Breaking the double-edged sword of effort/trying hard: Developmental equilibrium and longitudinal relations among effort, achievement, and academic self-concept.

Ever since the classic research of Nicholls (1976) and others, effort has been recognized as a double-edged sword: while 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 (REMs) over the first 4 years of secondary school (grades 5–8), relating effort, achievement (test scores and school grades), ASC, and ASC × Effort interactions for a representative sample of 3,144 German students (Mage = 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 nonsignificant 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 × 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.

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 (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, Marsh, Ciarrochi, Marshall, & Abduljabbar, 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 socioeconomic status.

In the present investigation we briefly review the extensive literature and the well-established support for the 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 Omelich (1979, 1984); Marsh (1984), and Dweck (2006; Dweck & Leggett, 1988; Mueller & Dweck, 1998).

REMs of Relations Between Achievement and ASC: What Is the Role of Effort?

REMs of ASC and Achievement
Although a growing body of research (e.g., Huang, 2011; Marsh, 2007b; Marsh, Kuyper, et al., 2014; Marsh 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 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 (Guay, Marsh & Boivin, 2003; Marsh & Craven, 2006; also see meta-analyses by Huang, 2011; Valentine, DuBois, & Cooper, 2004).

**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 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 a 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, a 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, while 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, De Fraine, Van Den Noortgate, and Van Damme (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 Omelich (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 (Covington, 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 signaled 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, Relich, Barnes, & Debus, 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 forced-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. emphasized 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 & Leggett, 1988; Mueller & 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, while 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, 1998). 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 × 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 versus 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 counterbalancing 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, 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 (Eccles, 2009; Marsh, 2007b; Marsh & O’Mara, 2008; Marsh et al., 2008; Murayama 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, Parker, Ciarrochi, and Heaven (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, Preacher, Selig, & Card, 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 ($N = 3,144$), 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-groups tests of invariance across the three school-type tracks, developmental equilibrium tests of the invariance of effects across four waves of data (Figure 1), and latent ASC × Effort interactions at each wave of the study.
Research Hypotheses

Hypothesis 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.

Hypothesis 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.

- 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.
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.

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 (EVT; e.g., Eccles, 2009; Guo, Marsh, Morin, Parker, & Kaur, 2015; Pekrun, 1993, 2006) 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.

ASC × 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 × 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.

- 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 potentially 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 (see Figure 1) across the four waves, we predict paths in support of $H_{2a} - H_{2d}$ to be similar from one wave to the next ($\text{Wave}_i \rightarrow \text{Wave}_{i+1}$ for $i = 1$ to 3, $i+1 = 2$ to 4) 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.

Hypothesis 3 (Robustness of Path Coefficients):

- 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 $H_{2a} - H_{2d}$ is similar across school types. We test this by treating school type as a multiple (between-groups) grouping variable and evaluating the invariance of parameter estimates across groups.

- Over gender. Girls typically have lower math self-concepts (MSCs) than boys (Else-Quest, Hyde, & Linn, 2010; Marsh, 1989) and these gender differences have been shown to generalize broadly over different countries in large cross-national studies such as Programme for International Student Assessment (PISA; Marsh, Hau, Artelt, Baumert, & Peschar, 2006; OECD, 2015) and Trends in International Mathematics and Science Study (Goldman & Penner, 2014). Hence, it is reasonable to ask whether relations between math self-concept (MSC), math achievement (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 $H_{2a} - H_{2d}$ is similar across gender. We test this by treating gender as a multiple (between-groups) 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 (Frenzel, Goetz, Pekrun, & Watt, 2010; Frenzel, Pekrun, Dicke, & Goetz, 2012; Marsh et al., in press; Murayama et al., 2013; Murayama, Pekrun, Suzuki, Marsh, & Lichtenfeld, in press; Pekrun et al., 2007; Pekrun, Lichtenfeld, Marsh, Murayama, & Goetz, in press). The data are from four measurement waves, spanning the first four years of secondary school (grades 5 to 8). Students ($N = 3,144$; 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; $N = 1,045$), middle-ability (intermediate; $N = 982$), or low-ability (vocational: $N = 1,117$) tracks.

**Measures**

**MSC and effort**

At each measurement wave the same two sets of items were used to assess MSC (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:

MSC (Wave 1: $\alpha = .876$; Wave 2: $\alpha = .895$; Wave 3: $\alpha = .893$; Wave 4: $\alpha = .910$); math effort (Wave 1: $\alpha = .780$; Wave 2: $\alpha = .809$; Wave 3: $\alpha = .801$; Wave 4: $\alpha = .805$).

**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, Kleine, Blum, & Pekrun, 2005; vom Hofe, Pekrun, Kleine, &
Goetz, 2002; 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 multimatrix sampling, with a balanced, incomplete block design (for details, see vom Hofe et al., 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 four 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, 2008–2014, 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 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 (Jeličić, 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 multigroup invariance, we tested models of the invariance of parameter estimates across groups (Millsap, 2011; Vandenberg & Lance, 2000) 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 confirmatory factor analysis (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 the 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 nonmatching constructs (i.e., the cross-paths in Figure 1) in the subsequent wave (see the Research Hypotheses section), 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 nonmatching 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\textsubscript{i} \rightarrow ASC\textsubscript{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\textsubscript{\text{Wave}i} \rightarrow ASC\textsubscript{\text{Wave}i+1} = ASC\textsubscript{\text{Wave}i+1} \rightarrow ASC\textsubscript{\text{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\textsubscript{\text{Wave}i} \rightarrow Effort\textsubscript{\text{Wave}i+1} = ASC\textsubscript{\text{Wave}i+1} \rightarrow Effort\textsubscript{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 \times Effort interaction. Methodologically, it is difficult to detect interaction effects in nonexperimental designs (Marsh, Wen, & Hau, 2004; Marsh, Wen, Hau, & Nagengast, 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 et al., 2004) incorporated into Mplus. The LMS approach takes into account the nonnormal 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, & Grayson, 2005; Marsh Hau, & Wen, 2004), 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 (Marsh et al., 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 cutoff 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 CFAs 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.

H1: 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-1C and 2A-
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 Models 2A-2C, 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 different school types) resulted in decrements in fit smaller than those needed to support the assumption of invariance. In Models 3A-3C we added the additional constraint that factor loadings be invariant over time. Here, there was almost no change in the fit, compared with 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 $H_1$, 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

<p>| Table 1: Goodness of Fit for Alternative Models of Academic Self-Concept, Effort, Test Scores, and School Grades Over Four Waves of Data and Three School-Track Types |</p>
<table>
<thead>
<tr>
<th>Model</th>
<th>Chi $df$</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
<th>Model description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirmatory factor analysis models of factorial invariance over groups and waves*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1A configural</td>
<td>4879</td>
<td>2903</td>
<td>.960</td>
<td>.953</td>
<td>.024</td>
</tr>
<tr>
<td>Model 1B metric</td>
<td>4047</td>
<td>2908</td>
<td>.960</td>
<td>.954</td>
<td>.024</td>
</tr>
<tr>
<td>Model 1C scalar</td>
<td>5120</td>
<td>3032</td>
<td>.957</td>
<td>.953</td>
<td>.025</td>
</tr>
<tr>
<td>Model 2A configural</td>
<td>3891</td>
<td>2724</td>
<td>.978</td>
<td>.973</td>
<td>.019</td>
</tr>
<tr>
<td>Model 2B metric</td>
<td>3868</td>
<td>2758</td>
<td>.978</td>
<td>.973</td>
<td>.018</td>
</tr>
<tr>
<td>Model 2C scalar</td>
<td>4015</td>
<td>2852</td>
<td>.976</td>
<td>.972</td>
<td>.019</td>
</tr>
<tr>
<td>Model 3A configural</td>
<td>3926</td>
<td>2796</td>
<td>.977</td>
<td>.972</td>
<td>.019</td>
</tr>
<tr>
<td>Model 3B metric</td>
<td>3944</td>
<td>2812</td>
<td>.977</td>
<td>.972</td>
<td>.019</td>
</tr>
<tr>
<td>Model 3C scalar</td>
<td>4092</td>
<td>2876</td>
<td>.975</td>
<td>.971</td>
<td>.019</td>
</tr>
<tr>
<td>Model 4 scalar</td>
<td>4031</td>
<td>3128</td>
<td>.969</td>
<td>.967</td>
<td>.021</td>
</tr>
</tbody>
</table>

Structural equation models: Tests of invariance of path coefficients over multiple school-type groups and waves |
| Model 5a | 4730 | 3984 | .974 | .971 | .019 | Path coeff. invar. over multiple school-type groups and waves |
| Model 5b | 4492 | 3104 | .971 | .969 | .020 | Path coeff. invar. over multiple school-type groups and waves |
| Model 5c | 4521 | 3092 | .971 | .968 | .020 | Path coeff. invar. over multiple school-type groups and waves |
| Model 5d | 4631 | 3140 | .969 | .967 | .020 | Path coeff. invar. over multiple school-type groups and waves |

Structural equation models: Tests of invariance of path coefficients over multiple gender groups and waves |
| Model 6a | 3060 | 1976 | .977 | .974 | .018 | Path coeff. invar. over multiple gender groups and waves |
| Model 6b | 3188 | 2037 | .976 | .974 | .018 | Path coeff. invar. over multiple gender groups and waves |
| Model 6c | 3377 | 2048 | .972 | .970 | .019 | Path coeff. invar. over multiple gender groups and waves |
| Model 6d | 3453 | 2072 | .971 | .969 | .020 | Path coeff. invar. over multiple gender groups and waves |

Structural equation models (grades only, test scores only) |
| Model 7a | 3984 | 2039 | .979 | .970 | .021 | Model 6d for grades only (excluding test scores) |
| Model 7b | 3799 | 2039 | .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 priori path coefficients include all test-occasion 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). Syntax for these models is presented in supplemental material (Section 4). * 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.
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).

Table 2
Factor Structure Across Four Waves of Data (W1-W4)

<table>
<thead>
<tr>
<th>Variables</th>
<th>W1</th>
<th>W2</th>
<th>W3</th>
<th>W4</th>
<th>W1</th>
<th>W2</th>
<th>W3</th>
<th>W4</th>
<th>W1</th>
<th>W2</th>
<th>W3</th>
<th>W4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math self-concept</td>
<td>.971</td>
<td>.971</td>
<td>.971</td>
<td>.971</td>
<td>.946</td>
<td>.946</td>
<td>.946</td>
<td>.946</td>
<td>.946</td>
<td>.946</td>
<td>.946</td>
<td>.946</td>
</tr>
<tr>
<td>Math effort</td>
<td>.802</td>
<td>.862</td>
<td>.862</td>
<td>.892</td>
<td>.841</td>
<td>.841</td>
<td>.841</td>
<td>.841</td>
<td>.841</td>
<td>.841</td>
<td>.841</td>
<td>.841</td>
</tr>
<tr>
<td>Math test scores</td>
<td>.825</td>
<td>.825</td>
<td>.825</td>
<td>.915</td>
<td>.775</td>
<td>.775</td>
<td>.775</td>
<td>.775</td>
<td>.775</td>
<td>.775</td>
<td>.775</td>
<td>.775</td>
</tr>
</tbody>
</table>

Factor loadings

<table>
<thead>
<tr>
<th>Items</th>
<th>Math Self-concept</th>
<th>Math effort</th>
<th>Math test scores</th>
<th>Math school grades</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC1</td>
<td>.971</td>
<td>.971</td>
<td>.971</td>
<td>.971</td>
</tr>
<tr>
<td>SC2</td>
<td>.802</td>
<td>.862</td>
<td>.862</td>
<td>.892</td>
</tr>
<tr>
<td>SC3</td>
<td>.825</td>
<td>.825</td>
<td>.825</td>
<td>.915</td>
</tr>
<tr>
<td>SC4</td>
<td>.868</td>
<td>.868</td>
<td>.868</td>
<td>.888</td>
</tr>
<tr>
<td>SC5</td>
<td>.853</td>
<td>.856</td>
<td>.855</td>
<td>.856</td>
</tr>
<tr>
<td>SC6</td>
<td>.688</td>
<td>.686</td>
<td>.686</td>
<td>.686</td>
</tr>
<tr>
<td>EFF1</td>
<td>.846</td>
<td>.846</td>
<td>.846</td>
<td>.846</td>
</tr>
<tr>
<td>EFF2</td>
<td>.841</td>
<td>.841</td>
<td>.841</td>
<td>.841</td>
</tr>
<tr>
<td>EFF3</td>
<td>.775</td>
<td>.775</td>
<td>.775</td>
<td>.775</td>
</tr>
<tr>
<td>EFF4</td>
<td>.653</td>
<td>.653</td>
<td>.653</td>
<td>.653</td>
</tr>
</tbody>
</table>

Factor correlations

REM of Causal Ordering of ASC, Effort, and Achievement

We begin this section with a preliminary evaluation of H3, the robustness of path coefficients over school-track type (between-groups 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-groups 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 H3a, indicating that the pattern of results is similar in the three school-type ability tracks, and supporting developmental equilibrium (H3c).

In a parallel set of analyses (Models 6A–6D, Table 1), we also evaluated the invariance of the results over gender (H3b). 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 H3b.

**ASC and achievement (H2a, 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 × 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 (H2b, 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 and 7A) after controlling for prior effort, ASC, and achievement. However, the corresponding paths from prior test scores to effort were nonsignificant (Models 6D and 7B). Furthermore, prior effort had no significant effect on subsequent achievement—either test scores or school grades.

**Effort and ASC (H2c, 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 (see Table 3) in which each achievement indicator is considered separately (Models 7A and 7B) or in combination (Model 6D). In contrast, the ASC → effort path is nonsignificant for five of the six models, but is significantly positive in Model 7B (test scores only, with no interaction).

**ASC × Effort interactions (H2d)**

Consistently with a priori predictions, the ASC × 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 (see 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 counterproductive effects.
Similarly, and also consistently with a priori hypotheses, the ASC × Effort interaction had significantly positive effects on subsequent school grades. Thus, while the first-order effect of effort on school grades was positive but nonsignificant, 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 × 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 at 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 Versus 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 & 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 achievement 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, Pohlmann, Köller, & Marsh, 2009, Möller, Retelsdorf, Köller, & Marsh, 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 nonsignificant 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 EVT 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 × 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 × 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 × 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 × 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 influence subsequent ASC but also had positive effects on students’ grades. This positive effect of the ASC × 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 achievement tracks considered here. Nevertheless, we also emphasize that while REMs provide stronger tests of temporal ordering than do mere correlations based on cross-sectional data, the results are still correlational. Thus, while 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, ASC, 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 two or three 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., Salili & 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 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 while 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 & Leggett, 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, Colpin, Van Damme, & Verschueren, 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, while 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 et al., 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


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, 15–21. 10.1037/0022-0663.82.1.15


beliefs and achievements in multiple domains. *Learning and Instruction, 35*, 16–32.
10.1016/j.learninstruc.2014.08.005

10.1348/000709910X503501


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, 306–315. 10.1016/0092-6566(76)90020-9


