Journal of Educational Psychology
2017, Vol. 109, No. 3, 425–438

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000144

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

Reinhard Pekrun
University of Munich and Australian Catholic University

Philip D. Parker
Australian Catholic University

Kou Murayama
University of Reading and Kochi University of Technology

Jiesi Guo and Theresa Dicke
Australian Catholic University

Stephanie Lichtenfeld
University of Munich

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 at Year 5) measured each year during the first 5 years of secondary school, was particularly strong. It featured 4 independent retention groups (different groups of students, each repeating 1 of the 4 first years of secondary school; total N = 103), with multiple posttest waves to evaluate short- and long-term effects, controlling for covariates (gender, age, socioeconomic status, primary school grades, IQ) and 1 or more sets of 10 outcomes collected prior to retention. Tests of developmental invariance demonstrated that the effects of retention (controlling for covariates and preretention 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, grade retention, social comparison

Supplemental materials: http://dx.doi.org/10.1037/edu0000144.supp

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, Willson, & Hughes, 2009). Allen et al. (2009) note that the use of retention as an educational intervention, particularly in the United States, 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 the Programme for International Student Assessment (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 et al., 2009; Marsh, 2016; Marsh et

This article was published Online First August 15, 2016.

Herbert W. Marsh, Institute for Positive Psychology and Education, Australian Catholic University and Faculty of Education, King Saud University; Reinhard Pekrun, Department of Psychology, University of Munich and Institute for Positive Psychology and Education, Australian Catholic University; Philip D. Parker, Institute for Positive Psychology and Education, Australian Catholic University; Kou Murayama, Department of Psychology, University of Reading and Research Unit of Psychology, Education & Technology, Kochi University of Technology; Jiesi Guo and Theresa Dicke, Institute for Positive Psychology and Education, Australian Catholic University; Stephanie Lichtenfeld, Department of Psychology, University of Munich.

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

Correspondence concerning this article should be addressed to Herbert W. Marsh, Institute for Positive Psychology and Education (IPPE), Australian Catholic University, 25 Barker Street, Strathfield NSW 2135. E-mail: Herb.Marsh@acu.edu.au

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 (Guay, Marsh, & Boivin, 2003; Marsh & Craven, 2006), 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, because of 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 (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. In addition, 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 & Tsai, 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 preexisting 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 preexisting 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, Hughes, Kwok, Puckett, & Cerda, 2013; Moser, West, & Hughes, 2012). Using propensity matching to match retained with nonretained (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 1 year to permit same-year-in-school comparisons (what we refer to as "offset" comparisons), the retention group experienced initially higher scores than the nonretained 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). Summarizing the results of these multiple publications, 10 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 1 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 those 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 who were accelerated 1 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 1 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 et al., 2015; Marsh, Kuyper, Morin, Parker, & Seaton, 2014), 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, preexisting 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 Years 1 to 4, secondary school starts at Year 5, and compulsory schooling ends at Year 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 years 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 2 years of secondary school. Hence, the majority of retention in our study appeared in Years 7 and 8, rather than Years 5 and 6.

The Structure of the Data

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 5 years of secondary school. Data was collected from a representative sample of 1,325 students from 42 schools starting the year before the beginning 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 5 years of secondary school (see Figure 1). We evaluated retention in each of four separate groups: those retained at Year 5, a different group of students retained at Year 6, and so forth, noting that no students were retained for more than 1 year (for a discussion of the German school system, tracking, and retention, see Section 1 of the online Supplemental Materials). The study design (see Figure 1) provides a particularly strong foundation for evaluating retention effects on the basis of multiple natural experiments using longitudinal data that provide multiple posttest 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, socioeconomic status [SES], primary school grades, IQ).

Our main focus is on the four dichotomous grouping variables (see Figure 1) representing those students who repeated a school year in each of the 4 years from 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 preretention waves of data to control for preexisting differences, and multiple postretention 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 sig-

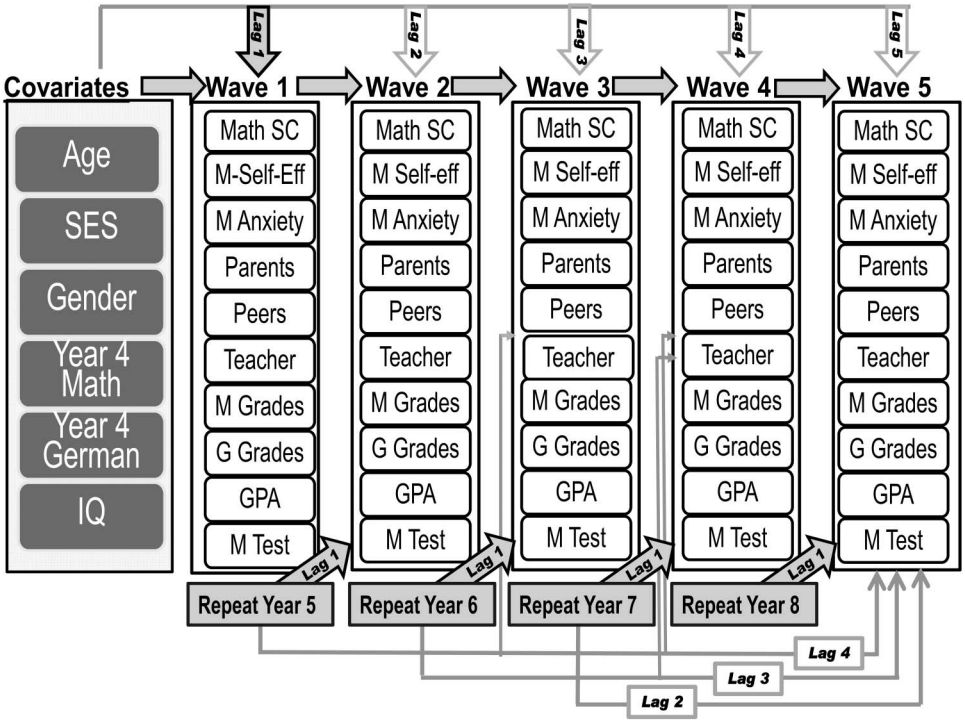

Figure 1. Waves 1 to 5 are the five yearly data collections in this longitudinal study. For students who repeated no grades, the data collections occurred during the first 5 years of secondary school (Years 5 to 9). The same set of 10 outcome variables was collected in each of the five waves. The six 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 to 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 to 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 to 5, Wave 1 variables are predictors of all variables in Waves 2 to 5, and so forth. Within each wave, all variables are correlated. For nonrepeating students, Waves 1 to 5 refer to Years 5 to 9 (the first 5 years of secondary school). Of the 1,325 students considered here, the numbers of students who repeated in each year were: Year 5, n = 10; Year 6, n = 12; Year 7, n = 35; Year 8, n = 45—a total of 103 students, or 7.8% of the total sample of 1,325 students. 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.

nificant others—parents (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 Section 2 of the online Supplemental Materials).

A Developmental Perspective: Developmental Equilibrium Hypothesis

A potentially important limitation of retention research is that it is mostly based on U.S. 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 & Matthews, 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 coun-

terbalancing effects of opposing forces. The application of equilibrium and related terms has a long history in psychological theorizing (Marsh et al., in press). Thus, for example, Marshall, Parker, Ciarrochi, and Heaven (2014) showed that a system of reciprocal effects between self-concept and social support had attained equilibrium by junior high school.

Here we test developmental equilibrium in relation to the invariance of retention effects in each of four separate 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 starts in year 5, although Years 5 and 6 are often considered part of primary schooling in U.S. 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, because of the increased *N* of the results. More importantly, it provides an apparently unique developmental perspective on the question of 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 and 1b)

Consistently with Marsh (2016), we predict that 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 2, 3, and 4 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). Nonsignificant effects at Lags 2–4 would indicate that Lag 1 effects were maintained, and 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 nonsignificant—that the initially positive effects of retention on MSC will be maintained.

Self-Efficacy and Anxiety (Hypotheses 2a and 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 favorable 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 and 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 favorable or are nonsignificant).

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 2 years, whereas 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, at least in part. 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 nonretained 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 nonretained 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 nonretained 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 nonrepeaters 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 nonretained 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 with test scores from nonrepeaters from the previous wave (when they were also in Year 5) who had studied the same curriculum. Similarly, for each postretention year, for all four retention groups, comparisons based on test scores (but not other outcomes) were "offset" by 1 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 1 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; Marsh et al., in press; Murayama, Pekrun, Lichtenfeld, & Vom Hofe, 2013; Murayama, Pekrun, Suzuki, Marsh, & Lichtenfeld, 2016; Pekrun et al., 2007; Pekrun, Lichtenfeld, Marsh, Murayama, & Goetz, in press), 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. The present investigation 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 = 11.75 years at Wave 1, *SD* = 0.7) were

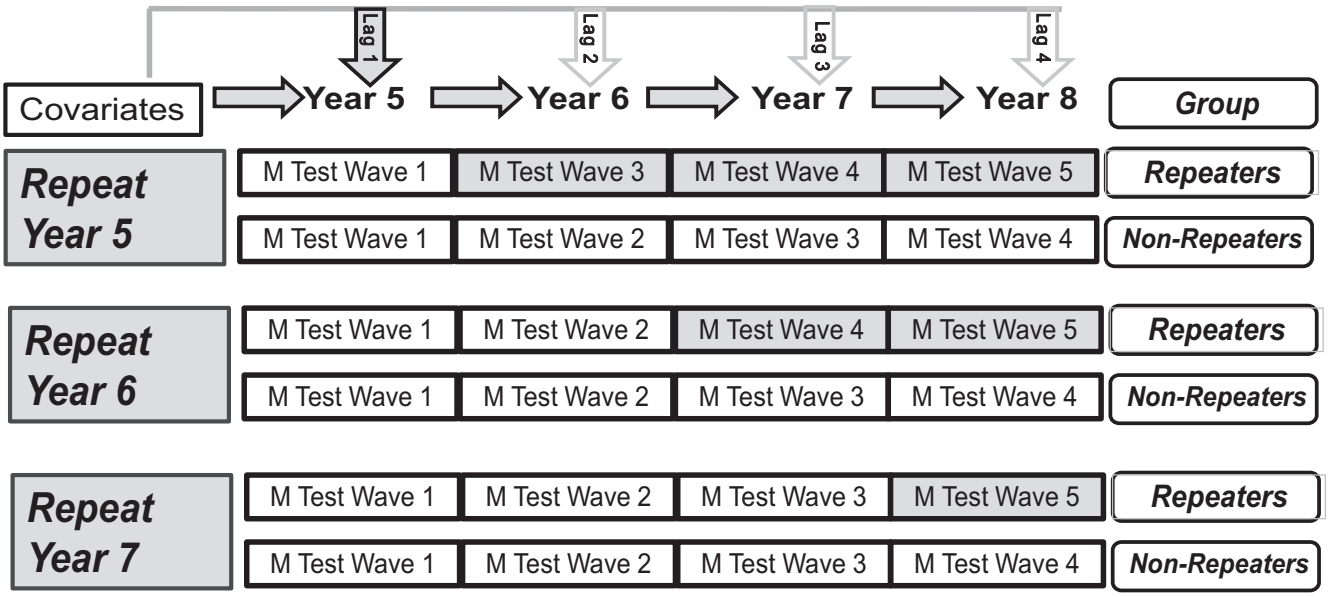

Figure 2. Offset comparisons for standardized math tests (M Tests) in Waves 1 to 5. Depicted is an alternative perspective on test scores in which retained students in each year following retention are compared with nonretained students from the previous wave. For example, math test scores for students repeating Year 5 in Wave 2 were compared with test scores of nonrepeating students when they also completed Year 5 (but in Wave 1 rather than Wave 2). Likewise, Year 6 (Wave 2) math test scores for nonrepeating students are compared with 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 postretention year (those shaded in gray for the repeater groups) for all four retention groups, comparisons based on test scores (but not other outcomes) were "offset" by 1 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 with the Year 9 scores for the nonrepeater 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 to 9, Year 5 variables are predictors of all variables in Years 6 to 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 to 8.

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 5 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 toward 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 very high at 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 4 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, present investigation the numbers of students who repeated in each year were as follows: Year 5 ($n = 10$); Year 6 ($n = 12$); Year 7 ($n = 35$); Year 8 ($n = 45$)—a total of 103 students, or 7.8% of the sample. The 103 repeating students did not differ significantly (all $ps > .05$) from the 1,222 nonrepeating students on gender (42% vs. 51% female); school type (43% Gymnasium, 23% Realschule, 23% Hauptschule vs. 40%, 30%, and 29%, respectively); age (11.7 vs. 11.8 years); or family SES (.01 vs. −.02).

In supplemental analyses, we evaluated potential biases associated with missing data after controlling for background variables (see "covariates" in Figure 1) and school type for the 10 outcomes in Year 5. More specifically, we evaluated the main effect of being included in the sample ("include" in online 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 online supplemental Table 2; the differences in outcomes for the repeating students compared with those who did not repeat Year 5); and the Repeat × Include Interaction ("IncldxRepeat" in online supplemental Table 2). This last parameter was of particular interest, as it explored whether the difference between repeating and nonrepeating 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 significantly in terms of school grades in other subjects, standardized test scores, or any of the other outcomes. Students had missing data over this 5-year span because of absences on the day of the data collection, and 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 nonrepeating 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 Section 1 of the online Supplemental Materials).

Measures

See Section 2 of the online Supplemental Materials 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 of these multi-item scales were based on self-report responses from students, using a 5-point-Likert scale: *not true at all, hardly true, somewhat true, mostly true, or completely true*. Across the five 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 Section 2 of the online Supplemental Materials).

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 5 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, Blum, & Pekrun, 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 nonrepeating 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. Student 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 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, 2008–2014, Version 7). We used the robust maximum likelihood

estimator, 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 (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, 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 4 years from 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 as 1 = retention and 0 = nonretention. 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 (ESs)—scaled so that higher scores reflect more favorable 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 nonretained 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 nonretained 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 equal to zero; 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 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 ESs (based on the means model) and the invariant ESs (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 online 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 nonsignificant effects reflect maintenance of the positive effects at Lag 1. Consistently with Hypothesis 1b, the *ES*s for Lags 2–4 were nonsignificant (maintenance of Lag 1 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, ESs (.359 for self-efficacy, .293 for anxiety; Table 1) were smaller than for MSC. Consistently with Hypothesis 2b, Lag 2–4 ESs were nonsignificant for both self-efficacy (maintenance of Lag 1 effects), although for anxiety effects there was a positive Lag 4 effect (a

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

10 outcomes	Invariant Lag 1 effects (ESs)		Invariant Lag 2 effects (ESs)		Invariant Lag 3 effects (ESs)		Invariant Lag 4 effects (ESs)	
	Effect size	*SE*	Effect size	*SE*	Effect size	*SE*	Effect size	*SE*
Math self-concept	**.597****	.094	.148	.116	−.113	.215	.405	.210
Math self-efficacy	**.359****	.084	.079	.122	−.155	.161	.128	.326
Math anxiety	**.293****	.092	.207	.117	−.100	.159	**.656****	.217
Parents	.173	.110	.008	.129	.277	.236	.336	.180
Peer	.023	.094	−.020	.154	.002	.203	.365	.270
Teacher	**.305****	.099	.149	.133	−.007	.166	.209	.194
Math grades	**1.010****	.119	−.033	.134	.077	.240	.396	.210
German grades	**.454****	.068	−.059	.117	−.025	.160	.191	.203
Grade Point Average	**.452****	.054	−.092	.080	.053	.110	−.187	.181
Math test	**−.188***	.076	−.143	.100	.059	.091	.222	.178
Total	**.348****	.042	.024	.059	.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 invariance model, effect sizes (ESs) were constrained to be invariant over the four retention groups. ESs 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. ESs that are statistically significant ($p < .05$) in relation to their standard errors (*SE*s) are in bold.
* $p < .05$. ** $p < .01$.

positive sleeper effect), even though Lag 2 and 3 effects were nonsignificant.

Relations with significant others (Research Questions 3a and 3b). Lag 1 ESs for the effects of student perceptions of positive teacher support were significantly positive (ES = .305), whereas the nonsignificant 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 (grade point average [GPA]). Lag 1 retention effects were significantly positive for all three measures of school grades (ESs = .452 to 1.010). The results were particularly large for math school grades (mean ES = 1.010), reflecting stronger controls for preexisting differences in math, because of 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 nonsignificant, 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 nonretained 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 nonsignificant, indicating that the small negative effects of retention on test scores were maintained (see 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 nonretained students in the previous wave (i.e., same-year-in-school comparisons). Thus, test scores for the retained groups were compared with those in nonretained 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 (see Table 2), Lag 1 ESs 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 six ESs (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 seven were significantly

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 1 Year (Same-Year-in-School Comparisons)

		Time (number of waves following retention)					
		Total effects			Direct effects		
Repeating group	Comparison	Lag 1	Lag 2	Lag 3	Lag 1	Lag 2	Lag 3
Repeat Year 5	Original	−.078 (.206)	−.076 (.175)	.034 (.149)	−.078 (.206)	−.152 (.102)	−.107 (.189)
	Offset	.101 (.110)	**.603** (.146)	.242 (.219)	.101 (.110)	**.442** (.146)	.024 (.156)
Repeat Year 6	Original	.022 (.143)	−.079 (.148)		.022 (.143)	−.193 (.152)	
	Offset	**.677** (.155)	**.371** (.151)		**.677** (.155)	−.022 (.157)	
Repeat Year 7	Original	**−.253** (.106)			**−.253** (.106)		
	Offset	.053 (.165)			.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 with nonrepeating 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 with math test scores for nonrepeating 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 1-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.

positive ($p < .01$). 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 Section 4 of the online Supplemental Materials). For the critical 40 Lag 1 effects (i.e., four retention groups × 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 were 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 2 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 nonsignificant, 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 preexisting differences in relation to math than there were for other school subjects—particularly those included in the GPA measure, for which controls in relation to some school subjects were limited. As noted earlier, residual preexisting differences are likely to advantage nonrepeating students; this potential bias was apparently larger for nonmath outcomes.

Retention Effects for Standardized Math Tests—Same Age Versus 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 nonretained 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 nonretained 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 1 year behind the nonretained 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 nonretained students in our study is likely to be much smaller than in studies that use tests specifically based on the curriculum covered by the nonretained 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 2 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 nonretained 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 nonretained students, in that the comparison group for evaluating retention (i.e., the nonretained 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 nonretained 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 1 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 preexisting 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 nonretained students in respect of materials that have only been taught to nonretained 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; Marsh et al., 2015; Marsh et al., 2014) 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, Marsh, De Fraine, Van Den Noortgate, & Van Damme, 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—at least in part.

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 (Carroll, 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, 1986; 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 (1986) 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-beliefs 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 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 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 preretention 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 preretention data, and evaluated postretention results for the same set of outcomes for up to 3 years following retention. Furthermore, uncontrolled preexisting differences between retained and nonretained students were likely to favor nonretained 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 for at least one of 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 United States 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 because 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*. New York, NY: Cambridge University Press.

Allen, C. S., Chen, Q., Willson, V. L., & 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, NY: McGraw-Hill.

Carroll, J. B. (1989). The Carroll Model: A 25-year retrospective and prospective view. *Educational Researcher, 18,* 26–31. http://dx.doi.org/10.3102/0013189X018001026

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,* 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. *The British Journal of Sociology, 30,* 415–441. http://dx.doi.org/10.2307/589632

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. http://dx.doi.org/10.1037/a0026895

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

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,* 124–136. http://dx.doi.org/10.1037/0022-0663.95.1.124

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

Heller, K. A., & Perleth, C. (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,* 225–236. http://dx.doi.org/10.3102/00346543054002225

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. http://dx.doi.org/10.1080/10705519909540118

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,* 166–182. http://dx.doi.org/10.3102/0162373710367682

Huguet, P., Dumas, F., Marsh, H., Wheeler, L., Seaton, M., Nezlek, J., . . . Régner, I. (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,* 156–170. http://dx.doi.org/10.1037/a0015558

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,* 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,* 420–437.

Jimerson, S. R., & Brown, J. A. (2013). Grade retention. In J. A. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 42–44). New York, NY: 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. http://dx.doi.org/10.3102/00346543060002265

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. http://dx.doi.org/10.1177/0165025407077757

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

Marsh, H. W., Abduljabbar, A. S., Morin, A. J. S., Parker, P., Abdelfattah, F., Nagengast, B., & Abu-Hilal, M. M. (2015). The big-fish-little-pond effect: Generalizability of social comparison processes over two age cohorts from Western, Asian, and Middle Eastern Islamic countries. *Journal of Educational Psychology, 107,* 258–271. http://dx.doi.org/10.1037/a0037485

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., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163. http://dx.doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Psychometrics: A festschrift to Roderick P. McDonald* (pp. 275–340). Hillsdale, NJ: Erlbaum.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indices and dangers in overgeneralizing Hu & Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Kuyper, H., Morin, A. J. S., Parker, P. D., & Seaton, M. (2014). Big-fish-little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction, 33,* 50–66. http://dx.doi.org/10.1016/j.learninstruc.2014.04.002

Marsh, H. W., Pekrun, R., Lichtenfeld, S., Guo, J., Arens, A. K., & Murayama, K. (in press). Breaking the double-edged sword of effort/trying hard: Developmental equilibrium and longitudinal relations among effort, achievement, and academic self-concept. *Developmental Psychology*.

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20,* 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. http://dx.doi.org/10.3102/00028312034004691

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. http://dx.doi.org/10.1111/cdev.12176

Micklewright, 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,* 603–621. http://dx.doi.org/10.1037/a0027571

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. http://dx.doi.org/10.1111/cdev.12036

Murayama, K., Pekrun, R., Suzuki, M., Marsh, H., & 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*. Advance online publication. http://dx.doi.org/10.1037/pspp0000079

Muthén, L. K., & Muthén, B. (2008–2014). *Mplus user's guide*. Los Angeles CA: Author.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U. S. 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*. Advance online publication.

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. http://dx.doi.org/10.1016/j.cedpsych.2010.10.002

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

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,* 152–174. http://dx.doi.org/10.1111/bjep.12028

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In R. Murnane & G. Duncan (Eds.), *Whither opportunity? Rising inequality and the uncertain life chances of low-income children* (pp. 91–116). New York, NY: Russell Sage Foundation Press.

Reynolds, A. J. (1992). Grade retention and school adjustment: An explanatory analysis. *Educational Evaluation and Policy Analysis, 14,* 101–121. http://dx.doi.org/10.3102/01623737014002101

Roderick, M. (1994). Grade retention and school dropout: Investigating the association. *American Educational Research Journal, 31,* 729–759. http://dx.doi.org/10.3102/00028312031004729

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. http://dx.doi.org/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. http://dx.doi.org/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.

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 fur Padagogik, 45*(Beiheft), 83–100.

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* [Computer software]. Retrieved from Australian Council for Educational Research website https://www.acer.edu.au/conquest

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

Received January 19, 2016
Revision received May 12, 2016
Accepted May 19, 2016 ■