trying hard may be considered a strong indicator of lack of ability, because this interpretation is congruent with their already low ASCs. Thus, negative effects of effort on ASC should be especially strong in these students. Overall, these assumptions imply that the proposed negative effect of effort on subsequent ASC, and the positive effect of effort on subsequent achievement, should be qualified by the positive effects of the ASC x effort interaction on these outcome variables. In terms of our REM model (see Figure 1), we posit a latent interaction between the latent ASC and effort factors and predict that this interaction has a positive effect on subsequent ASC and achievement—particularly school grades, for which effort is likely to be more influential than for test scores.

e. **Developmental Equilibrium.** Although only two measurement waves are required to test the temporal ordering of variables in REMs, typically it is recommended that three or more waves be considered (Marsh, Byrne, & Yeung, 1999; Marsh & Craven, 2006). Indeed, at least three waves—and preferably more—are required to test developmental equilibrium assumptions that the effects of one variable on another are consistent over time—a pattern of results with potential important developmental implications, and which also facilitates the summary of results. In support of developmental equilibrium, on the basis of tests of path invariance in the REM (Figure 1) across the four waves, we predict paths in support of predictions 2a–2d to be similar from one wave to the next ($Wave_i \rightarrow Wave_{i+1}$ for i+1 = 1 to 3) in terms of the pattern of statistical significance and the size of the path coefficients associated with the ASC, effort, and achievement factors. We test this by treating multiple waves as a within-group variable, evaluating the invariance of parameter estimates within each wave, and the invariance of paths leading from one wave to the next.

3. **Robustness of Path Coefficients**

   a. **Over multiple school types**. A salient feature of the German school system is that at an early age students are tracked into different school types, based largely on prior student achievement and school performance. Although this is not a major focus of the present investigation, we posit

that at least the pattern of results in support of predictions 2a–2d is similar across school types. We test this by treating school type as a multiple (between-group) grouping variable and evaluating the invariance of parameter estimates across groups.

b. **Over gender**. Girls typically have lower MSCs than boys (Marsh, 1989; Else-Quest, Hyde & Linn, 2010) and these gender differences have been shown to generalize broadly over different countries in large cross-national studies such as PISA (OECD, 2015; Marsh, Hau, et al., 2006) and TIMSS (Goldman & Penner, 2014). Hence, it is reasonable to ask whether relations between MSC, MACH and effort vary as a function of gender. Although this is not a major focus of the present investigation, we posit that at least the pattern of results in support of predictions 2a–2d is similar across gender. We test this by treating gender as a multiple (between-group) grouping variable and evaluating the invariance of parameter estimates across gender groups.

Method

Sample

The data come from the German Project for the Analysis of Learning and Achievement in Mathematics (PALMA) database, a large longitudinal study investigating the development of math achievement and its determinants during the secondary school years in the German federal state of Bavaria (see Frenzel, Goetz, Pekrun, & Watt, 2010; Frenzel, Pekrun, Dicke, & Goetz, 2012; Murayama, Pekrun, Lichtenfeld, & vom Hofe, 2013; Pekrun et al., 2007). The data are from four measurement waves, spanning the first four years of secondary school. Students (N = 3,421; 50% girls; mean age = 11.75 at Wave 1, SD = 0.68) are allocated to school tracks on the basis of their academic performance during the elementary school: either to the high-ability (academic; N = 1,187), middle-ability (intermediate; N = 1,048), or low-ability (vocational: N = 1,186) tracks.

Measures

Math self-concept and effort. At each measurement wave the same two sets of items were used to assess math self-concept (6 items, e.g., It is easy for me to learn in math) and effort (4 items, e.g., In math, I try hard to understand everything) using a 5-point-Likert scale: “not true”, “hardly true”, “a bit true”, “largely true”, or “absolutely true” (see Supplemental Materials for the wording of all the items; for evidence of validity see Pekrun et al., 2007, as well as the results in this study). The scales were demonstrated to be highly reliable measures at each of the four measurement waves: math self-concept (Wave1: α = .876;

Wave2: α = .895; Wave3: α = .893; Wave4: α = .910); math effort (Wave1: α = .780; Wave2: α = .809; Wave3: α = .801; Wave4: α = .805).

Math achievement. Students' math achievement was measured both in terms of school grades and on standardized achievement test scores. Math school grades were obtained from school documents based on students' report cards (school reports), received by students at the end of each school year. For ease of interpretation, we recoded the grades so that higher values represented higher achievement. The students also completed the PALMA Mathematical Achievement Test (vom Hofe, Pekrun, Kleine, & Goetz, 2002; vom Hofe, Kleine, Blum & Pekrun, 2005; also see Pekrun et al., 2007) at each wave. Using both multiple-choice and open-ended items, this test measures students' modeling and algorithmic competencies in arithmetic, algebra, and geometry. The test was constructed using multi-matrix sampling, with a balanced, incomplete block design (for details, see vom Hofe, Pekrun, Kleine, & Götz, 2002). Specifically, for each measurement point, students completed one of two parallel versions of the same test. The number of items increased with each wave, varying from 60 to 90 items across the five waves. Anchor items were included, to allow the linkage of the two different test forms, as well as the different measurement points. The achievement scores were scaled, using one-parameter logistic item response theory (Rasch scaling; Wu, Adams, Wilson, & Haldane, 2007). Additional analyses confirmed the unidimensionality and longitudinal invariance of the test scales (see Murayama et al., 2013).

Statistical Analyses

All analyses in the present investigation were done with Mplus (Muthén & Muthén, 2008–14, version 7.3), in which self-concept and effort were represented by latent factors estimated from multiple items. The instruments were administered once a year over four consecutive years (see Figure 1). As is typical in large longitudinal field studies, a substantial portion of the sample had missing data for at least one of the four occasions, due primarily to absence, to changing schools, or to problems in matching student responses across different waves. The breakdown of the number of waves of data completed by each student was: 1 (17.0%), 2 (27.1%), 3 (10.8%) and 4 (45.2%). All models were fitted using the robust Maximum Likelihood estimator (MLR) available in Mplus, in conjunction with the Mplus full-information-likelihood method (FIML) procedures to handle missing data. The FIML method has been shown to result in unbiased estimates for missing values, even in the case of a high level of missing values (Enders, 2010) and has been

further demonstrated to be an adequate method for dealing with missing data in longitudinal study designs (Jeličič, Phelps, & Lerner, 2009).

Invariance over multiple groups and multiple waves. Tests of measurement invariance evaluate the extent to which measurement properties generalize over multiple groups, situations, or occasions. Of particular substantive importance for longitudinal data in developmental and educational psychology is the evaluation of differences over time. Unless the underlying factors are measuring the same construct in the same way, and the measurements themselves are operating in the same manner across time, the comparison of parameter estimates is potentially invalid. One distinctive feature of longitudinal analysis is that it should normally include correlated uniquenesses between responses to the same item on different occasions (see Jöreskog 1979; Marsh 2007a, Marsh & Hau 1996). Without the inclusion of correlated uniquenesses, estimates of test-retest correlations and stability paths over time are likely to be positively biased, whereas the contributions of other constructs will be negatively biased in a way that is particularly critical in the interpretation of REMs. Importantly, there is no easy way to control for these biases in models based on manifest measures (i.e., responses to scale scores or single-item responses), rather than on fully latent REMs based on responses to individual items, such as those considered here. School-type track (high, medium, low) was treated as a multiple grouping variable, in order to test the generalizability of findings. Extending models of multi-group invariance, we tested models of the invariance of parameter estimates across groups (Millsap, 2011; Vandenberg & Lance, 1990) and over time (Widaman, Ferrer, & Conger, 2010).

In order to identify models, the traditional approach of fixing the first factor loading to 1.0 was used. However, in order to provide parameter estimates standardized to a common metric over the multiple waves, a preliminary CFA was done in which factor loadings were invariant over the multiple waves. The metric was identified by fixing to 1.0 the factor variances of constructs measured in Wave 1, instead of fixing the first factor loading to 1.0. In subsequent SEMs these standardized factor loadings were used to define the latent factors, fixing the first factor loading for each factor to the value obtained in CFAs in which the factor variances were fixed to be 1.0. In this way, all parameters were estimated in relation to a common metric across all four waves, and were standardized in relation to responses at Wave 1 (see Marsh, Morin, Parker & Kaur, 2014; Nagengast & Marsh, 2013; also see Supplemental Materials for Mplus syntax).

Invariance of path coefficients. Following preliminary tests of the invariance of factor loadings over time, our main interest is in the path coefficients relating the four constructs in each wave to the non-

matching constructs (i.e., the cross-paths in Figure 1) in the subsequent wave (see Research Hypotheses), hereafter referred to as lag-1 cross-paths. However, also included in each path model were paths leading from the same variable collected in earlier data waves (i.e., test-retest stability paths for matching variables—the horizontal paths in Figure 1). Thus, for example, ASC in Wave 4 was predicted by school grades, test scores, and effort from Wave 3 (lag-1 cross-paths), but also by ASC factors from Waves 1–3 (matching variables from lags 1–3, test-retest or stability paths). In this respect the REM is conservative, in that it shows the effects of non-matching variables (i.e., the effects of ASC on school grades, controlling for prior ASC, school grades, test scores, and effort), particularly in relation to studies that included only two or even three waves of data. However, although we focus on models with lag-2 and lag-3 stability paths, the results based on models with only lag-1 stability paths (summarized in the Supplemental Materials) indicate that the goodness of fit for these models is only marginally poorer, and that the critical cross-path estimates used to test REM hypotheses are substantially similar. The major difference in these models is that those with only lag-1 stability paths result in substantially larger test-retest stability paths than do models with lag-2 and lag-3 stability paths (which control for prior measures of the same constructs). Hence, the results suggest that this potentially important methodological issue is not particularly important in this particular study.

It is useful to evaluate the invariance of relations among the individual difference variables in SEMs of longitudinal data. Test-retest stability refers to the size of the test-retest (or autoregressive) paths for the same variable for two or more waves (e.g., $ASC_i \rightarrow ASC_{i+1}$ where i = Waves 1–3). For the present purposes, these test-retest paths are said to be invariant when lag-1 paths are equal over multiple waves (e.g. ASC $Wave_i \rightarrow$ ASC $Wave_{i+1}$ = ASC $Wave_{i+1} \rightarrow$ ASC $Wave_{i+2}$, assuming equal length intervals; Kenny, 1979). Particularly relevant for REMs, developmental equilibrium refers to the invariance over waves of cross-paths from one variable in one wave to another variable in the next wave (e.g., ASC $Wave_i \rightarrow$ effort $Wave_{i+1}$ = ASC $Wave_{i+1} \rightarrow effort_{i+2}$). Importantly, tests of developmental equilibrium require at least three and preferably more than three waves of data.

Latent interactions. REMs typically consider only first-order effects (sometimes referred to as "main" effects), but a central focus of the present investigation is on ASC-effort interaction. Methodologically, it is difficult to detect interaction effects in non-experimental designs (Marsh, Wen & Hau, 2004; Marsh, et al., 2013; Nagengast, et al., 2011; Trautwein et al., 2012; also see Appendix A in the Supplemental Materials for

more details). For the present purposes three interaction effects—one each for Waves 1–3—were constructed, based on latent ASC and effort factors and using the latent moderated structural equations (LMS) approach (Klein & Moosbrugger, 2000; Marsh, Wen, & Hau, 2004; 2004) incorporated into Mplus. The LMS approach takes into account the non-normal distribution of product terms, and does not need the specification of a complex sequence of constraints. However, traditional fit indices are not provided.

Goodness of fit. Given the known sensitivity of the chi-square test to sample size, to minor deviations from multivariate normality, and to minor misspecifications, applied SEM research generally focuses on indices that are sample-size independent (Hu & Bentler, 1999; Marsh, Balla & Hau, 1996; Marsh Hau & Wen, 2004; Marsh, Hau & Grayson, 2005), such as the Root Mean Square Error of Approximation (RMSEA), the Tucker-Lewis Index (TLI), and the Comparative Fit Index (CFI). Guidelines for fit are: TLI and CFI values greater than .90 and .95 are typically interpreted to reflect acceptable and excellent fits to the data, respectively. RMSEA values smaller than .08 or .06 for the RMSEA support acceptable and good model fits respectively. The chi-square difference test can be used for the comparison of two nested models, but it suffers from even more problems than does the chi-square test for single models (see Marsh, Hau, & Grayson, 2005). Thus, Cheung and Rensvold (2002) and Chen (2007) suggested that if the decrease in fit for the more parsimonious model is less than .01 for incremental fit indices like the CFI, or if the changes are less than .015 for RMSEA, then there is reasonable support for the more parsimonious model. However, we emphasize that these fit indices and cut-off values should be treated as only rough guidelines to be interpreted cautiously in combination with other features of the data.

Results

The results of the present investigation are presented in two parts. The preliminary focus is on confirmatory factor analyses (CFA) of the factor structure underlying responses to the 48 items (6 ASC, 4 effort, 2 achievement indicators for each of four waves of data). In the second part, the main focus is on the REMs of relations among ASC, effort, and achievement factors over the four waves of data.

Hypothesis 1: CFA Factor Structure and Relations Among ASC, Effort and Achievement

In a series of CFA models we tested the invariance of the factor structure over the multiple groups representing the three school types (high-, medium- and low-track schools). Consistently with traditional approaches to multiple group invariance (e.g., Marsh, Muthén, et al., 2009; Meredith, 1993), we compare configural (no invariance), metric (invariance of factor loadings), and scalar (invariance of intercepts)

models. In longitudinal analyses in which the same items are administered across multiple waves, it is critical to incorporate a priori correlated uniquenesses relating responses to the same items across multiple waves (Jöreskog, 1979; Marsh & Hau, 1996); failure to include them is likely to result in a poor fitting model and biased parameter estimates. In order to test this feature for these data we began by comparing models with and without correlated uniqueness (Models 1A-C and 2A-C in Table 1). Consistently with expectations, the a priori Model 2 with correlated uniquenesses provided a better fit. In relation to traditional indices of fit, the configural, metric and scalar tests based on Model 2A-C, respectively, all provided excellent fits to the data (e.g., all CFIs and TLIs > .97), and the imposition of demanding tests of factorial invariance over multiple groups (representing difference school type) resulted in decrements in fit smaller than those needed to support the assumption of invariance. In Model 3A-C we added the additional constraint that factor loadings be invariant over time. Here, there was almost no change in the fit, compared to those based on Model 2 (e.g., CFI = .977 vs. .978 for configural invariance based on Models 2A and 3A). Thus, there was support for the invariance of factor loadings over multiple waves as well as for scalar invariance over the multiple (school-track) groups.

Of particular interest in the present investigation, consistently with Hypothesis 1, ASC, effort, test scores, and school grades were all positively correlated within each of the four waves of data. Indeed, all correlations in Table 2 are positive. Within each of the four waves, correlations between ASC and achievement were substantial for both test scores (.439 to .513) and for school grades (.504 to .638). The corresponding correlations with effort, although smaller in size, were also highly significant from a statistical perspective, both for test scores (.130 to .230) and for school grades (.108 to .238). The correlations also demonstrate that within each wave, ASC and effort are positively correlated (.292 to .378).

Reciprocal Effects Models (REMs) of Causal Ordering of ASC, Effort and Achievement

We begin this section with a preliminary evaluation of Hypothesis 3, the robustness of path coefficients over school-track type (between-group tests of invariance) and multiple waves (within-group test of developmental equilibrium). In each of a series of four models (Models 5A–5D, Table 1), we begin with the factor structure in Model 3C (i.e., between-group scalar invariance of factor loadings and intercepts over the three school-type groups, and factor loading invariance over multiple waves). In these REMs, however, path coefficients were posited to represent relations among the four constructs (ASC, effort, test scores and school grades) over time (see Figure 1). The four models differed in terms of the invariance constraints,

varying from the least constrained model, with no invariance of path coefficients (Model 5A) to the most constrained model, with path coefficients constrained to be invariant over the multiple groups and across the multiple waves (Model 5D). Inspection of the corresponding fit indices indicates that even the fit of the most constrained Model 5D is very good (CFI = .969, TLI = .967, RMSEA = .020), and differs little from that of the least constrained Model 5A (ΔCFI= .005, ΔTLI = .004, ΔRMSEA = .001). These results provide support for the robustness of the effects posited in Hypothesis 3a, indicating that the pattern of results is similar in the three school-type ability tracks, and supporting developmental equilibrium Hypothesis 3C).

In a parallel set of analyses (Models 6A–6D, Table 1), we also evaluated the invariance of the results over gender (Hypothesis 3b). The results again indicate that the fit of even the most constrained Model 6D (paths invariant over gender and multiple waves) is very good (CFI = .971, TLI = .969, RMSEA = .020), and differs little from that of the least constrained Model 6A (ΔCFI= .006, ΔTLI = .005, ΔRMSEA = .002). These results support the robustness of the effects in relation to gender, as posited in Hypothesis 3b.

ASC and achievement (Hypothesis 2a, Table 3): Consistently with predictions and with a substantial body of research, across the three school-type ability tracks and the four waves, there is a discernible and clear pattern of reciprocal positive effects relating the ASC both to school grades and to test scores, as well as paths relating both of these measures of achievement to ASC (Model 6D). In Models 7A and 7B, REMs were tested separately for school grades (Model 7A) and for test scores (Model 7B). Although the effects are somewhat stronger when each of the achievement indicators is considered separately, the pattern of results is consistent across Models 7A (grades only), 7B (test scores only) and Model 6D (grades and test scores)—with and without ASC-by-Effort interaction (which is discussed below). Because there is support for the invariance of the paths over waves (the four years, within-group invariance) and multiple (school-track) groups, support for the conclusions is robust.

Effort and achievement (Hypothesis 2b, Table 3): Here the pattern of results is more complicated. Consistently with a priori predictions, prior school grades had a positive effect on subsequent effort (Models 6D &7A) after controlling for prior effort, ASC and achievement. However, the corresponding paths from prior test scores to effort were non-significant (Models 6D & 7B). Furthermore, prior effort had no significant effect on subsequent achievement—either test scores or school grades.

Effort and ASC (Hypothesis 2c, Table 3): Consistently with the logic of the double-edged sword, prior effort had a small but statistically significant negative effect on subsequent ASC, after controlling for the

effects of prior ASC and achievement. This result is consistent over the six models (Table 3) in which each achievement indicator is considered separately (Models 7A & 7B) or in combination (Model 6D). In contrast, the **ASC➔effort** path is non-significant for 5 of the 6 models, but is significantly positive in Model 7B (test scores only, with no interaction).

ASC-by-Effort interactions (Hypothesis 2d). Consistently with a priori predictions, the ASC-by-Effort interaction had statistically significantly positive effects on ASC. This interaction was consistent across models with grades only, test scores only, and both grades and test scores (Table 3). As illustrated in Figure 2, and in support of predictions based on the double-edged sword hypothesis, the effects of prior effort on subsequent ASC were negative for low-ASC students but increasingly more positive (or less negative) for students with more positive ACSs. In support of the developmental equilibrium of this pattern of results, the first-order and interaction effects are invariant across the four waves of data, covering the first four years of secondary schooling in Germany. In support of the robustness of the findings, the results—including support for developmental equilibrium—are also invariant over the three school types (high-, medium- and low-track schools). In summary, although effort is a double-edged sword for low-ASC students, this is not the case for high-ASC students. Hence, ASC appears to be a critical feature in breaking the double-edged sword and undermining its counter-productive effects.

Similarly, and also consistently with a priori hypotheses, the ASC-by-Effort interaction had significantly positive effects on subsequent school grades. Thus, whilst the first-order effect of effort on school grades was positive but non-significant, the effect of effort was more positive for students who also had positive ASCs. However, although this is not entirely surprising, neither effort nor ASC-by-Effort interaction had statistically significant effects on subsequent test scores. As noted earlier, such effects were expected to be more substantial for school grades, which reflect student effort to a greater extent than do standardized test scores on low-stakes tests, which have no further implications.

Discussion

In the present investigation, we developed new theoretical perspectives, in combination with current statistical approaches, to revisit the implications of the double-edged sword of effort and trying hard, in academic settings. More specifically, we evaluated the reciprocal effects relating ASC, effort and achievement over the first four years of secondary schooling for a large representative sample of German students. Based on the integration of different theoretical and methodological perspectives, we derived a

diverse set of a priori predictions, some extending well-established findings but others offering apparently new theoretical perspectives.

Developmental Equilibrium

In the present investigation, support for developmental equilibrium indicates that the pattern of results was similar across all four waves of data considered here. This support has important substantive implications, demonstrating that the results are consistent and stable over a potentially extended developmental period: here, the first four years of secondary school. Methodologically, support for developmental equilibrium facilitates interpretation and presentation of the results, provides a more parsimonious model, and results in statistically stronger tests of a priori predictions. However, because tests of developmental equilibrium require as least three or more waves of data, and most REM studies of ASC consider only two waves, apparently no previous research has considered ASC, effort, school grades, and standardized achievement test scores in latent REMs based on as many waves of data as are considered here—potentially important design features for future research.

Correlational vs. REM Results

Our findings provide two different perspectives on the relations of effort with ASC and achievement. With respect to correlations within each wave of the study, effort was positively correlated with ASC and with both indicators of achievement. Although ASC within each wave was more strongly related to test scores, and particularly school grades, than to effort, the positive correlations between effort and ASC were highly significant from a statistical perspective and at least moderate in size (*r*s of .292 to .378 within each of the four waves). These results are apparently consistent with previous research, which suggests that effort/success attributions are positively related to ASC, even if less so than are attributions to ability (Marsh, Cairns, et al. 1984). Further, research shows that students choose ability over effort but would prefer to have both (Covington and Omelich, 1979; 1984), and underlines the value placed on effort by teachers, parents and students themselves. However, a very different perspective is evident in the REMs of the effects of prior effort on subsequent ASC, after controlling for prior measures of these variables. Here, consistently with the double-edged sword premise, for low-ASC students the effect of prior effort on subsequent ASC was consistently negative across waves and school tracks. Hence, even though prior school grades did have a positive effect on subsequent effort as well as ASC, and school grades were highly correlated with ASC, prior effort had a negative effect on subsequent ASC. The juxtaposition of these two sets of results highlights

why it is important to evaluate relations among variables with longitudinal REMs, rather than simply relying on correlations based on a single wave of cross-sectional data.

Symmetrical Pattern of Reciprocal Effects: Achievement and ASC

Consistently with a substantial body of research we found that, after controlling for prior outcomes, ASC and achievement are reciprocally related: higher prior ASC led to higher subsequent achievement and higher prior achievement led to higher subsequent ASC. Although this is clearly consistent with previous research, our results were more robust than in most previous research. Thus, for example, we extended the typical REM to provide tests of developmental equilibrium that are only possible when there are at least three or more waves of data, and demonstrated that the paths relating achievement and ASC from one wave to the next were invariant over the four waves considered here. In further support of the generalizability of the results, the paths in support of the REM for achievement and ASC were consistent over school-type tracks, which are such a distinctive feature of the German system, as well as gender.

The pattern of reciprocal effects was also consistent over two quite different indicators of achievement—standardized test scores and school grades—whether considered separately, or together within the same model. Test scores provide a common metric upon which to base conclusions, whereas school grades are likely to be idiosyncratic to particular settings, and to be heavily influenced by grading-on-a-curve, particularly in settings like the German system, which incorporates tracking according to ability at the school or class level. However, school grades provide a more salient, local source of feedback to students about their accomplishments, and are likely to be more influenced by individual student characteristics such as effort and ASC than are the low-stakes standardized tests such as those used here (i.e., tests that are not part of the routine assessment procedures used in schools, that have no implications for students, and in which students do not even receive feedback on their performance). Consistently with other research (e.g., Möller, et al., 2009, 2011), support for the reciprocal effects of achievement and ASC was found both with school grades and with test scores considered separately, as well as in combination. Importantly, the reciprocal effects involving school grades can only partly be explained in terms of test scores, and vice-versa, thus demonstrating the importance of considering both as separate constructs.

Asymmetrical Pattern of Reciprocal Effects: Achievement, Effort and ASC

As shown by a large body of research, as well as in the present investigation, support for reciprocal effects involving ASC typically results in a symmetrical pattern of results. Thus, prior ASC has a positive

effect on subsequent achievement, and prior achievement has a positive effect on subsequent ASC. However, juxtaposing (a) classic work on the double-edged sword in relation to effort, (b) subsequent work in expectancy value theory and self-concept research, and (c) current approaches to testing REMs, we found support for an apparently new theoretical prediction: that there would be an asymmetrical pattern of effects such that prior school grades had positive effects on subsequent effort, but prior effort had non-significant or negative effects on subsequent grades and ASC. Consistently with the double-edged sword premise, we predicted, and found, that effort has a negative effect on ASC: having to work hard to achieve implies lesser ability, and undermines ASC. In contrast, consistently with expectancy value theory and self-concept research, we also predicted and found some support for the prediction that prior ASC had a positive effect on subsequent effort. Bringing together these two theoretical perspectives in combination with the REMs is apparently a new theoretical contribution to expectancy-value theory and ASC research that has practical implications for better understanding the double-edged sword. Nevertheless, we note that the positive effects of ASC on effort were not entirely consistent across different models in the present investigation and were not found by Pinxten et al. (2014); this suggests the need for further research.

ASC-by-Effort Interactions: Breaking the Double-Edged Sword

Following from the theoretical premise posited by Jagacinski and Nicholls (1990), we reasoned that the paths relating prior effort to subsequent ASC and achievement would depend on prior levels of ASC, such that effort has a more positive effect on subsequent outcomes when prior ASC is high. However, apparently there have been no previous attempts to test these predictions; nor is there even a clear understanding of design and methodology issues as to how to pursue the question. Extending typical REMs in order to test this prediction, we added tests of the latent interaction between effort and ASC for each wave of data. Consistently with a priori predictions, the effects of effort were more positive (or less negative) when ASC was higher. This ASC-by-Effort interaction was statistically significant for subsequent ASC and for school grades. These results suggest that having a high ASC breaks the double-edged sword, at least to some extent, in that for students with high ASCs the effects of effort are positive.

The effect of effort-by-ASC interaction on subsequent ASC, a critical feature of our study, was similar in models that included test scores, school grades or both. However, whereas the effort-by-ASC interaction had a significant effect on subsequent school grades, neither effort nor its interaction with ASC had any significant effect on test scores. Although this is consistent with our earlier discussion of the

distinction between these two indicators of achievement, it is possible that these results would have been different if test scores had been based on high-stakes tests (e.g., tests that were critically important to students, that were based on content readily available to students, that students were both able and encouraged to prepare for, and tests on which students received feedback). A relevant extension, beyond the scope of the present investigation, would be to evaluate the generalizability of the findings across different types of standardized tests.

It is also noteworthy that the interaction of ASC and effort not only influenced subsequent ASC but also had positive effects on students' grades. This positive effect of the ASC x Effort interaction implies that students achieve more when they have high ASC and invest more effort, suggesting that there is a synergy between these two factors in boosting achievement. Thus, the findings imply that high ASC not only can undo negative effects of effort on subsequent ASC, but adds to the positive effects of effort on subsequent achievement, suggesting that it's best to have both high ASC and high effort. This finding not only has theoretical implications in relation to the double-edged sword, but also has practical implications, in that classroom teachers need to reinforce the synergy between these two constructs.

Strengths, Limitations and Directions for Further Research

Several features of the present investigation support the robustness of the findings. In particular, support for developmental equilibrium indicates that the pattern of results was similar across all four waves of data, as well as the three school-type ability tracks considered here. Nevertheless, we also emphasize that whilst REMs provide stronger tests of temporal ordering than do mere correlations based on cross-sectional data, the results are still correlational. Thus, whilst it is appropriate to hypothesize the temporal ordering of effects, and to evaluate models of temporal ordering and developmental equilibrium, their support is limited to empirical evidence that is consistent with the hypotheses, without ruling out alternative interpretations of the findings. For this reason, the results are referred to here as path coefficients, temporal ordering effects, predictive effects, or simply effects, rather than "causal" effects. Nevertheless, our findings do have implications for theory, social policy, intervention, and practice that will lead to further investigation, and that are consistent with growing concerns about the appropriateness of randomized control trials as the best way to build knowledge about social policy (Schorr, 2016).

From a developmental perspective, we expanded the theoretical and statistical rationale for tests of developmental equilibrium, with our longitudinal data covering the potentially turbulent early-to-middle

adolescent period. More specifically, we found support for developmental equilibrium as the invariance of effects across four waves of data, based on the assumption that the self-system had attained a developmental balance in relation to predictions from each of our theoretical models of relations among effort, academic self-concept, and achievement. In extending tests of developmental equilibrium we offered new developmental perspectives that should apply to developmental research more generally, including the need for longitudinal data based on more than just 2 or 3 waves, and stronger tests of the statistical assumptions underlying the models. Indeed, we suggest that formal tests of developmental equilibrium should be a useful addition to many developmental studies where an historical focus on statistical tests of the null hypothesis had led to prioritizing developmental differences and change, rather than developmental continuity—even when there is strong support for developmental equilibrium, as in the present investigation.

We also note the limitations of reliance on self-report, such that shared method effects might have inflated relations between the key constructs. In particular, effort was only measured by student self-reports, so that it would be interesting to consider alternative measures such as teacher reports, observation measures or student diary studies. However, we also note that, given the focus on students' self-perceptions, self-report measures are probably the most valid measure of how students perceive their effort, even if external observers have different perspectives.

We note as a potential limitation of our study that it was based only on responses by German students, and that the results need to be replicated in other countries, including East Asian countries where there might be differences in perceptions of ability and effort (e.g., Sali & Hau,1994). We have no reason to assume that our results would not generalize to other countries, and at least some evidence based on cross-national PISA studies supports this assumption (e.g., Marsh, 2016; Nagengast & Marsh, 2013). However, given that our study was based on students in early secondary school years, tests of the generalizability of the results to other student populations, school types, age groups, and countries are clearly warranted.

There are also some key features of the present investigation that it will be important for future research to consider. In particular, it is imperative that outcome measures be assessed with psychometrically sound, multi-item instruments, and with latent models based on multiple indicators that control for complex measurement error (i.e., measurement error within each construct, but also the inevitable correlated uniquenesses when the same items are used for multiple waves). In relation to this caveat, we note that whilst ASC and effort were based on multiple indicators, both measures of achievement (school grades and test

scores) were based on single indicators. Of particular relevance, although it is possible to test REMs based on only two waves of data, there are critical advantages to having more waves of data. Here, for example, not only was there support for developmental equilibrium, which requires at least three waves of data, but controlling for test-retest stability on the basis of more than one lag provided a much stronger test of the robustness of the REM cross-paths, which is the critical feature of REM studies.

It is also important to note another limitation of the present investigation, which might provide directions for future research. We focused specifically on math constructs. Although we have no reason to believe that our conclusions are specific to the math domain (given that support for the REM of reciprocal relations between ASC and achievement generalizes over different domains; Marsh, 2007b; Valentine et al., 2004), it is important to test the generalizability of the results to other domains as well. Because our research was based on adolescent students, it is also relevant to test the generalizability of the results to younger children. Particularly in the light of classic research by Nicholls (1976, 1978) about the age and developmental level at which children can differentiate between ability and effort as separate predictors of academic outcomes, it is entirely possible that for young children, effort has a positive effect on ASC.

Although this is beyond the scope of the present investigation, there are theoretically sound reasons to explore other potential moderators of the path coefficients relating effort, ASC, and achievement. In particular, Dweck (2006; Dweck & Legget, 1988; Mueller & Dweck, 1998) reported that children who have a growth mindset (i.e., a belief about the incremental nature of intelligence, as opposed to a fixed entity belief), are more likely to attribute failure outcomes in particular to a lack of effort. Hence, it might be expected that students with a growth mindset, effort would be less likely to have negative consequences for ASC, particularly following failure. Similarly, for students who have a mastery orientation, as opposed to a performance orientation, failure might be expected to have less negative consequences for ASC (Elliot, Murayama & Pekrun, 2011). Similar predictions might also result from other theoretical models, such as self-determination theory (Ryan & Deci, 2000), which distinguishes between autonomous and controlled motivation, and the two-factor theory of passion (Vallerand, 2015), which distinguishes between harmonious and obsessive passion. Thus, growth mindset, mastery goal orientation, autonomous motivation orientation or harmonious passion as individual difference variables at the level of the student (or with teachers who promote a growth mindset, mastery goals, autonomous motivation, harmonious passion, or a positive ASC at the level of the classroom), might be expected to moderate or mediate the longitudinal effects of effort, ASC,

and achievement on each other. We note however, that performance orientations at the level of the student and the classroom were also posited to moderate the negative effects of school or class-average achievement on ASC (the big-fish-little-pond effect), but rigorous tests of this hypothesis have found no support for it (Cheng, McInerney, & Mok, 2014; Wouters et al., 2015; see review by Marsh, Martin, Yeung, & Craven, in press). More generally, Hattie (2012, p. 46) emphasized that:

> When students invoke learning rather than performance strategies, accept rather than discount feedback, set benchmarks for difficult rather than easy goals, compare their achievement to subject criteria rather than with that of other students, develop high rather than low efficacy to learning, and effect self-regulation and personal control rather than learned helplessness in the academic situation, then they are much more likely to realize achievement gains and invest in learning. These dispositions can be taught; they can be learned.

Hattie's proposals, although not specifically focused on the double-edged sword per se, each suggest a potential moderator of the negative effect of effort on ASC. Thus, the effect of effort on ASC might be less negative (or even positive) when students invoke learning/mastery strategies, embrace constructive feedback, focus on difficult goals, use criterion references and self-improvement to gauge success, have positive self-perceptions of competence, and develop self-regulation and personal control strategies in the face of adversity. However, rigorous research along the line of the Wouters et al. (2015) study is needed, to test these speculations.

The focus of our study was on ASC, but we note that there are myriad constructs designed to measure self-perceptions of competence (e.g., self-efficacy, agency, outcome expectancy, confidence, etc.) that might also be relevant. However, we also note the possibility of jingle-jangle fallacies (Marsh, Craven, Hinkley, & Debus, 2003), where two scales with similar names might measure different constructs, whilst two scales with apparently dissimilar labels might measure similar constructs. Thus, we caution researchers who argue for the theoretical distinctiveness of alternative constructs purporting to measure self-perceptions of competence in a particular academic domain, to empirically test such claims by including measures of the alternative constructs in their research.

Summary

In summary, we replicated and extended the well-established pattern of reciprocal effects between ASC and achievement. Although support for the REM is well-established in the research literature, our study

extended previous research, demonstrating the consistency of the results over multiple waves (developmental equilibrium) and different ability tracks. Thus, across a wide range of ages and different school types, catering to students of different ability levels, teachers need to develop the academic skills of their students, nurture positive ASCs in students, and reinforce the connection between these constructs. Although the importance of positive ASCs is acknowledged at all levels of schooling, teachers are not particularly effective at enhancing self-concept, particularly at the secondary level, where an increasing emphasis on getting good marks on standardized tests might not be supportive of positive ASCs.

Integrating this REM research with the classic double-edged sword of effort in relation to ASC, we extended the traditional ASC-REMs to include effort, and derived new theoretical predictions and methodological approaches to test the pattern of reciprocal effects. Consistently with our juxtaposition of different theoretical perspectives, we predicted and found an asymmetrical pattern of reciprocal effects associated with effort and ASC. Whereas prior ASC had a positive effect on subsequent effort (which is consistent with EVT and ASC research), prior effort had a negative effect on subsequent ASC (consistent with the double-edged sword premise). However, on the basis of a previously untested theoretical premise, we also predicted and found that for students with initially higher ASCs, the effect of prior effort on subsequent ASC was more positive, thereby providing a key to breaking the double-edged sword. Although the importance of ASC is already well-established, what is novel in these findings is that something that teachers already understand to be important in itself, also seems to be a solution to an entirely different problem. However, past research suggests that most classroom teachers might not be very good at promoting and maintaining positive self-concepts in their students (e.g., Craven, Marsh & Debus, 1991; Hattie, 2012), given that the systemic focus is more on good test scores as the putative basis of accountability (for students, teachers and school systems more generally). Hence, it is important to demonstrate yet another reason why a more concerted effort should be made to promote self-concept within education settings. Although our primary focus was theoretical and methodological, the results also have practical implications for teachers, parents, and students themselves, in terms of understanding the double-edged sword of effort and trying hard: One apparent way to break the vicious cycle of effort reduction and self-handicapping strategies is to promote and maintain a positive ASC.

References

Calsyn, 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.

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.

Cheng, R. W-y., McInerney, D. M., & Mok, M. M. C. (2014). Does Big-Fish-Little-Pond Effect always exist? Investigation of goal orientations as moderators in the Hong Kong context. *Educational Psychology*. http://dx.doi.org/10.1080/01443410.2014.898740

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

Covington, M. V. (1984). The motive for self-worth. In R. Ames & C. Ames (Eds.). *Research on motivation in education* (pp. 77–131). Academic Press, Orlando.

Covington, M. V. (1992). *Making the grade: A self-worth perspective on motivation and school reform.* Cambridge: Cambridge University Press.

Covington, M. (1998). *The will to learn: A guide for motivating young people.* New York, NY: Cambridge University Press.

Covington, M. V., & Omelich, C. L. (1979). Effort: The double-edged sword in school achievement. *Journal of Educational Psychology*, *71*, 169–182.

Covington, M. V., & Omelich, C. L. (1984). An empirical examination of Wiener's critique of attribution research. *Journal of Educational Psychology*, *76*, 1214–1225.

Craven, R. G., Marsh, H. W., & Debus, R. (1991). Effects of internally focused feedback and attributional feedback on the enhancement of academic self-concept. *Journal of Educational Psychology*, *83*, 17–26.

Dweck, C. S. (1991). Self-theories and goals: Their role in motivation, personality, and development. In R.A. Dienstbier (Ed). *Perceptions on motivation: Nebraska Symposium on Motivation*. Vol 38. Lincoln: University of Nebraska Press.

Dweck, C. S. (2006). *Mindset*. New York: Random House.

Dweck, C. S., & Legget, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95,* 256–273. doi:10.1037/0033-295x.95.2.256.

Diener, C. I., & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy and achievement cognitions following failure. *Journal of Personality and Social Psychology, 36,* 451–462. doi:10.1037/0022-3514.36.5.451

Eccles, J. S. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist*, *44*, 78–89. doi:10.1080/00461520902832368

Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3 × 2 achievement goal model. *Journal of Educational Psychology*, *103*(3), 632–648. http://dx.doi.org/10.1037/a0023952

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*, 103-127. http://dx.doi.org/10.1037/a0018053

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

Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. G. (2010). Development of mathematics interest in adolescence: Influences of gender, family, and school context. *Journal of Research on Adolescence*, *20*, 507–537.

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. doi: 10.1037/a0026895

Goetz, T., Frenzel, A. C., Hall, N. C., & Pekrun, R. (2008). Antecedents of academic emotions: Testing the internal/external frame of reference model for academic enjoyment. *Contemporary Educational Psychology*, *33*, 9–33.

Goldman, A. D. & Penner, A. M. (2014). Exploring international gender differences in mathematics self-concept International. Journal of Adolescence and Youth, DOI: 10.1080/02673843.2013.847850

Guo, J., Marsh, H. W., Parker, P., Morin, A. J. S. & Kaur, G. (2015). Directionality of the Associations of High School Expectancy-Value, Aspirations, and Attainment: A Longitudinal Study. A*merican Educational Research Journal*, *52*(2), 371–402. doi:10.3102/0002831214565786

Guay, F., Larose, S., & Boivin, M. (2004). Academic Self-concept and Educational Attainment Level: A Ten-year Longitudinal Study. *Self and Identity*, *3*, 53–68. doi: 10.1080/13576500342000040

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.

Hattie, J. (2012). *Visible Learning for Teachers: Maximizing Impact on Learning*. New York: Routledge.

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.

Jagacinski, C. M., & Nicholls, J. G. (1990). Reducing effort to protect perceived ability: They'd do it but I wouldn't. *Journal of Educational Psychology*, *82*(1), 15–21. http://dx.doi.org/10.1037/0022-0663.82.1.15

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.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal investigations. In J. R. Nesselroade & B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303–351). New York: Academic Press.

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457–474.

Levpušček, M. P., Zupančič, M., & Sočan, G. (2013). Predicting achievement in mathematics in adolescent students: The role of individual and social factors. *The Journal of Early Adolescence*, *33*, 523–551. doi:10.1177/0272431612450949

Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development*, *31*, 357–365.

Marsh, H. W. (1984). Relations among dimensions of self-attribution, dimensions of self-concept, and academic achievements. *Journal of Educational Psychology*, *76*, 1291–1308.

Marsh, H. W. (1987). The Big-Fish-Little-Pond Effect on Academic Self-concept. *Journal of Educational Psychology*, *79*, 280–295.

Marsh, H. W. (1989). Age and sex effects in multiple dimensions of self-concept: Preadolescence to Early-adulthood. *Journal of Educational Psychology*, *81*, 417–430.

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. (2007a). Application of confirmatory factor analysis and structural equation modeling in sport and exercise psychology. In G. Tenenbaum & R. C. Eklund (Eds), *Handbook of sport psychology* (3rd ed). (pp. 774–798). Hoboken, NJ, US: John Wiley & Sons Inc.

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

Marsh, H. W. (2016). Cross-cultural generalizability of year in school effects: Negative effects of acceleration and positive effects of retention on academic self-concept. *Journal of Educational Psychology, 108*, 256-273.http://dx.doi.org/10.1037/edu0000059

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315–353). Hillsdale NJ: Erlbaum.

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., Cairns, L., Relich, J., Barnes, J., & Debus, R. L. (1984). The relationship between dimensions of self-attribution and dimensions of self-concept. *Journal of Educational Psychology*, *76*(1), 3–32. http://dx.doi.org/10.1037/0022-0663.76.1.3

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., Craven, R. G., Hinkley, J. W., & Debus, R. L. (2003). Evaluation of the Big-Two-Factor Theory of academic motivation orientations: An evaluation of jingle-jangle fallacies. *Multivariate Behavioral Research, 38,* 189-*224.*

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., Artelt, C., Baumert, J., Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing. 6,* 311–360.

Marsh, H. W., Hau, K-T., & Grayson, D. (2005). Goodness of Fit Evaluation in Structural Equation Modeling. In A. Maydeu-Olivares & J. McArdle (Eds.) *Psychometrics: A Festschrift to Roderick P. McDonald* (pp. 275–340). Hillsdale, NJ: Erlbaum.

Marsh, H. W., Hau, K-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indices and dangers in overgeneralizing Hu & Bentler's (1999) findings. *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., Lüdtke, O., Nagengast, B., Trautwein, U., Jansen, M., Abduljabbar, A. S., & Abdelfattah, F. (2015). Dimensional comparison theory: Paradoxical relations between self-beliefs and achievements in multiple domains. *Learning and Instruction*, *35*, 16–32.

Marsh, H. W. & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology*, *81*, 59–77. doi: 10.1348/000709910X503501

Marsh, H. W. & Martin, A. J., Yeung, A. S. & Craven, R. G. (in press). Competence Self-Perceptions: A Cornerstone of Achievement Motivation and the Positive Psychology Movement. In A. J. Elliot (Ed.), *Handbook of competence and motivation*. New York: Guilford Publications.

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10,* 85–110. doi:10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*(3), 439–476. doi: 10.1080/10705510903008220

Marsh, H. W., & O'Mara, A. (2008). 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*(4), 542–552. http://dx.doi.org/10.1177/0146167207312313

Marsh, H. W., & O'Mara, A. J. (2010). Long-term total negative effects of school-average ability on diverse educational outcomes: Direct and indirect effects of the big-fish-little-pond effect. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 24*(1), 51–72.

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20*(3), 319–350.

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

Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9*, 275–300.

Marsh, H. W., Wen, Z., Hau, K. T., & Nagengast (2013). Structural equation models of latent interaction and quadratic effects. In G. Hancock & R. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (2nd ed.) (pp. 267–308). New York: Information Age Publishing.

Marsh, H. W. & Yeung, A. S. (1997). Coursework selection: The effects of academic self-concept and achievement. *American Educational Research Journal, 34*, 691–720.

Marshall, S. L., 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.

Martin, A. J. (2010). *Building Classroom Success: Eliminating Academic Fear and Failure.* New York: Continuum International Publishing Group.

Martin, A., Marsh, H. W., & Debus, R. (2001). Self-handicapping and defensive pessimism: Exploring a model of predictors and outcomes from a self-protection perspective. *Journal of Educational Psychology, 93*, 87–102.

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

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543.

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.

Mueller, C. M. & Dweck, C. S. (1998). Intelligence praise can undermine motivation and performance. *Journal of Personality and Social Psychology 75*, 33–52. doi:10.1037/0022-3514.75.1.33

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.

Muthén, L. K., & Muthén, B. (2008–14). *Mplus user's guide*. Los Angeles CA: Muthén & Muthén.

Nagengast, B., & Marsh, H. W. (2013). Motivation and engagement in science around the globe: Testing measurement invariance with multigroup SEMs across 57 countries using PISA 2006. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A Handbook of International Large–Scale Assessment Data Analysis* (pp. 317–344). Chapman & Hall, CRC Press.

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T., & Trautwein, U. (2011). Who took the "×" out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science*, *22*(8), 1058–1066. doi:10.1177/0956797611415540

Nicholls, J. G. (1976). Effort is virtuous, but it's better to have ability: Evaluative responses to perceptions of effort and ability. *Journal of Research in Personality, 10*(3), 306–315. http://dx.doi.org/10.1016/0092-6566(76)90020-9

Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child Development, 49*(3), 800–814.http://dx.doi.org/10.2307/1128250

Nicholls, J. G. (1979). Quality and equality in intellectual development: The role of motivation in education. *American Psychologist, 34*(11), 1071–1084. http://dx.doi.org/10.1037/0003-066X.34.11.1071

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review*, *91*, 328–346.

Nicholls, J. G. (1989). *The competitive ethos and democratic education*. Cambridge: Harvard University Press.

OECD (2015), The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence, PISA. OECD Publishing. http://dx.doi.org/10.1787/9789264229945-en

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

Parker, P. D., Marsh, H. W., Ciarrochi, J., Marshall, S., & Abduljabbar, A. S. (2014). Juxtaposing math self-efficacy and self-concept as predictors of long-term achievement outcomes. *Educational Psychology*, *34*(1), 29–48. doi:10.1080/01443410.2013.797339

Pekrun, R. (1993). Facets of students' academic motivation: A longitudinal expectancy-value approach. In M. Maehr & P. Pintrich (Eds.), *Advances in motivation and achievement, 8*, (pp. 139–189). Greenwich, CT: JAI Press.

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.

Piaget, J. & Cook, M. (Trans), (1952). The origins of intelligence in children. New York: WW Norton & Co.

Pinxten, M., De Fraine, B., Van Damme, J., & D'Haenens, E. (2010). Causal ordering of academic self-concept and achievement: Effects of type of achievement measure. *British Journal of Educational Psychology, 80*(4), 689–709. doi:10.1348/000709910X493071

Pinxten, M., Marsh, H. W., De Fraine, B., Van Den Noortgate, W., & Van Damme, J. (2014). Enjoying mathematics or feeling competent in mathematics? Reciprocal effects on mathematics achievement and perceived math effort expenditure. *British Journal of Educational Psychology*, *84*(1), 152–174. doi:10.1111/bjep.12028

Ryan, R. M. & Deci, E. L. (2000) Self-determination theory and the facilitation of intrinsic motivation, social development and well-being. *American Psychologist, 55*, 68–78.

Salili, F., & Hau, K.-T. (1994). The effect of teachers' evaluative feedback on Chinese students' perception of ability: A cultural and situational analysis. *Educational Studies*, *20*(2), 223-236. http://dx.doi.org/10.1080/0305569940200206

Schorr, L. B. (2016). Reconsidering evidence: What it means and how we use it. Stanford Social Innovation Review. http://ssir.org/articles/entry/reconsidering_evidence_what_it_means_and_how_we_use_it

Trautwein, U., Lüdtke, O., Marsh, H. W., & Nagy, G. (2009). Within-school social comparison: How students perceive the standing of their class predicts academic self-concept. *Journal of Educational Psychology, 101*(4), 853–866. http://dx.doi.org/10.1037/a0016306

Trautwein, U., Lüdtke, O., Roberts, B. W., Schnyder, I., & Niggli, A. (2009). Different forces, same consequence: Conscientiousness and competence beliefs are independent predictors of academic effort and achievement. *Journal of Personality and Social Psychology, 97*, 1115–1128. doi:10.1037/a0017048

Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy–value theory: A latent interaction modeling study. *Journal of Educational Psychology, 104*, 763–777. doi:10.1037/a0027470

Trautwein, U., Schnyder, I., Niggli, A., Neumann, M., & Lüdtke, O. (2009). Chameleon effects in homework research: The homework-achievement association depends on the measures used and the level of analysis chosen. *Contemporary Educational Psychology, 34*, 77–88.

Vallerand, R. J. (2015). *The psychology of passion: A dualistic model*. Oxford University Press.

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.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–69.

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*, Beiheft, *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.

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.

Wouters, S., Colpin, H., Van Damme, J., & Verschueren, K. (2015). Endorsing achievement goals exacerbates the big-fish-little-pond effect on academic self-concept. *Educational Psychology, 35*(2), 252–270. http://dx.doi.org/10.1080/01443410.2013.822963

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: Generalised item response modeling software*.

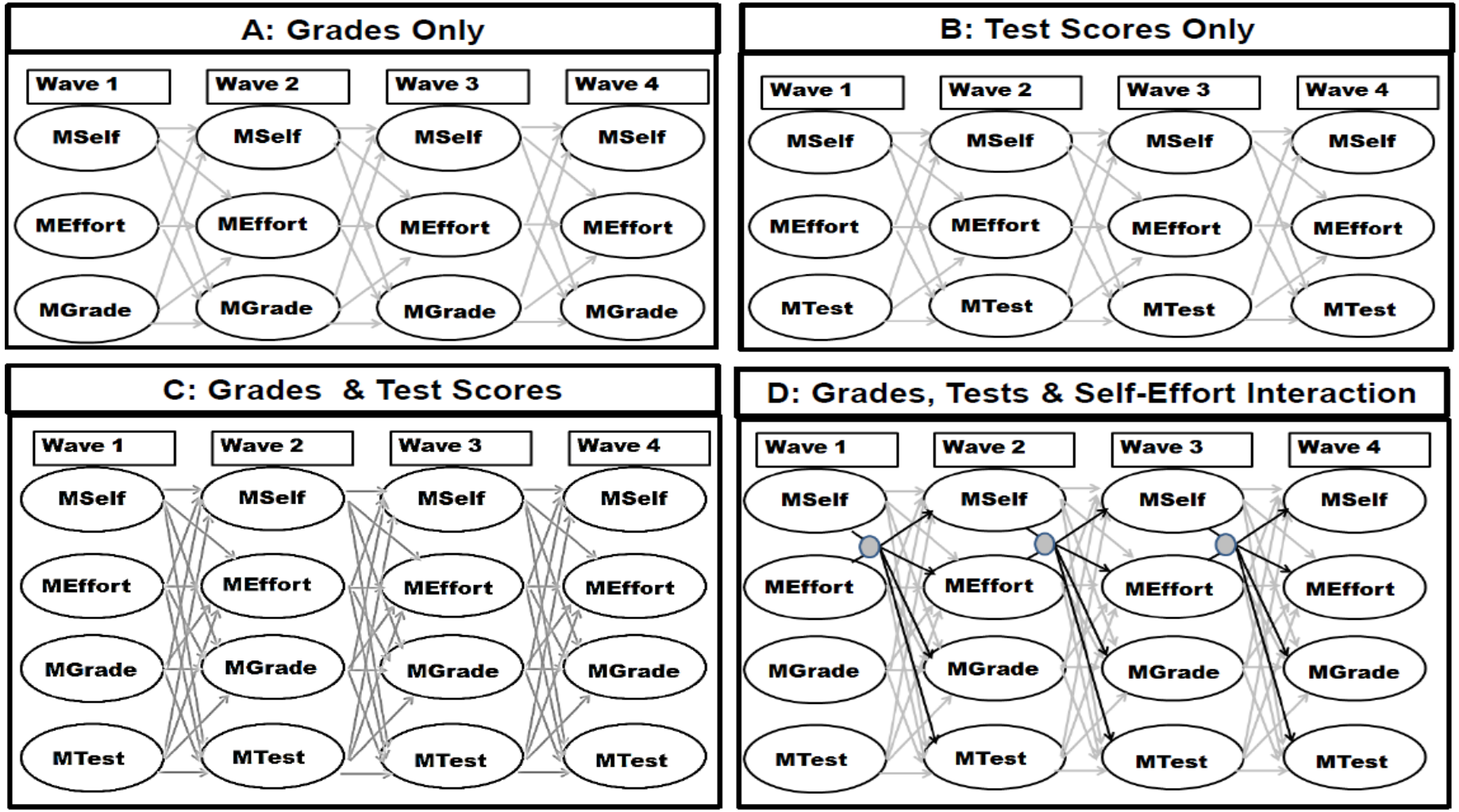

Figure 1. Alternative structural equation models evaluating reciprocal effects among math self-concept (MSelf), math effort (MEffort), school grades in mathematics (MGrade) and standardized test scores in mathematics (MTest). In Figure 1D the latent interactions between math self-concept and math effort (represented by the *grey* circle) at each wave are added to the model (black paths represent interaction effects). See parameter estimates in Table 2.

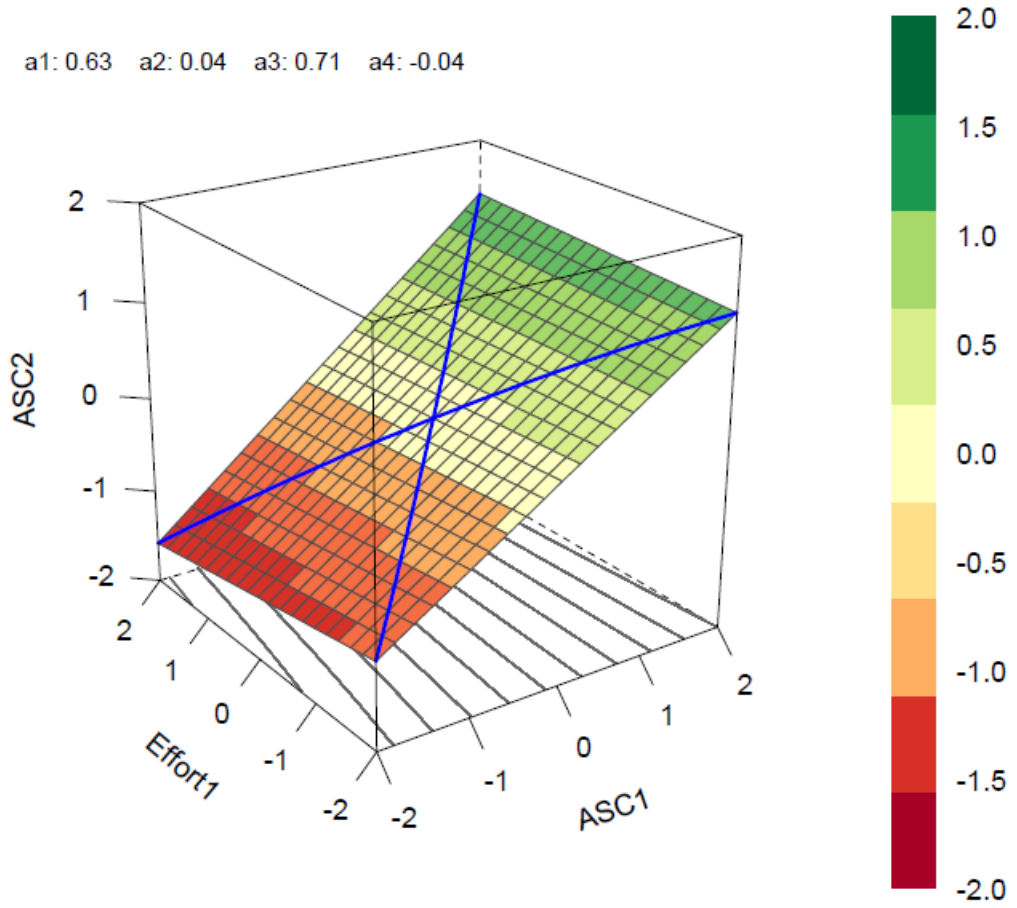

Figure 2. Interaction effects relating Effort, academic self-concept (ASC) and their interaction at Wave 1 to ASC at Wave 2. In support of predictions based on the double-edged sword hypothesis, the effects of prior effort on subsequent ASC are negative for low-ASC students but increasingly more positive (or less negative) for students with more positive ACSs. In support of the developmental equilibrium of this pattern of results, the first-order and interaction effects are invariant across the four waves of data covering the first four years of secondary school. In support of the robustness of the findings, the results—including support for developmental equilibrium—are also invariant over three school-types: high-, medium- and low-track schools based on achievement and school performance in primary school.

Table 1

Goodness of Fit for Alternative Models of Academic Self-concept (ASC), Effort, Test Scores and School Grades Over Four Waves of Data and Three School-Track Types

	Chi-	Df	CFI	TLI	RMSEA	Model Description
Confirmatory Factor Analysis Models of Factorial Invariance over Groups and Waves						
Model 1A Configural	4879	2904	.960	.953	.024	No Correlated Uniqueness; No Invariance
Model 1B Metric	4947	2968	.960	.954	.024	M1A + factor loading invariance over groups
Model 1C Scalar	5120	3032	.957	.953	.025	M1B + intercept invariance over groups
Model 2A Configural	3801	2724	.978	.973	.019	M1A + Correlated Uniqueness; No Invariance
Model 2B Metric	3868	2788	.978	.973	.018	M2A + factor loading invariance groups
Model 2C Scalar	4015	2852	.976	.972	.019	M2B + intercept invariance over groups
Model 3A Configural	3926	2796	.977	.972	.019	M2A + invariance over time
Model 3B Metric	3944	2812	.977	.972	.019	M3A + factor loading invariance over groups
Model 3C Scalar	4092	2876	.975	.971	.019	M3B + intercept invariance over groups
Model 4 Scalar	4631	3128	.969	.967	.021	M3C + all factor variances constrained to be 1.0 and factor correlations invariant over groups
Structural Equation Models: Tests of Invariance of Path Coefficients over Multiple School-Type Groups & Waves						
Model 5a [c]	4230	2984	.974	.971	.019	Path Coeff no Invar over Multiple Groups/Waves
Model 5b	4492	3104	.971	.969	.020	Path Coeff Invar over Multiple Groups
Model 5c	4521	3092	.971	.968	.020	Path Coeff Invar over Multiple Waves
Model 5d [c]	4631	3140	.969	.967	.020	Path Coeff Invar over Multiple Groups/Waves
Structural Equation Models: Tests of Invariance of Path Coefficients over Multiple Gender Groups & Waves						
Model 6a	3060	1976	.977	.974	.018	Path Coeff no Invar over Multiple Groups/Waves
Model 6b	3188	2037	.976	.974	.018	Path Coeff Invar over Multiple Groups
Model 6c	3377	2048	.972	.970	.019	Path Coeff Invar over Multiple waves
Model 6d	3453	2072	.971	.969	.020	Path Coeff Invar over Multiple Groups/Waves
Structural Equation Models[b] (Grades only, Test Scores only)						
Model 7A	3944	2639	.970	.967	.021	Model 6d for Grades Only (excluding test scores)
Model 7B	3759	2639	.974	.972	.020	Model 6d for Test Scores Only (excluding grades)

Note. Chi = chi-square; df = degrees of freedom; CFI = Comparative fit index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation. FL = factor loadings. Mplus syntax for the selected Models is presented in the Supplemental Materials.

[a] A priori path coefficients include all test-retest stability paths relating measures of the same construct from different waves (i.e., lag-1 to lag-3 paths for Wave 4 factors) but only lag-1 cross-paths relating different constructs. [b] Path coefficients, as well as factor loadings, are constrained to be equal across the four waves of data; factor loadings, path coefficients and intercepts are constrained to be invariant across the three school-type groups. [c] Syntax for these models presented in Supplemental Materials (section 4).

Table 2

Factor Structure Across Four Waves of Data (W1–W4)

	Math Self-concept				Math Effort				Math Test Scores				Math School Grades			
	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4
Items							Factor loadings									
SC1	.971	.971	.971	.971												
SC2	.802	.802	.802	.802												
SC3	.825	.825	.825	.825												
SC4	.868	.868	.868	.868												
SC5	.856	.856	.856	.856												
SC6	.686	.686	.686	.686												
EFF1					.846	.846	.846	.846								
EFF2					.841	.841	.841	.841								
EFF3					.775	.775	.775	.775								
EFF4					.653	.653	.653	.653								
Test									1	1	1	1				
Grades													1	1	1	1
Factors							Factor Correlations									
Self-Concept (SC)																
W1SC																
W2SC	.683															
W3SC	.554	.672														
W4SC	.557	.641	.750													
Effort (Eff)																
W1EFF	.378	.206	.138	.144												
W2EFF	.169	.292	.171	.134	.476											
W3EFF	.136	.206	.326	.223	.317	.516										
W4EFF	.125	.180	.232	.351	.286	.449	.560									
Test Scores (Tst)																
W1Tst	.451	.427	.381	.383	.130	.062	.074	.056								
W2Tst	.401	.439	.414	.409	.099	.139	.126	.136	.664							
W3Tst	.363	.418	.444	.454	.074	.103	.182	.133	.604	.676						
W4Tst	.377	.444	.458	.513	.090	.149	.173	.230	.591	.691	.737					
School Grades (Grd)																
W1Grd	.504	.477	.402	.413	.123	.076	.118	.148	.546	.514	.522	.518				
W2Grd	.423	.576	.426	.446	.076	.108	.131	.148	.497	.509	.519	.517	.684			
W3Grd	.304	.392	.561	.503	.062	.104	.168	.189	.422	.490	.525	.538	.548	.598		
W4Grd	.320	.346	.436	.638	.070	.066	.160	.238	.353	.415	.500	.527	.514	.540	.642	

Note. In this common metric model (M4, Table 1) factor loadings are invariant over multiple waves and the three school-type groups; factor covariances are invariant over the three school types, and factor variances are constrained to have a mean of 1.0 across the three school types (See Supplemental Materials for Mplus syntax). Across the four waves, Math self-concept was measured with the same six items (MSC1–MSC6) and math effort was measured with the same four items (MEff1–MEff4), whilst math test scores and school grades (marks) were based on a single score for each wave.

Table 3

Selected Path Coefficients for Alternative Models of Reciprocal Effects Between Academic Self-Concept, Effort, Standardized Test Scores, School Grades and ASC-by-Effort Interaction

Dependent Variable	**Grades & Test Scores (Model 6D)**		**Grades Only (Model 7A)**		**Test Scores Only (Model 7B)**	
Independent Variable	No interaction	With Interaction	No interaction	With Interaction	No interaction	With Interaction
TMSelf$_i$		274411.188				
MSelf$_{i-1}$	.568/.020	.593/.020	.586/.020	.659/.018	.611/.019	.677/.019
MEffort$_{i-1}$	-.029/.013	-.025/.013	-.029/.013	-.040/.014	-.030/.013	-.029/.014
MSelfxEff$_{i-1}$		.031/.011		.036/.013		.036/.012
MTEST$_{i-1}$	.112/.013	.104/.011			.131/.011	.083/.011
Mgrd$_{i-1}$	.075/.015	.066/.014	.117/.013	.092/.011		
MEffort$_i$						
MSelf$_{i-1}$	.010/.020	.002/.021	.009/.020	-.006/.018	.040/.018	.021/.018
MEffort$_{i-1}$	.456/.021	.467/.021	.456/.021	.595/.022	.456/.021	.539/.021
MSelfxEff$_{i-1}$		-.022/.015		.008/.014		.009/.015
MTEST$_{i-1}$	-.007/.016	-.002/.019			.019/.015	.019/.015
MGrade$_{i-1}$	.079/.017	.080/.017	.076/.016	.049/.013		
MGradeT$_i$						
MSelf$_{i-1}$	.048/.014	.047/.015	.089/.014	.105/.018		
MEffort$_{i-1}$	.003/.013	.017/.013	.003/.013	.005/.018		
MSelfxEff$_{i-1}$		.044/.012		.056/.017		
MGrade$_{i-1}$	.451/.015	.453/.015	.514/.015	.511/.015		
MTEST$_{i-1}$	.189/.013	.189/.013				
MTEST$_i$						
MSelf$_{i-1}$	.063/.013	.065/.014			.137/.013	.150/.015
MEffort$_{i-1}$	.003/.011	.005/.012			-.006/.011	.005/.015
MSelfxEff$_{i-1}$		-.001/.011				.012/.015
MGrade$_{i-1}$	.164/.012	.164/.012				
MTEST$_{i-1}$	.458/.013	.460/.013			.505/.013	.502/.013

Note. Presented here are the four dependent (outcome) variables for Wave$_I$ (I = 2–4) and the ratio of the path coefficients over the standard errors for the corresponding predictor variables for Wave $_{i-1}$. Path coefficients are invariant over waves (see Figure 1). If the ratio of the path coefficient over its standard error is greater than 2, the path coefficient is statistically significant ($p < .05$).

Supplemental Materials

Supplemental Materials 1: Item wording for the multi-item academic self-concept and effort constructs considered here

2. Supplemental Materials 2: Sampling Design and Missing Data

3. Supplemental Table 1: Selected Path Coefficients For Alternative Models of Reciprocal Effects Between Academic self-concept, Effort, Standardized Test Scores, School Grades and ASC-by-Effort Interaction. Based on models with only lag-1 stability coefficients.

4. Supplemental Materials 3: Examples of Mplus syntax for:

- **Model 3C (Table 1).**
- **Model 5a (Table 1)**
- **Model 5d (Table 1)**

References in Supplemental Materials

Supplemental Materials 1: Item wording for the multi-item academic self-concept and effort constructs considered here

Math Self-concept Items Administered in Each Wave

- In math, I am a talented student.
  In Mathematik bin ich ein begabter Schüler.
- It is easy for me to understand things in math.
  Es fällt mir leicht, in Mathematik etwas zu verstehen.
- I can solve math problems well.
  Mathe-Aufgaben kann ich gut lösen.
- It is easy to me to write tests/exams in math.
  Es fällt mir leicht, Mathe-Schulaufgaben/Proben zu schreiben.
- It is easy to me to learn something in math.
  Es fällt mir leicht, etwas für Mathematik zu lernen.
- If the math teacher asks a question, I can answer it correctly most of the time.
  Wenn der Mathe-Lehrer eine Frage stellt, weiß ich meistens die richtige Antwort.

Math Effort Items Administered in Each Wave

- In math, I invest much effort to understand everything.
  In Mathe gebe ich mir viel Mühe, alles zu verstehen.
- In math, I try my best to do everything as well as possible.
  In Mathe versuche ich sehr, alles so gut wie möglich zu machen.
- I do my math homework as well as I can.
  Ich mache meine Mathe-Hausaufgaben so gut wie möglich.
- When we take a math test I give all my effort.
  Wenn wir eine Mathe-Schulaufgabe/Probe schreiben, hole ich das letzte aus mir raus.

Supplemental Materials 2: Sampling Design and Missing Data

Sampling Design

At the first assessment (grade 5), the sample comprised 2,070 students from 42 schools (49.6% female, mean age = 11.7 years; 37.2% lower-track school students, 27.1% intermediate-track school students, and 35.7% higher-track school students). In each subsequent year, the study not only tracked the students who had participated in the previous assessment(s), but also included those students who had not yet participated in the study, but had become students of PALMA classrooms at the time of the assessment (for more details on sampling procedures, see Pekrun et al., 2007). This sampling strategy resulted in the following sample sizes for the subsequent years: 2,059 students in grade 6 (237 new students; 50.0% female, mean age = 12.7 years); 2,397 students at grade 7 (631 new students; 50.1% female, mean age = 13.7 years); 2,410 students at grade 8 (206 new students; 50.5% female, mean age = 14.8 years). Across all assessments (i.e., grades 5 to 8), a total of 3,421 students (49.7% female) took part in the study. The breakdown of the number of waves of data completed by each student was: 1 (17.0%), 2 (27.1%), 3 (10.8%) and 4 (45.2%). We note, however, that much of the apparently large level of missing data is due to the inclusion of new students in subsequent waves who were not included in earlier waves. Indeed, of the 2,070 students in Wave 1, the number of waves completed was 1 (215, 10%), 2 (256, 12%), 3 (178, 9%) and 4 (1421, 69%). Thus, more than 2/3 of the students participating in Wave 1 participated in all four waves.

Missing Data

Due to the longitudinal design of the study, there is a certain proportion of missing data. Given the structure of participation in the study, listwise deletion methods would fail to make use of the information from those participants who did not take part in all of the assessments. This issue is critical for making full use of the PALMA dataset, as a substantial number of students started to participate in the study after the first assessment. In addition, listwise deletion produces unbiased estimates only under the assumption that the missing data occur completely at random, which is not realistic in many cases (Enders, 2006). Accordingly, as noted in the main text, we applied the full information maximum likelihood method (FIML; see Schafer & Graham, 2002). This approach uses the likelihood function, which considers individual patterns of missing data, and which has the important advantage of rendering results that make use of the full information from the entire sample. Moreover, as compared to listwise deletion, FIML provides unbiased parameter estimates under a considerably weaker assumption: namely, that missingness is not dependent on the missing values themselves (Enders, 2006).

Indeed, although it has been common practice in the past to retain only participants who provided data for all the multiple waves—which technically represents the listwise deletion of participants who participated in a single wave—there is now an emerging consensus within the statistical community that the best data-analytic method for dealing with missing data follows a simple yet fundamental principle: Use all of the available data (Newman, 2009, p. 11). In longitudinal studies, this consensus translates into the need to include all participants, irrespective of completeness (e.g., Bollen & Curran, 2004; Sinha, Laird, & Fitzmaurice, 2010; Enders, 2010; Graham, 2009; Hedeker & Gibbons, 2006).

Not surprisingly, in our study, missing data patterns were consistent across variables within each wave, in that missingness was largely a function of the wave rather than the variable: That is, there was very little missing data within each wave for students who participated in that wave. The standard approach to missing data is FIML, based on the same assumption of missing at random (MAR). We also note that the critical issue is not whether the results are similar for those with and without missing data, but rather the missingness process itself. Indeed, except under very restrictive situations, there is no reason to expect that listwise deletion (i.e., results based on those with no missing data) would provide meaningful results or an appropriate basis of comparison for more appropriate methods. Importantly, the appropriateness of the assumptions underlying FIML is substantially enhanced by the availability of four waves of data, in that if missingness for any variable at $wave_i$ is a function of a variable at $wave_i$, then this is likely to be picked up through the

availability of the same variable at $wave_k$ ($k \neq i$).

More specifically, as emphasized in classic discussions of missing data (e.g., Newman, 2014), under the MAR assumption that is the basis of FIML, missingness is allowed to be conditional on all variables included in the analyses, but does not depend on the values of variables that are missing. In a longitudinal panel design, this implies that missing values can be conditional on the values of the variable collected in a different wave. This makes it unlikely that MAR assumptions are violated, as the key situation of not MAR is when missingness is related to the variable itself. Hence, having so many waves of parallel data provides strong protection against this violation of the MAR assumption.

3. Supplemental Table 1

Selected Path Coefficients For Alternative Models of Reciprocal Effects Between Academic Self-Concept, Effort, Standardized Test Scores, School Grades and ASC-by-Effort Interaction. Based on Models with Only lag-1 Stability Coefficients

Dependent Var	**Grades & Test Scores (Model 8)**		**Grades Only (Model 7A)**		**Test Scores Only (Model 7B)**	
Independent Var	No interaction	With Interaction	No interaction	With Interaction	No interaction	With Interaction
TMSelf$_i$						
MSelf$_{i-1}$	.665/.017	.734/.015	.698/.017	.725/.016	.689/.015	.764/.014
MEffort$_{i-1}$	-.038/.013	-.037/.012	-.038/.013	-.043/.013	-.038/.013	-.048/.013
MSelfxEff$_{i-1}$		.030/.011		.027/.011		.033/.012
MTEST$_{i-1}$	.130/.012	.087/.012			.146/.011	.086/.011
Mgrd$_{i-1}$	.062/.015	.047/.012	.111/.013	.095/.013		
MEffort$_i$						
MSelf$_{i-1}$	.008/.020	-.001/.019	.007/.020	-.005/.020	.039/.018	.017/.018
MEffort$_{i-1}$	.535/.019	.645/.017	.535/.019	.552/.019	.532/.019	.630/.018
MSelfxEff$_{i-1}$		.019/.016		-.026/.016		.010/.015
MTEST$_{i-1}$	-.008/.016	-.057/.013			.016/.015	-.020/.014
MGrade$_{i-1}$	.080/.017	.075/.015	.076/.016	.082/.016		
MGradeT$_i$						
MSelf$_{i-1}$	.060/.014	.087/.018	.113/.014	.119/.014		
MEffort$_{i-1}$	.002/.013	.002/.017	-.003/.013	.005/.013		
MSelfxEff$_{i-1}$		.039/.015		.041/.012		
MGrade$_{i-1}$	.492/.015	.484/.015	.577/.014	.572/.014		
MTEST$_{i-1}$	.217/.013	.207/.013				
MTEST$_i$						
MSelf$_{i-1}$	.091/.014	.110/.017			.176/.013	.197/.015
MEffort$_{i-1}$	-.008/.011	-.007/.015			-.020/.011	-.020/.015
MSelfxEff$_{i-1}$		.010/.014				.009/.015
MGrade$_{i-1}$	.186/.013	.184/.012				
MTEST$_{i-1}$	.552/.013	.544/.013			.611/.012	.604/.013

Note. Presented here are the four dependent (outcome) variables for Wave $_I$ (I = 2–4) and the ratio of the path coefficients over the standard errors for the corresponding predictor variables for Wave $_{i-1}$. Path coefficients are invariant over waves (see Figure 1). If the ratio of the path coefficient over its standard error is greater than 2, the path coefficient is statistically significant ($p < .05$). This model differs from Model 3 in the main text in that only lag-1 stability coefficients (i.e., matching variables only in the path from the immediately preceding wave) are included. The pattern and size of the paths are similar to those in Table 3. The main difference is that the lag-1 stability coefficients here are systematically higher, because no lag-2 or lag-3 stability paths are included in the model.

Supplementary Materials 4: Examples of Mplus syntax

Model 3C (Table 1).

TITLE: **Model 3C (Table 1).**

```
missing all (-9);
USEVARIABLES ARE
   MSC1_1 MSC2_1 MSC3_1 MSC4_1 MSC5_1 MSC6_1
   MSC1_2 MSC2_2 MSC3_2 MSC4_2 MSC5_2 MSC6_2
   MSC1_3 MSC2_3 MSC3_3 MSC4_3 MSC5_3 MSC6_3
   MSC1_4 MSC2_4 MSC3_4 MSC4_4 MSC5_4 MSC6_4
   eff1_1,eff2_1,eff3_1,eff4_1
   eff1_2,eff2_2,eff3_2,eff4_2
   eff1_3,eff2_3,eff3_3,eff4_3
   eff1_4,eff2_4,eff3_4,eff4_4
   Mach_1 Mach_2 Mach_3 Mach_4
   MGrd5 MGrd6 MGrd7 MGrd8
   schltypM;
GROUPING IS SCHLTYPM  (1 = high 2=med 3=low);
ANALYSIS:
ESTIMATOR=MLR;
PROCESSORS = 4;
model:
   T1MSC BY MSC1_1-MSC6_1(flsc1-flsc6);
   T2MSC BY MSC1_2-MSC6_2(flsc1-flsc6);
   T3MSC BY MSC1_3-MSC6_3(flsc1-flsc6);
   T4MSC BY MSC1_4-MSC6_4(flsc1-flsc6);
   T1eff BY eff1_1-eff4_1(fLEf1-fLEf4);
   T2eff BY eff1_2-eff4_2(fLEf1-fLEf4);
   T3eff BY eff1_3-eff4_3(fLEf1-fLEf4);
   T4eff BY eff1_4-eff4_4(fLEf1-fLEf4);
   ach_1   by Mach_1    ;    Mach_1 @0;
   ach_2   by Mach_2    ;    Mach_2 @0;
   ach_3   by Mach_3    ;    Mach_3 @0;
   ach_4   by Mach_4    ;    Mach_4 @0;
   Grd5   by  MGrd5    ;     MGrd5 @0;
   Grd6   by  MGrd6    ;     MGrd6 @0;
   Grd7   by  MGrd7    ;     MGrd7 @0;
   Grd8   by  MGrd8    ;     MGrd8 @0;
   MSC1_1-MSC6_1 pwith MSC1_2-MSC6_2  ;
   MSC1_1-MSC6_1 pwith MSC1_3-MSC6_3  ;
   MSC1_1-MSC6_1 pwith MSC1_4-MSC6_4  ;
   MSC1_2-MSC6_2   pwith MSC1_3-MSC6_3  ;
   MSC1_2-MSC6_2   pwith MSC1_4-MSC6_4  ;
   MSC1_3-MSC6_3   pwith MSC1_4-MSC6_4  ;
   eff1_1-eff4_1 pwith eff1_2-eff4_2  ;
   eff1_1-eff4_1 pwith eff1_3-eff4_3  ;
   eff1_1-eff4_1 pwith eff1_4-eff4_4  ;
   eff1_2-eff4_2   pwith eff1_3-eff4_3  ;
   eff1_2-eff4_2   pwith eff1_4-eff4_4  ;
   eff1_3-eff4_3   pwith eff1_4-eff4_4  ;
OUTPUT: svalues TECH1; stdyx; tech4; sampstat; mod;
```

Model 5a (Table 1)

```
TITLE: Model 5a (table 1)

      missing all (-9);

         USEVARIABLES ARE
    MSC1_1 MSC2_1 MSC3_1 MSC4_1 MSC5_1 MSC6_1
     MSC1_2 MSC2_2 MSC3_2 MSC4_2 MSC5_2 MSC6_2
     MSC1_3 MSC2_3 MSC3_3 MSC4_3 MSC5_3 MSC6_3
     MSC1_4 MSC2_4 MSC3_4 MSC4_4 MSC5_4 MSC6_4
    eff1_1,eff2_1,eff4_1,eff7_1
    eff1_2,eff2_2,eff4_2,eff7_2
    eff1_3,eff2_3,eff4_3,eff7_3
    eff1_4,eff2_4,eff4_4,eff7_4
    Mach_1 Mach_2 Mach_3 Mach_4
     MGrd5 MGrd6 MGrd7 MGrd8
    ;
   GROUPING IS SCHLTYPM  (1 = high 2=med 3=low);
       define:
    MACH_1  =   MACH_1  /(((0.803 + 0.554 + 0.693)/3) **.5);
    MACH_2  =   MACH_2  /(((0.742 + 0.489 + 0.517)/3) **.5);
    MACH_3  =   MACH_3  /(((0.643 + 0.469 + 0.601)/3) **.5);
    MACH_4  =   MACH_4  /(((0.562 + 0.496 + 0.486)/3) **.5);
    MGRD5   =   MGRD5   /(((0.876 + 0.872 + 1.034)/3) **.5);
    MGRD6   =   MGRD6   /(((0.968 + 0.917 + 0.989)/3) **.5);
    MGRD7   =   MGRD7   /(((1.138 + 0.835 + 0.978)/3) **.5);
    MGRD8   =   MGRD8   /(((0.960 + 0.949 + 1.092)/3) **.5);
       ANALYSIS:
     ESTIMATOR=MLR;
     PROCESSORS = 4;
          model:
      t1msc BY msc1_1@0.97123 (flsc1);
      t1msc BY msc2_1*0.80469 (flsc2);
      t1msc BY msc3_1*0.82591 (flsc3);
      t1msc BY msc4_1*0.86844 (flsc4);
      t1msc BY msc5_1*0.85587 (flsc5);
      t1msc BY msc6_1*0.68621 (flsc6);
      t2msc BY msc1_2@0.97123 (flsc1);
      t2msc BY msc2_2*0.80469 (flsc2);
      t2msc BY msc3_2*0.82591 (flsc3);
      t2msc BY msc4_2*0.86844 (flsc4);
      t2msc BY msc5_2*0.85587 (flsc5);
      t2msc BY msc6_2*0.68621 (flsc6);
      t3msc BY msc1_3@0.97123 (flsc1);
      t3msc BY msc2_3*0.80469 (flsc2);
      t3msc BY msc3_3*0.82591 (flsc3);
      t3msc BY msc4_3*0.86844 (flsc4);
      t3msc BY msc5_3*0.85587 (flsc5);
      t3msc BY msc6_3*0.68621 (flsc6);
      t4msc BY msc1_4@0.97123 (flsc1);
      t4msc BY msc2_4*0.80469 (flsc2);
      t4msc BY msc3_4*0.82591 (flsc3);
      t4msc BY msc4_4*0.86844 (flsc4);
      t4msc BY msc5_4*0.85587 (flsc5);
      t4msc BY msc6_4*0.68621 (flsc6);
      t1eff BY eff1_1@0.84649 (flef1);
      t1eff BY eff2_1*0.84109 (flef2);
      t1eff BY eff4_1*0.77134 (flef3);
      t1eff BY eff7_1*0.64324 (flef4);
      t2eff BY eff1_2@0.84649 (flef1);
      t2eff BY eff2_2*0.84109 (flef2);
      t2eff BY eff4_2*0.77134 (flef3);
```

```
  t2eff BY eff7_2*0.64324 (flef4);
  t3eff BY eff1_3@0.84649 (flef1);
  t3eff BY eff2_3*0.84109 (flef2);
  t3eff BY eff4_3*0.77134 (flef3);
  t3eff BY eff7_3*0.64324 (flef4);
  t4eff BY eff1_4@0.84649 (flef1);
  t4eff BY eff2_4*0.84109 (flef2);
  t4eff BY eff4_4*0.77134 (flef3);
  t4eff BY eff7_4*0.64324 (flef4);

MSC1_1-MSC6_1 pwith MSC1_2-MSC6_2  ;
MSC1_1-MSC6_1 pwith MSC1_3-MSC6_3  ;
MSC1_1-MSC6_1 pwith MSC1_4-MSC6_4  ;

   MSC1_2-MSC6_2   pwith MSC1_3-MSC6_3  ;
   MSC1_2-MSC6_2   pwith MSC1_4-MSC6_4  ;

   MSC1_3-MSC6_3   pwith MSC1_4-MSC6_4  ;
  eff1_1-eff7_1 pwith eff1_2-eff7_2  ;
eff1_1-eff7_1 pwith eff1_3-eff7_3  ;
eff1_1-eff7_1 pwith eff1_4-eff7_4  ;
   eff1_2-eff7_2   pwith eff1_3-eff7_3  ;
   eff1_2-eff7_2   pwith eff1_4-eff7_4  ;
   eff1_3-eff7_3   pwith eff1_4-eff7_4  ;

T1MSC  with     T1eff MGrd5 Mach_1 ;
T2MSC  with     T2eff MGrd6 Mach_2;
T3MSC  with     T3eff MGrd7 Mach_3;
T4MSC  with     T4eff MGrd8 Mach_4;

Mach_1    with   T1eff MGrd5  ;
Mach_2    with   T2eff MGrd6 ;
Mach_3    with   T3eff MGrd7 ;
Mach_4    with   T4eff MGrd8 ;

 T1eff   with MGrd5  ;
 T2eff   with MGrd6 ;
 T3eff   with MGrd7 ;
 T4eff   with MGrd8 ;

T2MSC on T1MSC Mach_1 MGrd5 T1eff  ;
T3MSC on T2MSC Mach_2 MGrd6 T2eff
        T1MSC  ;
T4MSC on T3MSC Mach_3 MGrd7 T3eff
         T2MSC T1MSC ;

T2eff on T1MSC T1eff Mach_1 MGrd5 ;
T3eff on T2MSC T2eff Mach_2 MGrd6
        T1Eff  ;
T4eff on T3MSC T3eff Mach_3 MGrd7
        T2Eff T1Eff  ;
Mach_2 on T1MSC Mach_1 MGrd5 T1eff;
Mach_3 on T2MSC Mach_2 MGrd6 T2eff
        Mach_1  ;
Mach_4 on T3MSC Mach_3 MGrd7 T3eff
        Mach_2 Mach_1  ;

MGrd6 on T1MSC Mach_1 MGrd5 T1eff  ;
MGrd7 on T2MSC Mach_2 MGrd6 T2eff
        MGrd5  ;
MGrd8 on T3MSC Mach_3 MGrd7 T3eff
        MGrd6 MGrd5  ;
```

```
T1MSC-T1eff WITH T2MSC-T4eff@0;
T2MSC-T2eff WITH T3MSC-T4eff@0;
T3MSC-T3eff WITH T4MSC-T4eff@0;

T1MSC-T1eff WITH  Mach_1-MGrd8@0;
T2MSC-T2eff WITH  Mach_1-MGrd8@0;
T3MSC-T3eff WITH  Mach_1-MGrd8@0;

OUTPUT: svalues TECH1; stdyx; tech4; sampstat; mod;
```

Model 5d Table 1)

```
 missing all (-9);

    USEVARIABLES ARE
MSC1_1 MSC2_1 MSC3_1 MSC4_1 MSC5_1 MSC6_1
 MSC1_2 MSC2_2 MSC3_2 MSC4_2 MSC5_2 MSC6_2
 MSC1_3 MSC2_3 MSC3_3 MSC4_3 MSC5_3 MSC6_3
 MSC1_4 MSC2_4 MSC3_4 MSC4_4 MSC5_4 MSC6_4
eff1_1,eff2_1,eff4_1,eff7_1
eff1_2,eff2_2,eff4_2,eff7_2
eff1_3,eff2_3,eff4_3,eff7_3
eff1_4,eff2_4,eff4_4,eff7_4
Mach_1 Mach_2 Mach_3 Mach_4
 MGrd5 MGrd6 MGrd7 MGrd8
;

GROUPING IS SCHLTYPM  (1 = high 2=med 3=low);
   define:
 MACH_1  =   MACH_1  /(((0.803 + 0.554 + 0.693)/3) **.5);
 MACH_2  =   MACH_2  /(((0.742 + 0.489 + 0.517)/3) **.5);
 MACH_3  =   MACH_3  /(((0.643 + 0.469 + 0.601)/3) **.5);
 MACH_4  =   MACH_4  /(((0.562 + 0.496 + 0.486)/3) **.5);
 MGRD5   =   MGRD5   /(((0.876 + 0.872 + 1.034)/3) **.5);
 MGRD6   =   MGRD6   /(((0.968 + 0.917 + 0.989)/3) **.5);
 MGRD7   =   MGRD7   /(((1.138 + 0.835 + 0.978)/3) **.5);
 MGRD8   =   MGRD8   /(((0.960 + 0.949 + 1.092)/3) **.5);

   ANALYSIS:
  ESTIMATOR=MLR;
  PROCESSORS = 4;

      model:
  t1msc BY msc1_1@0.97123 (flsc1);
  t1msc BY msc2_1*0.80469 (flsc2);
  t1msc BY msc3_1*0.82591 (flsc3);
  t1msc BY msc4_1*0.86844 (flsc4);
  t1msc BY msc5_1*0.85587 (flsc5);
  t1msc BY msc6_1*0.68621 (flsc6);
  t2msc BY msc1_2@0.97123 (flsc1);
  t2msc BY msc2_2*0.80469 (flsc2);
  t2msc BY msc3_2*0.82591 (flsc3);
  t2msc BY msc4_2*0.86844 (flsc4);
  t2msc BY msc5_2*0.85587 (flsc5);
  t2msc BY msc6_2*0.68621 (flsc6);
  t3msc BY msc1_3@0.97123 (flsc1);
  t3msc BY msc2_3*0.80469 (flsc2);
  t3msc BY msc3_3*0.82591 (flsc3);
  t3msc BY msc4_3*0.86844 (flsc4);
  t3msc BY msc5_3*0.85587 (flsc5);
  t3msc BY msc6_3*0.68621 (flsc6);
  t4msc BY msc1_4@0.97123 (flsc1);
  t4msc BY msc2_4*0.80469 (flsc2);
  t4msc BY msc3_4*0.82591 (flsc3);
  t4msc BY msc4_4*0.86844 (flsc4);
  t4msc BY msc5_4*0.85587 (flsc5);
  t4msc BY msc6_4*0.68621 (flsc6);
  t1eff BY eff1_1@0.84649 (flef1);
  t1eff BY eff2_1*0.84109 (flef2);
  t1eff BY eff4_1*0.77134 (flef3);
  t1eff BY eff7_1*0.64324 (flef4);
  t2eff BY eff1_2@0.84649 (flef1);
```

```
  t2eff BY eff2_2*0.84109 (flef2);
  t2eff BY eff4_2*0.77134 (flef3);
  t2eff BY eff7_2*0.64324 (flef4);
  t3eff BY eff1_3@0.84649 (flef1);
  t3eff BY eff2_3*0.84109 (flef2);
  t3eff BY eff4_3*0.77134 (flef3);
  t3eff BY eff7_3*0.64324 (flef4);
  t4eff BY eff1_4@0.84649 (flef1);
  t4eff BY eff2_4*0.84109 (flef2);
  t4eff BY eff4_4*0.77134 (flef3);
  t4eff BY eff7_4*0.64324 (flef4);

MSC1_1-MSC6_1 pwith MSC1_2-MSC6_2  ;
MSC1_1-MSC6_1 pwith MSC1_3-MSC6_3  ;
MSC1_1-MSC6_1 pwith MSC1_4-MSC6_4  ;
   MSC1_2-MSC6_2   pwith MSC1_3-MSC6_3  ;
   MSC1_2-MSC6_2   pwith MSC1_4-MSC6_4  ;
   MSC1_3-MSC6_3   pwith MSC1_4-MSC6_4  ;
  eff1_1-eff7_1 pwith eff1_2-eff7_2  ;
eff1_1-eff7_1 pwith eff1_3-eff7_3  ;
eff1_1-eff7_1 pwith eff1_4-eff7_4  ;
   eff1_2-eff7_2   pwith eff1_3-eff7_3  ;
   eff1_2-eff7_2   pwith eff1_4-eff7_4  ;
   eff1_3-eff7_3   pwith eff1_4-eff7_4  ;

T1MSC  with     T1eff MGrd5 Mach_1 ;
T2MSC  with     T2eff MGrd6 Mach_2;
T3MSC  with     T3eff MGrd7 Mach_3;
T4MSC  with     T4eff MGrd8 Mach_4;

 Mach_1    with   T1eff MGrd5  ;
 Mach_2    with   T2eff MGrd6 ;
 Mach_3    with   T3eff MGrd7 ;
 Mach_4    with   T4eff MGrd8 ;

  T1eff   with MGrd5  ;
  T2eff   with MGrd6 ;
  T3eff   with MGrd7 ;
  T4eff   with MGrd8 ;
T2MSC on T1MSC Mach_1 MGrd5 T1eff (x1-x4) ;
T3MSC on T2MSC Mach_2 MGrd6 T2eff (x1-x4)
       T1MSC (x6);
T4MSC on T3MSC Mach_3 MGrd7 T3eff (x1-x4)
       T1MSC T2MSC (x6-x7);

T2eff on T1MSC T1eff Mach_1 MGrd5 (y1-y4);
T3eff on T2MSC T2eff Mach_2 MGrd6 (y1-y4)
       T1Eff (y6);
T4eff on T3MSC T3eff Mach_3 MGrd7 (y1-y4)
       T2Eff T1Eff (y6-y7);
Mach_2 on T1MSC Mach_1 MGrd5 T1eff (z1-z4);
Mach_3 on T2MSC Mach_2 MGrd6 T2eff (z1-z4)
       Mach_1 (z6);
Mach_4 on T3MSC Mach_3 MGrd7 T3eff (z1-z4)
       Mach_2 Mach_1 (z6-z7);

MGrd6 on T1MSC Mach_1 MGrd5 T1eff (w1-w4) ;
MGrd7 on T2MSC Mach_2 MGrd6 T2eff (w1-w4)
       MGrd5 (w6);
MGrd8 on T3MSC Mach_3 MGrd7 T3eff (w1-w4)
       MGrd6 MGrd5 (w6-w7);
```

```
T1MSC-T1eff WITH T2MSC-T4eff@0;
T2MSC-T2eff WITH T3MSC-T4eff@0;
T3MSC-T3eff WITH T4MSC-T4eff@0;

T1MSC-T1eff WITH  Mach_1-MGrd8@0;
T2MSC-T2eff WITH  Mach_1-MGrd8@0;
T3MSC-T3eff WITH  Mach_1-MGrd8@0;

OUTPUT: svalues TECH1; stdyx; tech4; sampstat; mod;
```

References in Supplemental Materials

Bollen, K. A., & Curran, P. J. (2004). Autoregressive Latent Trajectory (ALT) Models: A Synthesis of Two Traditions.*Sociological Methods & Research, 32*(3), 336-383. http://dx.doi.org/10.1177/0049124103260222

Enders, C. K. (2006). A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosomatic Medicine, 68*, 427-736. doi:10.1097/01.psy.0000221275.75056.d8

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

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Hedeker, D., & Gibbons, R. D. (2006). *Wiley Series in Probability and Statistics. Longitudinal data analysis.* Hoboken, NJ: Wiley-Interscience.

Newman, D. A. (2009). Missing data techniques and low response rates: The role of systematic nonresponse parameters. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 7-36). New York: Routledge/Taylor & Francis Group.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods, 17*(4), 372-411. http://dx.doi.org/10.1177/1094428114548590

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Sinha, S. K., Laird, N. M., & Fitzmaurice, G. M. (2010). Multivariate logistic regression with incomplete covariate and auxiliary information. *Journal of Multivariate Analysis, 101*(10), 2389-2397. http://dx.doi.org/10.1016/j.jmva.2010.06.010