Long-term positive effects of repeating a year in school: six-year longitudinal study of self-beliefs, anxiety, social relations, school grades, and test scores

Article

Accepted Version

Long-term Positive Effects of Repeating a Year in School: Six-Year Longitudinal Study of Self-Beliefs, Anxiety, Social Relations, School Grades, and Test Scores

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

Reinhard Pekrun, University of Munich, Germany

Philip D. Parker, Australian Catholic University

Kou Murayama, University of Reading, UK

Jiesi Guo, Australian Catholic University

Theresa Dicke, Australian Catholic University

Stephanie Lichtenfeld, University of Munich, Germany

Revised 10 May, 2016

Revised: 28 March, 2016

18 January, 2016

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

Abstract

Consistently with a priori predictions, school retention (repeating a year in school) had largely positive effects for a diverse range of 10 outcomes (e.g., math self-concept, self-efficacy, anxiety, relations with teachers, parents and peers, school grades, and standardized achievement test scores). The design, based on a large, representative sample of German students (N = 1,325, M age = 11.75 years) measured each year during the first five years of secondary school, was particularly strong. It featured four independent retention groups (different groups of students, each repeating one of the four first years of secondary school, total N = 103), with multiple post-test waves to evaluate short- and long-term effects, controlling for covariates (gender, age, SES, primary school grades, IQ) and one or more sets of 10 outcomes realised prior to retention. Tests of developmental invariance demonstrated that the effects of retention (controlling for covariates and pre-retention outcomes) were highly consistent across this potentially volatile early-to-middle adolescent period; largely positive effects in the first year following retention were maintained in subsequent school years following retention. Particularly considering that these results are contrary to at least some of the accepted wisdom about school retention, the findings have important implications for educational researchers, policymakers and parents.

Keywords: Math self-concept, achievement, retention, school retention, social comparison

Long-term Positive Effects of Repeating a Year in School: Six-Year Longitudinal Study of Self-Beliefs, Anxiety, Social Relations, School Grades, and Test Scores

Grade retention is the practice of requiring a student in a given grade or year in school to repeat the same grade level in the following year (Allen, Chen, Wilson, & Hughes, 2009). Allen et al. (2009) note that the use of retention as an educational intervention, particularly in the US, has fluctuated since the early 1900s, reaching a peak in the 1970s before declining in the 1980s and then increasing rapidly in the 1990s—apparently in response to the standards-based reform movement, following the publication of *A Nation at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education, 1983). Marsh (2016) also noted that, on the basis of international PISA data, there is substantial country-to-country variation in the use of retention.

Social Comparison Theory

Marsh (2016) evaluated the effects of de facto retention (starting school late or repeating a grade) on academic self-concept from the perspective of social comparison theory. Theoretical models such as social comparison theory, adaptation level theory, and range-frequency theory (e.g., Huguet, Dumas et al., 2009; Marsh, 2016; Marsh, Seaton, et al., 2008) posit that students compare their own academic accomplishments with those of their classmates, as one basis for academic self-concept formation. Thus, the academic accomplishments of classmates form a frame of reference or standard of comparison that students use to form their own academic self-concepts. Furthermore, there is a growing body of research showing that academic self-concept is reciprocally related to school-based performance measures (e.g., school grades on report cards) in particular, but also to standardized achievement test scores (Marsh & Craven, 2006; Guay, Marsh, & Boivin, 2003), and that academic self-concept might be even more important than achievement in predicting future academic choices (Marsh & Yeung, 1997).

In academic self-concept studies the frame of reference is typically defined in terms of the academic achievement of classmates. However, for a variety of reasons, such as acceleration,

or starting school at an early age, students can find themselves in classes with older, more academically advanced students, who might form a more demanding frame of reference than would same-age classmates. Similarly, due to starting school at a later age, or to being held back to repeat a grade, students can find themselves in classes with younger, less academically advanced students than would other students of the same age. In the present investigation our focus is on the effects of repeating a year in school on a diverse set of self-beliefs, self-perceptions of relations with significant others, school grades, and standardized test scores collected during the first five years of secondary school.

Time to Learn

Although not studied specifically in relation to retention, Bloom (1968, 1976) contended that weaker students merely need more time to learn materials than do stronger students, but that once learning is achieved, the differences between more and less able students diminish in terms of subsequent achievement, academic self-beliefs, and motivation to learn. Also, there is ample evidence that without appropriate intervention, small differences in achievement at any particular stage of education become larger over time, so that the gap between the more and less able students increases. This cumulative disadvantage has reciprocal effects with subsequent motivation, as well as achievement, creating a downward spiral (i.e., the Mathew Effect; Stanovich, 1986; Walberg, 1983). Hence, we hypothesize that because retained students have an extra year to learn the materials that originally led to their retention, they should be better able to learn those materials in the first year following retention and should also have more positive self-beliefs, giving them a stronger basis for learning new materials and for maintaining positive self-beliefs in subsequent school years.

Grade Retention Effects

Grade retention effects on achievement. Retention effects (i.e., repeating a year in school) have been studied extensively in relation to academic achievement (e.g., Alexander, Entwisle, & Dauber, 2003; Jimerson, 2001; but see Reynolds, 1992; Roderick, 1994; Roderick

& Engel, 2001). However, as emphasized by Jimerson and Brown (2013, p. 140), "because of potential short- and long-term effects that grade retention can have on student achievement and socioemotional outcomes, it remains a controversial topic in research and practice". Indeed, there is a general belief, supported by some research evidence, that retention has negative effects on academic achievement (e.g., Hattie, 2012). As emphasized by Allen et al. (2009), this negative view of retention is evident in a policy statement by the National Association of School Psychologists, which "urges schools and parents to seek alternatives to retention that more effectively address the specific instructional needs of academic underachievers" (p. 481).

However, critical design and methodological issues, such as the need for appropriate control groups and controlling for pre-existing differences—especially prior achievement, which is inevitably confounded with retention—dictate caution in reaching overarching conclusions such as these (Jimerson & Brown, 2013). Thus, on the basis of their meta-analysis of grade retention studies in which they controlled for study quality, Allen et al. (2009) reported that their results "challenge the widely held belief that retention has a negative effect on achievement" (p. 480). They found that studies showing negative effects of retention were largely limited to poor quality studies with insufficient control for pre-existing differences.

Consistently with the Allen et al. (2009) meta-analysis, a number of publications based on an ongoing longitudinal study challenge the view that retention has negative effects, or else show that negative effects in prior studies are likely the result of inadequate control for selection effects (Cham, Hughes, West, & Im, 2015; Im et al., 2013; Moser, West, & Hughes, 2012). Using propensity matching to match retained with non-retained (promoted) primary school students, Wu, West, and Hughes (2010) found that retention had short-term positive effects on school-belonging, teacher-rated engagement, and academic self-concept. In a follow-up to this study, Im et al. (2013) found that retained and promoted students, following transition to middle school, did not differ in terms of achievement, engagement, or school-belonging (although they did not report the follow-up measures of academic self-concept

considered in the earlier study; a focus of the present investigation). At Year 5, Moser et al. (2012) compared growth trajectories on math and reading achievement for propensity-matched students who had been retained or promoted in Year 1 of primary school. After shifting scores back one year to permit same-year-in-school comparisons (what we refer to as "offset" comparisons), the retention group experienced an initially higher scores than the non-retained group, assessed on the basis of Year 1 scores. However, the positive retention effects dissipated over time, such that by Year 5 there were no differences between the two groups. The authors also warned that retention effects on achievement might vary, depending on the nature of the measure, and noted that in Year 3 the retained students were more likely to pass a state accountability math test that was closely aligned to the school curriculum (Hughes, Chen, Thoemmes, & Kwok, 2010). Summarising the results of these multiple publications, ten years into this longitudinal research program, Cham et al. (2015) concluded that their ongoing research studies "have not supported the popular view within the educational literature that grade retention harms students' educational success. Instead, we have either found advantages for the retained group or have failed to reject the null hypothesis of no difference between the retained and promoted groups" (p. 18).

Cross-national comparisons. Marsh (2016) recently proposed a frame-of-reference model to evaluate the effects of relative year in school (e.g., being one school year ahead or behind same-age students) based on math constructs and using PISA data from 41 countries. Marsh showed that for countries participating in PISA, students typically are grouped into the same grade or year in school according to their age, rather than to their abilities in general or in particular school subjects. Thus, with the exception of students who start school early or late, those identified as gifted, or in need of remedial assistance, it is typical for students within the same class to be of a similar age. For example, based on nationally representative samples of 15-year-olds (total N = 276,165) from 41 countries (PISA 2003 data), 67% of the students were in their modal year in school for their country (Marsh, 2016). However, for nearly all countries,

there were 15-year-old students accelerated one or more years relative to their modal year in school (e.g., students in Years 11 or 12 when their modal or "age-appropriate" year group was Year 9 or 10), whereas others were in year groups one or more years behind their modal year group (e.g., students in Years 7 or 8 when their modal or "age-appropriate" year group was Year 9 or 10). Extending a model of social comparison theory, Marsh (2016) predicted a priori, and found, that the effects of de facto retention (starting school late or repeating a grade) on math self-concept (MSC) were consistently positive across the 41 countries. These positive effects of de facto retention were reasonably consistent across the 41 countries and individual student characteristics. Relative year in school seemed to be the critical variable. The critical finding for our purposes is that the positive effects on MSC were similar for students who started late or who had been retained previously.

Noting limitations and directions for further research, Marsh (2016) emphasizes that the cross-sectional nature of the PISA data precludes stronger longitudinal models. He argues, however, that for retained students, the uncontrolled, pre-existing differences leading to retention would be likely to negatively bias estimates of the positive effects of de facto retention, working against the hypothesized positive effects that he predicted and found. Similarly, the cross-sectional nature of the data precluded longitudinal models that more fully differentiated between de facto retention based on starting school at an older age, and grade retention. Particularly relevant to the present investigation, and from the perspective of educational policy, the reliance on cross-sectional PISA data precluded evaluation of the effects of retention on changes in academic achievement based either on school grades or on standardized test scores.

Rationale for A Priori Research Hypotheses and Research Questions

The German school system and grade retention. In Germany, elementary school spans Grade 1 to 4, secondary school starts at Grade 5, and compulsory schooling ends at Grade 9 in most states, including the state of Bavaria, where the present investigation was

conducted. There is no tracking in elementary school, but in most states, including Bavaria, students are placed into one of three tracks at the start of secondary school: lower-track schools (Hauptschule), medium-track schools (Realschule), and higher-track schools (Gymnasium), on the basis of their elementary school achievement. Grade retention is used in elementary school as well as across all secondary school tracks, and is based on students' achievement in main subjects. The number of repeated retentions per student is limited, and in the present investigation no students repeated more than one grade. We also note that in the German school system teachers are very reluctant to use retention in the first two years of secondary school. Hence, the majority of retention in our study appeared in Years 7 and 8, rather than Years 5 and 6.

In the present investigation we evaluate the effects of grade retention (repeating a school year) on a range of psychosocial and achievement outcomes (see Figure 1) for a single cohort of students as they progress through the first five years of secondary school. Data was collected from a representative sample of 1,325 students from 42 schools starting the year before the start of secondary school; Year 4 school grades in German and math, and then school grades, standardized achievement tests, and psychosocial variables for each of the subsequent five years of secondary school (see Figure 1). We evaluated retention in each of four separate groups: those retained at Year 5, the different group of students retained at Year 6, etc., noting that no students were retained for more than one year (for a discussion of the German school system, tracking, and retention see Supplemental Materials, Section 1). The study design (Figure 1) provides a particularly strong foundation for evaluating retention effects on the basis of multiple natural experiments using longitudinal data that provide multiple post-test waves to evaluate short- and long-term effects of retention and multiple pretest waves as controls for all outcomes as well as the covariates (gender, age, SES, primary school grades, IQ).

Our main focus is on the four dichotomous grouping variables (Figure 1) representing those students who repeated a school year in each of the four Years 5–8. For example, the lagged effects of repeating Year 5 are represented by the path from the grouping variable ("Repeat Year 5" in Figure 1) to outcomes in the immediate subsequent Wave 2 (Lag 1 effects), as well as all effects in the subsequent three waves (Lag 2–4 effects at Waves 3–5; Figure 1). Whereas most students are in Year 6 in Wave 2, the students repeating Year 5 are in Year 5 at Wave 2. It is important to emphasize that there are Lag 1 effects for each of the four retention groups. Thus (see Figure 1), there are separate estimates of Lag 1 effects for students repeating Years 5, 6, 7 and 8 (i.e., the effects of the first year following retention for each of the four retention groups). Similarly, different groups of students repeating Years 6, 7, and 8, have multiple pre-retention waves of data to control for pre-existing differences, and multiple post-retention waves to evaluate the short- and long-term effects of retention. This enables us not only to test these Lag 1 effects for each of the four separate groups, but also to test the consistency of these lagged effects across the four groups that span this potentially volatile early-to-middle adolescent period.

An intentionally diverse set of outcomes was considered, including: self-belief variables, the focus of the Marsh (2016) study; achievement measures, which have been the focus of most retention studies; anxiety, to represent the emotional response of students to retention; and student self-reports of relations with significant others—parents (parental assistance, academic assistance from parents), teachers (positive teacher support), and peers (peer appreciation of math). (Item wording and reliability estimates, as well as correlations among the multiple factors, are presented in the Supplemental Materials, Section 2).

A Developmental Perspective: Developmental Invariance Hypothesis

A potentially important limitation of retention research is that it is mostly based on US primary school students, and—even when longitudinal in terms of following-up the effects of retention over multiple school years—typically includes results based on retention in a single

school year (see Allen, et al., 2009; Holmes & Mathews, 1984; Jimerson, 2001). In this sense, the research lacks a developmental perspective. Here however, we introduce an apparently unique developmental equilibrium perspective, evaluating the consistency of the retention effects over the potentially volatile early-to-middle adolescent period on the basis of longitudinal data and multiple retention groups. 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. 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.

Here we test developmental equilibrium in relation to the invariance of retention effects in each of four year groups spanning this early-to-middle adolescent period. More specifically, we evaluate support for developmental invariance, based on the hypothesis that retention effects are the same for students retained in Years 5, 6, 7, and 8 (see Figure 1). In this sense, our study is longitudinal, in that it covers the entire early-to-middle adolescent period, but also because it evaluates retention for separate groups of students who had been retained in Years 5, 6, 7, and 8. The German secondary school system constitutes Years 5–9, although Years 5 and 6 are often considered part of primary schooling in US studies). Combining the effects of retention across these four groups partly compensates for the typically small sample sizes of retention groups based on retention in a single year, greatly increasing the robustness and statistical power, due to the increased *N* of the results. More importantly, it provides an apparently unique developmental perspective on the question whether the self-system has achieved a developmental balance in relation to the retention effects such that they are the same for students retained in Years 5–8.

Research Hypotheses and Questions: Retention Effects in Relation to Specific Outcomes

Math self-concept (MSC; Hypotheses 1a, 1b): Consistently with Marsh et al. (2016) we predict retention has positive effects on MSC in the first year following grade retention

(Lag 1), after controlling for covariates and outcomes from prior waves (Hypothesis 1a). Lag 2–4 effects are the direct effects of retention two, three, and four years respectively following retention, after controlling for Lag 1 effects as well as the effects of covariates and outcomes from the earlier waves. Positive effects at Lags 2–4 would indicate "sleeper effects" (new positive effects, in addition to the positive effects already observed). Non-significant effects at Lags 2–4 would indicate that Lag 1 effects were maintained, whilst negative effects at Lags 2–4 would indicate that Lag 1 effects were not fully maintained. We hypothesize (Hypothesis 1b) that the Lag 2–4 effects of retention will be small and largely non-significant—that the initially positive effects of retention on MSC will be maintained.

Self-efficacy and anxiety (Hypotheses 2a & 2b): Although the grounds for these a priori predictions are less clear, both of these variables are strongly related to MSC. On this basis we anticipate that the effects of retention will be favourable and similar in direction, although perhaps smaller in size, to those predicted for MSC (increased self-efficacy and reduced anxiety) at Lag 1 (Hypothesis 2a), and that these effects will be retained over time (Hypothesis 2b).

Relations with significant others (Research Questions 3a & 3b): Our study includes three variables associated with the positive interactions that students perceive having with significant others (parental assistance, positive teacher support, peer appreciation of math) in relation to math. We leave as research questions the direction of effects of retention on these outcomes at Lag 1 (Research Question 3a) and Lags 2–4 (Research Question 3b), but anticipate that the Lag 1 effects are at least not negative (i.e., are either favourable or are non-significant).

School grades, Lag 1 (Hypothesis 4a, Research Question 4b): In each year of our study, end-of-year school grades (i.e., school-based performance measures) were collected from school records. For the present purposes we focus on school grades in math, German (native language), and an average over other subjects. This latter might differ according to the student and year in school (e.g., English, other foreign language, biology, sport, and music).

Because retained students study the same materials in the year following retention, Lag 1 retention effects are predicted to be positive and substantial (Hypothesis 4a). An optimistic perspective is that positive Lag 1 effects on school grades are maintained or even increased in subsequent Lags 2–4. However, predicted positive effects at Lag 1 are based on studying the same material for two years, whilst Lag 2–4 retention effects are based on students studying new materials for a single year only. Hence, it is entirely possible that the positive effects at Lag 1 will not be fully maintained—that Lag 2–4 retention effects will be negative, offsetting the positive effects at Lag 1, in part at least. Thus we leave this as a research question, rather than a research hypothesis based on a priori predictions (Research Question 4b).

Standardized math test scores, same age comparisons (Research Questions 5a and b): In each year of our study, students completed a standardized math test. Although the tests were not specifically based on the school curriculum, in each year they contained a range of advanced materials suitable to the year in school for non-retained students in each wave of the study. Particularly as retained students have had a chance to learn more fully the materials that they have studied previously, an optimistic perspective would be that Lag 1 retention effects are positive for math test scores. However, because retained students are a year behind their non-retained classmates, they have not studied advanced materials covered in the curriculum that are included in the standardized math test and that have been studied by non-retained students. In this sense, the math test based on same-age comparisons might be considered "unfair" for retained students—at least in terms of inferring what students have learned, relative to the materials that they have actually studied. On the other hand, it could also be argued that the same-age comparisons accurately reflect the fact that repeaters lag behind non-repeaters in what they have studied. Hence, we leave this as a research question. Particularly given that Lag 1 retention effects on math test scores are left as a research question, there is no basis for predicting Lag 2–4 retention effects; these also are left as a research question.

Offset math test scores, Lag 1 same-year-in-school comparisons (Hypothesis 6a, Research Question 6b): An alternative perspective on test scores is to compare retained students in each year following retention with non-retained students from the previous wave when they were in the same year in school (see Figure 2). Thus, in this offset strategy (based on comparisons of the same year in school, or what Im et al. [2013, p. 361], refer to as "shifting back" scores), math test scores for retained students repeating Year 5 are compared to test scores from non-repeaters from the previous wave (when they were also in Year 5) who had studied the same curriculum. Similarly, for each post-retention year, for all four retention groups, comparisons based on test scores (but not other outcomes) were "offset" by one year, so that comparisons were based on students having completed the same year in school (see Figure 2). For these offset comparisons, we predict that the Lag 1 retention effects will be positive, and more positive than those based on the original (same-age) comparisons (test scores not offset by one year; presented in Research Question 5). However, similar to the logic based on school grades (see Research Question 4b), the predicted positive effects for test scores at Lag 1 might not be fully retained over Lags 2-4 and so that we leave this as Research Question 6b.

Method

Sample

Our data are based on the *Project for the Analysis of Learning and Achievement in Mathematics* (PALMA; Frenzel, Pekrun, Dicke, & Goetz, 2012; Murayama et al., 2013, 2016; Pekrun et al., 2007), a large-scale longitudinal study investigating the development of math achievement and its determinants during secondary school in Germany. The study was conducted in the German federal state of Bavaria and included five measurement waves spanning Years 5 to 9, in addition to school grades from the last year of primary school (Year 4). Data (1,325 students from 42 schools; 50% girls; mean age = 11.7 at Wave 1, *SD* = 0.7) were collected from the year before the start of secondary school (Year 4 school grades in

German and math) and school grades, standardized achievement tests, and psychosocial variables for each of the subsequent five years of secondary school (see Figure 1).

Sampling and assessments were conducted by the Data Processing and Research Center of the International Association for the Evaluation of Educational Achievement. The samples represented the typical student population in the state of Bavaria in terms of student characteristics such as gender, urban versus rural location, and SES (for details, see Pekrun et al., 2007). Students answered the questionnaire towards the end of each successive school year. All instruments were administered in the students' classrooms by trained external test administrators. Participation in the study was voluntary, parental consent was obtained for all students and the acceptance rate was a very high 91.8%. Surveys were depersonalized to ensure participant confidentiality.

Our central focus is on evaluating the effects of grade retention in each of the first four years of secondary school. Because grade retention is not a frequent occurrence, the numbers repeating are relatively small. Of the 1,325 students considered here who participated in all five waves of the study, the numbers of students who repeated in each year were: Year 5 (10); Year 6 (12); Year 7 (35); Year 8 (45)—a total of 103 students, or 7.8% of the sample. The 103 repeating students did not differ significantly (all *p*'s > .05) from the 1,222 nonrepeating students on gender (42% versus 51% female); school type (43% Gymnasium, 23% Realschule, 23% Hauptschule versus 40%, 30%, and 29%, respectively); age (11.7 versus 11.8 years); or family SES (.01 versus -.02).

In supplemental analyses we evaluated potential biases associated with missing data after controlling for background variables (see "covariates" Figure 1) and school type for the ten outcomes in Year 5. More specifically, we evaluated the main effect of being included in the sample ("include" in Supplemental Table 2; the difference between the 1,325 students in the final sample vs. the 745 students excluded because of missing data); main effect of repeat ("repeat" in Supplemental Table 2; the differences in outcomes for the repeating students

compared with those who did not repeat Year 5); and the repeat-by-include interaction ("IncldxRepeat" in Supplemental Table 2). This last parameter was of particular interest as it explored whether the difference between repeating and non-repeating students depended upon whether the students were included in the final sample. The effects of include were statistically significant for two of 10 outcomes; those students in the final sample had significantly higher math grades ($p < .01$) and German grades ($p < .05$) than students excluded because of missing data, but did not differ significant in terms of school grades in other subjects, standardized test scores or any of the other outcomes. Students had missing data over this five-year span due to absences on the day of the data collection, but also because families moved. However, we note that there are very strong controls for biases associated with these outcomes as each of the 10 outcomes was measured in each of the five waves of data. More importantly for present purposes, differences between repeating and continuing students did not depend upon whether the students were or were not included in the final sample. More specifically, differences between the repeating and non-repeating students on the 10 outcomes in Year 5 did not vary significantly as a function of missing data, thereby supporting the appropriateness of the analyses (see Supplemental Materials, Section 1).

Measures (see Supplemental Materials, Section 2 for more detail on measures)

Six psychosocial constructs. At each measurement wave the same set of items was used to assess MSC, math self-efficacy, math anxiety (Achievement Emotions Questionnaire-Mathematics, see Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011), and student perceptions of significant others—parents (Parental Assistance), teachers (Positive Teacher Support), and peers (Peer Appreciation of Math). All these multi-item scales were based on self-report responses from students, using a 5-point-Likert scale: “not true”, “hardly true”, “a bit true”, “largely true”, or “absolutely true”. Across the 5 waves and the six multi-item scales, the 30 coefficient alpha estimates of reliability were generally high (αs varying from .75 to .92; median $\alpha = .87$) and were consistent over the multiple waves. For ease of interpretation,

anxiety scores were reverse scored so that—consistently with other constructs—higher scores reflect more favorable outcomes. (Item wording and reliability estimates, as well as correlations among the multiple factors, are presented in the Supplemental Materials, Section 2).

Math achievement. Students' achievement was measured both in terms of school grades (from Year 4, the last year of primary school, and in Years 5–9, the first five years of secondary school) and standardized achievement test scores in math (Years 5–9). School grades were end-of-year final grades obtained from school records. Standardized math achievement was assessed by the PALMA Mathematical Achievement Test (vom Hofe, Kleine, Pekrun & Blum, 2005). Using both, multiple-choice and open-ended items, this test measures students' modeling and algorithmic competencies in arithmetic, algebra, and geometry. In each successive year, the test covered the same content areas, but the number and difficulty of the items increased in line with the year in school completed by non-repeating students; the number of items increased from 60 to 90 items across the five waves. The obtained achievement scores were scaled using one-parameter logistic item response theory (Rasch scaling; Wu, Adams, Wilson, & Haldane, 2007), and standardized in relation to Year 5 results (i.e., the first measurement point) to establish a common metric across the five waves.

Covariates. Students' school grades in math and German at the end of primary school (Year 4), gender, IQ, age, and SES served as covariates for the overall study. Students' IQ was measured using the 25-item nonverbal reasoning subtest of the German adaptation of Thorndike's Cognitive Abilities Test (Heller & Perleth, 2000). SES was assessed by parent report using the Erikson Goldthorpe Portocarero (EGP) social class scheme (Erikson, Goldthorpe, & Portocarero, 1979), which consists of ordered categories of parental occupational status; higher values represent higher social class.

Statistical analyses

All analyses were done with Mplus 7.3 (Muthén & Muthén, 2008–14, Version 7). We used the robust maximum likelihood estimator (MLR), which is robust against violations of normality assumptions. All analyses were based on manifest variables, using the complex design option to account for nesting of students within schools. As is typical in large longitudinal field studies, some students had missing data for at least one of the measurement waves, due primarily to absence or to changing schools. Because of the nature of the data analyses described below (particularly the "offset" comparison of math test scores), analyses were based on the 1,325 students who participated in all five waves. For this group, the relatively small amounts of missing data (less than 1% for each variable) were handled with Full Information Maximum Likelihood (FIML), the default option in Mplus.

The primary analysis was a "full-forward" structural equation model that is saturated, in the sense that all paths are estimated (see Figure 1). For example, covariates are predictors of all variables in Years 5–9, Year 5 variables are predictors of all variables in Years 6–9, and so forth. Within each wave, all variables were correlated. A specific focus is the four dichotomous grouping variables representing students who repeated a school year in one of the four Years 5–8. For example, a student repeating Year 5 is tested again in Year 5 (now in Wave 2 rather than Wave 1), and again in Years 6, 7, and 8 (in Waves 3–5). The effect of repeating Year 5 is represented by the path from the grouping variable ("Repeat Year 5") to outcomes in the immediate subsequent wave (Lag 1 effects), as well as all subsequent waves (effects at Lags 2–4). Similarly, different groups of students, repeating Years 6, 7, and 8, are each followed up in subsequent years, to test the effects of retention.

In order to facilitate interpretation of the results, all covariates and Year 5 outcomes were standardized ($M = 0$, $SD = 1$) across the entire sample. Outcomes for Years 6–9 were then standardized in relation to mean values of each construct in Year 5, so that measurement in relation to a common metric was retained. The four grouping variables representing retention were scored 1 = retention, 0 = non-retention. Hence, the unstandardized coefficients associated

with each of these variables represent the difference between the two groups in relation to Year 5 standard deviation units, after controlling for covariates and outcomes in all waves prior to retention for each of the retention groups—hereafter referred to as effect sizes (*ES*s)—scaled so that higher scores reflect more favourable outcomes. As noted earlier (see discussion of research questions, and Hypotheses 6 and 7), retention effects on standardized achievement tests were evaluated in relation to both, same-age comparisons (e.g., comparing results of retained Year 5 students with those of non-retained Year 6 students who are of a similar age) and same-year-in-school comparisons (e.g., comparing results of retained Year 5 students with non-retained students when they also were in Year 5—see Figure 2).

Preliminary Analyses: Evaluation of Developmental Invariance Hypothesis

The path model depicted in Figure 1 is a "full forward" structural equation model that is completely saturated with degrees-of-freedom (*df*) = 0; all paths relating variables in different waves are estimated, as are all correlations and correlated residuals relating variables within each wave. We evaluated two alternative models to summarize the retention effects. In the "means model" we used the model constraint option in Mplus to compute the mean effects size (*ES*) across the relevant retention groups for each outcome, along with the standard error and a test as to whether the mean was significantly different from zero. Thus, for example, the mean *ES* for MSC was the mean retention effect averaged across the four retention groups (i.e., students retained in Years 5, 6, 7, and 8). Importantly, this model is still saturated, in that it did not impose any constraints. However, it provides a much stronger, more robust test of the overall retention effects, in that the test of the mean across retention groups is based on a larger *N* than tests of each group separately, compensating in part for the small number of retained students in each retention group.

In order to more formally evaluate the invariance of retention effects, we next tested a "developmental invariance" model in which all lagged effects were constrained to be the same across the four retention groups. Thus, for example, Lag 1 retention effects for MSC were

constrained to be the same for the different groups of students who had been retained in Years 5, 6, 7, and 8, respectively. This highly constrained, parsimonious model imposed a total of 60 invariance constraints. Particularly given the large number of constraints, the fit of this model was remarkably good, providing strong support for the developmental invariance of retention effects across the four retention groups. Not surprisingly, the mean *ES*s (based on the means model) and the invariant *ES*s (based on the developmental invariance model) were similar, and both provided a parsimonious summary of the retention effects. For the present purposes we focus on results based on the statistically stronger developmental invariance model, but results for the means model—including the estimates for each of the year groups considered separately, as well as details about the fit of the developmental invariance—are presented in the Supplemental Materials (Section 4).

Results

Effects of Retention

Math self-concept (**Hypotheses 1a and 1b).** Consistently with Hypothesis 1a, the effects of retention on MSC in the first year following retention (invariant Lag 1 effects) were positive and statistically significant (*ES* = .597, Table 1). Lag 2–4 effects reflect the direct effect of the intervention after controlling for outcomes from all previous waves, including the Lag 1 effects; positive effects reflect "sleeper" effects, negative effects reflect a significant diminishing of the positive effects at Lag 1, and non-significant effects reflect maintenance of the positive effects at Lag 1. Consistently with Hypothesis 1b, the *ES*s for Lags 2–4 were non-significant (maintenance of T1 effects).

Self-efficacy and anxiety (Hypotheses 2a and 2b). Consistently with Hypothesis 2a, the effects of retention on these outcomes were significantly positive (noting that anxiety was reverse scored so that higher values reflect less anxiety). However, *ES*s (.359 for self-efficacy, .293 for anxiety; Table 1) were smaller than for MSC. Consistently with Hypothesis 2b, Lag 2–4 *ES*s were non-significant for both self-efficacy (maintenance of T1 effects), although for

anxiety effects there was a positive Lag 4 effect (a positive sleeper effect) even though Lag 2 and 3 effects were non-significant.

Relations with significant others (Research Questions 3a & 3b). Lag 1 *ES*s for the effects of student perceptions of Positive Teacher Support were significantly positive (*ES* = .305), whilst the non-significant Lags 2–4 effects indicated that these positive effects of retention were maintained in subsequent school years. There were no statistically significant effects (Lags 1–4) of retention for perceptions of Parental Assistance or Peer Appreciation of Math.

School grades (Hypothesis 4a and Research Question 4b). Retention effects were evaluated for end-of-year school grades for math and for German (required subjects), and an average grade over other subjects (GPA). Lag 1 retention effects were significantly positive for all three measures of school grades (*ES*s = .452 to 1.010). The results were particularly large for math school grades (mean *ES* = 1.010), reflecting stronger controls for pre-existing differences in math, due to the focus of the study on math (i.e., other outcomes, including test scores, were math-specific). Although we anticipated that the corresponding Lag 2–4 effects might be negative (but left this as a research question), these effects were all non-significant, demonstrating that the substantial positive effects of retention on school grades in the first year following retention were maintained in subsequent school years.

Standardized math tests, same age comparisons (Research Questions 5a and b). Retention effects were evaluated in relation to standardized achievement test scores collected in each year of the study. We anticipated that these Lag 1 effects based on same age comparisons might inappropriately disadvantage retained students (who had not studied some of the advanced materials covered by non-retained students), but left this as a research question. Indeed, Lag 1 effects for math test scores were significantly negative (*ES* = -.188), although the size of the effect was much smaller than the corresponding positive effect on

school grades (*ES* = +1.010). Lag 2–4 effects for test scores were non-significant, indicating that the small negative effects of retention on test scores were maintained (Table 1).

Standardized math tests, same-year-in-school comparisons (Hypothesis 6a and Research Question 6b). In an alternative perspective on test scores (see Figure 2 and Table 2), we compared test scores of retained students in each year following retention with those of non-retained students in the previous wave (i.e., same-year-in-school comparisons). Thus, test scores for the retained groups were compared to those in non-retained groups who had completed the same year in school and studied the same curriculum, but on the basis of data from one wave earlier. Because of the nature of the offset comparisons (see Table 1), these had to be conducted separately for retention groups in Years 5–7 (and were not possible for the "repeat Year 8" retention Group; see discussion in Table 2). Consistently with Hypothesis 6a (Table 2), Lag 1 *ES*s were more positive for these offset comparisons (based on the same year in school) than were those based on the same wave (same-age comparisons, evaluated in Research Question 5a). For these offset comparisons, all 6 *ES*s (based on total effects in Table 2) were positive (.053 to .677; *M* = .341) in favor of the retention group, and three were statistically significant. In summary, when test scores for retained students were compared with those of other students in the same year group, there were significantly positive effects of retention.

Summary of Results.

Given the persistent belief that retention has negative effects, the most important finding here is that in research based on a particularly strong and more appropriate design, the effects of retention were mostly positive, and almost none were significantly negative. Indeed, for the critical Lag 1 effects based on the first year following the intervention, only one of the 10 effects was significantly negative ($.05 < p < .01$), and 7 were significantly positive ($p < .05$). Averaged across the 10 outcomes, the mean of Lag 1 effects was statistically significant (.384). Evaluation of Lag 2–4 effects of retention demonstrate that these Lag 1 effects were

maintained, or in the case of anxiety, improved further in subsequent years. Although our focus has been on the invariant estimates across the four retention groups, it is also relevant to look at the results for each of the four groups separately (see Supplemental Materials, Section 4). For the critical 40 Lag 1 effects (i.e., four retention groups x 10 outcomes) based on the first year following the intervention, only one of the 40 effects was significantly negative ($.05 < p < .01$). Furthermore, none of the mean effects for any of the 10 outcomes averaged across the four retention groups was significantly negative. In contrast, 23 of 40 effects were significantly positive; the mean effects averaged across the four groups were significantly positive for 6 of 10 outcomes, as was the grand mean effect averaged across all outcomes (.384).

Consistently with Marsh (2016), the effects of retention on MSC were positive (*M* Lag 1 *ES* = .597), and the results were generally favorable for self-efficacy and anxiety. However, perhaps surprisingly, the results were even more positive for math school grades (*M* Lag 1 *ES* = 1.010); the retention effects were also positive for other school grade measures. Retention effects for relations with significant others were positive, but only student perceptions of teacher support were statistically significant.

Discussion, Limitations and Directions for Further Research

Developmental Equilibrium

The developmental perspective adopted here is apparently new in retention research, and has important implications. Consistently with the developmental equilibrium hypothesis, the largely positive effects of retention, and the maintenance of these effects, were highly consistent across different groups of students who had been retained in Years 5, 6, 7, and 8. Support for this hypothesis not only supports the robustness and consistency of the positive retention effects, but also indicates that the self-system has achieved equilibrium in relation to retention effects over this potentially volatile period. Because this is an apparently new strategy in retention research, it is important that future research tests the generalizability of these retention effects and extends to students of other ages.

Retention Effects for School Grades

The substantial Lag 1 effects in favor of retained students, particularly for math grades (*M ES* = 1.010) require further consideration. These Lag 1 effects might be argued to advantage the retained students unfairly, because they had studied the same curriculum for two consecutive years. However, this would not be the case for effects in subsequent years following retention (i.e., Lags 2–4). Hence, because of the finding that Lag 2–4 effects for math grades were non-significant, the initial positive Lag 1 effects were maintained in subsequent school years. The positive retention effects were larger for math school grades than for school grades in German, and the GPA based on other school subjects. However, this difference can be explained at least in part by the focus of this study on math, with the consequence that there were stronger controls for pre-existing differences in relation to math than there were for other school subjects—particularly those included in the GPA measure, where controls in relation to some school subjects were limited. As noted earlier, residual pre-existing differences are likely to advantage non-repeating students; this potential bias was apparently larger for non-math outcomes.

Retention Effects for Standardized Math Tests—Same Age Vs Same Year (Offset) Comparisons

Retention effects for math standardized test scores were the least positive, and were slightly negative when based on same-age comparisons (-.188, Table 1). However, these results apparently reflected—at least in part—an apparent unfairness in these comparisons, in the sense that retained students were being tested on advanced materials that they had not covered in their studies, whereas these materials had been covered by non-retained students. In an alternative strategy, we argued that retained student results should be compared with those of students who had completed the same year in school—what we refer to as offset (or same-year-in-school) comparisons. Thus, for example, results for the Year 5 retention group were compared with the results of students who had completed Year 5 in the previous wave, rather

than with the results for these same students after they had completed Year 6. For these offset comparisons, the total effects for the retention group were all positive (*M ES* = .341)—significantly so for three of six comparisons.

Interpretation of these results on the basis of standardized test scores is not straightforward. On the one hand, it might be argued that the same-age comparisons unfairly favored non-retained students, as they were taught materials covered in the test that had not been taught to the retained students. Furthermore, this same issue was present in all subsequent years (i.e., retained students were always one year behind the non-retained students). However, the standardized math test in our study focused on generic skills appropriate for the age groups, and was not specifically based on the school curriculum. This is similar to the rationale for PISA tests. Hence, the advantage for non-retained students in our study is likely to be much smaller than in studies that use tests specifically based on the curriculum covered by the non-retained students.

On the other hand, it might be argued that our offset comparisons unfairly advantage the retained students, who have been taught the same materials for two consecutive years. Again, this potential advantage would likely be even larger for a test that more closely reflected the curriculum—in this case, for the class completed by the retained students, rather than the non-retained students. However, even to the extent that such comparisons advantaged the retained students, this advantage would only be relevant for Lag 1 comparisons: in subsequent school years, previously retained students would only have been taught the new materials in a single school year. Hence, it is important to emphasize that for the offset comparisons, our results show that the positive effects of retention in the first year following retention (Lag 1 results) were maintained over subsequent school years (Lags 2–4). Furthermore, even the offset comparisons have a potential bias in favor of the non-retained students, in that the comparison group for evaluating retention (i.e., the non-retained students) is truncated, excluding all the poorest performing students who were originally part of that cohort (i.e., the retained students).

Hence, the offset comparisons provide important evidence for the benefits of retention, even for standardized test scores.

The offset approach used here, to test for the effects of retention on the basis of standardized test scores, is not the only strategy to circumvent potentially biased comparisons in favor of non-retained students. For example, an alternative approach might be to compare the results of retained students with those of their new classmates following retention (that is, those who, while in the same year in school are typically one year younger), rather than their former classmates, prior to retention. This approach would have the advantage of comparing retained students with a whole cohort of new students, rather than with a truncated cohort that excluded retained students, but would have the disadvantage that controlling for pre-existing differences might be more problematic. Although there is apparently no completely satisfactory solution to this problem, it is critical that future research provide reasonable controls in relation to potentially biased comparisons of retained and non-retained students in respect of materials that have only been taught to non-retained students. Similarly, systematic reviews and meta-analyses of the effects of retention need to distinguish results on the basis of how this issue is addressed in primary studies (see Allen, et al., 2009).

Potential Process Mechanisms to Explain Positive Retention Effects

Although they are beyond the scope of the present investigation, it is important to explore process mechanisms to explain the positive retention effects: these can be the basis of further research. The Marsh (2016) study, which was a starting point for the present investigation, used frame of reference models (e.g., social comparison theory) to predict positive effects of retention (and negative effects of acceleration) on academic self-concept. In this respect, the present investigation is consistent with previous findings. Furthermore, there is a growing body of research demonstrating that academic self-concept and achievement—particularly school grades, but also test scores—are reciprocally related (e.g., Marsh & Craven, 2006; Pinxten, et al., 2014). Relatedly, the fact that students do so much better, in terms of

school grades, after repeating a year in school, is likely to reinforce their MSC and psychological adjustment more generally. Hence, this theoretical rationale explains the results of the present investigation—in part at least.

Although apparently there have been no retention studies focusing mainly on the time required to master new materials, or on Matthew Effects, these theoretical perspectives appear to be relevant. There is clear theoretical and empirical evidence from mastery learning interventions that weaker students might merely need more time to master new material, material that can be mastered more quickly by stronger students (Caroll, 1989; Kulik, Kulik & Bangert-Drowns, 1990). There is also theoretical and empirical research on the Matthew Effect showing that without intervention, students who fall behind at any particular stage in schooling tend to fall behind even further in subsequent school years (e.g., Stanovich, 1976; Walberg & Tsai, 1983). According to Bloom (1976), if weak students are given sufficient time and resources to achieve mastery, the differences between more and less able students will diminish, and achieving mastery has potentially profound effects on positive self-beliefs and motivations to learn. Similarly, Stanovich (1976) argued that early intervention is critical, to break the vicious cycle created by Matthew Effects. Consistent with these theoretical and empirical perspectives, the fact that retained students had an extra year to learn the materials that had led to their retention not only helped them to learn those materials more effectively in the first year following retention, but also resulted in more positive self-belief and gave them a stronger basis for learning new materials in subsequent school years. Hence, retention can be seen as a potentially useful intervention to counter the negative consequences of failure to learn critical academic materials.

We also note that retained students tend to be more mature (i.e., a year older than their new classmates following retention). Indeed, it is curious that there seems to be widespread support for holding students back when they start school so that they are among the oldest in their class, rather than the youngest (also referred to as “academic red shirting”: see Gladwell,

2008), but the opposite view prevails in terms of holding students back by repeating a school year when they have not adequately mastered the materials (the so-called "old for grade" hypothesis; see Im, Hughes et al., 2013). However, Marsh (2016) argues that the advantage of being relatively older than classmates in terms of academic self-concept is similar for students who started late and those who repeat a year in school, and that this pattern of results has broad cross-national generalizability. Our results are consistent with those conclusions, but extend them in important new directions—particularly in relation to academic achievement and the long-term maintenance of short-term benefits of retention.

Limitations

A major limitation of the present investigation is the relatively small number of retained students, particularly for any given school year. Although this limitation is inherent in the nature of this research, it means that very large samples are needed to obtain even modest numbers of retained students. To some extent, our design compensated for this limitation by considering multiple retention groups. Relatedly, although the longitudinal design is clearly stronger than cross-sectional comparisons, and than comparisons based on just two waves of data for a single retention group, causal interpretations of correlational data should always be made cautiously. As noted by Allen et al. (2009), the most critical problem in making causal inferences about grade retention is the absence of randomized control trials that control for pre-retention differences, although they also note that: "For obvious reasons, random assignment of students to the 'treatments' of retention and promotion is neither feasible nor ethical" p. 481). Nevertheless, our design was particularly powerful in that we controlled for a strong set of covariates and a complete set of outcome variables for up to three waves of pre-retention data, and evaluated post-retention results for the same set of outcomes for up to three years following retention. Furthermore, uncontrolled pre-existing differences between retained and non-retained students were likely to favor non-retained students, thus working against our a priori hypotheses and supporting results in favor of retention. Importantly, the results were

consistent across multiple groups who had been retained in Years 5–8; this is consistent with our developmental equilibrium hypothesis.

Our study was based on students at the start of secondary school from a single German state, so there is clearly a need to replicate the results in different settings and with different age groups. We also note as a potential limitation the large number of students with missing data across the five waves of this longitudinal study. However, we do note that at least the positive effects of retention on academic self-concept results replicate and extend the results of Marsh (2016), which showed that the positive effects of retention generalize reasonably well across nationally representative samples of 15-year-olds from 41 different countries.

As emphasized by Reardon (2011), Parker, Jerrim, Schoon, and Marsh (2016), and many others, there is clear evidence of a steadily increasing gap between academically advantaged and disadvantaged students, particularly in the US but also in many other industrialized countries as well. There is also evidence (Micklewright & Schnepf, 2007) that the median achievement levels of countries as a whole are negatively related to the gap between the advantaged and disadvantaged. Hence, countries all over the world are trying to devise policies to decrease the gap. From this perspective, the strategic use of retention might be an effective strategy to counter this trend. However, we also note that there is an economic component of costs to the school system associated with retention and providing an extra year of schooling. There is also perhaps a "cost" to individual students in terms of potentially delaying their entry into the labor market. Hence, although this is obviously beyond the scope of our study, cost-benefit analyses would be needed to evaluate whether the costs are outweighed by the benefits.

Summary and Implications

Our results have important implications for educational researchers, but also for parents, teachers, and educational policymakers. Indeed, schools in different countries, and even in different geographic regions of the same country, use diverse strategies in relation to school retention, apparently without fully understanding the implications of these policy practices in

relation to a variety of psychosocial variables and academic achievement measures such as those considered here, which have long-term implications for academic choice and accomplishments. Particularly since the results of the present investigation are contrary to at least some accepted wisdom in relation to retention, as understood by parents and schools, there is a need for further research to more fully evaluate the generalizability and construct validity of the interpretations offered here. However, our results clearly refute any simplistic conclusion that retention is necessarily “bad”.

References

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (2003). *On the success of failure: A reassessment of the effects of retention in the primary school grades*. Cambridge University Press.

Allen, C., Chen, Q., Willson, V., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multi-level analysis. *Educational Evaluation and Policy Analysis, 31,* 480–499. http://dx.doi.org/10.3102/0162373709352239

Bloom, B. S. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.

Carroll, J. B. (1989). The Carroll Model: A 25-year retrospective and prospective view. *Educational Researcher*, *18,* 26–31.

Cham, H., Hughes, J. N., West, S. G., & Im, M. H. (2015). Effect of retention in elementary grades on grade 9 motivation for educational attainment. *Journal of School Psychology, 53*(1), 7–24. http://dx.doi.org/10.1016/j.jsp.2014.10.001

Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in 3 western European societies: England, France and Sweden. *British Journal of Sociology, 30,* 415–441.

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

Gladwell, M. (2008). *Outliers*. New York: Little, Brown and Company.

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology, 95*(1), 124–136.

Hattie, J. A. (2012). *Visible learning: A synthesis of 800+ meta-analyses on achievement.* Abingdon, England: Routledge.

Heller, K. A., & Perleth, Ch. (2000). Kognitiver Fähigkeitstest für 4.-12. Klassen, Revision (KFT 4-12+ R) [Cognitive Ability Test, revised version (KFT 4-12 + R)]. Göttingen, Germany: Hogrefe.

Holmes, C. T., & Matthews, K. M. (1984). The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. *Review of Educational Research, 54*(2), 225–236. http://dx.doi.org/10.2307/1170303

Hughes, J. N., Chen, Q., Thoemmes, F., & Kwok, O.-m. (2010). An investigation of the relationship between retention in first grade and performance on high stakes tests in third grade. *Educational Evaluation and Policy Analysis, 32*(2), 166–182.

Huguet, P., Dumas, F., Marsh, H. W., Regner, I., Wheeler, L., Suls, J., Seaton, M. & Nezlek (2009). Clarifying the role of social comparison in the Big-Fish-Little-Pond Effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology, 97*, 671–710.

Im, M. H., Hughes, J. N., Kwok, O.-m., Puckett, S., & Cerda, C. A. (2013). Effect of retention in elementary grades on transition to middle school. *Journal of School Psychology, 51*(3), 349–365. http://dx.doi.org/10.1016/j.jsp.2013.01.004

Jimerson, S. R. (2001), Meta-analysis of Grade Retention Research: Implications for Practice in the 21st Century. *School Psychology Review 30*(3), 420–437.

Jimerson, S. R. & Brown, J. A. (2013), Grade Retention. In J. A. Hattie & E. M. Anderman (Eds.), *International guide to student achievement*. New York: Routledge.

Kulik, C. L., Kulik, J. A., & Bangert-Drowns, J. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*, 265–299.

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.*

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*(2) 133–163.

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., et al. (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. http://dx.doi.org/10.1007/s10648-008-9075-6

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.

Mickelwright, J. & Schnepf, S. (2007). Inequalities in industrialised countries. In S. P. Jenkins & J. Micklewright (Eds). *Inequality and poverty re-examined* (pp. 129–145). Oxford, UK: Oxford University Press.

Moser, S. E., West, S. G., & Hughes, J. N. (2012). Trajectories of math and reading achievement in low achieving children in elementary school: Effects of early and later retention in grade. *Journal of Educational Psychology, 104*, 580–602.

Murayama, K., Pekrun, R., Suzuki, M., Marsh, H. and Lichtenfeld, S. (2016). Don't aim too high for your kids: parental over-aspiration undermines students' learning in mathematics. *Journal of Personality and Social Psychology*. http://centaur.reading.ac.uk/44843/

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.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: Government Printing Office.

Parker, P. D., Jerrim, J., Schoon, I., & Marsh, H. W. (2016). A Multination Study of Socioeconomic Inequality in Expectations for Progression to Higher Education: The Role of Between-School Tracking and Ability Stratification. *American Educational Research Journal*, Online First. doi:10.3102/0002831215621786

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology*, *36*, 36–48.

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.

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

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither opportunity*, 91–116.

Reynolds, A. J. (1992). Grade retention and school adjustment: An explanatory analysis. *Educational Evaluation and Policy Analysis*, *14*(2), 101–121.

Roderick, M. (1994). Grade retention and school dropout: Investigating the association. *American Educational Research Journal, 31*, 729–759.

Roderick, M. & Engel, M. (2001). The Grasshopper and the Ant: Motivational Responses of Low-Achieving Students to High-Stakes Testing. *Educational Evaluation And Policy Analysis, 23*, 197–227. doi: 10.3102/01623737023003197

Stanovich, K. E (1986). Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy. *Reading Research Quarterly, 21*, 360–407. doi:10.1598/rrq.21.4.1

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.

Walberg, H. J. & Tsai, S. 1983. Matthew effects in education. *American Educational Research Journal*, *20*, 359–373.

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

Wu, W., West, S. G., & Hughes, J. N. (2010). Effect of grade retention in first grade on psychosocial outcomes. *Journal of Educational Psychology*, *102*(1), 135–152. http://dx.doi.org/10.1037/a0016664

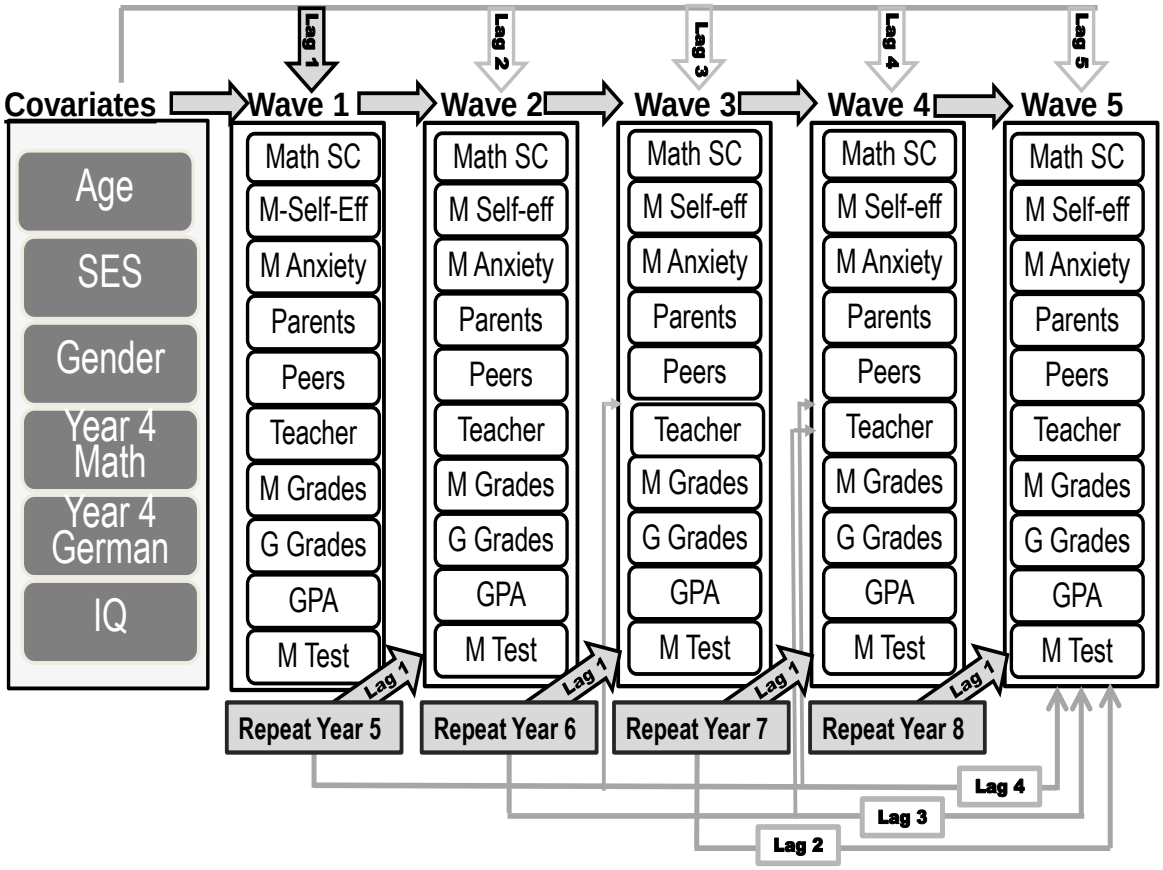

Figure 1. Waves 1–5 are the five yearly data collections in this longitudinal study. For students who repeated no grades, the data collections occurred during the first five years of secondary school (Years 5–9). The same set of 10 outcome variables was collected in each of the five waves. The six covariates are pretest control variables with paths leading from each covariate to all outcomes in Wave 1 (Lag 1 effects, as this is the immediate next wave), Wave 2 (Lag 2 effects), and so forth. Of specific interest are the four dichotomous grouping variables representing students who repeated a school year in each of the four Years 5–8. For example, a student repeating Year 5 is tested again in Years 5 (now in Wave 2 rather than Wave 1), 6, 7, and 8 (in Waves 3–5). The effect of repeating Year 5 is represented by the path from the grouping variable ("Repeat Year 5") to outcomes in the immediate subsequent wave (Lag 1 effect). The effects of repeating Year 5 are also evaluated in relation to outcomes in Wave 3 (Lag 2 effects, as the outcomes in Wave 3 are two waves following Wave 1), Wave 4 (Lag 3 effects), and Wave 5 (Lag 4 effects). Similarly, different groups of students repeating Years 6 ("Repeat Year 6"), Years 7 ("Repeat Year 7"), and Years 8 ("Repeat Year 8") are each followed in subsequent years to test the effects of repeating grades. For these subsequent groups, Lag 1 effects refer to the effects of repeating a grade on the immediate subsequent wave. For example, for the "Repeat Year 6" group, Lag 1 effects are in relation to outcomes in Wave 3, whereas for the "Repeat Year 7" group, Lag 1 effects are in relation to outcomes in Wave 4. The model depicted is a "full-forward" structural equation model that is saturated, in the sense that all paths are estimated. For example, covariates are predictors of all variables in Waves 1–5, Wave 1 variables are predictors of all variables in Waves 2–5, and so forth. Within each wave, all variables are correlated.

SES = socioeconomic status; Math SC = self-concept in math; M-Self-Eff = self-efficacy in math; M anxiety = anxiety in math; Parents = parents work with student in math; Peers = math is valued among peers; Teacher = positive reinforcement from teacher in math; M Grades = final year grade in math; G Grades = final year grade in German; GPA = average grade in other subjects; MTest = standardized math achievement test. For non-repeating students, Waves 1–5 refer to Years 5–9 (the first five years of secondary school). Of the 1,325 students considered here, the numbers of students who repeated in each year were: Year 5 (10); Year 6 (12); Year 7 (35); Year 8 (45)—a total of 103 students, or 7.8% of the total sample of 1,325 students.

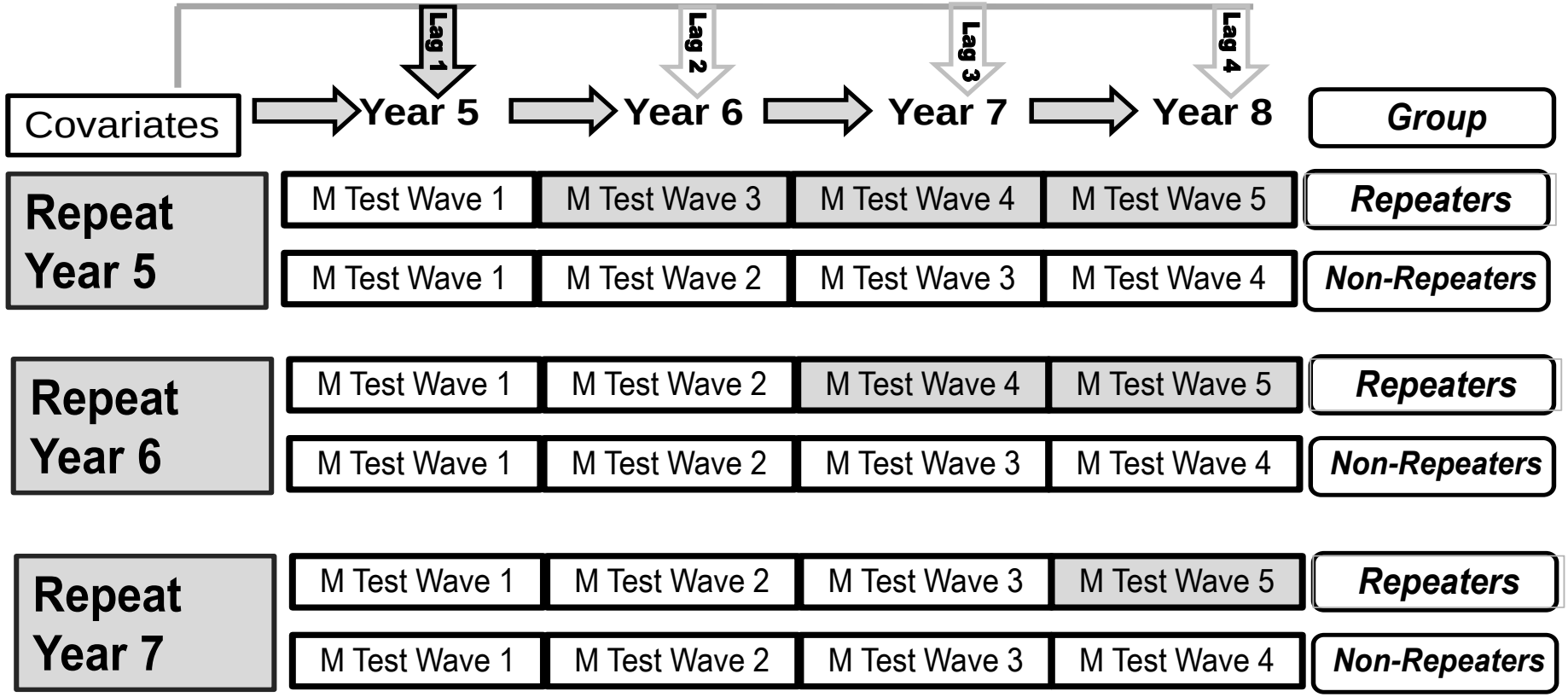

Figure 2. Offset comparisons for standardized math tests (M Tests) in Waves 1–5. Depicted is an alternative perspective on test scores in which retained students in each year following retention are compared with non-retained students from the previous wave. For example, math test scores for students repeating Year 5 in Wave 2 were compared to test scores of non-repeating students when they also completed Year 5 (but in Wave 1 rather than Wave 2). Likewise, Year 6 (Wave 2) math test scores for non-repeating students are compared to test scores from repeaters who have also just completed Year 6 (but in Wave 3 rather than Wave 2). In this way, math tests are based on the performances of students who have studied the same curriculum. Similarly, for each post-retention year (those shaded in grey for the repeater groups) for all four retention groups, comparisons based on test scores (but not other outcomes) were "offset" by one year, so that comparisons were based on students having completed the same year in school. Separate analyses were done for each retention group, except for the "repeat Year 8" retention Group, in which this offset strategy was not possible (i.e., there are no Year 9 scores for the retention group that can be compared to the Year 9 scores for the non-repeater group). In other respects, the offset analysis is like the "full-forward" structural equation model depicted in Figure 1, in that all the same covariates and outcomes are included (only the math test scores are "offset"); all covariates are predictors of all variables in Years 5–9, Year 5 variables are predictors of all variables in Years 6–9, and so forth. Again, the main focus of the present investigation is on the dichotomous grouping variables representing students who repeated a school year in one of the four Years 5–7.

Table 1
The Short- (Lag 1) and Long-Term (Lags 2–4) Effects Of Grade Retention Across Four Years of Secondary School

10 Outcomes	Invariant Lag 1 Effects (*ES*s)		Invariant Lag 2 Effects (*ES*s)		Invariant Lag 3 Effects (*ES*s)		Invariant Lag 4 Effects (*ES*s)	
	ES	*SE*	*ES*	*SE*	*ES*	*SE*	*ES*	*SE*
Math SC	**0.597****	.094	0.148	.116	-0.113	.215	0.405^	.210
M-Self-Eff	**0.359****	.084	0.079	.122	-0.155	.161	0.128	.326
M Anxiety	**0.293****	.092	0.207	.117	-0.100	.159	**0.656****	.217
Parents	0.173	.110	0.008	.129	0.277	.236	0.336^	.180
Peer	0.023	.094	-0.020	.154	0.002	.203	0.365	.270
Teacher	**0.305****	.099	0.149	.133	-0.007	.166	0.209	.194
M Grades	**1.010****	.119	-0.033	.134	0.077	.240	0.396^	.210
G Grades	**0.454****	.068	-0.059	.117	-0.025	.160	0.191	.203
GPA	**0.452****	.054	-0.092	.080	0.053	.110	-0.187	.181
M-Test	**-0.188***	.076	-0.143	.100	0.059	.091	0.222	.178
Total	**0.348****	.042	0.024	.059	0.027	.075	**.272****	.090

Note. Analysis based on Figure 1 (where variables are defined), a "full-forward" structural equation model that is saturated, in the sense that all paths are estimated and correlations within each wave are estimates. Based on support of developmental invariant model, effect sizes (*ES*) were constrained to be invariant over the four retention groups. *ES*s are the "direct effects" of repeating a grade on each outcome variable, controlling for covariates and all outcomes from prior waves. Lag 1 paths are those for the first year after repeating a grade, Lag 2 paths are the effects on the second year following grade retention, controlling for outcomes from all prior waves—including Lag 1 effects, and so forth. All outcome variables are standardized in relation to Year 5 (Wave 1) values. *ES*s that are statistically significant ($p < .05$) in relation to their standard errors (SE) are in bold.
** $p < .01$; * $p < .05$; ^ $p < .10$

Table 2

Comparison of Effects of Repeating a Year in School Based on the Original Math Tests (Same-Age Comparisons) and Math Tests Offset by one Year (Same-Year-in-School Comparisons)

		Time (number of waves following retention)					
		Total Effects			Direct Effects		
Repeating Group		Lag 1	Lag 2	Lag 3	Lag 1	Lag 2	Lag 3
Repeat Year 5	Original	-0.078 (.206)	-0.076 (.175)	0.034 (.149)	-0.078 (.206)	-0.152 (.102)	-0.107 (.189)
	Offset	0.101 (.110)	**0.603** (.146)	0.242 (.219)	0.101 (.110)	**0.442** (.146)	0.024 (.156)
Repeat Year 6	Original	0.022 (.143)	-0.079 (.148)		0.022 (.143)	-0.193 (.152)	
	Offset	**0.677** (.155)	**0.371** (.151)		**0.677** (.155)	-0.022 (.157)	
Repeat Year 7	Original	**-0.253** (.106)			**-0.253** (.106)		
	Offset	0.053 (.165)			0.053 (.165)		

Note. The analyses presented here are based on Figure 1 (where variables are defined) and on the analyses in Table 1, but differ in several important aspects. First, separate analyses were done for each of the four groups of repeaters. Second, as with the analyses in Table 1, outcomes following the repeated year are controlled for covariates and outcomes from all previous waves, and correlations within each wave are estimated. Most importantly, math standardized test scores (but none of the other outcomes) for repeating groups were offset by one wave, such that repeating students were compared to non-repeating students who had completed the same year in school (see Figure 2). Thus, for students who repeated Year 5, math test scores for Waves 3–5 (when they were in Years 6–8) were compared to math test scores for non-repeating students for Waves 2–4 (when they were also in Years 6–8). For each of the repeating groups, separate analyses are presented for the original math test scores and for one-year offset math test scores. Results are presented both for the total effect of retention (controlling for covariates and outcomes prior to retention) and for direct effects (controlling for covariates, outcomes prior to retention, and outcomes following retention—as in Table 1). Results involving Wave 5 are not presented, because the offset transformation for retention groups uses Wave 5 math test scores as Wave 4 (see Figure 2). Standard errors of each path are presented in parentheses, and statistically significant paths, $p < .05$, are presented in bold.

Supplemental Materials

Section 1: A Brief Summary of the German School System and Grade Retention

Section 2. Description of Psychosocial Outcome Variables Considered in the Present Investigation (Wording of the Items and Coefficient Alpha Estimates of Reliability) and Correlations Among the Variables Considered in the Present Investigation

Section 3: Comparison of Students with Data from all Five Waves With Those who had Missing Data

Section 4: Tests of Developmental Invariance: Theoretical Background and Statistical Tests

Section 5: Expanded Results for the Offset Comparisons Summarized in Table 2 of the Main Text

References Cited Only in Supplemental Materials

Section 1: A Brief Summary of the German School System and Grade Retention

In Germany, elementary school spans Grade 1 to 4 and secondary school starts at Grade 5 in most states, including the state of Bavaria, where the present investigation was conducted. The only exception is in Berlin, where secondary schooling starts in Grade 7. Compulsory schooling ends at Grade 9. There is no tracking in elementary school. In most states, including Bavaria, students are placed into one of three tracks at the start of secondary school, including lower-track schools (Hauptschule), medium-track schools (Realschule), and higher-track schools (Gymnasium), on the basis of their elementary school achievement. Grade retention is used in elementary school as well as across all secondary school tracks, and is based on students' achievement in main subjects. Across years and school tracks, these subjects include mathematics, German, students' first foreign language (e.g., English), and other subjects depending on grade level and track. Decisions about retention are made by teachers, and cut-off values are used for grade scores in these subjects. The number of repeated retentions per student is limited, and in the present investigation no students repeated more than one grade. In terms of using retention across all grade levels, the German school system is similar to the school system in the US, which also practices retention for low-achieving students across school years. We also note that in the German school system teachers are very reluctant to use retention in the first two years of secondary school. Hence, the majority of retention in our study came in Years 7 and 8, rather than Years 5 and 6.

For the 1,325 students considered in the present investigation, 536 (40.5%) were in the hi-track; 406 (30.6%) were in the med-track and 383 (28.9%) were in the low-track. For these 1,325 students, 1,222 did not repeat a grade, whereas 103 did repeat a grade in one of the four years: Year 5 (10); Year 6 (12); Year 7 (35); Year 8 (45)—a total of 103 students, or 7.8% of the sample. None of these students repeated more than one grade and none changed schools when they were retained. The group of 103 repeating students did not differ significantly from the 1,222 non-repeating students in terms of:

- Gender: repeating 42% female, 58% male; Non-repeating: 51% female; 49% = male. Chi-sq (df = 1, N = 1,325) = 3.01, p = .08).
- school type: repeating 43% hi-track; 34% = med-track; 23% low-track. Non-repeating: 40% hi-track; 30% = med-track; 29% low-track. Chi-sq (df = 2, N = 1,325) = 1.76, p = .41);
- Age: repeating 11.7 years, non-repeating 11.8 years. t-test (df = 1323) = -.61, p = .54.
- Family socioeconomic status (standardized M = 0, SD = 1): Repeating .01; non-repeating -.02. t-test (df = 1323) = -.33, p = .74.

Section 2. **Description of Psychosocial Outcome Variables Considered in the Present Investigation (Wording of the Items and Coefficient Alpha Estimates of Reliability) and Correlations Among the Variables Considered in the Present Investigation**

An intentionally diverse set of outcomes was considered, including the self-belief variables that were the focus of the Marsh (2016) study; achievement measures that have been the

focus of most retention studies; anxiety, to represent the emotional response of students to retention; and relations with significant others.

Math Psychosocial Outcome Variables (student self-report)
At each measurement wave the same set of items was used to assess math self-concept, math self-efficacy, math anxiety, perceived positive reinforcement from the teacher, perceived positive math reinforcement from the teacher, perceived math appreciation among peers, and perceived instructional support from parents. All these variables were based on self-report responses from students using a 5-point-Likert scale: "not true", "hardly true", "a bit true", "largely true", or "absolutely true". Across the 5 waves and six constructs, the 30 coefficient alpha estimates of reliability were generally high and consistent over the multiple waves. The actual items used to measure each of these constructs and coefficient alpha estimates of reliability are as follows:

Math self-concept (wave1: $\alpha = .876$; wave2: $\alpha = .895$; wave3: $\alpha = .893$; wave4: $\alpha = .910$; wave 5; $\alpha = .920$):

In math, I am a talented student.
It is easy for me to understand things in math.
I can solve math problems well.
It is easy to me to write tests in math.
It is easy to me to learn something in math.
If the math teacher asks a question, I usually know the right answer.

Math Self-efficacy (wave1: $\alpha = .858$; wave2: $\alpha = .860$; wave3: $\alpha = .878$; wave4: $\alpha = .876$; wave 5; $\alpha = .897$):

In math, I am sure to be able to solve even the most difficult tasks.
I am convinced that I can understand even the most difficult contents presented by our math teacher.
I am convinced that I can perform well in my math homework and on math tests.
I am convinced that I am able to master the skills taught in my math classes.

Math Test Anxiety (wave1: $\alpha = .860$; wave2: $\alpha = .866$; wave3: $\alpha = .867$; wave4: $\alpha = .870$; wave 5; $\alpha = .876$):

Before a math test I am very nervous.
When taking a math test I am tense and nervous.
Even before I take a math test I worry I could fail.
When taking a math test, I worry I will get a bad grade.
Before a math test I am so anxious that I would rather not take the test.
When I have an upcoming math test, I get sick to my stomach.
When taking a math test I am so anxious that I can't fully concentrate.

Positive Teacher Support (wave1: $\alpha = .778$; wave2: $\alpha = .792$; wave3: $\alpha = .752$; wave4: $\alpha = .816$); wave 5; $\alpha = .789$):

My math teacher praises me when I work hard at school.
My math teacher praises me when I get a good grade.
My math teacher is happy when I succeed in math.

Math Appreciation among peers (wave1: α = .852; wave2: α = .841; wave3: α = .832; wave4: α = .835; wave 5; α = .806):

Most students in my class think math is cool.
Most students in my class think that math is important.
Most students in my class think that math is fun.

Parental Assistance (wave1: α = .854; wave2: α = .885; wave3: α = .883; wave4: α = .896; wave 5; α = .891):

When I am at a loss in math, my parents help me.
When there is something I did not understand in a math test, I can ask my parents for advice.
When I made mistakes in my math homework, my parents can explain to me what I did wrong.
I can learn from my parents how to solve math problems.
When I get a poor grade in math, my parents discuss the test with me so that I do not make the same mistakes again.
My parents can explain math well.

Math achievement. Students' achievement was measured both in terms of school grades and of standardized achievement test scores (Years 5–9). School grades were end-of-year final grades obtained from school documents. School grades in German and math were available from Year 4 (the year prior to the start of secondary) and from Years 5–9 (the first four years of primary school) for all students as these are required subjects. In Years 5–9, end-of-school final grades from school records were also recorded for biology, music, sport, first foreign language, and second foreign language—depending on the subjects that a student completed. In Germany, school grades range from 1 to 6, with 1 depicting the highest and 6 the lowest achievement. For ease of interpretation, we recoded the grades prior to all analyses so that higher scores represent higher achievement. For the present purposes, three measures of school grades were considered for each wave for math and German, and the mean of all other subjects was recorded.

Math achievement was additionally assessed by the PALMA Mathematical Achievement Test (vom Hofe, Pekrun, Kleine, & Götz, 2002). Using both multiple-choice and open-ended items, this test measures student' modeling competencies 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 et al., 2002). Specifically, for each measurement point, students filled out one of two parallel versions of the same test. In each successive year, the test covered the same content areas but the number and difficulty of the items increased according to year in school completed by non-repeating students; the number of items increased from 60 to 90 items across the five waves. The number of items increases with each wave, varying between 60 and 90 items across the five waves. We included anchor items to allow for the linkage of the two test forms and the five measurement points. The obtained achievement scores were scaled using one-parameter logistic item response theory (Rasch scaling; Wu, Adams, Wilson, & Haldane, 2007), with $M = 100$ and $SD = 15$ at grade 5 (i.e., the first measurement point). Additional analyses confirmed the unidimensionality and longitudinal invariance of the test scales (see Murayama et al., 2013).

Covariates. Students' gender, IQ, age, and SES measured at Year 5 served as covariates for the overall study. Students' IQ was measured using the 25 item nonverbal reasoning subtest of the German adaptation of Thorndike's Cognitive Abilities Test (Kognitiver Fähigkeitstest, KFT 4-12+R; Heller & Perleth, 2000). SES was assessed by parent report using the Erikson Goldthorpe Portocarero (EGP) social class scheme (Erikson, Goldthorpe, & Portocarero, 1979), which consists of ordered categories of parental occupational status; higher values represent higher social class.

Supplemental Table 1

Correlations Among the Variables Considered in the Present Investigation

	Covariates						Wave 2																		
Covariates																									
Age	1.0																								
SES	.061	1.0																							
Gender	.024	.044	1.0																						
Math4	.097	.261	.084	1.0																					
German4	.108	.303	-.137	.637	1.0																				
IQ	.028	.250	-.005	.537	.495	1.0																			
Wave 1																									
MSC1	-.023	.053	.230	.240	-.035	.147	1.0																		
SEFIC1	-.022	.057	.215	.164	-.029	.087	.813	1.0																	
AX1	-.024	.083	.148	.216	.076	.141	.484	.428	1.0																
PRNTIN1	.020	.195	-.101	.101	.143	.103	.157	.217	.067	1.0															
PERMVAL1	.000	-.032	.073	-.181	-.257	-.140	.206	.237	.066	.109	1.0														
TCHRFRC1	-.047	-.019	-.046	.025	.036	.033	.234	.269	.097	.245	.274	1.0													
MGRD1	-.019	.175	.039	.485	.335	.410	.414	.318	.342	.115	-.105	.091	1.0												
DGRD1	.004	.202	-.230	.367	.523	.342	.010	.005	.070	.160	-.162	.060	.446	1.0											
GPABLSM1	-.006	.210	-.165	.483	.520	.393	.118	.108	.137	.166	-.212	.085	.584	.600	1.0										
ZMTSTC1	.004	.253	.169	.657	.466	.573	.354	.283	.263	.121	-.094	.071	.575	.343	.432	1.0									
Wave 2																									
MSC2	.003	.038	.195	.233	-.031	.105	.605	.519	.380	.117	.134	.154	.407	.020	.125	.334	1.0								
SEFIC2	.010	.035	.196	.209	-.030	.080	.527	.524	.362	.141	.154	.157	.325	-.012	.101	.275	.842	1.0							
AX2	.022	.059	.123	.194	.005	.114	.364	.330	.578	.056	.056	.074	.284	.022	.107	.252	.526	.473	1.0						
PRNTIN2	.054	.147	-.086	.072	.089	.044	.103	.139	.113	.453	.112	.207	.071	.096	.112	.058	.167	.204	.099	1.0					
PERMVAL2	.024	-.042	.044	-.131	-.179	-.144	.138	.158	.062	.031	.393	.181	-.065	-.124	-.133	-.063	.245	.260	.115	.182	1.0				
TCHRFRC2	-.002	.009	-.072	.056	.048	.004	.162	.204	.098	.139	.181	.389	.119	.109	.112	.106	.296	.319	.146	.322	.297	1.0			
MGRD2	.004	.124	.042	.440	.255	.314	.367	.291	.299	.103	-.105	.080	.685	.375	.484	.490	.506	.412	.350	.090	-.027	.180	1.0		
DGRD2	.015	.173	-.247	.312	.476	.293	.016	.017	.073	.118	-.202	.056	.439	.683	.599	.293	.058	.001	.028	.067	-.132	.085	.456	1.0	
GPABLSM2	.038	.191	-.216	.375	.458	.298	.080	.096	.104	.173	-.153	.089	.492	.570	.749	.344	.105	.088	.085	.108	-.092	.146	.516	.631	1.0
ZMTSTC2	.044	.297	.108	.692	.542	.595	.296	.231	.238	.106	-.136	.061	.540	.374	.452	.775	.315	.264	.260	.104	-.064	.137	.486	.309	.379
Wave 3																									
MSC3	.016	.011	.216	.213	-.039	.128	.500	.418	.350	.082	.140	.096	.323	-.030	.087	.299	.634	.542	.425	.113	.159	.221	.362	.004	.109
SEFIC3	-.011	.029	.207	.207	-.018	.098	.452	.425	.341	.121	.149	.132	.268	-.033	.077	.281	.571	.552	.401	.137	.140	.234	.297	-.009	.092
AX3	.022	.050	.078	.191	.065	.151	.285	.234	.449	.053	.027	.060	.237	.056	.118	.238	.356	.308	.563	.103	.032	.162	.210	.032	.091
PRNTIN3	.064	.165	-.048	-.033	-.016	-.044	.108	.128	.096	.415	.097	.119	.044	.020	.042	.004	.147	.157	.103	.517	.108	.136	.026	.025	.074
PERMVAL3	-.024	-.074	.042	-.148	-.212	-.167	.149	.153	.065	.009	.297	.084	-.068	-.139	-.144	-.074	.178	.181	.129	.094	.377	.191	-.014	-.108	-.081
TCHRFRC3	-.045	-.026	-.031	-.044	-.030	-.047	.091	.088	.068	.092	.098	.202	-.001	.013	-.002	.014	.154	.149	.101	.153	.137	.319	.045	.039	.054
MGRD3	.051	.144	.070	.379	.241	.286	.278	.239	.249	.105	-.068	.050	.535	.305	.421	.415	.355	.297	.264	.111	-.009	.160	.583	.364	.452
DGRD3	.019	.149	-.275	.258	.415	.274	.004	.002	.101	.138	-.154	.045	.387	.566	.490	.259	.072	.031	.067	.119	-.095	.140	.373	.593	.530
GPABLSM3	.036	.183	-.232	.337	.408	.251	.082	.089	.111	.149	-.123	.062	.440	.499	.638	.283	.127	.099	.122	.134	-.061	.153	.438	.538	.696
ZMTSTC3	.068	.272	.073	.696	.528	.598	.271	.194	.228	.127	-.153	.040	.552	.364	.472	.743	.307	.265	.273	.086	-.103	.089	.511	.326	.398
Wave 4																									
MSC4	.016	.039	.211	.211	-.013	.124	.471	.419	.341	.070	.128	.081	.315	-.017	.090	.288	.557	.487	.358	.081	.111	.168	.362	-.001	.098
SEFIC4	.004	.062	.195	.207	.015	.121	.432	.428	.313	.115	.140	.107	.277	.015	.108	.273	.523	.511	.345	.126	.118	.207	.322	.003	.105
AX4	.027	.106	.086	.186	.093	.153	.246	.217	.410	.046	-.009	.040	.227	.067	.125	.239	.292	.243	.477	.034	-.046	.075	.235	.060	.090
PRNTIN4	.011	.182	.025	-.063	-.078	-.070	.089	.122	.093	.348	.126	.057	.007	-.027	-.026	-.014	.154	.172	.097	.463	.147	.140	.021	-.047	.011
PERMVAL4	-.034	-.012	.034	-.137	-.200	-.118	.153	.168	.055	.046	.317	.142	-.054	-.127	-.117	-.044	.196	.204	.117	.119	.381	.190	-.012	-.124	-.067
TCHRFRC4	-.044	.015	-.093	-.046	-.030	-.012	.109	.117	.052	.109	.101	.224	.060	.087	.052	.018	.158	.161	.118	.145	.076	.274	.099	.077	.091
MGRD4	.088	.092	.038	.269	.132	.184	.278	.207	.233	.048	-.043	.020	.483	.264	.346	.277	.299	.235	.234	.043	-.021	.082	.514	.306	.371
DGRD4	.049	.145	-.211	.224	.373	.244	.011	-.003	.070	.122	-.143	.057	.345	.523	.433	.239	.041	-.001	.065	.102	-.083	.103	.343	.541	.464
GPABLSM4	.058	.183	-.197	.294	.376	.197	.084	.077	.078	.157	-.116	.051	.354	.431	.576	.230	.102	.083	.085	.087	-.074	.092	.348	.466	.611
ZMTSTC4	.104	.304	.096	.715	.574	.598	.260	.204	.249	.121	-.134	.049	.524	.346	.457	.729	.295	.259	.259	.102	-.098	.136	.472	.293	.391
Wave 5																									
MSC5	.025	.058	.182	.190	-.014	.111	.439	.387	.300	.099	.120	.089	.317	.013	.125	.247	.502	.432	.346	.083	.124	.150	.368	.030	.138
SEFIC5	.014	.075	.147	.178	.005	.101	.400	.384	.288	.143	.133	.121	.296	.029	.143	.236	.458	.438	.319	.131	.118	.185	.320	.031	.148
AX5	.054	.079	.059	.173	.103	.156	.181	.167	.343	.038	.007	.038	.177	.059	.100	.175	.207	.176	.397	.034	-.028	.056	.191	.053	.106
PRNTIN5	.000	.167	.036	-.081	-.115	-.079	.073	.088	.060	.301	.105	.048	-.012	-.057	-.021	-.020	.090	.113	.061	.391	.093	.058	-.020	-.076	-.025
PERMVAL5	-.017	-.017	.057	-.096	-.162	-.068	.089	.133	.043	.028	.190	.060	-.027	-.082	-.063	-.041	.183	.180	.118	.043	.275	.095	-.016	-.095	-.056
TCHRFRC5	.007	-.006	-.137	-.078	-.029	-.064	.039	.065	.005	.112	.095	.132	.006	.069	.043	-.039	.073	.067	.019	.107	.106	.154	-.003	.051	.060
MGRD5	.081	.068	-.012	.243	.117	.194	.268	.193	.212	.048	-.040	.016	.451	.234	.337	.301	.329	.269	.250	.065	-.025	.117	.516	.262	.350
DGRD5	.038	.138	-.244	.236	.361	.260	.026	.026	.068	.148	-.159	.051	.306	.475	.410	.203	.065	.037	.047	.119	-.115	.112	.302	.511	.428
GPABLSM5	.039	.161	-.170	.328	.379	.224	.076	.068	.094	.091	-.128	-.005	.354	.416	.551	.261	.101	.069	.076	.061	-.071	.069	.360	.468	.600
ZMTSTC5	.097	.315	.106	.699	.546	.579	.266	.214	.232	.149	-.132	.042	.532	.350	.456	.710	.286	.245	.254	.087	-.077	.107	.462	.300	.372
REPY5W2	-.021	-.063	.000	-.154	-.101	-.124	-.046	-.028	-.049	-.015	.058	.008	-.167	-.158	-.218	-.138	.030	.016	-.075	.027	.050	.047	-.024	-.074	-.070
REPY6W3	.047	-.032	.048	-.045	-.054	-.015	-.028	-.006	-.037	-.018	.128	-.012	-.116	-.094	-.145	-.035	-.015	.013	-.017	-.045	.039	-.021	-.137	-.151	-.147
REPY7W4	.032	.003	.028	-.023	.009	.004	-.068	-.055	-.032	-.054	.006	-.027	-.118	-.099	-.099	-.061	-.115	-.083	-.099	-.054	-.012	-.048	-.115	-.109	-.123
REPY8W5	-.018	.032	.021	-.034	-.008	-.010	-.023	-.004	-.052	-.003	.016	.002	-.088	-.069	-.074	-.002	-.025	-.007	-.030	-.009	-.014	.003	-.106	-.092	-.129
Mean	11.907	.000	.000	.000	.000	.000	.001	-.002	.025	.006	-.033	-.032	.156	.122	.062	.132	-.201	-.194	.003	-.294	-.338	-.172	.047	.117	.039
Var	0.213	1.000	1.000	1.000	1.000	1.000	0.967	0.992	0.981	0.962	0.993	1.007	0.902	0.949	0.436	1.018	1.192	1.193	1.019	1.172	0.891	1.061	0.973	0.848	0.414

Continued

	Wave 3										Wave 4									
Wave 3																				
MSC3	1.0																			
SEFIC3	.843	1.0																		
AX3	.533	.489	1.0																	
PRNTIN3	.203	.232	.075	1.0																
PERMVAL3	.313	.286	.128	.141	1.0															
TCHRFRC3	.259	.267	.164	.218	.284	1.0														
MGRD3	.500	.422	.346	.082	.070	.112	1.0													
DGRD3	.080	.068	.108	.061	-.070	.040	.435	1.0												
GPABLSM3	.160	.138	.145	.110	-.050	.079	.496	.591	1.0											
ZMTSTC3	.316	.287	.289	.025	-.093	.004	.461	.311	.391	1.0										
Wave 4																				
MSC4	.683	.604	.422	.168	.213	.154	.441	.046	.129	.307	1.0									
SEFIC4	.633	.620	.392	.191	.213	.193	.392	.055	.137	.292	.867	1.0								
AX4	.404	.345	.638	.057	.042	.056	.302	.100	.127	.271	.525	.493	1.0							
PRNTIN4	.213	.218	.125	.612	.212	.152	.123	.026	.060	-.035	.242	.280	.090	1.0						
PERMVAL4	.248	.242	.124	.172	.474	.157	.064	-.081	-.048	-.084	.297	.295	.084	.269	1.0					
TCHRFRC4	.170	.183	.135	.209	.201	.321	.110	.090	.105	.036	.267	.296	.150	.231	.260	1.0				
MGRD4	.397	.339	.265	.096	.054	.060	.630	.325	.427	.367	.546	.474	.370	.102	.091	.226	1.0			
DGRD4	.067	.044	.086	.063	-.061	.041	.387	.628	.544	.285	.079	.076	.111	.042	-.071	.097	.427	1.0		
GPABLSM4	.133	.102	.127	.079	-.035	.039	.405	.484	.712	.360	.139	.151	.137	.058	-.043	.119	.427	.506	1.0	
ZMTSTC4	.316	.291	.278	.028	-.077	.010	.470	.311	.393	.838	.347	.348	.288	.005	-.066	.056	.376	.305	.359	1.0
Wave 5																				
MSC5	.590	.497	.348	.137	.207	.129	.425	.062	.173	.275	.695	.612	.432	.182	.258	.202	.494	.109	.165	.302
SEFIC5	.537	.504	.319	.171	.208	.157	.387	.077	.194	.267	.639	.642	.388	.227	.273	.223	.437	.116	.186	.301
AX5	.329	.285	.507	.038	.035	.071	.273	.076	.118	.229	.397	.349	.627	.072	.059	.104	.300	.101	.124	.252
PRNTIN5	.138	.135	.076	.553	.132	.100	.035	-.029	.000	-.056	.141	.158	.069	.670	.184	.143	.038	-.019	-.019	-.057
PERMVAL5	.197	.179	.103	.115	.350	.103	.038	-.055	-.036	-.069	.186	.189	.085	.155	.401	.162	.055	-.074	-.031	-.035
TCHRFRC5	.120	.107	.062	.131	.144	.248	.040	.095	.107	-.052	.119	.135	.081	.155	.136	.253	.095	.092	.110	-.038
MGRD5	.364	.297	.231	.084	.053	.067	.591	.323	.384	.362	.464	.408	.315	.082	.087	.168	.674	.361	.377	.378
DGRD5	.040	.000	.051	.086	-.116	.030	.350	.550	.468	.252	.057	.058	.086	.041	-.117	.080	.350	.574	.436	.281
GPABLSM5	.129	.109	.091	.070	-.072	-.002	.391	.464	.638	.370	.161	.149	.140	.010	-.044	.083	.413	.467	.684	.376
ZMTSTC5	.283	.259	.257	.000	-.090	-.003	.459	.311	.385	.805	.324	.319	.269	.000	-.076	.005	.356	.310	.355	.864
REPY5W2	.046	.010	-.048	.042	.077	.035	-.063	-.094	-.143	-.148	-.009	-.017	-.047	.043	.061	.019	-.091	-.090	-.101	-.135
REPY6W3	.026	.001	.042	-.028	.046	-.009	-.026	-.041	-.047	-.048	-.005	.001	.010	-.037	.000	-.031	-.062	-.079	-.060	-.045
REPY7W4	-.126	-.115	-.121	-.043	-.042	-.098	-.241	-.174	-.189	-.070	-.024	-.050	-.034	-.003	-.016	.006	.000	-.070	-.056	-.091
REPY8W5	-.063	-.057	-.085	-.003	-.051	.018	-.188	-.148	-.155	-.037	-.103	-.080	-.087	-.069	-.021	-.083	-.281	-.190	-.197	-.073
Mean	-.449	-.432	-.034	-.622	-.674	-.303	-.176	-.043	-.094	1.169	-.412	-.364	.112	-.902	-.666	-.251	-.151	.018	-.158	1.851
Var	1.124	1.125	0.981	1.222	0.771	1.003	0.995	0.854	0.411	1.228	1.213	1.179	0.945	1.351	0.780	1.107	1.048	0.826	0.510	1.501

Continued

Continued

	Wave 5													Repeat
Wave 5														
MSC5	1.0													
SEFIC5	.874	1.0												
AX5	.469	.433	1.0											
PRNTIN5	.153	.204	.047	1.0										
PERMVAL5	.259	.265	.033	.170	1.0									
TCHRFRC5	.234	.242	.112	.155	.204	1.0								
MGRD5	.596	.548	.369	.037	.064	.137	1.0							
DGRD5	.119	.109	.114	.003	-.086	.067	.405	1.0						
GPABLSM5	.208	.209	.127	-.040	-.078	.082	.461	.521	1.0					
ZMTSTC5	.313	.310	.253	-.046	-.034	-.001	.362	.283	.363	1.0				
REPY5W2	.014	-.014	.012	.048	.063	.023	-.043	-.062	-.121	-.134	1.0			
REPY6W3	-.010	-.015	.023	.017	.020	-.009	-.011	-.078	-.085	-.034	-.008	1.0		
REPY7W4	-.040	-.048	.009	.011	-.017	.022	-.088	-.054	-.065	-.086	-.015	-.016	1.0	
REPY8W5	-.025	-.032	-.026	-.021	-.039	-.011	-.059	-.073	-.074	-.085	-.016	-.018	-.031	1.0
Mean	-.475	-.454	.062	-1.139	-.664	-.231	-.126	.102	.021	2.176	.008	.009	.027	.034
Var	1.271	1.250	0.929	1.382	0.662	0.940	1.176	0.868	0.453	1.697	0.007	0.009	0.026	0.033

Note. Correlations among factors used in the present investigation. N = 1,325. Waves 1–5 are the five yearly data collections in this longitudinal data. For students who repeated no grades, the data collections occurred during the first five years of secondary school (Years 5–9). The five covariates were treated as control variables: age, SES (parents' socioeconomic status) gender (1 = female, 2 = male), math4 (math grade in year 4, the last year of primary school), german4 (German grade in year 4), and IQ. The same set of 10 outcome variables was collected in each of the five waves: MSC (math self-concept), sefic (math self-efficacy), ax (math anxiety), prntin (parental academic assistance), permval (peer appreciation of math, tchrfrc (positive teacher support), mgrd (math school grade), dgrd (german school grade), gpablsm (average grade in other core subjects), zmtstc (standardized math test).

Section 3: Comparison of Students with Data from all Five Waves with Those who had Missing Data

In the supplemental analyses presented in this section, based on the 10 outcome variables considered in Year 5, we evaluated differences between repeating and non-repeating groups for the 1,325 students who had complete data for all five waves, with those based on the 745 who had not participated in all five waves beyond the first wave, controlling for pretest variables (gender, age, SES, primary school grades, IQ). For the purposes of these analyses, we evaluated the effects of:

- Repeat: first-order (main) effects of differences between students who have and have not repeated in Year 5 (wave 1);
- Include: first-order (main) effects of differences between the 1,325 students who had complete data for all five waves, and the 745 who had not participated in all five waves beyond the first wave;
- Repeat-by-Include Interaction: a test of whether the effects of being repeated varied as a function of being included (i.e., there being missing data following Year 5);
- Effects of pretest variables.

However, the primary interest is in the Repeat-by-Include Interaction, a test of whether the effect of repeating a grade varied as a function of being excluded from the analysis due to missing data following the first wave.

Next we tested the effects of constraining all 10 interaction terms simultaneously to be zero, which resulted in a non-significant difference [chi-square (df = 10, N = 2070) = 8.267, p = .603]. In summary, differences between the repeat and non-repeat groups did not vary significantly as a function of missing data.

Supplemental Table 2

Comparison of Students with Data from all Five Waves With Those who had Missing Data

	Set of 10 Outcome Variables									
	Math Self concept	Math Self-Effic	Math Anxiety	Math Grade	German Grade	Other School Grades	Math Test Scores	Parent Inter-actions	Peer Inter-actions	Teacher Inter-actions
Predictor Variables										
Main Effects										
Include	-.012	-.007	.005	-.037	.005	-.085	.148**	.111*	.083	.064
Repeat	-.017	.008	.104**	-.059	-.043	.014	-.112**	-.098**	-.166**	-.072**
Interaction Effect										
IncludeXrepeat[a]	.143	-.020	-.718	.504	.702	.010	.013	-.296	-.207	.222
Pretest Covariates										
Age	-.007	.005	.039	.054*	.031	-.025	-.029	-.036	-.023	-.016
SES	-.071**	-.062**	.059**	-.135**	-.017	.015	-.098**	-.065**	-.072**	-.046**
Sex	.180**	.172**	-.115**	-.120**	.045	-.040	.010	-.168**	-.144**	.116**
MGrdYr4	.419**	.308**	-.278**	.054	.001	.023	.486**	.160**	.308**	.407**
GGrdYr4	-.179**	-.099**	.036	.013	-.185**	.069	.141**	.508**	.324**	.038
School-Type	.277**	.240**	-.182**	.043	.064	.104**	.461**	.359**	.263**	-.041
IQ	.133**	.081**	-.100**	.032	-.034	.031	.260**	.140**	.150**	.307**

Note. Repeat = students repeating Year 5 (wave 1). Include = students who had complete data for all five waves. IncldxRepeat = Repeat-by-Include Interaction

* $p < .05$; ** $p < .01$.

[a] In an additional model, we tested the effects of simultaneously holding all 10 interaction effects to be exactly zero. The chi-square difference ($df = 8$) = 8.203 was not statistically significant ($p > .05$), supporting the contention that the differences between the repeating and non-repeating students on the 10 outcomes in Year 5 did not vary significantly as a function of missing data, thereby supporting the appropriateness of the analyses

Section 4: Tests of Developmental Invariance: Theoretical Background and Statistical Tests

Theoretical Background for Tests of Developmental Invariance

In this section we provide further discussion of the developmental equilibrium hypothesis that posits the consistency of the retention effects over early-to-middle adolescence on the basis of longitudinal data and multiple retention groups across this critical developmental period. 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. 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. We also note that support for such tests of developmental equilibrium facilitates the 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).

Here we test developmental equilibrium in relation to the invariance of retention effects in each of four year groups spanning this early-to-middle adolescent period, based on the assumption that the self-system has attained a developmental balance in relation to retention. More specifically, we evaluate support for developmental invariance (Hypothesis 7), based on the a priori hypothesis that retention effects are the same for students retained in Years 5, 6, 7 and 8. In pursuit of this aim we briefly discuss the evaluation of goodness of fit and then discuss two alternative models.

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 structural equation model 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 typically are 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. However, we emphasize that these fit indices and cut-off values should be treated only as rough guidelines, to be interpreted cautiously in combination with other features of the data.

Alternative Models

Means Model. The model depicted in Figure 1 is a "full-forward" structural equation model: all paths relating variables in different waves are estimated, as are all correlations and correlated residuals relating variables within each wave. In this sense the model is saturated, such that degrees-of-freedom (*df*) = 0 and the necessarily "perfect fit" is not particularly meaningful, even though the parameter estimates are interpretable. The critical parameter estimates are the paths from the four dichotomous retention variables (e.g., "Repeat Year 5" in Figure 1) to outcomes in all subsequent waves. In this application there is a total of 40 Lag 1 paths (i.e., 10 outcomes x 4 retention groups), 30 Lag 2 paths (noting that there are no Lag 2 paths possible for the "Repeat Year 8" in Figure 1), 20 Lag 3 paths (for students repeating Years 5 and 6), and 10 Lag 4 paths (only available for students repeating Year 5). The most important of these retention effects are the 40 Lag 1 estimates, but the high-order lagged effects (lags 2–4) also provide valuable information about whether Lag 1 effects have been maintained.

We took two complementary approaches to summarize the retention effects. In the "means model" we used the model constraint option in Mplus to compute the mean effects size (*ES*) across the relevant retention groups for each outcome, along with the standard error and a test as to whether the mean was significantly different from zero. Thus, for example, the mean *ES* for math self-concept (MSC) was the mean retention effect averaged across the four retention groups (i.e., students retained in Years 5, 6, 7 and 8). Importantly, this model is still saturated, in that it did not impose any constraints. However, it provides a much stronger, more robust test of the overall retention effects, in that the test of the mean across retention groups is based on a much larger N than are tests of each group separately, compensating in part for the small number of retained students in each retention group. These results are also heuristic in that they provide comparisons of the results for each group considered separately, as well as the mean across groups.

Developmental Invariance Model. The means model is heuristic, but—except by informal inspection—does not test the developmental invariance model (Hypothesis 7) that the retention effects are the same across the four retention groups that span the early-to-middle adolescent period considered here. In order to test this model more formally, we next tested the developmental invariance model, in which all lagged effects were constrained to be the same across the four retention groups. Thus, for example, Lag 1 retention effects for MSC were constrained to be the same for students retained in Years 5, 6, 7 and 8. This highly constrained, parsimonious model imposed a total of 60 invariance constraints: 30 for Lag 1 effects (i.e., the 10 outcomes x 4 retention = 40 retention *ES*s were represented by 10 *ES*, one for each outcome across the four retention groups); 20 for Lag 2 retention effects and 10 for Lag 3 retention effects. Particularly given the large number of constraints, the goodness of fit of this model was remarkably good: Chi-square = 81.73 (*df* = 60), CFI = .999; TLI = .982; RMSEA = .017. The excellent fit of this model provides strong support for the developmental invariance Hypothesis 7. Not surprisingly, the mean *ES*s (based on the means model) and the invariant *ES*s (based on the developmental invariance model) are very similar, and both provide a more parsimonious summary of the retention effects. However, due in part to the fact that the number of retained students varies across the four year groups, the *ES*s based on the two models are not the same. For the present purposes the juxtaposed results based on both models are presented in the main text, but the estimates for each of the year groups considered separately, as well as for the means (based on the means model) and the invariant estimates (based on the developmental invariance model) are presented below.

Supplemental Table 3

The Short- (Lag 1) and Long-Term (Lags 2–4) Effects Of Grade Retention in Each of Four Years of Secondary School

Outcomes	Repeated Year 5		Repeated Year 6		Repeated Year 7		Repeated Year 8		Total Across All Year Groups			
	Lag 1 (wave2 Yr6)		Lag 1 (wave3 Yr7)		Lag 1 (wave4 Yr8)		Lag 1 (wave5 Yr9)		Mean Lag 1		Invariant Lag 1	
	ES	*SE*	*ES*	*SE*	*ES*	*SE*	*ES*	*SE*	*ES*	*SE*	*ES*	*SE*
MSC	**1.091**	.382	0.419	.228	**0.574**	.121	**0.573**	.117	**0.664**	.125	**0.597**	.094
M Slf-Eff	**0.752**	.377	-0.001	.195	**0.321**	.139	**0.418**	.129	**0.372**	.114	**0.359**	.084
Anxiety	-0.384	.427	**0.602**	.209	**0.313**	.133	**0.343**	.124	0.219	.124	**0.293**	.092
Parents	0.537	.331	-0.083	.219	**0.389**	.186	-0.016	.147	0.207	.122	0.173	.110
Peer	0.154	.612	0.174	.283	0.111	.128	-0.089	.123	0.088	.177	0.023	.094
Teacher	0.786	.481	-0.001	.306	**0.499**	.178	0.162	.130	**0.362**	.151	**0.305**	.099
M-Grade	**1.392**	.377	**0.698**	.263	**1.114**	.180	**0.905**	.138	**1.027**	.146	**1.010**	.119
G-Grade	**0.723**	.150	**0.638**	.178	**0.396**	.154	**0.354**	.120	**0.528**	.072	**0.454**	.068
GPA	**0.746**	.180	**0.502**	.129	**0.410**	.082	**0.356**	.096	**0.503**	.06	**0.452**	.054
M-Test	-0.079	.205	0.018	.144	**-0.254**	.107	-0.211	.111	-0.131	.085	**-0.188**	.076
Total	**0.572**	.209	**0.297**	.069	**0.387**	.061	**0.279**	.052	**0.384**	.063	**0.348**	.042
	Lag 2 (wave 3 Yr7)		Lag 2 (wave 4 Yr8)		Lag 2 (wave 5 Yr9)				Mean Lag 2		Invariant Lag 2	
MSC	**0.634**	.204	-0.228	.237	0.123	.170			0.176	.112	0.148	.116
M Slf-Eff	0.151	.248	-0.141	.182	0.104	.169			0.038	.123	0.079	.122
Anxiety	0.151	.393	-0.051	.185	**0.327**	.162			0.142	.145	-0.207	.117
Parents	0.324	.216	**-0.530**	.247	0.099	.163			-0.036	.122	0.008	.129
Peer	0.275	.456	-0.427	.240	0.024	.195			-0.043	.177	-0.020	.154
Teacher	-0.058	.300	-0.203	.176	0.327	.179			0.022	.134	0.149	.133
M-Grade	0.015	.337	-0.168	.248	0.003	.166			-0.050	.143	-0.033	.134
G-Grade	-0.271	.237	-0.256	.187	0.185	.138			-0.114	.108	-0.059	.117
GPA	**-0.417**	.149	-0.002	.093	0.048	.097			-0.124	.080	-0.092	.080
M-Test	-0.152	.101	-0.202	.153	-0.118	.149			-0.157	.090	-0.143	.100
Total	0.065	.154	**-0.221**	.094	0.112	.070			-0.014	.060	0.024	.059
	Lag 3 (wave 4 Yr8)		Lag 3 (wave 5 Yr9)						Mean Lag 3		Invariant Lag 3	
MSC	-0.366	.315	0.051	.275					-0.158	.218	-0.113	.215
M Slf-Eff	-0.314	.218	-0.029	.249					-0.172	.159	-0.155	.161
Anxiety	-0.042	.119	0.215	.262					0.087	.145	0.100	.159
Parents	-0.077	.301	0.527	.296					0.225	.229	0.277	.236
Peer	0.011	.268	-0.037	.263					-0.013	.202	0.002	.203
Teacher	0.032	.232	-0.068	.255					-0.018	.157	-0.007	.166
M-Grade	-0.301	.312	**0.611**	.192					0.155	.172	0.077	.240
G-Grade	-0.007	.235	-0.133	.308					-0.070	.166	-0.025	.160
GPA	0.135	.178	-0.067	.124					0.034	.097	0.053	.110
M-Test	-0.004	.157	0.085	.159					0.041	.089	0.059	.091
Total	-0.093	.106	0.115	.091					0.068	.870	0.027	.075
	Lag 4 (wave 5 Yr9)											
MSC	0.405	.209							0.405	.209	0.405	0.209
M Slf-Eff	0.137	.325							0.137	.325	0.137	0.325
Anxiety	**0.670**	.222							**0.670**	.222	**0.670**	0.222
Parents	**0.346**	.177							**0.346**	.177	**0.346**	0.177
Peer	0.356	.269							0.356	.269	0.356	0.269
Teacher	0.212	.196							0.212	.196	0.212	0.196
M-Grade	0.408	.210							0.408	.210	0.408	0.210
G-Grade	0.188	.212							0.188	.212	0.188	0.212
GPA	-0.199	.181							-0.199	.181	-0.199	0.181
M-Test	0.223	.177							0.223	.177	0.223	0.177
Total	**0.275**	.090							**0.275**	.090	**0.275**	0.090

Grand total									**0.179**	.032	**0.179**	0.032

Note. Analysis based on Figure 1 (where variables are defined), a "full-forward" structural equation model that is saturated, in the sense that all paths are estimated and correlations within each wave are estimates. Effect sizes (*ES*) are the "direct effects" of repeating a grade (i.e., the four dichotomous grouping variables representing students who repeated a school year in each of the four Years 5–8, coded as 1, compared to non-repeating, continuing students (coded 0) on each outcome variable, controlling for covariates and all outcomes from prior waves. Lag 1 paths are those for the first year after repeating a grade, Lag 2 paths are the effects on the second year following grade retention, controlling for outcomes from all prior waves—including Lag 1 effects, and so forth. All outcome variables are standardized in relation to Year 5 (Wave 1) values. *ES*s that are statistically significant ($p < .05$) in relation to their standard errors (SE) are in bold.

Section 5: Expanded Results for the Offset Comparisons Summarized in Table 2 of the Main Text

Supplemental Table 5 (expanded version of material in Table 2 of main text)

Comparison of Effects of Repeating a Year in School Based on Original Math Tests (for Students in a Different Year in School) and Math Tests Offset by one Year (for Students Having Completed the Same Year in School)

	Original Math Test		Offset Math Test		Original Math Test		Offset Math Test		Original Math Test		Offset Math Test	
	Repeat Year 5 Lag 1 (wave2 Yr 6)				Repeat Year 6 Lag 1 (wave3 Yr 7)				Repeat Year 7 Lag 1 (wave4 Yr 8)			
MSC	**1.090**	0.382	**1.087**	0.382	**0.407**	0.229	**0.406**	0.230	**0.584**	0.122	**0.585**	0.122
M Slf-Eff	0.750	0.377	0.748	0.377	-0.003	0.195	-0.004	0.195	**0.330**	0.138	**0.330**	0.138
Anxiety	-0.386	0.427	-0.388	0.427	**0.600**	0.207	**0.599**	0.207	**0.314**	0.134	**0.314**	0.134
Parents	0.538	0.331	0.537	0.331	-0.090	0.216	-0.077	0.216	**0.397**	0.188	**0.399**	0.188
Peer	0.154	0.613	0.154	0.613	0.170	0.282	0.167	0.281	0.115	0.127	0.115	0.127
Teacher	0.787	0.481	0.787	0.481	0.001	0.304	0.001	0.306	**0.503**	0.177	**0.503**	0.177
M-Grade	**1.393**	0.377	**1.391**	0.377	**0.697**	0.226	**0.697**	0.226	**1.120**	0.183	**1.117**	0.182
G-Grade	**0.724**	0.188	**0.726**	0.188	**0.644**	0.177	**0.644**	0.176	**0.398**	0.154	**0.396**	0.154
GPA	**0.747**	0.180	**0.747**	0.180	**0.510**	0.129	**0.510**	0.129	**0.406**	0.082	**0.404**	0.082
M-Test	-0.078	0.205	0.101	0.110	0.022	0.143	0.677	0.155	**-0.253**	0.106	0.053	0.165
	Repeat Year 5 Lag 2 (wave3 Yr 7)				Repeat Year 6 Lag 1 (wave4 Yr 8)				Repeat Year 7 Lag 1 (wave4 Yr 8)**			
MSC	**1.159**	0.223	**1.157**	0.223	0.035	0.232	0.034	0.232	**0.490**	0.151	**0.490**	0.151
M Slf-Eff	**0.578**	0.272	**0.576**	0.272	0.032	0.197	0.032	0.197	**0.397**	0.160	**0.397**	0.16
Anxiety	-0.071	0.341	-0.072	0.341	0.302	0.237	0.302	0.237	**0.491**	0.157	**0.491**	0.157
Parents	**0.560**	0.212	**0.560**	0.212	-0.389	0.225	-0.389	0.225	0.302	0.179	0.308	0.179
Peer	0.428	0.511	0.428	0.512	-0.324	0.239	-0.325	0.239	0.058	0.170	0.057	0.154
Teacher	0.310	0.343	0.310	0.343	-0.170	0.177	-0.172	0.177	**0.478**	0.151	**0.481**	0.15
M-Grade	0.692	0.424	0.690	0.424	0.224	0.212	0.225	0.212	**0.471**	0.139	**0.468**	0.142
G-Grade	0.197	0.197	0.199	0.197	0.140	0.162	0.149	0.162	**0.376**	0.171	**0.374**	0.169
GPA	-0.019	0.166	-0.019	0.166	**0.331**	0.128	**0.331**	0.128	0.249	0.101	0.248	0.1
M-Test	-0.076	0.175	**0.603**	0.146	-0.079	0.148	**0.371**	0.151	-0.193	0.171	-0.193	0.129

Supplemental Table 5 (continued)

	Repeat Year 5 Lag 3 (wave4 Yr 8)				Repeat Year 6 Lag 3 (wave5 Yr 9)**			
MSC	0.434	0.273	0.432	0.273	0.125	0.296	0.125	0.296
M Slf-Eff	0.315	0.197	0.313	0.197	0.003	0.290	0.003	0.290
Anxiety	-0.051	0.257	-0.051	0.257	0.370	0.287	0.370	0.287
Parents	0.452	0.344	0.452	0.344	0.284	0.290	0.284	0.290
Peer	0.323	0.259	0.321	0.260	-0.067	0.277	-0.067	0.277
Teacher	0.304	0.219	0.304	0.219	-0.030	0.292	-0.030	0.292
M-Grade	0.141	0.287	0.139	0.287	0.759	0.225	0.753	0.228
G-Grade	0.123	0.224	0.125	0.224	0.086	0.293	0.060	0.288
GPA	0.197	0.144	0.198	0.144	0.111	0.171	0.112	0.169
M-Test	0.034	0.149	0.242	0.219	0.107	0.171	0.107	0.171

	RepeatYear5Lag 4(wave5Yr9)**			
MSC	**0.847**	0.226	**0.845**	0.225
M Slf-Eff	0.428	0.331	0.427	0.331
Anxiety	**0.554**	0.250	**0.554**	0.250
Parents	**0.558**	0.236	**0.555**	0.236
Peer	0.485	0.296	0.484	0.296
Teacher	0.251	0.154	0.251	0.154
M-Grade	**0.682**	0.300	**0.698**	0.292
G-Grade	0.352	0.161	0.365	0.155
GPA	0.004	0.194	0.013	0.196
M-Test	0.048	0.193	0.049	0.194

Note. Analyses are based on Figure 1 (where variables are defined) and analyses in Table 1, but differ in several important aspects. First, separate analyses were done for each of the four groups of repeaters. Second, as with the analyses in Table 1, outcomes following the repeated year were controlled for covariates and outcomes from all previous waves, and correlations within each wave were estimated. However, Lag 1 and Lag 2 effects (first and second years following the grade retention) were not controlled in the estimates of Lag 2 and Lag 3 effects. In this sense the Lag 2 and Lag 3 effects were the "total effects" of the grade retention intervention, rather than the "direct effects" in Table 1. Most importantly, math standardized test scores (but none of the other outcomes) for repeating groups were offset by one wave, such that repeating students were compared to non-repeating students who had completed the same year in school (see Figure 2). Thus, for students who repeated Year 5, math test scores for Waves 3–5 (when they were in Years 6–8) were compared to math test scores for non-repeating students for Waves 2–4 (when they were also in Years 6–8). For each of the repeating groups, separate analyses are presented for the original math test scores and for one-year offset math test scores. Results differ primarily for math test scores (highlighted in bold boxes), as only math test scores were offset. Standard errors of each path are presented (and statistically significant paths, $p < .05$, are presented in bold).

References Cited Only in Supplemental Materials

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.

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., 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., 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., 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.

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.