Investigating the association between the big fish little pond effect and grading on a curve : A large-scale quasi-experimental study

Fleischmann, Moritz, Hübner, Nicolas, Marsh, Herbert W., Trautwein, Ulrich and Nagengast, Benjamin

Fleischmann, M., Hübner, N., Marsh, H. W., Trautwein, U. and Nagengast, B. (2021). Investigating the association between the big fish little pond effect and grading on a curve : A large-scale quasi-experimental study. *International Journal of Educational Research*, *110*, Article 101853. https://doi.org/10.1016/j.ijer.2021.101853

Fleischmann, M., Hübner, N., Marsh, H. W., Trautwein, U., & Nagengast, B. (2021). Investigating the Association between the Big Fish Little Pond Effect and Grading on a Curve: A Large-Scale Quasi-Experimental Study. International Journal of Educational Research, 110, 101853–. https://doi.org/10.1016/j.ijer.2021.101853

Investigating the Association between Grading on a Curve and the Big Fish Little Pond Effect: A Large-Scale Quasi-Experimental Study

Abstract

Equally able students have a lower academic self-concept in high-achieving classrooms, the big fish little pond effect (BFLPE). Grading on a curve—providing the best grades to the best students in the class and the worst grades to the worst students—has been speculated to contribute to the BFLPE. However, empirical evidence for this assumption is not conclusive as it stems from correlational studies. We tested the association between the BFLPE and grading on a curve with a natural experiment from the 1970s in which Swedish municipalities were free to abolish grading (*N* = 9,104). The BFLPE did not differ between nongraded and graded students. Our results suggest that students engage in social comparisons independent of whether or not they are graded.

Introduction

Academic self-concepts are students' self-perceptions of their competence in academic domains (Marsh et al., 2016). They have been found to have high power for predicting subsequent academic achievement (see Huang, 2011; Valentine et al., 2004, for meta-analyses) as well as academic aspirations and choices (e.g., Guo et al., 2015). A long tradition of research has suggested that academic self-concept is impacted by social comparison processes, as is evident from the negative effect of the average level of achievement in educational contexts (e.g., school or classroom) on individuals' self-concept after controlling individual achievement. This finding is referred to as the big fish little pond effect (BFLPE; Marsh, 1987).

Not just academic self-concept but also teacher-assigned grades are assumed to be subject to such frame-of-reference effects (Hübner et al., 2020). Equally able students receive worse grades in high achieving classrooms, a practice also referred to as grading on a curve (Trautwein et al., 2006). Because teacher-assigned grades are of great importance for academic self-concept formation (e.g., Marsh & Craven, 1997; Skaalvik & Skaalvik, 2002), early research theorized that grading on a curve contributes to the BFLPE by providing relative class ranking information (e.g., Marsh, 1987). Empirical evidence for this assumption stems from studies that additionally controlled the BFLPE for teacher-assigned what substantially reduced the contextual effect (e.g., Marsh, 1987; Marsh & Rowe, 1996). However, the design-based challenge of such a traditional mediation approach is that it provides a weak basis for testing the causal association between the BFLPE and grading on a curve. Ideally, large samples of schools would be randomly assigned to conditions where one group of students received grades, and the other did not. However, because of logistical requirements and ethical considerations, it is not surprising that there are no large-scale experimental field studies with random assignment.

To address this research gap, we evaluated a unique natural quasi-experiment in Sweden, which took place in the 1970s. Study participants attended elementary school during a time period in which municipalities were free to decide whether to keep or abolish grading. To our knowledge, the present investigation is the first to use a natural experiment to examine the association between the BFLPE and grading on a curve. By comparing nongraded and graded students, our study provides a much stronger test than any previous research of the widely accepted but untested assumption of grading on a curve contributing to the BFLPE. A thorough examination of the association between the BFLPE and grading on a curve is important not for the theory of academic self-concept formation but also for educational policy and practice.

The Association Between the BFLPE and Grading on a Curve

Basically, it is assumed that teachers assign the best grades to the best students in their class, the worst grades to the worst students in their class, and place the others somewhere in-between (Cizek et al., 1995). This grading practice is referred to as grading on a curve. Empirical evidence for teachers' tendency to grade on a curve comes from qualitative work (e.g., McMillan et al., 2002) but also empirical studies showing that students' standardized achievement varied across educational environments, whereas this was not the case for grades (e.g., Dompnier et al., 2006). Moreover, regressing teacher-assigned grades on individual and context achievement typically reveals a negative contextual effect in that equally able students have lower grades in high-achieving educational environments (e.g., Neumann et al., 2011; Trautwein et al., 2006). Because teacher-assigned grades are highly predictive for domain-specific academic self-concept (e.g., Marsh & Craven, 1997; Skaalvik & Skaalvik, 2002), already early research theorized that grading on a curve contributes to the BFLPE (e.g., Marsh, 1987). In other words, the BFLPE might be reinforced because equally able students receive lower grades in high-achieving learning environments, and this, in turn, results in a lower academic self-concept. According to this assumption, the BFLPE is not only due to an active social comparison process in which students engage in comparisons with classmates but also a passive comparison process by which students are compared with each other by their teacher. This idea was aso supported by a study by Trautwein et al. (2008), who investigated frame-of-reference effects on physical activity self-concept at two measurement points. At T1, when students had not received grades, the BFLPE was smaller than at T2 when grading was introduced. Generally, the BFLPE could potentially be reinforced not only by the provision of grades but also by the expectation of receiving class-referenced grades. The expectation of receiving written grades has also been linked to enhanced competition in educational contexts in qualitative research (Covington, 2000; Elliot & Moller, 2003; Kohn, 1999; Pulfrey et al., 2011; Romanowski, 2004). In particular, the expectation of receiving class-referenced grades that strongly reflect the relative position of an individual student's level of achievement in the classroom—as opposed to criterion- or self-referenced grades—have been theorized to foster competition (Schinske & Tanner, 2014; Seymour & Hewitt, 1997). In turn, an increase in competition is theoretically expected to promote interest in social comparison (Ruble & Frey, 1991).

To this date, researchers have investigated the association between grading on a curve and the BFLPE by applying traditional mediation analysis (e.g., Baron & Kenny, 1986; MacKinnon, 2012). They added teacher-assigned grades to the BFLPE model as an additional predictor variable, thus controlling the contextual effect for school grades. By controlling the frame-of-reference effect for teacher-assigned grades, they investigated whether equally able students who are provided with equal grades still have lower academic self-concepts in high-achieving classes. Trautwein et al. (2006) found that such an approach reduced the negative direct effect of class achievement on self-concept by about 50%. On a theoretical level, these results suggest that grading on a curve may contribute to the BFLPE by explicitly providing students with information regarding their relative class ranking. Following Marsh and colleagues (Marsh, 1987; Marsh & Rowe, 1996; also see Marsh & Seaton, 2015), Trautwein et al. (2006) concluded that the BFLPE is partly explained by grading on a curve and even raised the critical question: “Would we still find a BFLPE if no school grades were assigned?” (p. 802).

The major design-based challenge of previous studies investigating the association between grading on a curve and the BFLPE is the low informative value of traditional mediation models. Thus, previous research has not been able to determine whether grading on a curve reinforces the BFLPE of whether the two frame-of-reference effects coexist without being (causally) related to each other. Indeed, the authors of previous studies acknowledged that the results of traditional mediation models were only weak evidence for a causal relationship between the two frame-of-reference effects in the sense that grading on a curve reinforces the BFLPE (e.g., Marsh et al., 2007; Marsh, Kuyper, et al., 2014). One must also critically examine the ambiguous relationship between the two frame-of-reference effects because studies have shown that the BFLPE shrinks in a similar way when controlling for a measure of class rank (Dijkstra et al., 2008; Huguet et al., 2009). This means that the traditional mediation model might be “controlling within-class social comparison processes rather than class marks per se that is the reason why BFLPEs are substantially reduced when class marks are controlled” (Marsh, Kuyper, et al., 2014, p. 61). On a general level, the design-based challenge posed by the traditional mediation approach has been recognized by previous studies, which argued for stronger designs to investigate the association between grading on a curve and the BFLPE (e.g., Marsh, Kuyper, et al., 2014). These studies always called for the disentanglement of the confounding effects of these two processes as a fruitful direction for further research. However, a closer investigation of the association between grading on a curve and the BFLPE is still pending. To our knowledge, no study has investigated the association between grading on the curve and the BFLPE with designs other than the traditional mediation approach. Answering this unresolved issue is of great theoretical and practical relevance, as it contributes to the theory of academic self-concept formation. In addition, it is also of great practical relevance, as grading practices may be one factor that can be used to manipulate the BFLPE.

The Present Study

The present study uses a unique Swedish data set from 1980 when grading practices in elementary school varied between students due to a school reform. This reform gave municipalities the option to either abolish or keep providing written grades and report cards. To our knowledge, this is the only available data set in which grading was quasi-experimentally manipulated. Therefore, these data offer an unprecedented opportunity to evaluate the mechanisms behind the BFLPE. Thereby, our study provides a much stronger test than any previous research of the widely accepted but untested assumption of grading on a curve reinforcing the BFLPE. The study addresses three research questions:

Research Question 1: Did teachers in municipalities that continued to provide written grades and report cards grade on a curve? Research Question 1 is an important preliminary analysis because the assumption that grading reinforces the BFLPE depends on the provision and expectation of class-referenced grades.

Research Question 2: Is there support for the BFLPE in the present sample? This research question is aimed at replicating the well-known BFLPE finding. Moreover, it serves as a validation that the measures used for this study (see Method section) were appropriate for calculating frame-of-reference effects on academic self-concept.

Research Question 3: Did the size of the BFLPE differ between students who attended schools in municipalities that provided school grades and those that had abolished grading? The results for this third research question are at the core of the present article because they will provide evidence for whether grading reinforces the BFLPE.

Method

Study Background and Design: The Swedish Grading Reform in the 1970s

In the 1970s, Swedish children entered elementary education at the age of 7. They were assigned to schools based on predefined catchment areas determined by their residence and were not allowed to choose a different learning institution. Elementary education, in which class composition did not change, included Grades 1 to 6. Every class was typically taught by the same teacher from Grade 1 to the middle of Grade 4 when another teacher took over for the rest of elementary education (Klapp, 2015; Sjögren, 2010).

Until the 1968/1969 school year, students were provided with written grades and report cards in the core subjects of mathematics, Swedish, and English at the end of Grades 3 and 6. Beginning with the 1969/1970 school year, municipalities were free to decide to abolish grading. The reform made schools gradually abandon the practice of providing written grades in the 1970s before grading was finally abolished in the 1982/1983 school year throughout Sweden.

Generally, arguments for the shift in the grading policy were strongly influenced by the idea that providing grades promotes unhealthy competition between students and fosters inequalities in educational outcomes by encouraging high-performers and discouraging low-performers (Sjögren, 2010).

Data

The analyses were based on data coming from the Swedish “Evaluation through follow-up study” (ETF Study; Härnquist, 2000). For the present investigation, we used data from the first measurement occasion of the third ETF cohort (born in 1967) in spring 1980 when students were in Grade 6 of elementary education. This cohort is of special interest because these children attended elementary school during the reform window described above (from the 1974/1975 school year to the 1979/1980 school year), in which municipalities were free to decide to abolish grading. These data have already been used to investigate the effects of grading on student achievement. More specifically, Klapp (2015; see also Klapp et al., 2014) found the grading reform to differentially affect student achievement and Klapp (2017) found these effects to be mediated by self-concept. These previous studies did not use academic self-concept as an outcome and did not address (directly or indirectly) any of our research questions (e.g., the association between grading on a curve and the BFLPE). Generally, sampling from the third ETF cohort was conducted by means of a multistage sampling procedure in which a stratified sample of 29 municipalities was drawn in a first step, and school classes from these municipalities were drawn in a second step. The total sample consisted of *N* = 9,104 students who were nested in 421 classes from 138 schools. In the data, each school contained an average of *M* = 3.05 (*SD* = 2.73) classes and each class an average of *M* = 21.62 (*SD* = 6.55) students. A total of 49.14% of the sample was female, and students were on average *M* = 12.85 (*SD* = 0.33) years old. A total of 4,656 students were not graded, whereas the other 4,448 students received grades (for more information on the grading variable, see Appendix A). It is important to note that in spring 1980 when participants were measured, students in grading municipalities had not yet received their Grade 6 report cards. As municipalities were free in their decision to abolish grading, we compared nongraded and graded students with regard to the independent variables and covariates. We did not find differences between subgroups in any of these variables (table B in appendix B). In sum, this quasi-experimental design allows for the strongest test of the untested assumption of grading on a curve reinforcing the BFLPE.

Instruments

Domain-specific academic self-concept. Domain-specific academic self-concept was measured with items that were presented along with pictures and had to be answered with no or yes. For mathematics self-concept, the item was: “The girl in the picture thinks she is good at sums. Do you think you are good at sums?” Reading self-concept was the only reverse-scored item, which asked: “The boy in the picture thinks he is bad at reading. Do you think you are bad at reading?” Spelling self-concept was measured with: “The boy in the picture thinks he is good at spelling. Do you think you are good at spelling?” Moreover, general academic self-concept was assessed with: “The boy in the picture thinks he does well in school. Do you think you do well in school?” Research has shown the reliability and validity of single-item measures to be acceptable when the measure is homogenous and clearly defined (Gardner et al., 1998). As a consequence, single-item measures have been successfully used for measuring a variety of psychological constructs (Postmes et al., 2013; Wanous et al., 1997).

Domain-specific academic achievement. Domain-specific academic achievement was measured with standardized national tests. The standardized tests consisted of items from different subcategories (see Appendix C for a detailed description). As ETF data does provide total points within each of the subcategories, we calculated a sum score comprising total points from all subcategories. Reliability between the subcategories in, as measured by Cronbach’s Alpha, was $\alpha = .89$ in math, $\alpha = .85$ in Swedish, and $\alpha = .93$ in English.

To further assess the measurement quality of self-concept and achievement scales, we closely inspected their interrelations (table D in appendix D). As expected, domain-specific self-concept measures were strongly correlated with their respective achievement variables (math: r = .40, reading: r = .31, spelling: r = .34, general: r = .35). These correlations are nearly identical to those reported in a meta-analysis by Möller et al. (2009), who reanalyzed 69 datasets and found average correlations between math self-concept and math achievement of r = .37 and verbal self-concept and verbal achievement of r = .34. These results empirically support findings by Gogol et al. (2014), who found nearly identical relations within a nomological network for single-item measures as compared to multi-item scales, thereby further supporting the reliability and validity of our single-item self-concept measures.

Domain-specific teacher-assigned grades in grade 6. Domain-specific teacher-assigned grades in Grade 6 were retrieved from school administrative data. Grades were delivered on a scale from 1 to 5, with 5 representing the highest grade.

Covariates. As covariates, we used students' age, sex, SES (based on parents' occupations), and cognitive abilities (the mean of the total number of points scored on the verbal opposite ability test, the spatial ability test, and the inductive ability test).

Analyses

Analyses were run in Mplus 8 (Muthén & Muthén, 1998-2017). We took a multilevel structural equation modeling approach in which we explicitly modeled the individual as well as the class level. We did not explicitly model the school level as research has shown the class to be the pivotal frame-of-reference for academic self-concept formation (Marsh et al., 2014). But we controlled for the dependency of observations at the school level using a design-based correction of standard errors and fit statistics (implemented with the Mplus command TYPE = TWOLEVEL COMPLEX). Because domain-specific academic self-concepts were assessed with a binary variable (e.g., Do you think you are good at sums? No/Yes), we used multilevel linear probability models (Breen et al., 2018). In contrast to logistic regression, linear probability models directly model the probability of choosing a binary category, thus facilitating parameter interpretation. Further, linear probability models allow the comparison of parameters across different models in contrast to logistic regression, where the error variance is fixed (Mood, 2010). As robustness checks, we additionally analyzed all models with multilevel logistic regression models.

The proportions of missing values for model variables are presented in can be found in Table D in Appendix D. In all statistical models, full maximum likelihood estimation (FIML) was used to account for missing values (Enders, 2010; Graham, 2009). In the contextual effect models, all continuous predictor variables were standardized, and class-average achievement was calculated on the basis of standardized individual-level measures.

Results

Descriptive Statistics

Descriptive statistics for the total student sample are reported in Table D in Appendix D. Class- and school-level proportions of variance for self-concept were low. By contrast, variation in achievement on the class level was larger (between $V\ P_{cla} = .08$ for Swedish and $V\ P_{cla} = .12$ for general achievement), whereas variation at the school level was low (between $V P_{sch} = .01$ for Swedish and $V P_{sch} = .02$ for math and general). These low school-level proportions of variance show that next to the theoretical reasons presented above, there were no empirical reasons for explicitly modeling the school level. Descriptive statistics presented separately for the nongraded and graded student samples can be found in the supplementary material (Tables S1 and S2). Correlations between self-concept and achievement measures were similar across nongraded and graded students. As expected, the proportions of variance for grades in the sample of graded students were relatively low for math ($V\ P_{cla} = .04$; $V\ P_{sch} = .02$) and Swedish ($V\ P_{cla} = .02$; $V\ P_{sch} = .01$). Additionally, grades were strongly correlated with the respective achievement measures ($r = .85$ for math and $r =.86$ for Swedish).

Research Question 1: Did teachers in municipalities that continued to provide written grades and report cards grade on a curve?

To answer Research Question 1, we took the complete set of graded students and regressed grades on the covariates and achievement as well as class achievement. The results can be found in Table 1. In math, class achievement negatively predicted grades when the other variables were controlled for (b = -0.31, p < .001). In other words, an increase in class achievement by one standard deviation was associated with a decrease in grades by 0.31 standard deviations. Equally able students had worse grades in high-achieving classes and vice versa. Such frame-of-reference effects were also found for Swedish (b = -0.26, p < .001) and English grades (b = -0.34, p < .001). Generally, the results suggest that teachers in the municipalities that did not abolish grading, graded on a curve.

Research Question 2: Is there support for the BFLPE in the present sample?

Research Question 2 asked whether the BFLPE could be found in the total sample. To answer Research Question 2, we took the total student sample and regressed self-concept on the covariates, achievement, class achievement, and grading. Results from these multilevel linear probability models are presented in Tables 2 and 3. In all four domains, individual achievement positively predicted self-concept (math: b = 0.20; reading: b = 0.16; spelling: b = 0.25; general: b = 0.17; all ps < .001). This means that an increase of one standard deviation in academic achievement was associated with a 20, 16, 25, and 17 percentage point increase in the probability of stating that one was good at the respective domain. Grading negatively predicted general self-concept (b = -0.05, p = .001). This means that graded students had a 5 percentage point lower probability of stating they were good at school. In all four domains, class achievement negatively predicted self-concept (math: b = -0.10, p < .001; reading: b = -0.05, p = .003; spelling: -0.08, p < .001; general: b = -0.09, p < .001). This means that an increase of one standard deviation in class achievement was associated with a 10, 5, 8, and 9 percentage point decrease in stating that one is good at the respective domain. The fact, that the BFLPE could be found in all these domains gives further evidence for the reliability and validity of our single-item academic self-concept measures. These BFLPEs were also found in the respective logistic regression analyses (see Tables S3 and S4 in the supplemental online materials).

Research Question 3: Did the size of the BFLPE differ between students who attended schools in municipalities that provided school grades and those that had abolished grading?

Research Question 3 investigated whether the BFLPE differed between nongraded and graded students (i.e., Was the frame-of-reference effect reinforced by providing class-referenced grades?). Research Question 3 represents the main research question of the present paper. It builds on previous correlational work that suggested grading on a curve to contribute to the BFLPE by providing relative class-ranking information. As grading in our study was quasi-experimentally manipulated, our study provides a much stronger test than any previous research on the assumption of class-referenced grading reinforcing the BFLPE. To answer Research Question 3, we extended the statistical model from Research Question 2 and additionally modeled the interaction between grading and class achievement. The results from these multilevel linear probability models are presented in Tables 2 and 3. None of the interactions between the grading dummy and class achievement were significantly different from zero (math: $b = -0.07$, $p = .150$; reading: $b = 0.01$, $p = .676$; spelling: $b = 0.00$, $p = .962$; general: $b = -0.03$, $p= .467$). Thus, the BFLPEs did not differ between nongraded and graded students. These results were the same in the respective logistic regression analyses (see Tables S3 and S4 in the supplemental online materials).

Discussion

Previous studies found the BFLPE to be mediated by teacher-assigned grades and followingly argued that the BFLPE is driven by class-referenced grades that provide relative class ranking information. However, as these studies did not experimentally manipulate grading practices in the field, they were limited regarding their internal validity concerning the assumption grading on a curve drives the BFLPE. In the current study, we built on this research and evaluated a unique natural experiment in Sweden. Study participants attended elementary school during a period of time in which municipalities were free to decide to either keep or abolish the provision of written grades and report cards in elementary education. We found no differences in the size of the BFLPEs between nongraded and graded students. Our results support the contention that the grading on a curve does not reinforce the BFLPE. By comparing nongraded and graded students, our study provides a much stronger test than any previous research of the untested assumption of class-referenced grades reinforcing the BFLPE.

Along with sobering results from BFLPE moderation studies, our investigation suggests that social comparisons underlying the BFLPE happen spontaneously because students tend to inevitably rank order themselves in educational environments (see also Marsh et al., 2020). For example, students may make these comparisons when talking about homework with peers or based on their classmates' classroom participation. Such a conception is supported by classical social comparison theory, which views social comparison as a universal human drive (cf. Festinger, 1954). Also, more recent evolutionary approaches to social comparison view the tendency to compare oneself with others as a largely immutable aspect of human behavior (e.g., Frank, 2011).

The evolutionary approach to social comparison has implications for educational practice. It has repeatedly been argued that class-referenced grading encourages social comparisons in the classroom, thus negatively affecting student outcomes (e.g., Covington, 2000; Elliot & Moller, 2003; Kohn, 1999; Pulfrey et al., 2011; Romanowski, 2004). The evolutionary approach to social comparison suggests that the grading controversy might be less important than believed because students compare themselves with one another anyway, independent of grade provision. Grading opponents might also argue that grading increases the self-concept of high achievers by providing them with positive performance feedback. Such a practice would decrease the self-concept of low achievers because this group of students receives negative performance feedback, thus amplifying inequalities in educational outcomes. As reported above, we found no differential grading effects for low and high achievers, supporting the idea that students rank order themselves in educational environments independent of whether they receive written grades.

Limitations

Our study is unique because we used a natural quasi-experiment to gain a deeper understanding of the BFLPE. Typically, such field experiment studies are based on data that were not collected with the primary aim of answering the research question under investigation. Such a practice usually leads to some limitations, which was also the case for our study.

First, it is possible that teachers in both non-grading and grading municipalities conducted continuous classroom assessments. No information exists about whether these tests resulted in qualitative or quantitative (e.g., grade-like) performance feedback. On the other hand, our study showed that the abolishment of highly salient social comparison information such as class-referenced written grades and report cards would probably not be able to alter the BFLPE. Additionally, whereas grades that were given in Grade 3 provided relative performance feedback, grades from Grade 6 only were able to contribute to increased classroom competition because students had not received their report cards when they completed the academic self-concept instrument. These issues do not have any consequences for the processing of our primary research question, which asked about the effects on the BFLPE from a school reform that abolished written grades and report cards because they were assumed to induce unhealthy competition. Concerning this question, we can indeed say that the reform did not affect the BFLPE. On a theoretical level, we raised the question of whether grading reinforces the BFLPE. This question indeed could not be answered conclusively with the present study design for the abovementioned reasons. Because of the complexity of educational field research (e.g., the experimental manipulation of grading practices is virtually impossible), we argue that our study is one very important puzzle piece in testing the nature of social comparisons that underlie the BFLPE.

Another limitation of the present study is related to measurement issues. Academic self-concept was measured with items that referred to a sex-specific comparison target (e.g., “The girl in the picture thinks she is good at sums. Do you think you are good at sums? Yes/No”). In the other self-concept items, the target of comparison was a boy. On the one hand, one can argue that based on prevailing stereotypes (e.g., boys are better at math), participants may have reacted differently to the items, thus resulting in an unreliable measure of our outcome. On the other hand, the item-specific target of comparison was the same for boys and girls, and sex was included as a covariate in our analyses. In supplementary analyses, we tested for sex differences in domain-specific academic self-concept. In line with the literature (e.g., Marsh & Hattie, 1996; Watt & Eccles, 2008), boys had higher self-concept in math, whereas girls showed higher spelling and reading self-concepts. We interpret these results as indicating that the sex-related item format did not limit the validity of our self-concept items. Further, academic self-concept was measured with the help of binary single-item scales that asked whether students “are good” at the respective domains. As argued in the method section, we assumed that the single-item measures would be sufficient for measuring schematic, unidimensional, and subjective constructs such as academic self-concept. In an additional robustness check, we also constructed a multi-item general self-concept variable by averaging the four domain-specific self-concept indicators (Table S6). Again, we found no BFLPE differences between nongraded and graded students.

Conclusion and Future Prospects

Our study investigated the association between grading on a curve and the BFLPE by exploiting a natural experiment—namely, an grade abolishment school reform—and compared BFLPEs of non-graded and graded students. We found no BFLPE differences in four domains (math, reading, spelling, and general). These results are in line with an evolutionary approach to social comparison and suggest that students might compare with each other independent of the provision of class-referenced grades.

The present study took advantage of a natural experiment within the context of a unique educational reform, an opportunity unlikely to be available again in the near future. Indeed, investigating whether grading reinforces the BFLPE would ideally be tested by conducting a randomized controlled field trial. Given that it seems nearly impossible to randomly vary grading practices in the field, this issue cannot be resolved in a single study but has to be approached from different angles. Our study provides very good conditions from an internal validity perspective, with limitations concerning the treatment and measurement issues as described above. In the future, deeper insights into the association between the BFLPE and grading practices can be investigated by analyzing grading reforms with the help of cohort-control designs. These cohort comparisons may overcome some of the present limitations but will yield other drawbacks such as the confounding of grading and cohort effects. Because the assumption that grading reinforces the BFLPE is mainly based on the idea that grades to a certain extent are class-norm referenced, the issue can also be approached by comparing BFLPEs in class- and population-referenced grading systems.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037//0022-3514.51.6.1173

Breen, R., Karlson, K. B., & Holm, A. (2018). Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology*, *44*(1), 39–54. https://doi.org/10.1146/annurev-soc-073117-041429

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. A. (1995). Teachers' assessment practices: preparation, isolation, and the kitchen sink. *Educational Assessment*, *3*(2), 159–179. https://doi.org/10.1207/s15326977ea0302_3

Covington, M. V. (2000). Goal theory, motivation, and school achievement: an integrative review. *Annual Review of Psychology*, *51*, 171–200. https://doi.org/10.1146/annurev.psych.51.1.171

Dompnier, B., Pansu, P., & Bressoux, P. (2006). An integrative model of scholastic judgments: pupils’ characteristics, class context, halo effect and internal attributions. *European Journal of Psychology of Education*, *21*(2), 119–133. https://doi.org/10.1007/BF03173572

Elliot, A. J., & Moller, A. C. (2003). Performance-approach goals: good or bad forms of regulation? *International Journal of Educational Research*, *39*(4-5), 339–356. https://doi.org/10.1016/j.ijer.2004.06.003

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Frank, R. H. (2011). *The Darwin economy: Liberty, competition, and the common good.* Princeton University Press.

Gardner, D. G., Cummings, L. L., Dunham, R. B., & Pierce, J. L. (1998). Single-item versus multiple-item measurement scales: an empirical comparison. *Educational and Psychological Measurement*, *58*(6), 898–915. https://doi.org/10.1177/0013164498058006003

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., & Preckel, F. (2014). "My questionnaire is too long!": the assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology*, *39*(3), 188–205. https://doi.org/10.1016/j.cedpsych.2014.04.002

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

Guo, J., Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2015). Directionality of the associations of high school expectancy-value, aspirations, and attainment. *American Educational Research Journal*, *52*(2), 371–402. https://doi.org/10.3102/0002831214565786

Huang, C. (2011). Self-concept and academic achievement: a meta-analysis of longitudinal relations. *Journal of School Psychology*, *49*(5), 505–528. https://doi.org/10.1016/j.jsp.2011.07.001

Hübner, N., Wagner, W., Hochweber, J., Neumann, M., & Nagengast, B. (2020). Comparing apples and oranges: curricular intensification reforms can change the meaning of students' grades! *Journal of Educational Psychology*, *112*(1), 204–220. https://doi.org/10.1037/edu0000351

Klapp, A. (2015). Does grading affect educational attainment? A longitudinal study. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 302–323. https://doi.org/10.1080/0969594X.2014.988121

Klapp, A. (2017). Does academic and social self-concept and motivation explain the effect of grading on students' achievement? *European Journal of Psychology of Education*, *33*(2), 355–376. https://doi.org/10.1007/s10212-017-0331-3

Klapp, A., Cliffordson, C., & Gustafsson, J.-E. (2014). The effect of being graded on later achievement: Evidence from 13-year olds in swedish compulsory school. *Educational Psychology*, *36*(10), 1771–1789. https://doi.org/10.1080/01443410.2014.933176

Kohn, A. (1999). *Punished by rewards*. Houghton Mifflin.

MacKinnon, D. (2012). *Introduction to Statistical Mediation Analysis*. Taylor & Francis.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W., & Craven, R. (1997). Academic self-concept: Beyond the dustbowl. In G. D. Phye (Ed.), *Handbook of classroom assessment* (pp. 131–198). Academic Press.

Marsh, H. W., & Hattie, J. (1996). Theoretical perspectives on the structure of self-concept. In B. A. Bracken (Ed.), *A Wiley-Interscience publication. Handbook of self-concept: Developmental, social, and clinical considerations* (pp. 38–90). Wiley.

Marsh, H. W., Kuyper, H., Morin, A. J., 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. https://doi.org/10.1016/j.learninstruc.2014.04.002

Marsh, H. W., Martin, A. J., Yeung, A. S., & Craven, R. (2016). Competence self-perceptions. In C. Dweck & D. Yaeger (Eds.), *Handbook of competence and motivation* (pp. 85–115). Guilford Press.

Marsh, H. W., Parker, P. D., Guo, J., Pekrun, R., & Basarkod, G. (2020). Psychological comparison processes and self-concept in relation to five distinct frame-of-reference effects: Pan-human cross-cultural generalizability over 68 countries. *European Journal of Personality*, *3*(2), 180–202. https://doi.org/10.1002/per.2232

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, *95*(4), 203–213. https://doi.org/10.1080/00220670209596593

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research*, *79*, 1129–1167. https://doi.org/10.3102/0034654309337522

Mood, C. (2010). Logistic regression: why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*(1), 67–82. https://doi.org/10.1093/esr/jcp006

Muthén, L. K., & Muthén, B. (1998-2017). *Mplus user's guide*. Muthén & Muthén.

Neumann, M., Trautwein, U., & Nagy, G. (2011). Do central examinations lead to greater grading comparability? a study of frame-of-reference effects on the university entrance qualification in Germany. *Studies in Educational Evaluation*, *37*(4), 206–217. https://doi.org/10.1016/j.stueduc.2012.02.002

Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: reliability, validity, and utility. *The British Journal of Social Psychology*, *52*(4), 597–617. https://doi.org/10.1111/bjso.12006

Pulfrey, C., Buchs, C., & Butera, F. (2011). Why grades engender performance-avoidance goals: the mediating role of autonomous motivation. *Journal of Educational Psychology*, *103*(3), 683–700. https://doi.org/10.1037/a0023911

Romanowski, M. H. (2004). Student obsession with grades and achievement. *Kappa Delta Pi Record*, *40*(4), 149–151. https://doi.org/10.1080/00228958.2004.10516425

Ruble, D., & Frey, K. (1991). Changing patterns of comparative behavior as skills are acquired: A functional model of self-evaluation. In J. Suls & T. A. Wills (Ed.), *Social comparison: Contemporary theory and research* (pp. 79–113). Hillsdale.

Schinske, J., & Tanner, K. (2014). Teaching more by grading less or differently). *CBE Life Sciences Education*, *13*(2), 159–166. https://doi.org/10.1187/cbe.CBE-14-03-0054

Seymour, E., & Hewitt, N. M. (1997). Talking about leaving: why undergraduates leave the sciences. *Choice Reviews Online*, *34*(10), 34-5652-34-5652. https://doi.org/10.5860/CHOICE.34-5652

Sjögren, A. (2010). *Graded children - Evidence of long-run consequences of school grades from a nationwide reform*. IFAU – Institute for Labour Market Policy.

Skaalvik, E. M., & Skaalvik, S. (2002). Internal and external frames of reference for academic self-concept. *Educational Psychologist*, *37*(4), 233–244. https://doi.org/10.1207/S15326985EP3704_3

Trautwein, U., Gerlach, E., & Lüdtke, O. (2008). Athletic classmates, physical self-concept, and free-time physical activity: a longitudinal study of frame of reference effects. *Journal of Educational Psychology*, *100*(4), 988–1001. https://doi.org/10.1037/0022-0663.100.4.988

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*(4), 788–806. https://doi.org/10.1037/0022-0663.98.4.788

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement. A meta-analytic review. *Educational Psychologist*, *39*(2), 111–133. https://doi.org/10.1207/s15326985ep3902_3

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: how good are single-item measures? *Journal of Applied Psychology*, *82*, 247–252. https://doi.org/10.1037/0021-9010.82.2.247

Watt, H. M. G., & Eccles, J. S. (Eds.). (2008). *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences*. American Psychological Association.

Table 1

Results from Contextual Effects Models with Teacher-Assigned Grades as the Outcome

	Math				Swedish				English			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	.79	.02	[.75, .82]	< .001	.80	.02	[.76, .83]	< .001	.77	.02	[.74, .81]	< .001
Class achievement	-.31	.04	[-.39, -.23]	< .001	-.26	.06	[-.39, -.14]	< .001	-.34	.04	[-.42, -.26]	< .001

Note. Analyses were conducted with the graded student sample ($N = 4{,}448$). Outcomes are grades in the respective domain. Achievement and class achievement resemble standardized achievement scores in the respective domains. All analyses are controlled for age, sex, SES, and cognitive abilities.

Table 2

Results from Contextual Effects Models with Math and Reading Self-Concept as the Outcome

	Math								Reading							
	Model 1				Model 2				Model 1				Model 2			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	.20	.01	[.18, .22]	< .001	.20	.01	[.18, .22]	< .001	.16	.01	[.14, .17]	< .001	.16	.01	[.14, .17]	< .001
Grading	-.03	.01	-[.05, .00]	.051	-.02	.01	-[.05, .00]	.074	-.01	.01	-[.03, .01]	.348	-.01	.01	-[.03, .01]	.348
Class achievement	-.10	.02	-[.14, -.05]	< .001	-.05	.04	-[.12, .02]	.172	-.05	.02	-[.08, -.02]	.003	-.06	.02	-[.10, -.01]	.014
Grading x Class Achievement					-.07	.05	-[.15, .02]	.150					.01	.03	-[.05, .08]	.676

Note. The table contains results from multilevel linear probability analyses. Outcomes are dichotomous self-concept items in the respective domain (e.g., Do you think you are good at sums? 0 for No and 1 for Yes). Achievement and class achievement resemble standardized achievement scores in the respective domains. Grading is a dichotomous variable (0 for nongraded and 1 for graded). All analyses are controlled for age, sex, SES, and cognitive abilities.

Table 3

Results from Contextual Effects Models with Spelling and General Self-Concept as the Outcome

	Spelling								General							
	Model 1				Model 2				Model 1				Model 2			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	.25	.01	[.23, .26]	< .001	.25	.01	[.23, .26]	< .001	.17	.01	[.15, .20]	< .001	.17	.01	[.15, .20]	< .001
Grading	-.02	.01	-[.05, .00]	.055	-.02	.01	-[.05, .00]	.055	-.05	.02	-[.08, -.02]	.001	-.05	.02	-[.08, -.02]	.001
Class achievement	-.08	.02	-[.12, -.05]	< .001	-.08	.03	-[.13, -.03]	.001	-.09	.02	-[.14, -.05]	< .001	-.08	.03	-[.13, -.03]	.003
Grading x Class Achievement					.00	.04	-[.07, .07]	.962					-.03	.04	-[.11, .05]	.467

Note. The table contains results from multilevel linear probability analyses. Outcomes are dichotomous self-concept items in the respective domain (e.g., Do you think you are good at sums? 0 for No and 1 for Yes). Achievement and class achievement resemble standardized achievement scores in the respective domains. Grading is a dichotomous variable (0 for nongraded and 1 for graded). All analyses are controlled for age, sex, SES, and cognitive abilities.

Appendix A

The data contains information about whether students were graded in Grade 6. This information was derived from the grade variables that were based on school administrative data. When every student in a municipality had missing data on the grade variables, the students from the respective municipality were identified as nongraded students. When a majority of students in a municipality had nonmissing values on the grade variables (note that in the graded municipalities some students had “real” missing values on grade variables), students from the respective municipalities were identified as graded students. When students in our sample were not graded in Grade 6, the probability was very high (about 77%) that they were not graded in Grade 3 (see Table A).

Table A

Introduction of the Grading Reform

School year	Cohort in year	Percentage of municipalities that abolished grading in Grade 3	Percentage of municipalities that abolished grading in Grade 6
1974/1975	1	9.09	5.35
1975/1976	2	18.18	9.09
1976/1977	**3**	**34.22**	14.44
1977/1978	4	57.75	25.67
1978/1979	5	67.91	35.29
1979/1980	**6**	75.94	**44.39**

Note. The information in this table was retrieved from Sjögren (2010). Sjögren (2010) showed that in the 1976/1977 school year (when the present cohort was in Grade 3), approximately 34% of the municipalities had abolished grading in Grade 3. In the 1979/1980 school year (when our cohort was in Grade 6), approximately 44% of the municipalities had abolished grading in Grade 6. This means that 77% (34.22/44.39) of the municipalities that had abolished grading in the 1979/1980 school year (when our cohort was in Grade 6) had already abolished grading in Grade 3 (when our cohort was in Grade 3). Thus, when the students in our sample were not graded in Grade 6, the probability was high that they had not been graded in Grade 3.

Appendix B

Table B

Mean Differences in Model Variables between Nongraded and Graded Students

	b	*p*
Math self-concept	-0.01	.414
Spelling self-concept	-0.01	.578
Reading self-concept	-0.02	.081
General self-concept	-0.04	.003
Math achievement	0.00	.999
Swedish achievement	0.03	.449
General achievement	-0.04	.403
Age	-0.03	.304
Sex	-0.01	.152
Ses	0.07	.073
Cognitive ability	0.07	.062

Note. Mean differences were calculated by regressing the respective outcomes on the grading dummy (0 for nongraded and 1 for graded). Continuous outcomes were standardized.

Appendix C

The standardized mathematics test consisted of items from different subcategories (e.g., percentage ability or geometry ability). The standardized Swedish language test contained items from six subcategories (e.g., reading or spelling). The standardized English language test contained items from four subcategories (e.g., vocabulary or listening). Additionally, we constructed a measure of general academic achievement by averaging the math, Swedish, and English achievement scores.

Appendix D

Table D

Descriptive Statistics for the Total Sample

	Mis	*M*	*SD*	VP_{cla}	VP_{sch}	1	2	3	4	5	6	7	8	9	10
1. Math self-concept	0.11	0.69	0.46	.03	.01										
2. Reading self-concept	0.10	0.80	0.40	.01	.01	.16									
3. Spelling self-concept	0.10	0.67	0.47	.02	.00	.10	.31								
4. General self-concept	0.13	0.67	0.47	.03	.02	.48	.30	.28							
5. Math achievement	0.44	50.05	15.55	.10	.02	.40	.16	.12	.34						
6. Swedish achievement	0.39	66.44	18.26	.08	.01	.22	.31	.34	.33	.70					
7. General achievement	0.38	70.32	18.66	.12	.02	.27	.28	.30	.35	.82	.91				
8. Age	0.00	12.85	0.33	.01	.00	.00	-.02	-.03	-.02	-.05	-.05	-.06			
9. Sex	0.00	0.49	0.50	.00	.00	-.09	.03	.13	-.03	-.02	.17	.14	-.03		
10. SES	0.05	2.28	0.67	.08	.03	-.10	-.09	-.05	-.12	-.26	-.28	-.29	.04	-.02	
11. Cognitive abilities	0.10	22.84	5.82	.09	.01	.32	.17	.12	.28	.75	.71	.73	-.07	.02	-.25

Note. Descriptive statistics were based on the total sample ($N = 9{,}104$). Descriptive statistics were estimated using full information maximum likelihood estimation (FIML). The column *Mis* contains proportions of missing values. Note that achievement variables were often missing for whole classes (170 classes in math, 150 classes in Swedish, and 155 classes in general). The self-concept variables are binary such that 0 indicates that the student stated that he/she was not good and 1 that he/she was good at the respective domain. The sex variable is binary, with 0 for male and 1 for female. VP_{cla} is the proportion of class-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model. VP_{sch} is the proportion of school-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model.

Supplementary Material

Table S1

Descriptive Statistics for the Nongraded Student Sample

	M	*SD*	VP_{cla}	VP_{sc}	1	2	3	4	5	6	7	8	9	10
1. Math self-concept	0.69	0.46	.02	.02										
2. Reading self-concept	0.80	0.40	.01	.02	.17									
3. Spelling self-concept	0.68	0.47	.01	.01	.08	.30								
4. General self-concept	0.69	0.46	.03	.02	.50	.30	.24							
5. Math achievement	50.09	15.16	.09	.03	.39	.15	.11	.33						
6. Swedish achievement	66.11	18.18	.10	.02	.20	.28	.33	.32	.69					
7. General achievement	70.76	18.70	.13	.05	.24	.25	.28	.34	.81	.90				
8. Age	12.86	0.34	.01	.00	-.02	-.02	-.04	-.04	-.06	-.06	-.06			
9. Sex	0.50	0.50	.00	.00	-.11	.00	.12	-.05	-.03	.16	.14	-.04		
10. Ses	2.26	0.68	.12	.05	-.12	-.08	-.04	-.14	-.28	-.30	-.30	.03	-.03	
11. Cognitive ability	22.64	5.80	.10	.02	.30	.16	.12	.28	.73	.70	.71	-.08	.01	-.27

Note. Descriptive statistics were based on the nongraded student sample (*N* = 4,656). Descriptive statistics were estimated using full information maximum likelihood estimation (FIML). The self-concept variables are binary with 0 indicating that the student stated that he/she was not good and 1 that he/she was good at the respective domain. The sex variable is binary with 0 for male and 1 for female. VP_{cla} is the proportion of class-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model. VP_{sch} is the proportion of school-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model.

Table S2

Descriptive Statistics for the Graded Student Sample

	M	*SD*	VP_{cla}	VP_{sc}	1	2	3	4	5	6	7	8	9	10	11	12
1. Math self-concept	0.68	0.47	.03	.00												
2. Reading self-concept	0.79	0.40	.01	.00	.15											
3. Spelling self-concept	0.66	0.47	.02	.00	.12	.32										
4. General self-concept	0.65	0.48	.03	.01	.46	.30	.31									
5. Math achievement	50.17	15.66	.10	.02	.42	.17	.12	.35								
6. Swedish achievement	66.96	18.29	.06	.01	.25	.32	.35	.35	.71							
7. General achievement	70.29	18.61	.08	.01	.30	.30	.30	.37	.82	.92						
8. Age	12.85	0.33	.01	.00	.02	-.03	-.03	.01	-.04	-.04	-.05					
9. Sex	0.48	0.50	.00	.00	-.08	.05	.13	-.02	-.01	.19	.15	-.02				
10. Ses	2.31	0.65	.04	.02	-.08	-.10	-.06	-.10	-.24	-.27	-.27	.05	-.01			
11. Cognitive ability	23.04	5.85	.08	.01	.35	.18	.13	.30	.76	.71	.74	-.07	.03	-.24		
12. Grade math	3.19	1.00	.04	.02	.43	.17	.13	.36	.85	.68	.74	-.02	.05	-.23	.70	
13. Grade Swedish	3.14	0.94	.02	.01	.26	.30	.35	.36	.65	.86	.82	-.04	.27	-.25	.63	.68

Note. Descriptive statistics were based on the graded student sample (*N* = 4,448). Descriptive statistics were estimated using full information maximum likelihood estimation (FIML). The self-concept variables are binary with 0 indicating that the student stated that he/she was not good and 1 that he/she was good at the respective domain. The sex variable is binary with 0 for male and 1 for female. VP_{cla} is the proportion of class-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model. VP_{sch} is the proportion of school-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model.

Table S3

Results from Logistic Regression Contextual Effects Models with Math and Reading Self-Concept as the Outcome

	Math								Reading							
	Model 1				Model 2				Model 1				Model 2			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	1.09	.07	[.95, 1.23]	< .001	1.09	.07	[.95, 1.24]	< .001	1.09	.06	[.97, 1.20]	< .001	1.09	.06	[.97, 1.20]	< .001
Age	.05	.03	[-.02, .11]	.167	.05	.03	[-.02, .11]	.162	-.03	.03	[-.09, .03]	.323	-.03	.03	[-.09, .03]	.323
Sex	-.49	.07	[-.63, -.35]	< .001	-.49	.07	[-.63, -.35]	< .001	-.24	.07	[-.38, -.10]	.001	-.24	.07	[-.38, -.10]	.001
SES	.00	.03	[-.05, .06]	.922	.01	.03	[-.05, .06]	.841	-.06	.03	[-.12, .01]	.076	-.06	.03	[-.12, .01]	.076
Cognitive abilities	.08	.06	[-.03, .20]	.152	.08	.06	[-.03, .19]	.169	-.29	.05	[-.38, -.19]	< .001	-.29	.05	[-.38, -.19]	< .001
Grading	-.10	.08	[-.26, .07]	.251	-.09	.08	[-.25, .08]	.304	-.03	.07	[-.17, .11]	.683	-.03	.07	[-.17, .11]	.687
Class achievement	-.56	.16	[-.87, -.26]	< .001	-.23	.24	[-.71, .25]	.343	-.35	.09	[-.53, -.17]	< .001	-.35	.11	[-.57, -.13]	.002
Grading x Class Achievement					-.51	.28	[-1.06, .03]	.065					.01	.17	[-.31, .34]	.940

Note. The table contains results from logistic regression analyses. Outcomes are dichotomous self-concept items in the respective domain (e.g., Do you think you are good at sums? 0 for No and 1 for Yes). Achievement and class achievement resemble standardized achievement scores in the respective domains. Grading is a dichotomous variable (0 for nongraded and 1 for graded).

Table S38

Results from Logistic Regression Contextual Effects Models with Spelling and General Self-Concept as the Outcome

	Spelling								General							
	Model 1				Model 2				Model 1				Model 2			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	1.31	.05	[1.21, 1.42]	< .001	1.31	.05	[1.21, 1.42]	< .001	.93	.07	[.79, 1.06]	< .001	.92	.07	[.79, 1.06]	< .001
Age	-.05	.03	[-.11, .00]	.057	-.05	.03	[-.11, .00]	.057	.02	.03	[-.04, .07]	.528	.02	.03	[-.04, .07]	.528
Sex	.22	.06	[.11, .33]	< .001	.22	.06	[.11, .33]	< .001	-.46	.06	[-.58, -.34]	< .001	-.46	.06	[-.58, -.34]	< .001
SES	.07	.03	[.01, .13]	.021	.07	.03	[.01, .13]	.021	-.05	.03	[-.11, .01]	.084	-.05	.03	-[.11, .01]	.088
Cognitive abilities	-.58	.05	[-.67, -.49]	< .001	-.58	.05	[-.67, -.49]	< .001	.10	.05	[.00, .20]	.058	.10	.05	[.00, .20]	.056
Grading	-.09	.06	[-.21, .03]	.129	-.09	.06	[-.21, .03]	.134	-.31	.09	[-.49, -.13]	.001	-.31	.09	[-.49, -.13]	.001
Class achievement	-.38	.07	[-.53, -.24]	< .001	-.40	.09	[-.56, -.23]	< .001	-.48	.11	[-.69, -.27]	< .001	-.41	.14	[-.68, -.13]	.003
Grading x Class Achievement					.02	.12	[-.22, .26]	.867					-.15	.19	[-.52, .22]	.422

Note. The table contains results from logistic regression analyses. Outcomes are dichotomous self-concept items in the respective domain (e.g., Do you think you are good at sums? 0 for No and 1 for Yes). Achievement and class achievement resemble standardized achievement scores in the respective domains. Grading is a dichotomous variable (0 for nongraded and 1 for graded).

Table S39

Results from Linear Probability Models Investigating Differential Grading Effects for Low and High Achievers on Math and Reading Self-Concept

	Math				Reading				Spelling				General			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	.20	.01	[.17, .23]	< .001	.15	.01	[.13, .17]	< .001	.24	.01	[.22, .26]	< .001	.17	.01	[.14, .19]	< .001
Age	.01	.01	-[.01, .02]	.270	-.01	.01	-[.02, .00]	.197	-.01	.01	-[.02, .00]	.041	.00	.01	-[.01, .01]	.665
Sex	-.08	.01	-[.10, -.05]	< .001	-.03	.01	-[.05, -.01]	.002	.04	.01	[.02, .06]	< .001	-.08	.01	-[.10, -.06]	< .001
SES	.00	.01	-[.01, .01]	.549	-.01	.00	-[.01, .00]	.174	.01	.01	[.00, .02]	.017	-.01	.01	-[.02, .00]	.162
Cognitive abilities	.01	.01	-[.01, .03]	.303	-.04	.01	-[.05, -.03]	< .001	-.11	.01	-[.12, -.09]	< .001	.02	.01	[.00, .04]	.067
Grading	-.03	.01	-[.05, .00]	.057	-.01	.01	-[.03, .01]	.337	-.03	.01	-[.05, .00]	.054	-.05	.02	-[.08, -.02]	.001
Class achievement	-.09	.02	-[.14, -.05]	< .001	-.05	.02	-[.08, -.02]	.003	-.08	.02	-[.12, -.05]	< .001	-.09	.02	-[.13, -.05]	< .001
Grading x Achievement	-.01	.01	-[.03, .02]	.687	.02	.01	[.00, .04]	.069	.01	.01	-[.01, .03]	.195	.01	.01	-[.02, .03]	.514

Note. The table contains results from multilevel linear probability analyses. Outcomes are dichotomous self-concept items in the respective domain (e.g., Do you think you are good at sums? 0 for No and 1 for Yes). Achievement and class achievement resemble standardized achievement scores in the respective domains. Grading is a dichotomous variable (0 for nongraded and 1 for graded). As grading might accentuate self-concept differences between low and high achievers, we conducted additional analyses in which we modeled the interaction between grading and individual achievement. The interaction was not significantly different from zero in any of the domains. Hence, low and high achievers did not differ in effects of grading on academic self-concept.

Table S40

Results from Contextual Effects Models with Multi-Item General Self-Concept as the Outcome

	General							
	Model 1				Model 2			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	.15	.01	[.13, .16]	<.001	.15	.01	[.13, .16]	<.001
Grading	-.03	.01	[-.05, -.01]	.009	-.03	.01	[-.05, -.01]	.008
Class achievement	-.07	.01	[-.10, -.04]	<.001	-.06	.02	[-.11, -.02]	.003
Grading x Class Achievement					-.01	.03	[-.07, .05]	.739

Note. The table contains results from multilevel linear probability analyses. Grading is a dichotomous variable (0 for nongraded and 1 for graded). All analyses are controlled for age, sex, SES, and cognitive abilities