Zitiervorschlag: Hübner, N., Wagner, W., Hochweber, J., Neumann, M., & Nagengast, B. (2019). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grades! Journal of Educational Psychology. Advance online publication. <http://dx.doi.org/10.1037/edu0000351>

Zur Verfügung gestellt auf PHIQ:

PHIQ-DOI: <https://doi.org/10.18747/PHSG-coll3/id/378>
Original-DOI: <http://dx.doi.org/10.1037/edu0000351>

Dokumentart: Journal Article

Version: accepted version

Copyright-Hinweis: © American Psychological Association, 2019. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission. The final article is available, upon publication, at: <http://dx.doi.org/10.1037/edu0000351>

Lizenz: Alle Rechte vorbehalten

Comparing Apples and Oranges: Curricular Intensification Reforms can Change the Meaning of Students' Grades!

Nicolas Hübner, Wolfgang Wagner, Jan Hochweber, Marko Neumann, and Benjamin Nagengast

©American Psychological Association, 2019. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission. The final article is available, upon publication, at: <http://dx.doi.org/10.1037/edu0000351>

Author Note

Nicolas Hübner, Hector Research Institute of Education Sciences and Psychology, University of Tübingen. Wolfgang Wagner, Hector Research Institute of Education Sciences and Psychology, University of Tübingen. Jan Hochweber, Institute for Educational Assessment, University of Teacher Education, St. Gallen. Marko Neumann, German Institute for International Educational Research, Frankfurt a.M./Berlin. Benjamin Nagengast, Hector Research Institute of Education Sciences and Psychology, University of Tübingen.

This research was funded in part by the state government of Baden-Württemberg (Az: 33-7532.20/735) and supported by the Postdoctoral Academy of Education Sciences and Psychology of the Hector Research Institute of Education Sciences and Psychology, Tübingen, funded by the Baden-Württemberg Ministry of Science, Research and the Arts.

COMPARING APPLES AND ORANGES

Nicolas Hübner is a postdoctoral researcher at the Hector Research Institute of Education Sciences and Psychology, member of the Postdoctoral Academy of Education Sciences and Psychology and member of the LEAD Graduate School and Research Network (GSC 1028), funded by the Excellence Initiative of the German federal and state governments. This paper uses data from the National Educational Panel Study (NEPS): Additional Study Thuringia, *doi:10.5157/NEPS:TH:2.0.0*. From 2008 to 2013, NEPS data were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS has been implemented by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Correspondence concerning this manuscript should be addressed to Nicolas Hübner, Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Europastraße 6, 72072 Tübingen, Germany. Email: nicolas.huebner@uni-tuebingen

COMPARING APPLES AND ORANGES

Abstract

Teacher-assigned grades provide important information that is used by universities and colleges to make admission decisions and by employers to make hiring decisions. Besides grades, the results of standardized achievement tests are frequently used for student selection and allocation. However, correlations between the two achievement measures are far from perfect, and researchers have argued that this discrepancy can be at least partially attributed to norm-referenced grading, which is based on the composition of performance in a class. In this study, we investigated the results of a curricular intensification reform, which introduced changes in the distribution of student performance by making enrollment in certain courses mandatory. We investigated whether the reform resulted in changes in the relationship between standardized achievement-test results and teacher-assigned grades. We analyzed cohort control design data from two large representative samples of students from two German states (Baden-Württemberg: $N = 5,574$; Thuringia: $N = 2,202$) before and after the reform. Results indicated that students who received a certain grade before the reform (e.g., a grade of A, B, C, or D) differed in their standardized test performance from students who received the same grade after the reform. Furthermore, in math, course-level-specific reform effects on the association between grades and standardized test performance were found to vary between groups of students receiving good and poor grades. Implications for educational policy and school reforms and suggestions for grading are discussed.

Keywords: curricular intensification, reform, standardized achievement, grades

COMPARING APPLES AND ORANGES

Educational Impact and Implications Statement

This study suggests that reforms that change the achievement-related composition of students in upper secondary school courses can lead to changes in teacher-assigned grades. Changes in the achievement-related composition of students can occur, for instance, when mandatory enrollment is introduced for specific courses on an advanced course level for all students (e.g., curricular intensification reforms). We found substantial differences in the course-specific standardized test performance of students who had received a certain grade (e.g., a grade of A, B, C, or D) before the reform (in basic or advanced courses) compared with students who received the same grade after the reform (in core courses) in mathematics and English. We attributed these differences to teachers' norm-referenced grading and corresponding reference-group effects. From a broader perspective, the results indicate that reforms can lead to changes in student grading, which in turn might foster unfair and unreliable student selection.

COMPARING APPLES AND ORANGES

Comparing Apples and Oranges: Curricular Intensification Reforms can Change the Meaning of Students' Grades!

In recent decades, many countries have attempted to increase students' attainment and achievement levels, especially in subjects such as mathematics and languages. In the United States, this movement began in 1983 with the publication of the "A Nation at Risk report" (The National Commission on Excellence in Education, 1983) and peaked at the beginning of the new millennium. Policy reforms such as the No Child Left Behind (NCLB) Act were enacted to ensure the proficiency of all students in core subjects such as reading and math (Hess & Petrilli, 2006). Although each U.S. state was allowed to use its own definition of proficiency, these initial reforms were followed by the introduction of standards-based reforms, which consist of core components such as the rigorous standardized testing of students and the test-based accountability of schools (Ravitch, 2011; Swanson & Stevenson, 2002).

This movement can be understood as a reaction to a shared desire to increase economic growth, labor productivity, and economic competitiveness (Sahlberg, 2006). It is based on the idea that students' competencies and economic growth are positively related, which has been repeatedly documented (e.g., Hanushek & Woessmann, 2008, 2010). In pursuing this agenda, many countries have introduced curricular intensification (CI) reforms (e.g., the US or Germany; Domina, McEachin, Penner, & Penner, 2015; Volante, 2016) to increase student competencies in the core subjects. Prime examples of such movements in the US are the "Algebra for All" program in Chicago or efforts to enroll all eighth graders in algebra in California (e.g., Nomi & Raudenbush, 2016; Penner, Domina, Penner, & Conley, 2015). A prominent example of a CI reform in Germany is the upper secondary education reform, which introduced mandatory enrollment standards in specific core subjects in upper secondary schools in almost all German states (Trautwein & Neumann, 2008).

COMPARING APPLES AND ORANGES

Although the effects of such reforms on achievement measures and motivation have been widely investigated (Allensworth, Nomi, Montgomery, & Lee, 2009; Domina et al., 2015; Hübner et al., 2017; Trautwein, Neumann, Nagy, Lüdtke, & Maaz, 2010; Wagner, Rose, Dicke, Neumann, & Trautwein, 2014), comparatively little attention has been paid to the question of how CI reforms might affect teacher-assigned grades and what these grades mean (e.g., Domina et al., 2015; Hübner et al., 2017; Nomi & Raudenbush, 2016). School grades and standardized achievement-test performance are central predictors of important life outcomes such as socioeconomic success (Strenze, 2007), college and university students' GPA and institutional retention (Koretz et al., 2016; Richardson, Abraham, & Bond, 2012; Robbins et al., 2004), and postschool choices (Parker et al., 2012). Furthermore, they comprise a central part of admission/employment criteria for colleges, universities, and employers and provide an important foundation for students' academic self-concept (e.g., Clinedinst, Koranteng, & Nicola, 2015; Koretz et al., 2016; Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007; Robinson & Monks, 2005; Rojstaczer & Healy, 2012; Thorsen & Cliffordson, 2012). Grades and test scores have been found to paint different pictures with respect to individual student achievement, indicated by a far from perfect correlation between the two achievement measures (Borghans, Golsteyn, Heckman, & Humphries, 2016; Dickinson & Adelson, 2015; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Neumann, Trautwein, & Nagy, 2011; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006; Willingham, Pollack, & Lewis, 2002), and this discrepancy has been attributed at least to some extent to norm-referenced grading (e.g., Marsh et al., 2007; Trautwein et al., 2006).

Thus, in this study, we took a closer look at the CI reforms that were implemented in two German states. These reforms led to the detracking of students into core courses and thereby introduced changes in the achievement level and variability of students in these courses. We investigated whether these CI reforms subsequently affected students' grades and

COMPARING APPLES AND ORANGES

the relationships between their grades and standardized achievement-test performance before and after the CI reforms.

Grades, Standardized Test Performance, and the Frame of Reference

The correlation between grades and test scores has repeatedly been documented to be far from perfect (e.g., Marsh et al., 2005; Trautwein et al., 2006). Willingham et al. (2002) suggested that various sources contribute to the discrepancies between grades and test scores, for instance, situational differences (e.g., changes in motivation over time and across contexts) or systematic (e.g., variation in grading standards) and unsystematic errors (e.g., unreliability in grades and test scores). Historically, grades and standardized test performance have differed in their definition and purpose: Grades can be seen as composite measures, whereby multiple pieces of information (e.g., results in multiple classroom assessments, previous grades, behavior and engagement of the student in the class) are used by teachers (Brookhart et al., 2016) to provide a summary measure of students' achievement. In turn, standardized test performance is typically assumed to reflect a unidimensional construct that is highly dependent upon students' cognitive abilities. Most importantly, standardized test performance is often described as a more "objective" measure compared with grades because psychometric criteria are fulfilled (e.g., objectivity of the scoring procedure, [approximate] unidimensionality), and therefore, students can be reasonably compared across classrooms on a uniform metric (e.g., Bowers, 2011; Trautwein et al., 2006). Research on grading standards has suggested that teachers make use of different frames of reference when assigning grades (e.g., Neumann et al., 2011). Most important in this regard are the distinctions between self-referenced grading, criterion-referenced grading, and norm-referenced grading.

In self-referenced grading, teachers compare a student's achievement with that same student's previous achievement to judge the student's performance. In this case, the achievement of other students in the class or learning group is not relevant for the judgment.

COMPARING APPLES AND ORANGES

Criterion-referenced grading involves a comparison between a student's achievement and a specific learning goal. This type of grading is often used when students must meet specific objectives to get credit. However, few teachers rely exclusively on self-referenced or criterion-referenced grading. In fact, teachers typically make use of a variety of nonachievement-related criteria when assigning grades, such as students' interest or effort, socioeconomic status, or inappropriate behavior (Guskey, 2006; Hochweber, Hosenfeld, & Klieme, 2014; Westphal et al., 2016; Zimmermann, Schütte, Taskinen, & Köller, 2013), and teachers commonly make use of norm-referencing when they assign grades (e.g., Cizek, Fitzgerald, & Rachor, 1995; Marsh et al., 2007). Norm-referenced grading (also referred to as "grading on a curve") is based on the idea that students are first sorted (explicitly or implicitly) by their achievement. Comparably good students are then assigned As or Bs, comparably bad students are assigned Ds or Fs, and moderately performing students are graded somewhere in between (e.g., Trautwein et al., 2006). Students with equal levels of standardized achievement are likely to be assigned very different grades in classes with high- versus low-performing reference groups (e.g., Marsh, 1987; Marsh et al., 2007), and this reference group effect is mentioned a great deal in the literature on the Big-Fish-Little-Pond Effect (BFLPE; e.g., Marsh et al., 2008; Trautwein et al., 2006). In line with these findings, Neumann et al. (2011) showed that the reference group effect can have an impact on students' math grades: In a sample of $N = 3,526$ upper secondary school students, aggregated standardized achievement-test performance at the school level had a negative effect on students' course grades, independent of the course level (advanced or basic course), after standardized achievement-test performance at the individual student level was controlled for.

Thus, changing the composition of the course (e.g., by means of tracking or detracking) should typically be accompanied by a shift in the achievement-related rank order of students in the course. Such composition changes should affect the process of grading if

COMPARING APPLES AND ORANGES

grades are assigned to students by comparing and rank-ordering individual achievement (e.g., Brookhart, 2008, 2015; Schinske & Tanner, 2014; Trautwein et al., 2006).

Grading, Standardized Test Performance, and Curricular Intensification

In the face of recent school reform efforts dedicated to increasing student achievement (Hess & Petrilli, 2006; Organisation for Economic Co-operation and Development [OECD], 2015), many countries have implemented mandatory course enrollment policies that require students to enroll in more challenging courses. Such policies are typically referred to as curricular intensification (CI) reforms. Crosnoe and Benner (2015) defined CI as “the push to have more secondary school students enroll in higher-level coursework and persist further in their coursework” (p. 16). CI typically involves the detracking of students (Domina & Saldana, 2012; Sørensen, 1970), which means that students are no longer allowed to freely choose or drop specific courses (e.g., basic or advanced courses) but are instead mandated to enroll in core courses (e.g., Hübner et al., 2017; Hübner, Wagner, Nagengast, & Trautwein, 2018). The results of research on the effects of detracking have been mixed. Whereas some studies have found positive effects, others have found no or mixed effects on achievement measures or motivation (e.g., Domina et al., 2015; Hübner et al., 2017; Nomi & Raudenbush, 2016). When students are detracked, not only does CI typically change the academic requirements of a course (e.g., curriculum standards, instructional time, or the relative importance of grades in a certain subject for GPA; cf. Domina & Saldana, 2012; Hübner et al., 2018), but the achievement-related composition of the students in the course is also affected (e.g., Hübner et al., 2017; Nagy, Neumann, Trautwein, & Lüdtke, 2010). Because the average level of achievement in a class (the direct reference group) and the variability in that achievement level are essential for grading on a curve (Brookhart, 2015; Marsh, 1987; Trautwein et al., 2006), these changes to the composition of the class can lead to changes in teacher-assigned grades.

COMPARING APPLES AND ORANGES

The “Double-Dose Algebra” reform in Chicago (e.g., Nomi & Raudenbush, 2016) provided evidence that reforms can foster such reference group effects. This reform introduced algebra courses that offered additional support for students who were performing below the national median. In the course of introducing this double dose, schools introduced different tracking strategies: Some schools decided to group the low-performing students together in regular algebra classes, whereas other schools did not. Nomi and Raudenbush (2016) investigated the Chicago reform effects that were related to student tracking. They showed that placing students who performed below the national median in homogeneous low-skilled classrooms had little to no effect on their standardized achievement-test performance, whereas placing them in heterogeneous classrooms substantially increased their standardized test performance (0.22 SDs). Simultaneously, students placed in low-skilled classrooms were assigned higher grades (0.37 SDs) than students placed in heterogeneous classrooms.

As outlined, few studies have investigated how a CI reform might change the meaning of any particular grade (e.g., a grade of A, B, C, or D) before versus after the reform with respect to its ability to predict standardized achievement-test performance. The comparability of grades before versus after a reform is especially important at the end of upper secondary school in order to guarantee the comparability of grades when used for employee selection or college/university admission. Furthermore, if students exhibit similar standardized achievement-test performance but receive different grades before versus after a reform, this might lead to systematic disadvantages for students from specific cohorts with respect to their career prospects.

The Reform of Upper Secondary School in Germany

Between 2001 and 2012, 11 of the 16 states in Germany reformed their upper secondary school systems (Trautwein & Neumann, 2008). The states reduced course choice options and required students to take specific core subjects (e.g., math, a foreign language,

COMPARING APPLES AND ORANGES

and natural science) that were taught at a level that was equivalent to what had been the advanced course level before the reform. As such, these reforms provide a perfect opportunity for testing the effects of CI on grading practices.

Although different regulations in different states can result in variability across school systems (e.g., in the total number of years taken to graduate or the average number of hours spent per week in advanced and basic courses), the principles of the reforms were quite similar across the German states. Each reform was dedicated to two specific goals (The Standing Conference of the Ministers of Education and Cultural Affairs [KMK], 1995): First, the reform was expected to increase the average level of student achievement, mainly owing to the implementation of core subjects on an advanced level, which was expected to have a positive impact on students who would have chosen basic courses before the reform (the majority of students were enrolled in basic courses before the reform). Second, the reform was meant to increase the comparability of final examinations and corresponding grade point averages between schools and states by increasing the focus on specific subjects (Trautwein & Neumann, 2008).

In this article, we considered data from two states, namely, Thuringia and Baden-Württemberg, because of the strong comparability of the reforms in the two states, the availability of high quality data from these states, and a lack of other data sets of comparable quality and size from other states. Before the reforms, students in these two states chose two advanced courses at the beginning of upper secondary school, each taught for 5 hr (Baden-Württemberg) or 6 hr (Thuringia) per week. The two advanced courses were chosen for the entire span of upper secondary school and were part of students' final examinations at the end of upper secondary school. All students had to take a variety of additional basic courses taught for a limited number of hours per week. Students chose two of these basic-level courses to be part of their final examinations. Students from Thuringia had to enroll in math

COMPARING APPLES AND ORANGES

as a mandatory basic course (if it had not been chosen as an advanced course) for 4 hr per week, whereas students from Baden-Württemberg had to enroll in math as a basic course (if it had not been chosen as an advanced course) for 3 hr. The reforms were implemented in 2009 in Thuringia and in 2002 in Baden-Württemberg.

After each reform, students' choices were reduced: All students had to participate in mandatory core courses (equivalent to the advanced course level from before the reform) in the three subjects of mathematics, German, and one foreign language. These subjects were each taught for 4 hr per week. Students had to choose two additional core courses, which were also taught for 4 hr per week on an advanced-course level (e.g., Hübner et al., 2017). Similar to before the reforms, students still had to participate in several other subjects in addition to these five subjects on a basic-course level during upper secondary school.

Present Study and Research Questions

As outlined in the Grades, Standardized Test Performance, and the Frame of Reference section above, teachers often engage in norm-referenced grading (Trautwein et al., 2006). Therefore, all else being equal, different students with similar standardized achievement-test performance might receive different grades if the reference group a teacher is relying on changes (e.g., Neumann et al., 2011; Nomi & Raudenbush, 2016). At the same time, the average level of achievement of the specific reference group (i.e., a student's classmates) in the mandatory core courses should be lower than in advanced courses and higher than in basic courses (where self-selection had previously been in play).

In this study, we analyzed representative data from two German states (Thuringia and Baden-Württemberg) that enacted an upper secondary school reform, which introduced mandatory course enrollment in German, mathematics, and one foreign language. Because no standardized achievement test was administered for German, we could not test for effects in this subject. Also, although there were other courses the students could have taken as their

COMPARING APPLES AND ORANGES

foreign language, the large majority of students chose English as their foreign language (more than 95% in both studies). We therefore chose to focus our analyses on mathematics and English. Our overarching research question was: Are there differences in the standardized achievement-test performance of students who received certain grades before the reform (i.e., in basic or advanced courses) when compared with students who received comparable grades after the reform (i.e., in core courses)? We broke this overarching research question down into three more specific questions (Research Questions 1 to 3):

Research Question 1: Are there differences in the predicted standardized achievement-test performance in English and mathematics of students with an average grade in the corresponding course before versus after the reform? To address this question, we first investigated the relationship between grades and standardized achievement-test performance in math and English for the different course groups (before the reform: basic and advanced courses; after the reform: core courses). In line with previous research outlined above, we expected the detracking of students to lead to differences in the relationship between teacher-assigned grades and standardized achievement-test performance before versus after the reform. Students who received a certain grade (e.g., a grade of A, B, C, or D) in a core course were expected to have a lower standardized achievement-test performance when compared with students who received the same grade in an advanced course but a higher standardized achievement-test performance when compared with students who received the same grade in a basic course (Research Question 1a). In addition, we investigated whether reducing the course choice options from two types of courses (basic vs. advanced) to one type of course (core) changed the extent to which different grades assigned by teachers could be used to reliably distinguish between different levels of ability in students (Research Question 1b).

Research Question 2: Are there differences in the predicted standardized achievement-test performance when comparing courses before versus after the reform for specific grade

COMPARING APPLES AND ORANGES

groups (D to A)? To address this question, we extended the first analysis by looking closely at specific grade groups that had grades ranging from low to high grades (Grade Groups D to A). We expected that, in general, the results found in the first step would be replicated here.

Research Question 3: If there are differences between the course-specific predicted standardized achievement-test performance for different grade groups before versus after the reform, do they vary for the different grade groups? To address this question, we specified and tested Grade Group $\times$ Course Level interactions for high and low grade groups (As and Ds) on standardized achievement-test performance. We expected that the predicted standardized test performance for students earning high grades in core courses would be closer to the performance predicted for students earning high grades in advanced courses compared with the performance predicted for students earning high grades in basic courses (i.e., the absolute difference in predicted test performance would be smaller for advanced courses vs. core courses compared with basic courses vs. core courses). Similarly, we expected that the predicted standardized test performance for students earning low grades in core courses would be closer to the performance predicted for students earning low grades in basic courses compared with the performance predicted for students earning low grades in advanced courses (i.e., the absolute difference in predicted test performance would be smaller for basic courses vs. core courses compared with advanced courses vs. core courses).

Finally, we used a post hoc test to investigate the potential “hidden tracking” of students: Although students were officially grouped into core courses, the schools could have used information from late lower secondary school to group these students into separate core courses according to their previous achievement. We assumed that if the schools had used “hidden-tracking” strategies, achievement-related measures would show large proportions of variance on the course level after the reform (i.e., a large amount of variance would be explained by differences between courses after the reform).

Method

Descriptions of the Study and Sample

We used data from two German studies in two different states: The Transformation of the Secondary School System and Academic Careers (TOSCA) study (Köller, Watermann, Trautwein, & Lüdtke, 2004; Trautwein et al., 2010) and the Additional Study Thuringia (Blossfeld, Rossbach, & Maurice, 2011; Wagner et al., 2011) from the National Educational Panel Study (NEPS), included in the Scientific Use File 2.0.0. Both data sets contain representative data from one cohort before the reform and a different cohort after the reform, collected at the end of upper secondary school. This design is referred to as a cohort control design (e.g., Shadish, Cook, & Campbell, 2002) and is assumed to provide the foundation for a natural experiment. In Baden-Württemberg, a total of 88 general upper secondary schools (Gymnasium) participated at both time points, Cohort 1 (Time 1): $N = 2,772$ (age: $M = 19.5$ years); Cohort 2 (Time 2): $N = 2,802$ (age: $M = 19.3$ years). In Thuringia, 30 schools participated at both time points, Cohort 1 (Time 1): $N = 1,316$ (age: $M = 18.4$ years); Cohort 2 (Time 2): $N = 886$ (age: $M = 18.3$ years). In Thuringia, students graduated after 8 years, whereas students in Baden-Württemberg graduated after 9 years (until 2012). However, both groups of students were required to spend a similar minimum number of hours in school during their years of schooling (at least 265 hr across 8 or 9 years). In line with the different dates on which the reforms were implemented, the data in Thuringia were assessed in 2010 (before the reform) and in 2011 (after the reform). Data from Baden-Württemberg were assessed in 2002 (before the reform) and 2006 (after the reform). All data were collected at the end of upper secondary school, in the final semester of the year in which the students graduated. All standardized tests were administered in the last semester of upper secondary school. In both states, identical tests were administered before and after the reform.

COMPARING APPLES AND ORANGES

Instruments

Math achievement. The mathematics test in the TOSCA study was the *Advanced Mathematics Test*, which was based on test items from the Third International Mathematics and Science Study (TIMSS; Mullis et al., 1998). About two thirds of all items were administered in a multiple-choice format, whereas the other items were open-ended questions. The items were administered in a multimatrix design. Therefore, students worked on a subset of items (one of four booklets: 10, 28, 29, or 31 items with 10 overlapping items; Watermann, Nagy, & Köller, 2004) rather than on all 68 items. Item response theory (IRT) was used to scale all of the items so that student achievement before and after the reform could be meaningfully compared and so that the multimatrix design could be adequately accounted for. Intraclass correlations (ICCs) were calculated as the proportion of variance at the school level for the subsamples of students in basic, advanced, or core courses. The ICCs were: $\rho = .06, p < .001$ (basic course); $\rho = .06, p = .003$ (advanced course); $\rho = .04, p < .001$ (core course). According to the specified Partial Credit Model, the item difficulties showed a large coverage of the ability distribution ranging from -2.30 to 3.37. The weighted mean squares (WMNSQ) ranged from 0.90 to 1.25. The weighted likelihood estimator (WLE) reliability was .67 (Warm, 1989).

The mathematics test in the Additional Study Thuringia focused on *mathematical literacy*, which is also a focus of the Programme for International Student Assessment (PISA; e.g., OECD, 2004). Overall, students had to work on 35 items for 30 min (each student worked on a subset of 19 to 21 items per booklet). We applied a scaling strategy that was similar to the one applied for the math test in Baden-Württemberg. The ICCs were: $\rho = .09, p = .008$ (basic course); $\rho = .22, p = .006$ (advanced course); $\rho = .15, p = .116$ (core course). For the math test in Thuringia, the item difficulties ranged from -2.57 to 1.18, and the WMNSQ ranged from 0.89 to 1.12. The reliability of the test was satisfactory (WLE reliability = .67;

COMPARING APPLES AND ORANGES

e.g., Nunnally, 1978; Organisation for Economic Co-operation and Development, 2017). In both data sets, the eigenvalues from exploratory factor analyses using categorical indicators (IRT-EFAs) suggested a dominant first dimension (Thuringia: $\lambda_1 = 8.15$, $\lambda_2 = 1.87$, $\lambda_1 / \lambda_2 = 4.36$; Baden-Württemberg: $\lambda_1 = 12.62$, $\lambda_2 = 2.27$, $\lambda_1 / \lambda_2 = 5.57$). On the basis of these results and recent research on unidimensionality and the interpretability of psychological instruments (Gustafsson & Åberg-Bengtsson, 2010), we interpreted the math tests as essentially unidimensional (e.g., Raykov & Pohl, 2013).

English achievement. In the TOSCA study in Baden-Württemberg, a short version of the Test of English as a Foreign Language (TOEFL) was used to assess students' English competencies. The test consists of a total of 79 items from three subscales: Listening and Comprehension (28 items; Cronbach's $\alpha = .79$), Structure and Writing Expression (23 items, Cronbach's $\alpha = .75$), and Vocabulary and Reading Comprehension (28 items; Cronbach's $\alpha = .77$). The latent intercorrelations between the three factors were high ($r_s \geq .85$). Thus, we considered a unidimensional IRT model for the TOEFL test (e.g., Cohen, 1988; Organisation for Economic Co-operation and Development, 2012). The ICCs for the standardized English achievement test were: $\rho = .06$, $p = .006$ (basic course); $\rho = .07$, $p = .004$ (advanced course); $\rho = .05$, $p < .001$ (core course). The specified unidimensional one-parameter logistic IRT (1PL IRT) model revealed item difficulties ranging from -4.62 to 1.88 with a WMNSQ ranging from 0.85 to 1.18. The reliability of this score was high (WLE reliability = .86).

In the Additional Study Thuringia, an English reading test was administered with 33 items that were developed by the Institute for Educational Quality Improvement (IQB; Rupp, Vock, Harsch, & Köller, 2008). Students were given 30 min to work on 21 items, which were administered in a multiple-choice or multiple-matching format (National Educational Panel Study [NEPS], 2011). The scaling strategy was similar to the ones applied for the previous tests. The ICCs for standardized English achievement were: $\rho = .12$, $p = .008$ (basic course); ρ

COMPARING APPLES AND ORANGES

$= .10, p = .049$ (advanced course); $\rho = .08, p < .001$ (core course). The item difficulties for the unidimensional one-parameter logistic IRT model ranged from -2.48 to 0.50, and the WMNSQ ranged from 0.85 to 1.46/1.20. Note that the item with a WMNSQ of 1.46 was excluded from further analyses because it also had a negative item discrimination, indicating that students with lower ability had a higher probability of correctly solving the item. The reliability of this test was satisfactory (WLE reliability = .76).

All WMNSQ values fell within the fit statistic cutoffs (0.5 to 1.5) that Wright and Linacre (1994) considered to be “productive for measurement” (p. 370). Even when considering more restrictive cutoffs (e.g., 0.75 to 1.33; Embretson & Reise, 2000), the WMNSQ values of all items fit well. Values above 1 indicate a lower discrimination (of a specific item) than suggested by the model, whereas values below 1 indicate a higher discrimination (of a specific item) than suggested by the model. On the basis of these results, we decided to use the unidimensional one-parameter logistic IRT model to fit the data.

Grades. In the TOSCA study, data on grades came from the first semester of the final school year (i.e., 13.1; this notation means “Grade 13, Semester 1”) in mathematics and English. These grades were based on written exams and oral participation in the course and ranged from 0 (*worst achievement*) to 15 (*best achievement*) points. In the Additional Study Thuringia, data from all four semesters (11.1 [i.e., Grade 11, Semester 1] to 12.2 [i.e., Grade 12, Semester 2]) in math (all $rs \geq .75$) and English (all $rs \geq .79$) in upper secondary school were available for both cohorts and were strongly correlated. Because students remain with the same student group for all four semesters, we decided to average these grades. Robustness checks using models considering only information from Grade 12.1 (i.e., Grade 12, Semester 1) in Thuringia revealed comparable results. We additionally created specific grade groups (D to A) so that we could better capture the potential nonlinear trends in the data and test differences in the relationships for different grades (see Research Question 2). The grade

COMPARING APPLES AND ORANGES

groups were created as follows: Group D: < 6.5 points, Group C: ≥ 6.5 and < 9.5 points, Group B: ≥ 9.5 and < 12.5 points, Group A: ≥ 12.5 points. This taxonomy is comparable to the traditional grading metric used in the German upper secondary school system. We decided to place all students with grades lower than 6.5 points in one grade group because there were only a few students in the lowest levels of the grade distribution.

Additional measures. In addition, for post hoc robustness tests, we also considered final examination grades and a cognitive ability test. Final examination grades were based on centrally administered and evaluated tests at the end of upper secondary school for all students in the state and ranged from 0 (*lowest achievement*) to 15 (*highest achievement*) points. Furthermore, we considered a revised version of the cognitive skills test (CAT) for Grades 4 to 12 to measure cognitive ability (Heller & Perleth, 2000). The CAT consists of a verbal dimension (20 items), a quantitative dimension (20 items), and a figural-spatial dimension (24 items) and is based on the idea that intelligence is a multifaceted construct, covered by the Lorge-Thorndike-Intelligence-Test (e.g., NEPS, 2011).

Statistical Analysis

We first computed correlations between the standardized test scores and grades for English and mathematics across all students as well as descriptive statistics for these measures for each course. Then we tested for statistically significant differences between the scores and grades collected before the reform and those collected after the reform. Next, we specified multiple indicators multiple causes (MIMIC) models (e.g., Jöreskog & Goldberger, 1975) for each state using item response theory (IRT). In these models, standardized achievement-test performance was modeled using a unidimensional one-parameter logistic (1 PL) IRT model (measurement model) and predicted by the subject-specific grade (structural model) to avoid biased estimates due to unreliability, which would be the case if, for example, point estimates of achievement (e.g., WLEs) were used instead of latent variables. We estimated these models

COMPARING APPLES AND ORANGES

with cohort-specific structural models (multiple group) and constant measurement models across groups using a multiple-group IRT model implemented in Mplus 7.4 (Muthén & Muthén, 1998-2012). Multiple groups were built using information about the course (basic, advanced, and core) as well as the grade group (D to A). The main analyses proceeded in three steps.

To address Research Question 1, we estimated multiple-group models for three different groups (course-level groups: basic and advanced courses before the reform [first cohort] and core courses afterward [second cohort]). We did this separately for each subject and each state using grand-mean-centered grades as the predictor. Then we compared the predicted standardized achievement-test performance for an average grade (intercepts) between the three resulting groups using the model constraint option in Mplus (Research Question 1a). In order to investigate whether the reform changed the extent to which different grades assigned by teachers can be used to reliably distinguish between different levels of ability in students, we compared the amount of variance explained in standardized achievement-test performance between the course systems before and after the reform. To do this, we specified a multiple-group model that was comparable to the one used in the first step of the analysis (containing the three groups of basic, advanced, and core courses). In this model, we used the model constraint option to test for statistical significance in the differences in explained variance between all students before and after the reform.¹ Again, we estimated all models separately for each subject and state (Research Question 1b).

¹ The coefficient of determination for the group after the reform was calculated as a new parameter by dividing the explained variance (i.e., the squared regression weight multiplied by the variance of grades) by the total variance (i.e., the sum of explained and residual variance). For the groups before the reform, we calculated a combined coefficient of determination (i.e., variance explained across both groups), reflecting the R^2 from a multiple regression with grades and a dummy-coded variable for course level (0 = basic course, 1 = advanced course) and their interaction as predictors. The calculation was based on the variance explained by grades and by the mean difference between the two (sub)groups (i.e., basic and advanced levels), each weighted by the relative probability of group membership (reflecting the different group sizes), divided by the total variance (i.e., the sum of explained variance and weighted residual variance). The difference between the two coefficients of determination (before the reform, after the reform) was tested against the null hypothesis of a difference of zero.

COMPARING APPLES AND ORANGES

To answer Research Question 2, we specified multiple-group models for 12 different groups (all combinations of Course Level $\times$ Grade Group) and compared the average standardized achievement-test performance predicted for the four different grade groups across the three different courses.

We addressed Research Question 3 by estimating the models we used to answer Research Question 2 (again, separately for each subject and each state), but we added Course Level $\times$ Grade Group interactions for very good (As) and very bad (Ds) grades to check for potential differences in predicted standardized achievement-test performance, depending on the specific grade group and the course level.

To test for “hidden tracking,” we used student-, course-, and school-level data (final examination grades, fluid cognitive ability, and standardized achievement-test performance) that were available in the Thuringian data set. We specified three-level models (students nested in courses, courses nested in schools) and estimated the proportions of variance on the different levels for each of these variables. We ran the analyses separately for the different courses.

To ensure that we would be able to meaningfully interpret the coefficients, the metric of the latent achievement variable was transformed to $M = 50$ and $SD = 10$. We used the Benjamini-Hochberg procedure to control the false discovery rate within subjects and states (Benjamini & Hochberg, 1995).

Item analyses and selection. The quality of the standardized achievement tests was assessed beforehand with regard to reliability and differential item functioning (DIF) between cohorts. We specified 1PL IRT models when items were dichotomous, whereas we used the Partial Credit Model (Masters, 1982) when items were polytomous. We used the *TAM* package (Kiefer, Robitzsch, & Wu, 2017) in R (R Development Core Team, 2016) for both models. Furthermore, we specified two-parameter logistic (2PL) IRT models to check for

COMPARING APPLES AND ORANGES

negative item discriminations, which would suggest that students with lower average competence had a higher probability of correctly solving this item. Therefore, negative item discriminations might be an indicator of incorrect coding or poor item quality. Reliability estimates for weighted likelihood estimates (WLEs; Warm, 1989), item difficulties, WMNSQs, and DIF were computed with a unidimensional 1PL IRT model in Conquest 2.0 (Wu, Adams, Wilson, & Haldane, 2007). To analyze DIF, we applied a multiple-group IRT analysis with varying difficulties across subgroups at the item level. We identified the models by constraining the mean of the item difficulties to zero. To do this, in a first step, we specified two models, one model that included the items and the main effect of group (Model 1: item - wave) and a second model that also included the Item x Wave interaction (Model 2: item - wave + item*wave; for partial credit models, this statement also includes item step parameters). DIF is indicated in Model 2 if there are differences in the estimated item difficulties between the two groups. According to Pohl and Carstensen (2012), DIF effect sizes can be differentiated into > 1 logit (very strong DIF), > 0.6 but ≤ 1.0 logit (noteworthy), > 0.4 but ≤ 0.6 logits (considerable but not serious), and ≤ 0.4 logits (not considerable). In a second step, we compared the difference between Models 1 and 2 as an omnibus test of DIF across all items on the specific test.

Cluster structure and survey weights. Observations of students from the same school cannot be treated as independent because they are more similar to each other than they are to students from different schools. Ignoring the cluster structure usually leads to an underestimation of standard errors (Snijders & Bosker, 2012). To address the clustered data structure (students were nested within schools), we adjusted the standard errors by applying a design-based correction as implemented in Mplus 7.4 (Muthén & Muthén, 1998-2012), which takes the multilevel structure into account by employing a “sandwich” estimator (see e.g., Asparouhov, 2005; Muthén & Satorra, 1995). To do this, we followed the recommendations

COMPARING APPLES AND ORANGES

of McNeish, Stapleton, and Silverman (2017), who suggested that population-averaged methods should not rely on assumptions that are inherent to the specification of random effects in hierarchical linear modeling. We also considered survey weights to establish the representativeness of our results for the population of students in each state in upper secondary schools.

Missing values. Missing values are a common problem in social science research, and there are different approaches for handling them. Recent research has suggested that approaches such as full information maximum likelihood (FIML) or multiple imputation (MI) are, with a few exceptions, superior to traditional approaches such as listwise or pairwise deletion (e.g., Graham, 2009; Schafer & Graham, 2002). A particular advantage of modern approaches is a smaller estimation bias, even if assumptions such as the multivariate normal distribution or missing at random are violated (e.g., Enders, 2001, 2010; Newman, 2003). In this study, we therefore employed FIML estimation (e.g., Graham, 2009), which is theoretically expected to lead to the same results as MI if the number of imputations is large enough (Graham, Olchowski, & Gilreath, 2007).

Results

Preliminary Analyses

Descriptive statistics. First, we looked closely at the descriptive statistics. As Table 1 shows, there were considerable correlations between subject-specific grades and latent standardized achievement-test performance in both states when the data from before and after the reforms were combined. In Thuringia, standardized math-test performance was correlated $r = .48$ ($p < .001$) with math grade, and standardized test performance in English was correlated $r = .62$ ($p < .001$) with English grade. The correlations in Baden-Württemberg revealed a comparable pattern. However, math grade and math achievement had a slightly stronger correlation, $r = .64$ ($p < .001$), which might be attributed to the lower curricular

COMPARING APPLES AND ORANGES

validity of the mathematical literacy test in the NEPS in Thuringia (e.g., Mullis et al., 1998; Weinert et al., 2011). In addition, we computed correlations between grades and standardized test performance for all three course types and tested for potential differences. As Table 2 shows, the association between grades and standardized test performance was statistically significantly higher in all advanced courses compared with basic courses (e.g., $r_A = .52$ vs. $r_B = .30$, $p < .001$, in math in Thuringia). Furthermore, the association between grades and standardized test performance in math was statistically significantly higher in core courses compared with basic courses (Thuringia: $r_B = .30$ vs. $r_C = .53$, $p = .001$; Baden-Württemberg: $r_B = .54$ vs. $r_C = .75$, $p < .001$). These results suggest that the alignment between grades and standardized test performance, at least in math, was higher after the reform. Overall, grades tended to capture students' academic abilities as measured by standardized tests better after the reform.

Missing data. Next, we looked closely at the rates of missing data on our standardized achievement tests. After taking into account planned missingness, the missing data rate was below 6% per item for 33 of the 35 items from the math test in Thuringia. For the other two items, the rates of missing data were around 18% and 10%. All items from the English test in Thuringia had rates of missing data below 4%. In Baden-Württemberg, the *Test of English as a Foreign Language* had rates of missing data below 1% on all items. For the math test, none of the items had rates of missing data above 9%. The percentages of missing English and math grades were generally small ($< 2\%$). As explained in the statistical analysis section, we used FIML estimation to treat missingness at the item level.

Differential item functioning. Then we tested for DIF. The DIF analysis revealed that most items did not show considerable DIF. For the math test in Thuringia, we found “not considerable” DIF ranging from -0.26 to 0.29 logits. Furthermore, the deviance test was not statistically significant, $\chi^2(41) = 22.65$, $p = .991$. For the English test, we found a similar

COMPARING APPLES AND ORANGES

picture of “not considerable” DIF ranging from -0.22 to 0.30 logits. Again, the omnibus test for DIF was not statistically significant, $\chi^2(33) = 27.07, p = .757$.

For the math test in Baden-Württemberg, the majority of items showed “not considerable” DIF (≤ 0.4 logits). However, we found that three items had very strong DIF (> 1.0 logits) and two items had noteworthy DIF (> 0.6 but ≤ 1.0). Furthermore, the omnibus test was statistically significant here, $\chi^2(91) = 996.95, p < .001$. For the English test, again, most items showed “not considerable” DIF and only seven items showed considerable but not serious DIF. The omnibus test for DIF was statistically significant, $\chi^2(79) = 518.38, p < .001$, but Chi-square test results are strongly influenced by sample size (e.g., Thompson & Green, 2006).

Excluding items with very strong and noteworthy DIF did not change the results with respect to the direction of the effect or statistical significance: All differences between the coefficients in the multiple-group model for the 12 different groups with and without the items showing DIF were smaller than or equal to $d = 0.03$. For the sake of clarity, we decided to report the results of models that included all test items. This decision was based on the finding that the DIF seemed to vary unsystematically because some items showed DIF that favored students before the reform but other items favored students after the reform, without showing a clear pattern for specific content areas or for types of items that favored one cohort or the other.

Grades and Standardized Achievement-Test Performance

Next, we computed the means and standard deviations for grades and standardized test performance in mathematics and English for the different courses before and after the reform in each state (see Table 3). In Thuringia, before the reform, 47.1% of all students were enrolled in advanced courses in math and 32.9% in English. In Baden-Württemberg, 37.5% of all students were enrolled in advanced courses in math and 47% in English.

COMPARING APPLES AND ORANGES

Course-specific achievement-test performance. For the achievement tests, we found a comparable pattern across states. Standardized achievement-test performance was the lowest in basic courses before the reform. In math in Thuringia, for instance, students in basic courses achieved $M = 43.35$ ($SD = 8.40$) points, whereas students in advanced courses, on average, performed statistically significantly higher, $M = 54.93$, $SD = 9.27$, $d = -1.09$, $p < .001$. Achievement-test performance in core courses after the reform ($M = 50.34$, $SD = 9.91$) was statistically significantly lower in comparison with advanced courses ($d = -0.47$, $p < .001$) but statistically significantly higher when compared with basic courses ($d = 0.54$, $p < .001$). Similar patterns were found for standardized test performance in English in Thuringia and for both subjects in Baden-Württemberg (see Table 3).

Course-specific grade level. In contrast to the standardized test performance, grades presented a slightly different picture (see Table 3). In Baden-Württemberg, basic course grades in mathematics ($M = 8.01$, $SD = 3.59$) were statistically significantly lower when compared with advanced course grades in math ($M = 9.85$, $SD = 3.08$, $d = -0.54$, $p < .001$). Similarly, basic course grades in English ($M = 9.13$, $SD = 2.86$) were statistically significantly lower when compared with advanced course grades in English ($M = 9.67$, $SD = 2.70$, $d = -0.19$, $p < .001$). When we looked at differences between basic and advanced courses for students' achievement-test performance and grades, the differences were considerably smaller for grades than for standardized test performance, a finding that supports our expectation about the impact of the reference group on grades. In addition, we found differences between grades in core courses in mathematics ($M = 8.27$, $SD = 3.64$) and advanced courses in mathematics ($d = -0.53$, $p < .001$) as well as between grades in core courses in English ($M = 9.25$, $SD = 2.92$) and advanced courses in English ($d = -0.15$, $p = .001$). However, grades were not statistically significantly different between basic and core courses in mathematics ($d = -0.08$, $p = .078$) or in English ($d = -0.04$, $p = .342$). The differences in English grades in

COMPARING APPLES AND ORANGES

Thuringia were comparable to the patterns found for Baden-Württemberg. For math in Thuringia, however, average grades were not statistically significantly different between advanced ($M = 9.24$, $SD = 3.04$) and core courses ($M = 8.94$, $SD = 3.12$, $d = -0.10$, $p = .087$), whereas grades from basic courses ($M = 8.21$, $SD = 3.02$) were statistically significantly different from grades in core courses ($d = -0.24$, $p < .001$) and advanced courses ($d = -0.34$, $p < .001$).

Standardized Achievement for Students Earning Certain Grades

Course-specific achievement predicted by grades (Research Question 1a). After these first basic analyses, we investigated differences between predicted achievement-test performance in the three different courses for an average grade. In line with our predictions, we found statistically significant differences between all courses (all p s ranging from $p < .001$ to $p = .027$) and for both subjects in both states, indicating that for students with an average course grade, their standardized test performance differed across courses (see Table 4).

Differentiability of achievement-test performance (Research Question 1b). In the next step, we looked closely at the differentiability of student achievement-test performance before versus after the reform. Our models explained 44% of the variance in students' standardized achievement-test performance in Thuringia in English before and 35% after the reform ($|\Delta R^2| = .09$, $p = .136$), however, this difference in explained variance was not statistically significant. We found a similar pattern for math in Thuringia, where 37% of the variance was explained before and 28% was explained after the reform ($|\Delta R^2| = .09$, $p = .084$). In Baden-Württemberg, we found a comparable pattern (mathematics before: 59%, after: 56%, $|\Delta R^2| = .04$, $p = .170$; English before: 44%, after: 40%, $|\Delta R^2| = .04$, $p = .243$). These results indicate that the prediction of standardized test performance from grades, or the distinction between students' grades given their standardized test performance and course

COMPARING APPLES AND ORANGES

information, was not statistically significantly different before the reform in comparison with after the reform.

Course-specific achievement for specific grade groups (Research Question 2). In order to obtain a more coherent picture, we then investigated effects on course-specific grade groups. The results of these analyses are displayed in Tables 5 and 6 and Figure 1. As Figure 1 shows, we found comparable patterns for the two states in mathematics, and these findings were in line with our expectations. Before the reform, and similar to the findings displayed in Table 4, there was a considerable difference in the standardized achievement-test performance between students with comparable grades in advanced and basic courses. In Baden-Württemberg, these differences in mathematics were statistically significant across all different grade groups: For instance, before the reform, students in Grade Group B, on average, scored 58.3 points in advanced courses, whereas they scored only 46.4 points in basic courses ($p < .001$). After the reform, when all students had to participate in a mandatory core course in mathematics, the average standardized test performance in Grade Group B (55.3 points) was statistically significantly higher than in basic math courses ($p < .001$), whereas it was statistically significantly lower when compared with advanced math courses ($p < .001$). As Figure 1 shows, this pattern was comparable to the pattern found in Thuringia (see Table 5).

For English, we found a comparable picture in one state (Baden Württemberg, see Figure 1 and Table 6). All differences between courses were statistically significant here (all ps ranging from $p < .001$ to $p = .002$). However, in Thuringia, differences in achievement-test performance were statistically significant only for Grade Group C between the basic ($M = 44.4$) and advanced courses ($M = 48.4$, $p = .001$) and for Grade Group B between the core ($M = 52.4$) and advanced courses ($M = 55.9$, $p = .012$) and between the basic ($M = 50.9$) and

COMPARING APPLES AND ORANGES

advanced courses ($p < .001$). Results on the differences between different grade groups within course groups can be found in the notes of Tables 5 and 6.

Interactions between Grade Groups and Course Levels (Research Question 3).

Next, we also checked for Grade Group $\times$ Course Level interactions for Grade Groups D and A and found statistically significant effects in mathematics but not in English (see Figure 1). In both states, differences in the standardized test performance between the core and basic courses were statistically significantly smaller for Grade Group D compared with Grade Group A in math (Δ_{TH} : 6.6 points, $p = .015$; Δ_{BW} : 8.9 points, $p < .001$). Furthermore, in Baden-Württemberg, there was also a statistically significantly larger difference between core course achievement test performance and advanced course achievement test performance in Grade Group D compared with Grade Group A ($\Delta = -3.1$ points, $p = .016$; see Tables 5 and 6). These results suggest that for students in low grade groups (Ds) in core courses, their predicted standardized achievement-test performance in math was closer to the performance predicted in basic courses. For students in high grade groups (As) in core courses, the predicted standardized test performance in math was closer to the performance predicted in advanced courses.

Hidden tracking (post hoc analysis). Finally, we tested for whether there had been a “hidden tracking” of students into courses. As Table 7 shows, there were no statistically significant proportions of variance on the course level after the reform in English or in math for any of the achievement-related measures we investigated. These results are in line with our assumption that students had not been rigorously sorted into higher and lower achieving core courses on the basis of their achievement.

Discussion

In this study, we investigated how the reform of upper secondary school in two German states introduced changes in the meaning of teacher-assigned grades. School grades

COMPARING APPLES AND ORANGES

in upper secondary school are a major criterion for student selection at colleges and universities and are relevant for later employment. In spite of the societal relevance of grades for student allocation and selection, various studies have indicated that school grades and standardized achievement do not always have a high level of correspondence, as indicated by their far from perfect correlations. Research has suggested that one central factor that might produce such differences is variations in grading standards (Guskey, 2006; Hochweber et al., 2014; Willingham et al., 2002), for instance, norm-referenced grading (e.g., Trautwein et al., 2006).

Overall, the results of this study were in line with our expectations. We found statistically significant differences in the standardized achievement-test performance of students who had similar grades in advanced, basic, and core courses in mathematics. Students in core courses after the reform had a better standardized test performance, on average, than basic course students who had a similar course grade. Furthermore, in line with our expectations, comparisons of standardized achievement-test performance between advanced courses (before the reform) and core courses (after the reform) revealed the opposite pattern: Here, the average achievement-test performance of students in the advanced courses was statistically significantly higher than the performance of students in the core courses who had a similar grade.

The differences between the standardized math achievement-test performance of students in core and basic courses who had similar grades were more pronounced in the high grade groups (those who got As) compared with the low grade groups (those who got Ds). On the basis of this finding, we conclude that grading in the advanced courses before the reform was closer to that in the core courses for high grade groups (those who got As). Grading in the basic courses before the reform was more similar to the grading in the core courses for low grade groups (those who got Ds). Finally, differentiation across different levels of student

COMPARING APPLES AND ORANGES

achievement-test performance, with respect to the variance explained by grades and course level (only before the reform) compared with grades after the reform (all students in core courses), did not differ.

Implications for Research, Policy, and Practice

The findings of this study have several implications for research, policy, and practice. First, our results are in line with an increasing amount of literature on potentially unintended side effects of policy reforms in general and CI reforms more specifically (e.g., Gross, Booker, & Goldhaber, 2009; Hübner et al., 2017). Similar to Nomi and Raudenbush's (2016) findings, our results suggest that reforms, which change the composition of students in courses, can also have an impact on teacher-assigned grades. Our study further expanded Nomi and Raudenbush's findings to the end of upper secondary school, where school grades are an important outcome measure for third parties (e.g., employers or universities) that rely on grades for selection purposes as an indicator of students' abilities (e.g., Clinedinst et al., 2015; Koretz et al., 2016). Our results suggest that, in order to interpret grades as an indicator of students' abilities more meaningfully, the introduction of reforms and their effects should be monitored carefully by the different stakeholders who make use of grades as ability indicators. These results are also important for research that focuses on the transition of students to university or vocational training after upper secondary school because if students with lower achievement-test performance get grades that are similar to those of students with higher achievement-test performance because the grades were influenced by different reference groups or student cohorts, this might have an impact on the transition process and students' subsequent achievements.

Second, in our study, for students in low grade groups (Ds) in core courses, their standardized test performance in math was closer to the standardized test performance of students in basic courses. For students in high grade groups (As) in core courses, their

COMPARING APPLES AND ORANGES

standardized test performance in math was closer to the standardized test performance of students in advanced courses. On the one hand, this might suggest that, after the reform, teachers adapted their grading to some extent to resemble the full range of grading before the reform when the range of grades was larger (see Figure 1). Our results suggest that such adaptations occur more often in mathematics, perhaps because grading standards are easier to adapt in math in comparison with other subjects (e.g., points for objectively correct/incorrect steps on a math test). On the other hand, the reference group effect may have been stronger in math courses because of stronger compositional changes in the reference groups.

In English, grading standards might be more difficult to adapt, and the compositional change in reference groups was not as strong as it was in math. This was supported by findings on standardized test performance differences between advanced and basic courses between the two subjects (see Table 3), which revealed that the test performance differences between advanced and basic courses were considerably larger in math than in English in both states. In addition, this assumption was supported by differences in the proportions of variance on the course level in Thuringia in mathematics versus English: Whereas we found statistically significant proportions of variance at the course level before the reform in math ($\hat{\rho} = .12, p = .017$ in basic courses for final examination grades and for two CAT measures before the reform: numeric: $\hat{\rho} = .11, p = .001$; figural $\hat{\rho} = .10, p < .001$), we did not find these in English for any of the measures. On the basis of these findings, it seems more likely that differences between the results for low and high grade groups between subjects result from differential changes in the composition of students than from differences between teachers in adapting to the changes between subjects. In the future, researchers should address these differential changes more thoroughly and should use a design that will allow them to disentangle the two effects (differential adaptation of teachers by subject vs. differential changes in the composition of subjects and potential interactions between the two).

COMPARING APPLES AND ORANGES

Third, the findings from this study can be further integrated into discussions of the challenges of constructing effective educational policy reforms in general. Effects of reforms are known to be difficult to anticipate even when they are planned carefully (e.g., Gross et al., 2009), and current research has suggested that reforms that have only positive effects and no negative side effects are an exception (e.g., Domina & Saldana, 2012). Our study highlights the advantages of systematically evaluating policy reforms to identify intended and unintended side effects. Investing in rigorous reform monitoring on a national or state level could further contribute to the acceptance of school reforms among teachers as the “ultimate enactors of any change effort” (Porter, Fusarelli, & Fusarelli, 2015, p. 5), reduce the normative parts of reform efforts, and increase sustainable knowledge about effective and ineffective reform characteristics.

Finally, our results point to the importance of reconsidering the advantages of standards-based references to judge student achievement. As outlined, the differences that we found are strongly related to grading on a curve, which strongly relies on the composition of the students in a class, independent of the enactment of educational reforms. Of course, implementing more standardized grading systems would, on a large scale, involve great effort (e.g., improving teacher training and training for current teachers).

Limitations and Future Prospects

There are some limitations that should be mentioned before further prospects are outlined. First, the data we considered were based on cross-sectional cohort control designs (Shadish et al., 2002). Here, students from a cohort that went through the system before the policy reform were compared with a different cohort of students that went through the system after the reform. On the basis of these data, which were assessed in upper secondary school, we could not consider variables from lower secondary school, which might have helped us identify potential selection effects. However, in line with previous research that used these

COMPARING APPLES AND ORANGES

data, we found no considerable differences in observed background variables between the student cohorts (Hübner et al., 2017; Hübner et al., 2018).

Next, we used standardized achievement tests to provide an objective measure of student achievement, but a closer look at the psychometric properties of the instruments indicated that their reliability was not high in all cases. We addressed this issue by using IRT models in all analyses so that the lack of reliability would not bias the estimates. In addition, identical instruments were not used in the two assessments (NEPS and TOSCA), and thus, comparisons between states should be made with caution. However, the results were similar between states/instruments, and thus, this consistency offers support for the generalizability, robustness, and significance of our findings.

An assumption that is essential when working with missing data is that the data are at least conditionally missing at random (i.e., the MAR assumption). Unfortunately, this assumption cannot be tested formally (e.g., Enders, 2010), but we have evidence that suggests that missing values did not have a severe impact on our results and that MAR was reasonable to assume in this case. First, the missing values were rather low overall on the student and school levels. Second, we used a cohort control design, and previous research demonstrated that differences in the covariates between the data from before and after the reform were small (Hübner et al., 2017; Hübner et al., 2018). In addition, in Thuringia, previous research also considered data from the federal statistical office and did not find any major differences in the sample statistics (Hübner et al., 2018). This may be an indication of unsystematic, or at least comparably systematic nonresponse in cohorts before and after the specific reform. In both cases, no large bias for the differences between cohort-specific estimates would be expected. Third, we considered unit nonresponse in sampling weights in Thuringia (Aßmann & Schönberger, 2012), and the results in math were comparable between the two states. Thus, although we could not explicitly test the MAR assumption, we believe these arguments all

COMPARING APPLES AND ORANGES

strongly support the idea that missing values did not have a severe impact on our results and that the MAR assumption held in our data.

An important component of our analysis is related to the assumption that, after the reform, in the core courses, students were not grouped according to their previous achievement (e.g., in terms of “advanced core courses” and “basic core courses”). Our results for Thuringia showed that none of the proportions of variance on the course level were statistically significant after the reform in English or in math. Overall, these results support our assumption that students were not rigorously sorted into higher and lower achieving core courses. However, we could not formally test whether this assumption also held in Baden-Württemberg because course affiliation was not tracked in the TOSCA studies. Future research should focus more strongly on such potential “hidden” implementation effects in order to adequately identify the specific changes implemented by educational policy reforms. In the same vein and to support the testing of hidden tracking mechanisms, future studies should investigate whether the compositional differences across classrooms moderate relationships between grades and standardized achievement-test performance also using rich longitudinal non-reform data. On the basis of our findings, three suggestions for future research are apparent.

First, considering additional information on potential course choices after reforms in future studies might help researchers obtain a deeper understanding of the effects of detracking. Such information might be especially useful for studying the effects of reforms on individual course choice behavior and could help determine the strength of the reference group effect. However, the validity of such self-reported, counterfactual information (i.e., the extent to which such a measure could provide a valid measure of actual course enrollment in the absence of the reform) remains unclear.

COMPARING APPLES AND ORANGES

Second, findings for the CI reforms outlined here should also be tested in the context of other reforms. This is important to do to increase knowledge about how reforms affect student achievement and factors that are related to school achievement such as self-concept and interest. Comparable results might arise, particularly if the reforms introduce changes in tracking procedures and the composition of students in courses.

Third, the topic of the relationship between teacher-assigned grades and standardized test scores can be integrated into a larger debate about the alignment and foundations of the two measures (e.g., Willingham et al., 2002). The basic goal of the present research was not to identify the “better” measure because both of them come with specific advantages and disadvantages (Hochweber et al., 2014). From a more general perspective, this study focused on the fact that grades at the end of any phase in school (elementary, lower or upper secondary school) have a fundamental function for transition decisions with respect to societal selection and allocation, for instance, for university or for later employment. Therefore, even without discussing whether grades are the best basis for such decisions, the impact of educational reforms should be monitored more carefully because such reforms can lead to systematic changes in the expected standardized test performance of students who earned the same grade in different cohorts. Unfortunately, very few studies have investigated the impact of policies on grades, and studies that have linked CI policies with reference group effects are rare.

As a final note, it is important to investigate novel ways to implement standardized grading strategies and increase research on how teachers consolidate knowledge on students’ gender, socioeconomic status, or other factors when assigning grades. Otherwise, grading will continue to vary unsystematically between students with similar standardized test performance depending on the teacher and the composition of student performance in a course

COMPARING APPLES AND ORANGES

(e.g., Guskey, 2006; Hochweber et al., 2014; Westphal et al., 2016; Zimmermann et al., 2013).

COMPARING APPLES AND ORANGES

References

- Allensworth, E., Nomi, T., Montgomery, N., & Lee, V. E. (2009). College preparatory curriculum for all: Academic consequences of requiring algebra and English I for ninth graders in Chicago. *Educational Evaluation and Policy Analysis*, 31(4), 367–391.
<https://doi.org/10.3102/0162373709343471>
- Asparouhov, T. (2005). Sampling Weights in Latent Variable Modeling. *Structural Equation Modeling: a Multidisciplinary Journal*, 12(3), 411–434.
- Aßmann, C., & Schönberger, B. (2012). *Curricular Reform Study in Thuringia (TH) SUF Version 1.0.0: Data Manual [Supplement]: Weighting*. Bamberg: NEPS National Education Panel Study.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1), 289–300.
- Blossfeld, H.-P., Rossbach, H. G., & Maurice, J. von (Eds.). (2011). *Education as a lifelong process: The German National Educational Panel Study (NEPS)*. Zeitschrift für Erziehungswissenschaft: Sonderheft 14.
- Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences of the United States of America*, 113(47), 13354–13359.
<https://doi.org/10.1073/pnas.1601135113>
- Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation*, 17(3), 141–159. <https://doi.org/10.1080/13803611.2011.597112>
- Brookhart, S. M. (2008). *How to give effective feedback to your students*. Alexandria, Virginia: Association for Supervision and Curriculum Development. Retrieved from

COMPARING APPLES AND ORANGES

<http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=287379>

Brookhart, S. M. (2015). Graded achievement, tested achievement, and validity. *Educational Assessment, 20*(4), 268–296. <https://doi.org/10.1080/10627197.2015.1093928>

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., . . . Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803–848. <https://doi.org/10.3102/0034654316672069>

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

Clinedinst, M., Koranteng, A.-M., & Nicola, T. (2015). *State of college admission*. Arlington, VA: National Association for College Admission Counseling.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Crosnoe, R., & Benner, A. D. (2015). Children at school. In R. M. Lerner (Ed.), *Handbook of child psychology and developmental science* (pp. 1–37). Hoboken, New Jersey: Wiley. <https://doi.org/10.1002/9781118963418.childpsy407>

Dickinson, E. R., & Adelson, J. L. (2015). Choosing among multiple achievement measures: Applying multitrait-multimethod confirmatory factor analysis to state assessment, ACT, and student GPA data. *Journal of Advanced Academics, 27*(1), 4–22. <https://doi.org/10.1177/1932202X15621905>

COMPARING APPLES AND ORANGES

- Domina, T., McEachin, A., Penner, A., & Penner, E. (2015). Aiming high and falling short: California's eighth-grade algebra-for-all effort. *Educational Evaluation and Policy Analysis*, 37(3), 275–295. <https://doi.org/10.3102/0162373714543685>
- Domina, T., & Saldana, J. (2012). Does raising the bar level the playing field? Mathematics curricular intensification and inequality in American high schools, 1982-2004. *American Educational Research Journal*, 49(4), 685–708. <https://doi.org/10.3102/0002831211426347>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, 6(4), 352–370.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576. <https://doi.org/10.1146/annurev.psych.58.110405.085530>
- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science : the Official Journal of the Society for Prevention Research*, 8(3), 206–213. <https://doi.org/10.1007/s11121-007-0070-9>
- Gross, B., Booker, T. K., & Goldhaber, D. (2009). Boosting student achievement: The effect of comprehensive school reform on student achievement. *Educational Evaluation and Policy Analysis*, 31(2), 111–126. <https://doi.org/10.3102/0162373709333886>

COMPARING APPLES AND ORANGES

Guskey, T. R. (2006). Making high school grades meaningful. *The Phi Delta Kappan*, 87(9), 670–675.

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97–121). Washington, DC: American Psychological Association.

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3), 607–668.

<https://doi.org/10.1257/jel.46.3.607>

Hanushek, E. A., & Woessmann, L. (2010). Education and economic growth. In E. Baker, B. McGaw, & P. Peterson (Eds.), *International encyclopedia of education* (pp. 245–252). Amsterdam: Elsevier Academic.

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

Hess, F. M., & Petrilli, M. J. (2006). *No Child Left Behind primer*. New York: Peter Lang.

Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology*, 106(1), 289–300. <https://doi.org/10.1037/a0033829>

Hübner, N., Wagner, W., Nagengast, B., & Trautwein, U. (2018). Putting all students in one basket does not produce equality: Gender-specific effects of curricular intensification in upper secondary school. *School Effectiveness and School Improvement*. Advance online publication. <https://doi.org/10.1080/09243453.2018.1504801>

Hübner, N., Wille, E., Cambria, J., Oschatz, K., Nagengast, B., & Trautwein, U. (2017).

Maximizing gender equality by minimizing course choice options? Effects of obligatory

COMPARING APPLES AND ORANGES

- coursework in math on gender differences in STEM. *Journal of Educational Psychology*, 109(7), 993–1009. <https://doi.org/10.1037/edu0000183>
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351a), 631–639. <https://doi.org/10.1080/01621459.1975.10482485>
- Kiefer, T., Robitzsch, A., & Wu, M. (2017). TAM: Test Analysis Modules [Computer software]. Retrieved from <http://CRAN.R-project.org/package=TAM>
- Köller, O., Watermann, R., Trautwein, U., & Lüdtke, O. (Eds.). (2004). *Wege zur Hochschulreife in Baden-Württemberg: TOSCA - eine Untersuchung an allgemein bildenden und beruflichen Gymnasien* [Ways towards higher education entrance qualification in Baden-Württemberg: TOSCA – an investigation of general and vocational upper secondary schools]. Opladen: Leske und Budrich.
- Koretz, D., Yu, C., Mbekeani, P. P., Langi, M., Dhaliwal, T., & Braslow, D. (2016). Predicting freshman grade point average from college admissions test scores and state high school test scores. *AERA Open*, 2(4), 1-13. <https://doi.org/10.1177/2332858416670601>
- Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The Big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal*, 44(3), 631–669. <https://doi.org/10.3102/0002831207306728>
- 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., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The Big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, 20(3), 319–350. <https://doi.org/10.1007/s10648-008-9075-6>

COMPARING APPLES AND ORANGES

- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2), 397–416. <https://doi.org/10.1111/j.1467-8624.2005.00853.x>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>
- McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, 22(1), 114–140. <https://doi.org/10.1037/met0000078>
- Mullis, I. V.S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1998). *Mathematics and science achievement in the final year of secondary school: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267–316.
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.
- Nagy, G., Neumann, M., Trautwein, U., & Lüdtke, O. (2010). Voruniversitäre Mathematikleistungen vor und nach der Neuordnung der gymnasialen Oberstufe in Baden-Württemberg [Advanced math before and after the reform of general upper secondary school in Baden-Württemberg]. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke, & K. Maaz (Eds.), *Schulleistungen von Abiturienten. Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand* (pp. 147–180). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92037-5_6

COMPARING APPLES AND ORANGES

National Educational Panel Study (NEPS). (2011). *Curricular reform study in Thuringia - Main study 2009/10 (A70) - Students, 12th grade: Information on the competence test.*

Retrieved from <https://www.neps->

[data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/TH/1-0-0/C_A70_EN.pdf](https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/TH/1-0-0/C_A70_EN.pdf)

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>

Newman, D. A. (2003). Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques. *Organizational Research Methods*, 6(3), 328–362.
<https://doi.org/10.1177/1094428103254673>

Nomi, T., & Raudenbush, S. W. (2016). Making a success of "Algebra for All": The impact of extended instructional time and classroom peer skill in Chicago. *Educational Evaluation and Policy Analysis*, 38(2), 431–451. <https://doi.org/10.3102/0162373716643756>

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Organisation for Economic Co-operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD.

Organisation for Economic Co-operation and Development. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). (2015). *Education policy outlook 2015: Making reforms happen*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.

COMPARING APPLES AND ORANGES

Parker, P. D., Schoon, I., Tsai, Y.-M., Nagy, G., Trautwein, U., & Eccles, J. S. (2012).

Achievement, agency, gender, and socioeconomic background as predictors of postschool choices: A multicontext study. *Developmental Psychology*, 48(6), 1629–1642.

<https://doi.org/10.1037/a0029167>

Penner, A. M., Domina, T., Penner, E. K., & Conley, A. (2015). Curricular policy as a

collective effects problem: A distributional approach. *Social Science Research*, 52, 627–

641. <https://doi.org/10.1016/j.ssresearch.2015.03.008>

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report - Scaling the Data of the*

Competence Tests: (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Porter, R. E., Fusarelli, L. D., & Fusarelli, B. C. (2015). Implementing the common core:

How educators interpret curriculum reform. *Educational Policy*, 29(1), 111–139.

<https://doi.org/10.1177/0895904814559248>

R Development Core Team. (2016). *R: A language and environment for statistical computing*.

Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>.

Ravitch, D. (2011). *The death and life of the great American school system: How testing and choice are undermining education*. New York, N.Y.: Basic Books.

Raykov, T., & Pohl, S. (2013). Essential unidimensionality examination for multicomponent

scales: An interrelationship decomposition approach. *Educational and Psychological*

Measurement, 73(4), 581–600. <https://doi.org/10.1177/0013164412470451>

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university

students' academic performance: a systematic review and meta-analysis. *Psychological*

Bulletin, 138(2), 353–387. <https://doi.org/10.1037/a0026838>

COMPARING APPLES AND ORANGES

- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130(2), 261–288. <https://doi.org/10.1037/0033-2909.130.2.261>
- Robinson, M., & Monks, J. (2005). Making SAT scores optional in selective college admissions: A case study. *Economics of Education Review*, 24(4), 393–405. <https://doi.org/10.1016/j.econedurev.2004.06.006>
- Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940-2009. *Teachers College Record*, 114, 1–23.
- Rupp, A. A., Vock, M., Harsch, C., & Köller, O. (2008). *Developing standards-based assessment tasks for English as a first foreign language: Context, processes and outcomes in Germany. Standards-Based Assessment Tasks for English as a First Language: Vol. 1.* Münster: Waxmann.
- Sahlberg, P. (2006). Education reform for raising economic competitiveness. *Journal of Educational Change*, 7(4), 259–287. <https://doi.org/10.1007/s10833-005-4884-6>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177.
- 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>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Snijders, T.A.B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles: Sage.
- Sørensen, A. B. (1970). Organizational differentiation of students and educational opportunity. *Sociology of Education*, 43(4), 355–376. <https://doi.org/10.2307/2111838>

COMPARING APPLES AND ORANGES

- Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, 35(5), 401–426.
<https://doi.org/10.1016/j.intell.2006.09.004>
- Swanson, C. B., & Stevenson, D. L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis*, 24(1), 1–27. <https://doi.org/10.3102/01623737024001001>
- The National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington: Government Printing Office.
- The Standing Conference of the Ministers of Education and Cultural Affairs (KMK). (1995). *Weiterentwicklung der Prinzipien der gymnasialen Oberstufe und des Abiturs: Abschlussbericht der von der Kultusministerkonferenz eingesetzten Expertenkommission* [Future trends in upper secondary school at the Gymnasium and the Abitur: Final report of the expert commission deployed by the KMK]. Bonn: KMK.
- Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 119–169). Greenwich, CT: IAP.
- Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation*, 18(2), 153–172.
<https://doi.org/10.1080/13803611.2012.659929>
- 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>
- Trautwein, U., & Neumann, M. (2008). Das Gymnasium [The Gymnasium]. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer, & L. Trommer (Eds.), *Das Bildungswesen in der*

COMPARING APPLES AND ORANGES

Bundesrepublik Deutschland: Strukturen und Entwicklungen im Überblick (pp. 467–501).

Reinbek bei Hamburg: Rowohlt.

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O., & Maaz, K. (Eds.). (2010).

Schulleistungen von Abiturienten. Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand [Achievement of upper secondary school graduates. The newly ordered upper secondary school on the trial]. Wiesbaden: VS Verlag für Sozialwissenschaften.

Volante, L. (Ed.). (2016). *The intersection of international achievement testing and educational policy: Global perspectives on large-scale reform*. New York: Routledge.

Wagner, W., Kramer, J., Trautwein, U., Lüdtke, O., Nagy, G., Jonkmann, K., . . . Schilling, J. (2011). 15 Upper secondary education in academic school tracks and the transition from school to postsecondary education and the job market. *Zeitschrift Für Erziehungswissenschaft*, 14(S2), 233–249. <https://doi.org/10.1007/s11618-011-0196-1>

Wagner, W., Rose, N., Dicke, A.-L., Neumann, M., & Trautwein, U. (2014). Alle alles lehren - Schulleistungen in Englisch, Mathematik und den Naturwissenschaften vor und nach der Neuordnung der gymnasialen Oberstufe in Sachsen [Teaching everyone everything: Student achievement in English, mathematics and the natural sciences before and after Saxony's upper secondary school reform]. *Zeitschrift Für Erziehungswissenschaft*, 17(2), 345–369. <https://doi.org/10.1007/s11618-014-0492-7>

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450. <https://doi.org/10.1007/BF02294627>

Watermann, R., Nagy, G., & Köller, O. (2004). Mathematikleistungen in allgemein bildenden und beruflichen Gymnasien [Math achievement in the general and vocational upper secondary school]. In O. Köller, R. Watermann, U. Trautwein, & O. Lüdtke (Eds.), *Wege zur Hochschulreife in Baden-Württemberg: TOSCA - eine Untersuchung an allgemein bildenden und beruflichen Gymnasien* (pp. 205–283). Opladen: Leske und Budrich.

COMPARING APPLES AND ORANGES

- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). 5: Development of competencies across the life span. In H.-P. Blossfeld, H. G. Rossbach, & J. von Maurice (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). *Zeitschrift für Erziehungswissenschaft*: 14.
<https://doi.org/10.1007/s11618-011-0182-7>
- Westphal, A., Becker, M., Vock, M., Maaz, K., Neumann, M., & McElvany, N. (2016). The link between teacher-assigned grades and classroom socioeconomic composition: The role of classroom behavior, motivation, and teacher characteristics. *Contemporary Educational Psychology*, 46, 218–227. <https://doi.org/10.1016/j.cedpsych.2016.06.004>
- Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, 39(1), 1–37.
<https://doi.org/10.1111/j.1745-3984.2002.tb01133.x>
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370–371.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2: Generalised item response modelling software*. Camberwell: Australian Council for Educational Research.
- Zimmermann, F., Schütte, K., Taskinen, P., & Köller, O. (2013). Reciprocal effects between adolescent externalizing problems and measures of achievement. *Journal of Educational Psychology*, 105(3), 747–761. <https://doi.org/10.1037/a0032793>

Table 1

Correlations between Grades and Standardized Achievement-Test Performance in the two Federal States

Variable (State)	Math grade (TH)	English grade (TH)	Math grade (BW)	English grade (BW)
Math achievement (TH)	.48***	.32***	-	-
English achievement (TH)	.32***	.62***	-	-
Math achievement (BW)	-	-	.64***	.31***
English achievement (BW)	-	-	.25***	.62***

Note. Cohorts assessed before and after the reforms were combined. Standardized test performance was modeled as latent variables in these models instead of point estimates so that lack of reliability would not bias the estimates. Grades (uncategorized) were modeled as continuous predictors. TH = Thuringia; BW = Baden-Württemberg.

*** $p < .001$.

Table 2

Course-Specific Correlations between Standardized Achievement-Test Performance and Grades

Subject	Thuringia					
	r_A	p_{A_B}	r_B	p_{A_C}	r_C	p_{B_C}
Mathematics	.52	<.001	.30	.915	.53	.001
English	.61	.022	.61	.083	.59	.669
	Baden-Württemberg					
	r_A	p_{A_B}	r_B	p_{A_C}	r_C	p_{B_C}
Mathematics	.59	< .001	.54	.543	.75	< .001
English	.62	.018	.57	.253	.64	.080

Note. Index A = Advanced course; B = Basic course; Index C = Core course. p_{A_B} = p -value for the difference between the correlations in advanced and basic courses. p_{A_C} = p -value for the difference between the correlations in advanced and core courses. p_{B_C} = p -value for the difference between the correlations in basic and advanced courses.

Table 3

Descriptive Statistics for the Central Outcome Variables

Variable	Thuringia								
	Before the reform				After the reform				
	Advanced (A)		Basic (B)		Core (C)		<i>p</i> AB	<i>p</i> AC	<i>p</i> BC
Course	<i>N</i>	<i>M</i>	<i>N</i>	<i>M</i>	<i>N</i>	<i>M</i>			
Math grade	512	9.24 (3.04)	614	8.21 (3.02)	959	8.94 (3.12)	< .001	.087	< .001
English grade	369	10.11 (2.20)	719	8.89 (2.45)	924	9.62 (2.43)	< .001	.024	.002
Math ACH	512	54.93 (9.27)	614	45.35 (8.40)	959	50.34 (9.91)	< .001	< .001	< .001
English ACH	369	54.10 (9.79)	719	47.50 (8.81)	924	50.30 (9.32)	< .001	< .001	< .001
Baden-Württemberg									
Math grade	1017	9.85 (3.08)	1636	8.01 (3.59)	2800	8.27 (3.64)	< .001	< .001	.078
English grade	1078	9.67 (2.70)	1198	9.13 (2.86)	2479	9.25 (2.92)	< .001	.001	.324
Math ACH	1017	57.15 (9.49)	1636	43.75 (6.96)	2800	51.06 (9.31)	< .001	< .001	< .001
English ACH	1078	54.20 (9.77)	1198	46.6 (9.53)	2479	49.80 (9.67)	< .001	< .001	< .001

Note. Standardized test performance was transformed into scores with $M = 50$ and $SD = 10$. ACH = standardized achievement-test performance. N = Sample size; M = Mean; Standard deviations in parentheses. On the basis of results by Hübner, Wagner, Nagengast, and Trautwein (2018), we decided to exclude the specific group of basic course students in English after the reform because only a few students with a different language as the core subject chose English as a basic course after the reform. Compared with Trautwein et al. (2010), we considered data only from general upper secondary schools. We controlled the false discovery rate within subjects and countries (Benjamini & Hochberg, 1995).

Table 4

Standardized Achievement-Test Performance Predicted by Average Grades in Thuringia and Baden-Württemberg

Thuringia				
Course / Subject	Mathematics	p	English	p
Advanced (A)	54.3 _{BC}	$p_B < .001$	52.5 _{BC}	$p_B < .001$
Basic (B)	46.0 _{AC}	$p_C < .001$	48.7 _{AC}	$p_C = .027$
Core (C)	50.3 _{AB}	$p_A < .001$	50.1 _{AB}	$p_A = .002$
Baden-Württemberg				
Advanced (A)	54.8 _{BC}	$p_B < .001$	53.4 _{BC}	$p_B < .001$
Basic (B)	44.3 _{AC}	$p_C < .001$	47.0 _{AC}	$p_C < .001$
Core (C)	51.6 _{AB}	$p_A < .001$	50.0 _{AB}	$p_A < .001$

Note. Grades were given on a scale ranging from 0 (*lowest achievement*) to 15 points (*highest achievement*). Grades were modeled as continuous predictors. We controlled the false discovery rate within subjects and states (Benjamini & Hochberg, 1995). The metric of the latent variables was transformed to $M = 50$ and $SD = 10$ on the basis of pooled means and standard deviations of the latent variables. We also specified models that considered a group-specific quadratic trend. This did not have a severe impact on the results displayed here because only the p -value for the difference between core and basic courses in English in Thuringia changed and was no longer statistically significant ($p = .138$). Indices indicate two-sided statistically significant group differences between the advanced, basic, or core courses, respectively: Index _C = statistically significantly different from basic course; Index _A = statistically significantly different from advanced course; Index _C = statistically significantly different from core course.

Table 5

Standardized Achievement-Test Performance in Mathematics Predicted in Different Grade Groups in Thuringia and Baden-Württemberg

Thuringia								
Course / Group	D	p	C	p	B	p	A	p
Advanced (A)	48.2 _{BC}	$p_B < .001$	52.7 _{BC}	$p_B < .001$	56.8 _{BC}	$p_B < .001$	64.2 _B	$p_B < .001$
Basic (B)	41.8 _{AC}	$p_C = .014$	45.9 _A	$p_C = .080$	46.8 _{AC}	$p_C < .001$	50.5 _{AC}	$p_C < .001$
Core (C)	44.4 _{AB}	$p_A = .003$	47.7 _A	$p_A < .001$	52.5 _{AB}	$p_A = .004$	59.7 _B	$p_A = .062$
Baden-Württemberg								
Advanced (A)	49.4 _{BC}	$p_B < .001$	52.8 _{BC}	$p_B < .001$	58.3 _{BC}	$p_B < .001$	65.8 _{BC}	$p_B < .001$
Basic (B)	40.0 _{AC}	$p_C < .001$	43.0 _{AC}	$p_C < .001$	46.4 _{AC}	$p_C < .001$	50.6 _{AC}	$p_C < .001$
Core (C)	43.8 _{AB}	$p_A < .001$	49.7 _{AB}	$p_A < .001$	55.3 _{AB}	$p_A < .001$	63.3 _{AB}	$p_A = .012$

Note. Grades were given on a scale ranging from 0 (*lowest achievement*) to 15 points (*highest achievement*). Grade groups were created as follows: Group D: < 6.5 points, Group C: ≥ 6.5 and < 9.5 points, Group B: ≥ 9.5 and < 12.5 points, Group A: ≥ 12.5 points. We controlled the false discovery rate within subjects, states, and grade groups (Benjamini & Hochberg, 1995). When considering only grades from Grade 12 Semester 1 in Thuringia, all differences in grade group results remained comparable. However, all differences in Grade Groups A (core vs. advanced course now $p = .016$) and C (basic vs. core course now $p < .001$) were statistically significant in these analyses, and basic and core courses were not statistically significantly different in Grade Group D ($p = .143$). In Baden-Württemberg, all differences between different grade groups were statistically significant ($p < .001$). In Thuringia, all differences between all grade groups but one (basic course grade C vs. B, $p = .471$) were statistically significant (ranging from $p < .001$ to $p = .009$). The metric of the latent variables was transformed to $M = 50$ and $SD = 10$ on the basis of pooled means and standard deviations of the latent variables. Indices indicate two-sided statistically significant group differences between the advanced, basic, or core courses, respectively: Index _B = statistically significantly different from basic course; Index _A = statistically significantly different from advanced course; Index _C = statistically significantly different from core course.

Table 6

Standardized Achievement-Test Performance in English Predicted by Different Grade Groups in Thuringia and Baden-Württemberg

Thuringia								
Course / Group	D	<i>p</i>	C	<i>p</i>	B	<i>p</i>	A	<i>p</i>
Advanced	41.8	$p_B = .218$	48.4 _B	$p_B = .001$	55.9 _{BC}	$p_B < .001$	63.1	$p_B = .494$
Basic	40.4	$p_C = .242$	44.4 _A	$p_C = .055$	50.9 _A	$p_C = .110$	61.4	$p_C = .883$
Core	41.9	$p_A = .979$	46.5	$p_A = .168$	52.4 _A	$p_A = .012$	60.9	$p_A = .376$
Baden-Württemberg								
Advanced	44.9 _{BC}	$p_B < .001$	50.3 _{BC}	$p_B < .001$	56.5 _{BC}	$p_B < .001$	64.0 _{BC}	$p_B < .001$
Basic	39.6 _{AC}	$p_C = .002$	44.4 _{AC}	$p_C < .001$	48.9 _{AC}	$p_C < .001$	57.3 _{AC}	$p_C < .001$
Core	42.0 _{AB}	$p_A < .001$	46.8 _{AB}	$p_A < .001$	52.4 _{AB}	$p_A < .001$	61.1 _{AB}	$p_A = .001$

Note. Grades were given on a scale ranging from 0 (*lowest achievement*) to 15 points (*highest achievement*). Grade groups were created as follows: Group D: < 6.5 points, Group C: ≥ 6.5 and < 9.5 points, Group B: ≥ 9.5 and < 12.5 points, Group A: ≥ 12.5 points. We controlled the false discovery rate within subjects, states, and grade groups (Benjamini & Hochberg, 1995). When considering only grades from Grade 12 Semester 1 in Thuringia, all differences in grade group results remained comparable. However, in these analyses, advanced and basic courses in Grade Group C were statistically significant different ($p = .001$) as well as advanced and basic courses in Grade Group A ($p = .027$). In Baden- Württemberg, all differences between grade groups were statistically significant at $p < .001$. In Thuringia, all differences between grade groups were statistically significant (ranging from $p < .001$ to $p = .001$). The metric of the latent variables was transformed to $M = 50$ and $SD = 10$ on the basis of pooled means and standard deviations of the latent variables. Indices indicate two-sided statistically significant group differences between the advanced, basic, or core courses, respectively: Index _B = statistically significantly different from basic course; Index _A = statistically significantly different from advanced course; Index _C = statistically significantly different from core course.

Table 7

Proportions of Variance for the Different Measures on the Course Level

Variable	Math courses after the reform	
	ρ_{course}	p
CAT: verbal	.01	.894
CAT: numeric	.07	.070
CAT: figural	.04	.103
Final exam grade	.04	.281
Standardized test performance	.05	.110
	English courses after the reform	
	ρ_{course}	p
CAT: verbal	.01	.756
CAT: numeric	.02	.427
CAT: figural	.01	.685
Final exam grade	.01	.875
Standardized test performance	.02	.281

Note. ρ_{course} = proportion of variance on the course level.

Figure 1. Standardized achievement-test performance by grade group and course. Please refer to Table 4 and Table 5 for exact information on statistical significance of differences displayed for both states. Grade groups were built as: Group D: < 6.5 points, Group C: ≥ 6.5 and < 9.5 points, Group B: ≥ 9.5 and < 12.5 points, Group A: ≥ 12.5 points. The metric of the latent variable was transformed to $M = 50$ and $SD = 10$ on the basis of the pooled means and the standard deviation of the latent variable. TH = Thuringia; BW = Baden-Württemberg. AC = Advanced course; BC = Basic course; CC = Core course.