

DISSERTATION | DOCTORAL THESIS

Titel | Title

Teacher Grades for University Entrance Qualification – Standardization, Reform,
and Teachers’ Grade Decision-Making

verfasst von | submitted by

Mag.rer.nat. Vincent Stephan Schatz

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Doktor der Philosophie (Dr.phil.)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt |
Degree programme code as it appears on the
student record sheet:

UA 792 297

Dissertationsgebiet lt. Studienblatt | Field of
study as it appears on the student record
sheet:

Bildungswissenschaft

Betreut von | Supervisor:

Univ.-Prof. Dr. Nele Kampa | Assoz. Prof. Mag. Dr.
Marko Lüftenegger Privatdoz.

Table of Contents

Abstract	3
Kurzzusammenfassung	4
1. Introduction	5
2. Theoretical Framework.....	6
2.1 Defining Different Grading Approaches	6
2.2 Differentiating Between Teacher Grades and Standardized Test Scores	7
2.3 Determining the Degree of Standardization of the Austrian Matura	8
3. Focus and Structure of the Dissertation	9
4. Author Positionality	11
5. From Policy to Practice: Three Studies on Austria's Developing University Entrance System	12
5.1 Embedding the First Study: Non-Standardized Versus Standardized Teacher Grades	13
5.1.1 Matura Development Stage: Introduction of the Standardized Matura for Academic Schools (2014/15).....	13
5.1.2 About the First Dissertation Study	14
5.1.3 European Perspective: Evidence from Sweden	16
5.2 Embedding the Second Study: School Type, Grading Practices, and Response Behavior	19
5.2.1 Matura Development Stage: Introduction of the Standardized Matura for Vocational Schools	19
5.2.2 About the Second Dissertation Study	21
5.2.3 European Perspective: Evidence from Germany	23
5.3 Embedding the Third Study: The 2020 Grading Reform and Teachers Grade-Decision Making	26
5.3.1 Matura Development Stage: 2020 Grading Reform: Integrating Classroom Grades into Standardized Matura Grades.....	26
5.3.2 About the Third Dissertation Study	27
5.3.3 European Perspective: Evidence from Ireland	29
6. The Three Studies Forming my Dissertation	32
6.1 First Study – Extended Summary: Teacher Grades for University Entrance Qualification: Bias, But Not as We Expected.....	32
6.2 Second Study – Extended Summary: Self-Selection Bias in the Sensitive Topic of Teacher Grading: The Role of Subject and School Context	35
6.3 Third Study – Extended Summary: Grading After Reform: How Policy Change Affected Teachers' Grading for University Entrance Qualifications	37
7. Concluding Thoughts	40
8. References	45
9. Appendix	55
9.1 First Study.....	55
9.2 Second Study.....	80
9.3 Third Study	125

Abstract

This dissertation examines teachers' grading practices in the Matura, Austria's university entrance qualification at the end of secondary school. Teachers play a pivotal role in this system, as they themselves determine their students' eligibility for higher education. The dissertation traces three key stages in the development of the Matura assessment system over the past ten years, each addressed in a separate study.

The first study analyzed the introduction of standardized Matura examinations in 2015 at academic upper secondary schools. Longitudinal analyses revealed a very high level of alignment between non-standardized classroom grades and standardized Matura grades, as well as systematic grading differences between school types – two findings that call into question the added value of the costly standardized assessment system.

The second study examined the implementation of the same reform in 2016 at vocational upper secondary schools. Unintentionally, this study uncovered a school-type-moderated self-selection bias, which not only revealed the importance of school type as a contextual factor shaping grading practices, but also highlighted its relevance for teachers' response behavior, while at the same time illustrating the sensitivity of teacher grading as an educational research topic.

The third study focused on the 2020 grading reform, which allowed negative Matura examination results to be compensated by positive classroom grades. A longitudinal mixed-methods analysis revealed significantly higher classroom grades since 2020 and showed that teachers' grading behavior has since been shaped by systemic pressures, strategic grading decisions, and internal conflicts between empathy and fairness. According to teacher interviews, the reform has also largely undermined the standardized character of the Matura.

Overall, the dissertation highlights the socially and institutionally embedded nature of grading. Initially guided by the implicit assumption that comparability represents the ultimate goal of university entrance examinations, the dissertation concludes that this assumption conflicts with the inherent heterogeneity of the Austrian upper secondary education system and invites a reconsideration of the necessity of standardization and comparable educational qualifications.

Kurzzusammenfassung

Die vorliegende Dissertation untersucht die Notengebung bei der österreichischen Matura. Lehrer*innen kommt hier eine tragende Rolle zu, da sie selbst über die Hochschulreife ihrer Schüler*innen entscheiden. Die Dissertation zeichnet drei zentrale Entwicklungsschritte des Bewertungssystems der Matura der letzten 10 Jahre nach, die jeweils in einer eigenen Studie behandelt werden.

Die erste Studie analysierte die Einführung standardisierter Maturaprüfungen im Jahr 2015 an allgemeinbildenden Schulen. Längsschnittlich zeigte sich eine sehr hohe Übereinstimmung zwischen nicht-standardisierten Jahres- und standardisierten Maturanoten sowie systematische Benotungsunterschiede zwischen Schultypen – zwei Ergebnisse, die den Mehrwert des kostenintensiven standardisierten Systems infrage stellen.

Die zweite Studie beleuchtete die Umsetzung derselben Reform im Jahr 2016 an berufsbildenden Schulen. Unbeabsichtigt deckte diese Studie einen durch den Schultyp moderierten Selbstselektionsbias auf, was nicht nur die Bedeutung des Schultyps als kontextuellen Einflussfaktor auf die Notengebung unterstreicht, sondern auch dessen Relevanz für das Antwortverhalten von Lehrer*innen und zugleich die Sensibilität des Themas Leistungsbeurteilung als Gegenstand bildungswissenschaftlicher Forschung.

Die dritte Studie befasste sich mit der Notenreform 2020, die die Kompensation negativer Maturaprüfungen durch das Einfließen positiver Jahresnoten ermöglichte. Eine längsschnittliche Mixed-Methods-Analyse ergab signifikant bessere Jahresnoten seit 2020 und machte deutlich, dass das Notengebungsverhalten von Lehrer*innen seither von systemischen Kräften, strategischen Notenentscheidungen und inneren Konflikten zwischen Empathie und Fairness geprägt ist. Laut Lehrer*inneninterviews habe die Reform zudem den standardisierten Charakter der Matura weitgehend untergraben.

Die Dissertation stellt insbesondere den sozial und institutionell geprägten Charakter der Notengebung in den Fokus. Ursprünglich von der unbewussten Annahme geleitet, dass vergleichbare Ergebnisse das goldene Ziel von Hochschulreifeprüfungen seien, kommt die Dissertation zum Schluss, dass dieser Wunsch in Kontrast mit der inhärenten Heterogenität des österreichischen Oberstufensystems steht und lädt zu einem Umdenken hinsichtlich der Notwendigkeit von Standardisierung und vergleichbarer Bildungsabschlüsse ein.

1. Introduction

This dissertation investigates the role of teacher grading within the *Matura*, Austria's standardized university entrance qualification at the end of upper secondary school. Tracing the development of the Matura assessment system over the past decade, it provides an empirical account of three major stages of reform that changed teachers' roles as graders within the Matura system. Each stage of reform is examined through a separate study. The three stages include the transition from a norm-referenced to a standardized system in 2014/15 for academic upper secondary schools, the corresponding transition for vocational upper secondary schools in 2015/16, and finally the introduction of the current grading system following a grading reform implemented in 2020. Each study is further contextualized by drawing on evidence from another European country that has implemented similar steps or reforms, or encountered comparable challenges related to teacher grading in university entrance qualifications.

The theoretical basis for the dissertation is a set of definitions that provides clarification of the most prevalent grading approaches used in educational systems and highlights the distinctions among them: traditional grading, standardized grading, standardized testing, standards-based grading, and mastery grading. The current grading system of the Austrian Matura stretches the boundaries of these approaches and is therefore not easy to classify. Its complexity stems from the third major stage of reform in 2020, which merged two fundamentally different grading approaches (traditional and standardized grading) to determine final Matura grades.

The theoretical framework further guiding this dissertation builds on literature examining the composition of teacher-awarded grades, specifically drawing on Brookhart et al.'s (2016) framework of academic and non-academic achievement factors of grades, their divergence from standardized test scores (e.g., Bowers, 2011), and McMillan's (2003) model of teachers' grade decision-making.

As my three empirical studies will demonstrate, the Austrian Matura presents an interesting context for examining teachers' grading practices, particularly in light of the 2020 grading reform, which offers a unique opportunity to investigate the complex mechanisms underlying teachers' grading decisions, both in general and within the specific context of university entrance qualifications.

2. Theoretical Framework

2.1 Defining Different Grading Approaches

Grading practices in education can be broadly categorized into *standardized* and *non-standardized* grading approaches. Standardized grading refers to a structured system in which teachers follow uniform grading procedures to enhance consistency across graders, classrooms, and schools (Brookhart, 2016). Standardized *grading* is not to be confused with standardized *testing*, which usually involves externally developed and graded assessments designed to ensure grade reliability through objectively scored exam items (Haertel, 2006). Unlike standardized grading, which seeks to improve consistency within teacher-assigned grades, standardized testing removes teachers from the grading process to eliminate grader bias (Heyneman, 1988). A special form of standardized grading is *mastery grading*, which measures students' attainment of specific learning goals through a simple mastered versus not mastered evaluation (Guskey & Bailey, 2001).

Standardized grading is also not synonymous with *standards-based* grading (alternatively referred to as competency-based grading or proficiency-based grading), which evaluates students' performance and progress based on their ability to meet clearly defined learning standards and objectives (Link & Guskey, 2022). While it incorporates the standardized and centralized features of a standardized test, standards-based grading seeks to go beyond simply providing a final result; instead, it aims to serve as a tool to support learning by providing students with feedback on their progress against predetermined performance criteria (Swan et al., 2014; Tomlinson & McTighe, 2006).

In contrast, *non-standardized grading*—mostly referred to in the literature as *traditional grading* (Cain et al., 2022; Clemons, 2024; Koller, 2021)—describes a system in which teachers combine various forms of performance evidence, such as tests, quizzes, homework, participation, effort and attendance, into a single letter grade (Link & Guskey, 2022). This approach results in what has been described as hodgepodge grading (Brookhart, 1991; Cross & Frary, 1999), in which academic achievement is combined with other, non-academic factors, most often effort, behavior, and general attitude toward school (Brookhart et al., 2016).

Standardized and non-standardized grading systems are therefore not only distinct in terms of their assessment mechanism, but also in the achievement factors reflected in the resulting grades (Bowers, 2011). While standardized grading is based on the extent to which students

meet pre-defined academic performance criteria, non-standardized grading is based on a variety of academic and non-academic achievement factors to assess academic performance as well as to enhance student learning and motivation (Brookhart et al., 2016).

2.2 Differentiating Between Teacher Grades and Standardized Test Scores

A common research question in the field of teacher grading examines the consistency between traditional (non-standardized) teacher grades and externally assessed standardized test scores. For example, an early study by Carter (1952) found a correlation of $r = .52$ between teacher-assigned grades and standardized algebra test scores, controlling for intelligence in a sample of 235 high school students in Western Pennsylvania. A later comparison of high school GPA (usually non-standardized assessed grade point average) and standardized test scores by Pattinson et al. (2013) yielded similar results, with correlations of $r = .50$ for reading and $r = .58$ for Mathematics in a sample of more than 10,000 students across the U.S. Slightly higher correlations were reported by Duckworth et al. (2013), who found $r = .62 - .64$ for reading and Mathematics scores in relation to GPA in approximately 2,000 eighth- and ninth-grade students in New York City.

Although standards-based teacher grades are intended to measure the same construct as standardized test scores (i.e., academic achievement relative to predefined performance standards), their correlation with standardized test scores has been shown to be only moderate (e.g., Ross & Kostuch, 2011; Welsh et al., 2013), similar to that observed for traditional, non-standardized teacher grades. For example, Howley et al. (1999) found that standardized test scores accounted for about 50% of the variance in standards-based teacher grades, with the strength of this relationship varying across schools. Interviews revealed that some teachers in standards-based grading contexts still included non-academic achievement factors, such as attendance and participation, in their grades. As Brookhart et al. (2016) suggest, this practice may account for the only moderate correlation between standardized test scores and standards-based teacher grades. Moreover, due to these non-academic achievement factors, teacher grades (both standardized and non-standardized) tend to be higher than standardized measures (e.g., Hoffmann et al., 2022). This effect is particularly evident for girls (Gallagher, 1998; Guez et al., 2020) and low-achieving students (McMillan, 2003; Pilcher, 1994; Tierney et al., 2011).

The most important non-academic achievement factors considered by teachers include effort, participation, persistence, and behavior (Bowers, 2011; Sun & Cheng, 2014; Willingham et al., 2002). Teachers particularly rely on these factors for borderline students to prevent them from failing (Randall & Engelhard, 2010). Teachers' sense of fairness also plays a central role in grading decisions. Many teachers believe that students who demonstrate effort should not fail, even if they have not fully mastered the material (Bonner & Chen, 2021; Brookhart et al., 2016).

In summary, teacher grades differ from standardized test scores both in grade level and in the type of achievement they capture. Whereas standardized tests focus solely on academic achievement, teacher grades provide a multidimensional, holistic measure of students' overall abilities (Bowers, 2011).

2.3 Determining the Degree of Standardization of the Austrian Matura

According to the official designation by the Federal Ministry of Education (2025), the Austrian Matura is a standardized competency-oriented examination. While this title closely resembles what is referred to in the literature as competency-based (or standards-based) assessment, the grading of the Matura does not serve the purpose of providing feedback or supporting students' ongoing learning processes. Instead, Matura grades represent final results that, upon successful completion, qualify students for entry into all tertiary education institutions nationwide. It is the teachers, rather than external examiners, who assign the final grades. In line with Brookhart et al.'s (2016) terminology, the Matura can thus be classified as an exam based on standardized grading.

According to the Austrian Federal Institute for Educational Quality (BIFIE, 2013), the standardization of the Matura relies on three procedures to ensure comparability across subjects and years: task design based on subject didactics and test theory, expert panels determining difficulty with reference to the curriculum, and, where relevant, alignment with frameworks such as the CEFR. Each year, around 180 specially trained teachers create approximately 3,200 tasks, reviewed by 50 experts and piloted in schools, with the final selection being known to only two to four members from the Austrian Federal Institute for Educational Quality (Bayrhammer, 2024). One day before the written exams, tasks are delivered to schools in sealed envelopes, which principals open in classrooms on exam day to ensure simultaneous and transparent access for teachers and students.

However, the Matura is only partially standardized. It is built on three pillars (Bundesministerium für Bildung, 2025b): 1) standardized written examinations that are mandatory for all students in the subjects of German, Mathematics, and a foreign language (primarily English); depending on the school type, students must choose one or two additional subjects for their written examinations; 2) non-standardized oral examinations in which teachers develop the examination tasks themselves; 3) the submission of a "pre-scientific work," which is a research paper accompanied by a presentation in which students explore a chosen topic using scientific methods¹.

The Matura exam is prepared centrally by the Federal Ministry of Education (Bundeskanzleramt Österreich, 2025). Teachers are provided with standardized assessment criteria, including model solutions, to enhance the transparency and comparability of final grades (BIFIE, 2013). However, teachers grade their own students without anonymization and without any form of moderation by colleagues, teachers from other schools, or central examination boards.

3. Focus and Structure of the Dissertation

Because the practice of teachers grading their own students may compromise the comparability, credibility, and validity of Matura grades (Brown, 2024), this issue forms a central focus of the dissertation. Over the past decade, the Austrian Matura has undergone three major developmental stages, each altering the responsibilities of teachers in the grading process. The three studies forming my dissertation examined these stages, with each study primarily focusing on one stage while also providing insights into the others. Each study was guided by a research question that addressed both the corresponding developmental stage of reform and a specific gap in existing literature, drawing on each a distinct theoretical framework and contributing to the broader international research on teacher grading in general and in the context of university entrance qualifications in particular. Together, the three studies offer a promising foundation for further research on the Austrian Matura system.

The first research question corresponds to the 2015 Matura development stage, which introduced a standardized assessment system for academic upper secondary schools. For the first time in the history of the Matura, teachers were required to conduct two types of grading

¹ This pre-scientific work became optional in 2025, with students being allowed to opt for an additional exam instead.

simultaneously: non-standardized grading of students' ongoing classroom performance and standardized grading to assess the same students' readiness for university entrance qualifications. A substantial body of research has examined discrepancies between non-standardized teacher grades and standardized test scores, considering both their conceptual meaning (e.g., Willingham et al., 2002) and their statistical correlation (e.g., Duckworth et al., 2013). However, the existing literature largely focuses on contexts where standardized tests are scored by external bodies, excluding teachers from the grading process. To the best of my knowledge, no prior study has specifically examined a system in which the same teacher assigns both non-standardized report card grades (hereafter referred to as classroom grades) and standardized (Matura) grades, without anonymization. The Austrian Matura thus provides an ideal context to address this gap. Accordingly, the first overarching research question was as follows:

RQ 1: How do non-standardized classroom grades and standardized Matura grades relate to each other in a system where both are assigned by the same teacher?

The second research question corresponds to the extension of the standardized Matura system to vocational upper secondary schools in 2015/16, one year after its introduction in academic upper secondary schools. Evidence from Germany, which has a comparable dual-track school system, highlights the limited comparability of grades across school types (Köller et al., 1999; Watermann et al., 2013). Since the standardized Matura was introduced in vocational upper secondary schools to improve the comparability of final grades across school types, the second study examined differences in teacher grading between academic and vocational upper secondary schools, with the grade discrepancies identified in RQ1 serving as the primary outcome of interest. Thus, the second overarching research question of the dissertation was:

RQ 2: How do teachers' grading practices differ between academic and vocational upper secondary schools?

The third research question addressed the 2020 grading reform, which significantly changed the composition of final Matura grades and, with it, the teachers' responsibilities in their role as graders: they no longer assigned Matura grades solely based on standardized Matura exams but also influenced these exam grades through their non-standardized grading in the final school year. The third and, in my view, central study of the dissertation, therefore, investigated the effects of the 2020 grading reform on final classroom grades and changes in teachers' grade decision-making within this policy change. Drawing on literature on grade

inflation in recent years (e.g., Sanchez, 2023), the third study expanded existing knowledge on internal and external influences in teacher grading (McMillan, 2003) and on key sources of grader bias that threaten assessment credibility and validity (Brown, 2024), irrespective of whether grades are standardized or not. The third overarching research question was:

RQ 3: How have final classroom grades and teachers' grade decision-making changed since the 2020 grading reform?

4. Author Positionality

Before turning to the three studies that addressed these research questions, this section briefly outlines the perspective from which this dissertation was pursued. Research cannot be separated from the historical, political, and cultural lenses through which it is conducted (Staller, 2010). As Kuhn (1997, p. 113) has argued, "What a man sees depends both upon what he looks at and also upon what his previous visual-conceptual experience has taught him to see". In other words, every research project is shaped by implicit or explicit assumptions, perspectives, and frames of reference that are themselves embedded in broader social and historical contexts.

This dissertation was based on my implicit assumption that university entrance certificate grades should provide a comparable measure of academic performance among all students nationwide, regardless of their school type, class, teacher, or individual characteristics, to ensure a fair basis for university admission. Referring back to Kuhn (1997), I am uncertain whether this implicit assumption originated from my own personal beliefs or from the position promoted by the Austrian Ministry of Education of the Matura as a system that has to provide "fairness, the same conditions for all Matura students, and comparability of examination qualifications for higher education institutions" (Federal Ministry of Education, 2025). By explicitly acknowledging that this overarching policy goal informed my work on Austria's national assessment system, however, the dissertation recognizes its embeddedness in the current national policy rationale while also creating space to critically examine the extent to which these intended aims of Austria's standardized assessment system are realized in practice. Most important to me, my implicit assumptions became explicit and changed over the course of my continued engagement with my research topic; a development I will address again in the final section to maintain a chronological narrative.

5. From Policy to Practice: Three Studies on Austria's Developing University Entrance System

This section situates the three studies forming my dissertation within the political-educational context of Austria. As already briefly outlined, the studies are tracing key development stages in the Austrian Matura system, with each study addressing one of these stages by providing empirical insights into it. To further contextualize the findings, evidence from selected European countries that have implemented similar changes or have faced comparable challenges in teacher grading for university entrance qualifications is provided, allowing for a broader perspective.

Before turning to the first triad of i) Matura development stage, ii) brief summary of dissertation study, and iii) evidence from another European country, Table 1 provides an overview of key milestones in the historical evolution of the Matura, from its inception in 1849 to a comprehensive examination framework introduced in school year 1992/1993.

Table 1. Milestones in the Historical Development of the Austrian Matura

Year / Period	Key Development	Description
1849	Introduction of the Matura	Established through the <i>Organisationsstatut</i> (organizational statute) by Franz Serafin Exner and Hermann Bonitz; completion of an eight-year academic school (<i>Gymnasium</i>) ending with the <i>Maturitätsprüfung</i> (maturity examination).
1869	<i>Realschule</i> Reform	Reform introduced the Matura as the official final examination for <i>Realschulen</i> (less academically orientated upper secondary schools).
1878	Admission of Girls	Girls were admitted to sit for the Matura at <i>Gymnasien</i> ; women's higher education was gradually integrated into the secondary school system.
1908	Examination Reform	New regulations on examinations and grading; restructuring of the Matura.
1912	Reorganization of Girls' Schools	Six-year girls' schools were restructured into eight-year secondary schools modeled on boys' schools.
1924	Inclusion of a Written Thesis	The Matura included a written thesis (<i>Hausarbeit</i>) chosen by the candidate; oral exams were reduced to two.
1939–1945	World War II Period	Schools were aligned with militaristic objectives; many students left without completing the Matura, receiving a <i>Reifeklausel</i> (clause of academic maturity) instead.

Post-1945	Re-establishment	Postwar education policy returned to the 1935 curricula; Matura regulations were regularly adjusted, including special provisions for war-affected students.
1992/1993	Modern Examination Framework Introduced	New regulations introduced seven partial exams (written and oral) across at least four subjects. Mandatory written exams in German, a foreign language, and Mathematics were established. A <i>Fachbereichsarbeit</i> (subject-area thesis) replaced one written exam and formed part of the final oral examination.

Note. This table is developed based on the description of the historical evolution of the Matura in *Wien Geschichte Wiki* (2014).

5.1 Embedding the First Study: Non-Standardized Versus Standardized Teacher Grades

5.1.1 Matura Development Stage: Introduction of the Standardized Matura for Academic Schools (2014/15)

The next major step toward the current Matura system was the introduction of centralized, standardized written examinations in the core subjects of Mathematics, German, and English as a foreign language in the 2014/2015 academic year. This reform initially applied only to academic upper secondary schools, while vocational upper secondary schools were not yet included. Before this reform, teachers designed their own Matura examination tasks for their students. Since then, they receive centrally developed standardized tests (including model solutions) that are administered simultaneously nationwide. The centralization and standardization of the Matura aimed to ensure uniform competencies and equal conditions for all students, promote competence-oriented and objective assessment, enhance the comparability and transparency of results, strengthen the validity of final examinations, and support the cross-European comparability of qualifications within the European Qualifications Framework (EQF) (Bundesministerium für Bildung, 2025c).

While other European countries, such as Ireland (Lehane et al., 2025), Italy (Autonome Provinz Bozen Südtirol, 2025), and France (Ministère de l'Éducation nationale, 2025) assign the assessment of final grades to external bodies, Austria relies exclusively on the students' own teachers. Consequently, the standardization reform put Austrian teachers into a dual grader role: they conduct non-standardized, traditional grading when assessing students' classroom performance during the final school year and simultaneously evaluate the same students' performance on the standardized Matura examination—without anonymization or internal or external moderation.

5.1.2 About the First Dissertation Study

Many studies, particularly in the U.S., have examined the relationship between external test scores and either non-standardized (e.g., Brennan et al., 2001) or standardized (e.g., Welsh et al., 2013) teacher grades. However, I found no study that examined the relationship between non-standardized and standardized grades, especially when both are awarded by the same teacher. Austria provides a rare example of such a dual teacher grading system. Therefore, my first study longitudinally investigated the extent to which teachers' non-standardized traditional classroom grades correspond with their grading of the standardized Matura examination (RQ 1), and how this relationship varies across years, subjects, and school types. As the 2020 grading reform merged the two grades to form the overall Matura grade—each contributing 50%, whereas previously only the Matura exam grade determined the final result—I placed a particular focus on potential changes in that year (RQ 3). Specifically, I expected the Matura grades to align more closely with teachers' classroom grades than before the reform.

The theoretical background for my first study focuses on the reliability challenges of teacher grades compared to standardized test scores. Although teacher grades are highly reliable for ranking students within a classroom, they lack comparability across classrooms, schools, and federal states (Hübner et al., 2020; Neumann et al., 2009). Grading practices also vary according to students' abilities, with non-academic achievement factors influencing teacher grades more for low-performing students (McMillan, 2003; Pilcher, 1994; Tierney et al., 2011). In contrast, standardized examinations, particularly when supervised, are more resistant to collusion, cheating, and favoritism (Cheung, 2008). Objectively scored items show enhanced reliability, more consistent scoring, and reduced grader biases, partly because candidates are anonymous to the markers (Haertel, 2006; Heyneman, 1988). Empirical evidence further shows that test scores and classroom grades differ in achieved grade levels, with test scores being generally lower, especially for low-performing students (Arencibia Alemán et al., 2024; Hoffmann et al., 2022; Hübner et al., 2020; Neumann et al., 2011; Willingham et al., 2002).

The study analyzed 17,162 matched classroom and Matura grades from 2017 to 2023, collected from three academic and four vocational upper secondary schools in Vienna, Austria. Grades were awarded by 154 teachers across German, Mathematics, and English, covering the period before and after the 2020 grading reform. For each teacher and academic year, Cohen's d was calculated to quantify the standardized difference between classroom and

Matura grade distributions. An ANOVA was then conducted to examine the effects of year, subject, school type, and teacher gender on these discrepancies, as well as potential interaction effects.

Regarding the first overarching research question of my dissertation (RQ 1), one of the main findings of my first study is that, contrary to the large body of research demonstrating only moderate correlations between traditional grades and both standards-based teacher grades and standardized test scores, my Austrian data show a remarkably close alignment between traditional classroom grades and standardized Matura grades. This high level of alignment is particularly evident in the two language subjects German and English, and remains consistent across all years investigated (2017 to 2023). Notably, the 2020 grading reform, which formally merged the two grades, did not further strengthen this alignment. The only notable discrepancies between classroom and Matura grades occurred in Mathematics prior to 2019, until a special task force revised both the exam format and the grading scheme by simplifying text-based items, extending the time available for more complex tasks, and allowing the use of half points as well as a more lenient threshold for achieving a passing grade (Gsenger, 2019).

This result allows for several possible interpretations. It may indicate that teachers already base their classroom grading in the final school year on the standards-based criteria of the Matura, which could explain the high level of alignment observed. It could also suggest that providing an objective assessment of one's own students' pure academic performance is a particularly challenging task and that, despite standardized guidelines, non-academic achievement factors continue to play a role in Matura grades (Howley et al., 2000). However, the fact that this high alignment occurs primarily in the language subjects, as opposed to Mathematics, may also imply that in text-based assessments it is easier for teachers to consciously or unconsciously align grades with students' prior performance (Schröter et al., 2022b). My two subsequent studies provide additional insights into these interpretations, which will be discussed in detail in their respective sections.

My first study revealed two further findings. First, school type correlated significantly with the discrepancy between classroom and Matura grades (RQ 2): teachers from academic secondary schools tend to award significantly higher classroom grades relative to Matura grades, whereas teachers from vocational secondary schools award lower classroom grades relative to Matura grades. Second, the control variable of teacher gender also showed a

statistically significant effect, with male teachers awarding higher classroom grades in comparison to Matura grades than female teachers.

While the first effect of school type differences in the discrepancy between classroom and Matura grades was further examined in my second study and will be discussed there, I did not pursue the second (unexpected) finding concerning the significant effect of teacher gender. To better understand this result, I examined existing research on the potential influence of teacher gender on grading practices. Although extensive research has explored the effects of *students'* gender on teacher-awarded grades, standardized test scores, and the discrepancies between the two (e.g., Duckworth et al., 2015; Gallagher, 1998; Guez et al., 2020), studies examining the effect of *teacher* gender on grading practices remain scarce. A review article by Coenen et al. (2018) identified only eleven studies addressing this topic and highlighted two gender-related mechanisms in grading research: a) the “feminization” of the teaching profession, suggesting that a predominantly female school environment may advantage girls, and b) gender bias, implying that teachers may favor students of their own gender. However, studies from Sweden, Canada, Germany, and the Netherlands found no significant interaction between teacher and student gender, while results from India and the United States were mixed (Coenen et al., 2018). Further research in the U.S. (Ouazad, 2008) and Iran (Nowruzi et al., 2020) yielded partial evidence that teacher gender may influence grading patterns, with male teachers in some contexts awarding higher or more accurate grades compared to standardized test scores. More recently, Lievore and Triventi (2023) found that both male and female Mathematics teachers in Italy tended to assign higher grades to girls than to equally competent boys. The diversity of these studies, both in research focus and in their findings, suggests that teacher gender-related grading tendencies are complex and context-dependent, warranting further research to establish a more solid understanding of the potential effects of teacher gender on grading practices.

5.1.3 European Perspective: Evidence from Sweden

Another European example of a country that relies heavily on teacher grading for university entrance qualifications is Sweden. With a long-standing tradition of using teacher-awarded grades for higher education admission (Lundahl et al., 2017), Sweden, like Austria, implemented a standards-based grading system for this purpose, while keeping teachers responsible for the assessment itself. Introduced already in the mid-1990s, this standards-based system replaced the former norm-referenced approach to foster cooperation among students rather than competition (Gustafsson & Erickson, 2013).

In the current Swedish system, students must have successfully completed all courses of upper secondary school and are required to submit an individual project work to receive their *slutbetyg*, their higher education entrance certificate (Eurydice, 2024). From the very beginning of the introduction of standards-based assessments, teachers had to consult standardized national tests (*nationella prov*) for each subject provided by the Swedish National Agency for Education (*Skolverket*) to take them into account additionally to their awarded grades to ensure the comparability across teachers, classes and schools (Eurydice, 2025). However, no rules or guidelines on how these test results should impact decisions about individual students' grades were offered (Gustafsson & Erickson, 2013), and so it is in the discretion of the individual teacher how and to what extent the national test results are considered when determining the final grades (Dufaux, 2012). Researchers claim this grading mechanism as a “relatively loose strategy of taking test results into account” (Jönsson et al. 2021), and question its reliability (Swedish National Agency of Education, 2011).

Indeed, results from the first student cohort being graded by this mixture of classroom grades and test results revealed substantial differences between the grades achieved on the national tests and the final grades that the teachers awarded. In Mathematics, for example, 19% of students would have received a failing grade based on their test result, whereas only 5% received a final failing grade awarded by their teacher (Gustafsson & Erickson, 2013).

To determine interrater consistency on the standards-based teacher grades, the Swedish National Agency for Education explored 100 randomly chosen teacher graded tests by letting them being rescored by two or three independent raters (National Agency for Education, 2009). The results showed that teachers generally agreed on scores for English and Mathematics, but there was much less agreement on Swedish language tests. In grading essay tasks, for example, the interrater reliabilities varied between 0.86 and 0.93 for English as a foreign language, but only between 0.36 and 0.46 for Swedish.

An interesting Swedish study pursuing the question of how interrater reliability could be increased has been presented by Jönsson et al. (2021), by comparing an analytic assessment approach (grading of specific aspects of student work separately and combining them to an overall evaluation) with a holistic assessment approach (making an overall judgment of a student's work based on its general quality). The researchers report significantly higher interrater reliability in the analytic grading condition; however, this is only in the subject of

English. They further highlight that in both grading approaches (analytic and holistic), teachers only agreed in the rank order of the students, but the agreement concerning the exact grade was observed to be relatively low (52–67%) among all grading groups.

Despite these inconsistencies in reliability, Swedish teacher-awarded grades have been shown to be valid predictors of success in higher education. Cliffordson (2008) examined the predictive validity of university success in relation to both standards-based and norm-referenced teacher grades, as well as external test scores from the Swedish Scholastic Aptitude Test (SweSAT). The results indicated that teacher grades served as generally better predictors of university success than SweSAT scores, regardless of the grading system (standards-based and norm-referenced grading) used. In addition, the researchers found that teacher grades predicted university success more consistently across different fields of study, highlighting that the predictive strength of teacher grades did not vary much between disciplines. By contrast, SweSAT scores were less consistent predictors. Their ability to predict university success differed more strongly depending on the study program, showing that SweSAT scores were particularly poor predictors for students in technical and natural science programs. Comparing norm-referenced with standards-based teacher grades, it was found that the latter served as better predictors.

An interesting aspect of the Swedish system is that it offers the SweSAT as a second opportunity for students to gain university entrance. It is a paper-and-pencil, multiple-choice test designed and administered by the Swedish Council for Higher Education, used across all subjects to assess general abilities and knowledge (Stendlund et al., 2016). The test is administered twice a year, it is voluntary, and a test-taker can retake it as many times they wish (Cliffordson, 2004). Sweden's centralized admissions system requires that when a course or program has more applicants than spots, at least one-third of the students are admitted based on their SweSAT score, one-third on their upper secondary GPA, and up to one-third based on criteria set by individual tertiary education institutions (Stendlund et al., 2016). However, evidence indicates that this alternative route predominantly serves as a second chance for students with lower GPAs to access higher education entrance qualifications (Stendlund et al., 2016).

A very promising finding from Sweden for the Austrian context is that Matura grades, particularly since their standardization, may serve as good predictors of later success in higher

education—not despite, but because they are teacher-assigned.² Unfortunately, Austria does not provide a national test comparable to the SweSAT, which would allow an evaluation of the predictive accuracy of teacher-awarded Matura grades for university performance compared to standardized test results. However, it would be at least possible, similar to the approach taken by the Swedish National Agency for Education (2009), to investigate the extent of teacher rater bias in Matura results. This could be done by collecting a representative sample of already graded exam papers across different subjects and having them rescored anonymously by independent markers. By comparing the results across variables such as school type and teacher and student characteristics, future research within the Austrian context could dig deeper into identifying systematic grading patterns, potential sources of bias, and the extent to which teacher-assigned grades reflect comparable measures of students' academic achievement. Such research would offer valuable insights into the fairness, reliability, and validity of the Austrian Matura system, while also increasing teachers' awareness of unconscious bias in their grading practices.

5.2 Embedding the Second Study: School Type, Grading Practices, and Response Behavior

5.2.1 Matura Development Stage: Introduction of the Standardized Matura for Vocational Schools

A special feature of the Austrian secondary school system is the possibility for students to choose between attending an academic secondary school (Allgemeinbildende Höhere Schule, AHS) and a vocational secondary school (Berufsbildende Höhere Schule, BHS) to qualify for university entrance. Academic schools prepare students for university studies and end with the Matura graduation. Vocational schools focus on vocational training, but students also complete the Matura at the end of their schooling, thus having the same university entrance qualifications as academic school students. Vocational schooling lasts one year longer (5 years) than academic schooling does (4 years).

In the 2015/2016 school year, the standardized Matura was also implemented in vocational schools. However, this process proved to be considerably more complex than its earlier implementation in academic schools. While academic schools follow a largely uniform

² Similar evidence is also provided for lower tracks, showing that middle school teacher grades together with students' attendance are stronger predictors for high school performance than standardized test scores (Allensworth et al., 2014).

curriculum (Bundesministerium für Bildung, 2025a), the vocational school system includes multiple curricula across several school subtypes (Bundesministerium für Bildung, 2025e).³ For the standardized Mathematics Matura, this disciplinary diversity is managed so that all vocational school candidates are required to complete a common set of core tasks, regardless of their professional specialization. In addition, profession-specific sections are included to assess the particular knowledge relevant to each field (e.g., the application of mathematical principles to technical problem-solving in engineering-oriented programs).

Another distinction between the vocational school Matura and the academic school Matura is that in all vocational subtypes except the Higher Technical Secondary Vocational Schools (HTL) candidates may elect to take either Mathematics or English as the written examination, with the remaining subject being assessed orally.

This reflects a deliberate effort by the Austrian education system to accommodate diverse educational priorities and thereby foster curricular diversity, while simultaneously trying to ensure comparability of qualifications at the end of upper secondary school. My second dissertation study offers insights, among other findings, into how this balancing act is reflected in practice.

Before proceeding to the discussion of my second dissertation study, a few points should be noted regarding the educational background of students in both academic and vocational secondary schools: While many students in academic secondary schools have already followed an academic track in lower secondary education, where entry requirements in subjects such as German and Mathematics are needed, many students in vocational secondary schools come from non-academic lower secondary schools (middle schools). Specifically, 62% of students from lower academic tracks proceed to upper secondary AHS, whereas 33% enter BHS. In contrast, only 9% of non-academic lower secondary school graduates attend upper secondary academic schools, while 35% enroll in upper secondary vocational schools (BMBWF, 2021b). Overall, 61% of upper secondary students attend AHS and 39% attend

³ These subtypes of vocational schools include (a) Higher Technical Secondary Vocational Schools (Höhere Technische Lehranstalt, HTL), (b) Commercial Secondary Vocational Schools (Handelsakademie, HAK), (c) Secondary Vocational Schools for Human Services (Humanberufliche Schule, HUM; e.g., tourism, fashion, or social management), (d) Secondary Vocational Schools for Agriculture and Forestry (Höhere Lehranstalt für Landwirtschaft und Forstwirtschaft, HLFS), and (e) Secondary Vocational Schools for Early Childhood and Social Education (Bildungsanstalt für Elementarpädagogik / Bildungsanstalt für Sozialpädagogik, BAfEP/BASOP).

BHS (Statistik Austria, 2024b). These figures illustrate that the process of student selection is already in place before the start of upper secondary education.

5.2.2 About the Second Dissertation Study

The initial aim of my second study was to gain a deeper understanding of the grading differences between academic and vocational schools observed in the first study (RQ 2). Guided by the theoretical framework of my dissertation, the study examined whether teachers' grading beliefs regarding non-academic achievement factors (Bonner & Chen, 2021) could account for the observed discrepancies between non-standardized classroom grades and standardized Matura grades (e.g., Doz, 2023). Additionally, building on research highlighting the influence of systemic factors on individual teacher beliefs, a second aim was to investigate whether shared, collective beliefs within school types might serve as predictors of these grading differences (Fuller & Izu, 1986).

In this second study, however, I recognized that quantitative research processes do not always follow a straightforward, linear path. To illustrate this, a brief description of the data is necessary: My collected data comprised two components: first, the 17,162 matched classroom and Matura grades from 2017 to 2023 awarded by 154 teachers, which had already been used in the first study; and second, questionnaire responses from 56 of these 154 teachers, capturing their beliefs about non-academic achievement factors in grading. I matched these two data sources using an anonymized identification key.

Given the sensitivity of the topic of teacher grading, I conducted a robustness check for potential self-selection bias. I examined whether the 56 teachers who completed the questionnaire formed a representative subsample of the full dataset with respect to the variables of interest (DV: Cohen's d as the standardized discrepancy between classroom and Matura grades, drawn from the first study; IVs: year, school type, subject, and, most importantly, the interaction between school type and response status). This analysis examined (a) whether teachers who responded to the questionnaire differed in their grading behavior from nonrespondents and (b) whether school type moderated this relationship.

The result revealed a statistically significant interaction between school type and response status, indicating the presence of self-selection bias among participating teachers. The effect size of this interaction was the second largest in the model, suggesting that nonresponse was not randomly distributed with respect to the study's research question (grading differences among school types).

Due to this unexpected yet highly interesting research finding, and following an editor's suggestion, I subsequently developed an almost entirely new, methodologically focused manuscript that incorporated additional (qualitative) interview data collection and analysis. The revised study explicitly aimed to explore the mechanisms underlying the identified self-selection bias in the sensitive research context of teacher grading and its moderation by school type.

The final research design for my second paper employs a sequential explanatory mixed-methods approach (Ivankova et al., 2006). The quantitative phase documented the unexpected finding of self-selection bias, demonstrating that those teachers who completed the questionnaire on grading beliefs differed significantly in their grading behavior from nonrespondents. In the qualitative phase, I conducted interviews with 14 teachers representing the target group of nonrespondents to explore their perspectives on the school type moderated grading practices and response patterns identified in the quantitative analysis.

A closer look at the response patterns indicated that the mechanisms underlying self-selection bias varied not only across school types but also across vocational school subtypes and subjects: the effect was driven exclusively by Mathematics teachers in academic and technically oriented vocational schools. The interviews revealed that in technically oriented vocational schools, nonresponse was largely driven by Mathematics teachers' perceived lack of interest in the research topic (Saleh & Bista, 2017). They reported that the Matura is relatively easy for their students since its standardization, which diminished their motivation and perceived relevance of participating in a survey on the topic. Nonresponse was further associated with a pragmatic, practice-oriented professional mindset, which illustrates how teachers' attitudes and professional orientations can shape their willingness to engage in questionnaire-based research.

In academic schools, nearly all interview participants (including those from vocational schools) suggested that some nonresponding Mathematics teachers may have strategically awarded higher classroom grades to offset anticipated low exam performance, an opportunity opened by the 2020 grading reform. Because such strategically lenient grading was perceived as unprofessional and potentially stigmatizing, participants indicated that this may have reduced nonrespondents' willingness to disclose their grading beliefs. Among academic school teachers in particular, grading emerged as a highly sensitive issue, with nonresponse

associated with feelings of shame, concerns about privacy, reputational risk, and social desirability (Korkeila et al., 2001; Saleh & Bista, 2017).

From a methodological perspective, my findings highlight the importance of accounting for both topic sensitivity (Tourangeau et al., 2010) and professional context (Guillemot, 2025) when examining self-selection bias in educational research. They also demonstrate that such bias may remain undetected when analyses focus solely on differences between respondents and nonrespondents, as the self-selection bias only became apparent in the robustness check ANOVA after considering the moderating effect of school type.

Consistent with an incidental finding from the first study, the self-selection bias was driven exclusively by Mathematics teachers, whereas language teachers showed no such effect. Interview participants in both my second and third study suggested that aligning classroom and Matura grades may be easier in text-based subjects than in Mathematics, making strategic grading behavior more relevant and “attractive” for Mathematics teachers, who have less control over grading outcomes when assessments are standardized than, for instance, German teachers. These findings highlight both the higher stakes and the greater reluctance to participate among Mathematics teachers in academic schools, as evidenced by the second study. While the first study did not reveal any bias related to the 2020 reform (RQ 3), the second study tentatively suggested its possible presence (as it was indirectly linked to teachers’ responses to the survey), which was subsequently examined in greater depth in the third study.

Hence, the second study yielded insights beyond its original focus, providing a deeper understanding not only of grade differences but also of differences in both grading (RQ 2) and response behavior across school types. Moreover, the final version of my second study made an unanticipated contribution to the still limited literature on self-selection bias in questionnaire-based research on sensitive topics in education, which is otherwise often methodologically complex and challenging to investigate (Olsen, 2006).

5.2.3 European Perspective: Evidence from Germany

Germany provides another interesting example of the use of teacher-assigned grades for university entrance certification, particularly regarding the comparability of entrance exam grades across school types. Similar to the Swedish and Austrian way since 2020, it combines two distinct mechanisms of teacher grading to determine final grades: coursework grades from the last two years of school and scores from teacher-graded standardized national tests,

known as the *Abitur* examinations (*Abiturprüfungen*). Since 2017, Abitur task pools have been available for the core subjects German, English, French, and Mathematics, and from the 2025 examination year onward, these task pools also include Biology, Chemistry, and Physics (European Commission, 2024). The responsibility for awarding coursework grades and Abitur examination scores lies within the students' own teachers, but unlike the Swedish way, very strict guidelines are set by the Standing Conference of the Ministers of Education and Cultural Affairs (*Kultusministerkonferenz*, short *KMK*) for calculating final grades (Abitur Berlin Brandenburg, 2024).

Similar to Austria, the German upper secondary system provides both academic upper secondary schools (*Allgemeine Gymnasien*) and vocational upper secondary schools (*Berufliche Gymnasien*), with the latter offering students the opportunity to gain Abitur (the German pendant to the Austrian Matura) and additional vocational orientation, thus preparing them for higher education and the world of work (European Commission, 2024).

Additionally, many comprehensive schools allow students to pursue various qualifications, from basic school-leaving certificates to Abitur. Despite specific criteria for grading the Abitur exams, such as predefined performance criteria, correction guidelines, and evaluation sheets (European Commission, 2024), huge differences in Abitur grades have been observed between federal states, school types, and individual schools (Hübner et al., 2020; Neumann et al., 2009).

To shed light on possible mechanisms behind these discrepancies, Neumann et al. (2009) examined discrepancies between teacher-assigned coursework grades and externally assessed test scores in Mathematics (TIMSS) and English (TOEFL) across two German states. While English grades were mostly consistent, clear differences emerged in Mathematics: despite significant variations in external test performance, classroom grades were nearly identical across states. Since classroom grades heavily influence final Abitur grades, students with similar performance could receive different outcomes depending on their state. These differences were largely attributed to reference group effects. This effect was observed in both classroom grades and teacher-awarded standardized Abitur exam scores. Nonetheless, the standardized Abitur exams were found to be more closely aligned with external test scores than classroom grades (Neumann et al., 2009).

This result has been confirmed by a later study, with researchers concluding that standards-based teacher assessments provide a fairer and more objective evaluation of student

performance compared to classroom grades (Neumann et al., 2011). Nevertheless, research indicates that combining standardized Abitur exams with classroom grades offers a more comprehensive measure of student achievement, as this combination better predicts university success than either form of grading alone (Köller & Baumert, 2002; Nagy, 2007).

Further studies in Germany explored grading differences across various types of upper secondary schools. Köller et al. (1999) investigated the comparability of teacher grades between general upper secondary schools and upper secondary comprehensive schools. Their findings in Mathematics indicated a clear performance advantage on external tests for students from general upper secondary schools, which, however, did not correspond to differences in teacher classroom grades. Watermann et al. (2013) analyzed the comparability of classroom grades in Mathematics between general upper secondary schools and different vocational upper secondary schools in *Baden-Württemberg*. Although significant differences in Mathematics test performance were observed between general and vocational schools (specializing, for example, in economics, nutritional sciences, agricultural sciences, and social pedagogy), these differences were not reflected in teacher-awarded classroom grades.

For the Austrian context, Neumann et al.'s (2011) finding that standards-based teacher grades provide a more objective assessment than traditional classroom grades is particularly relevant, as it underscores the value of standardized Matura grades in addition to classroom assessments. Moreover, evidence that a combination of classroom and standardized grades predicts university success better than either type of grading alone (Köller & Baumert, 2002; Nagy, 2007) supports the rationale of Austria's current assessment model introduced with the 2020 grading reform.

At the same time, the evidence of grade discrepancies across school types highlights the need for similar research in Austria to evaluate the national comparability of final Matura grades. In particular, such research should focus on reference group effects, as emphasized in the German findings. Furthermore, as my second study demonstrates, vocational schools should be examined by subtype, as technically oriented schools appear to have students with higher Mathematics skills compared to non-technical vocational schools.

5.3 Embedding the Third Study: The 2020 Grading Reform and Teachers Grade-Decision Making

5.3.1 Matura Development Stage: 2020 Grading Reform: Integrating Classroom Grades into Standardized Matura Grades

The 2019/2020 academic year, like many other areas of life, was profoundly shaped by the extraordinary circumstances of the COVID-19 pandemic (Daumiller et al., 2023). To compensate for the loss of classroom instruction caused by school closures in the weeks preceding the Matura, the Austrian Ministry of Education introduced several exceptional measures for graduating students. The oral Matura examination became optional, the written component was limited to three subjects, the time allowed for each written exam was extended from four to six hours (RIS, 2021), and the external chairperson (usually a principal from another school) was removed as a mandatory member of the examination committee (Mychalewicz, 2024).

Most relevant to this dissertation is the 2020 grading reform, which will be described here in greater detail due to its central role in my third study: Starting in 2020, final Matura grades were no longer determined solely by exam performance. Instead, 50% of the final grade was derived from students' final-year classroom grade in the corresponding subject (RIS, 2021). Theoretically, this reform merged two fundamentally different grading approaches—non-standardized classroom grading and standardized exam grading—thereby challenging the original purpose of the standardized Matura. Practically, this meant, for instance, that a student with a classroom grade of 2 (good) and a Matura exam grade of 4 (sufficient) received a final Matura grade of 3 (satisfactory). In cases of a one-grade difference, the exam results prevailed. However, the reform also ensured that students with at least a classroom grade of 3 could no longer fail the Matura, as the inclusion of the classroom grade (3) prevents a final grade of 5 (leading automatically to a final Matura grade of 4). The reform also altered teachers' responsibilities as graders. By incorporating classroom grades into the calculation of final Matura grades, it made teachers' ongoing, non-standardized assessments directly influence standardized outcomes for the university entrance qualification.

While most of the exceptional measures for Matura candidates were withdrawn in subsequent years, the integration of classroom grades into final Matura grades remained. This has raised the question of whether the reform was truly intended as a temporary response to the pandemic (Austria Presse Agentur Science, 2021) or whether the extraordinary circumstances

of 2020 were leveraged to implement a lasting change to the grading system. Media coverage indicates that reforming, or even abolishing, the Matura has long been a contentious issue among political parties and educational stakeholders (Nimmervoll, 2024). Critics characterize the Matura as a relic of the past (Bischof, 2023) that promotes stressful, exam-focused teaching at the expense of future-oriented learning (Erdost & Cavallar, 2023), whereas conservative voices caution against a potential downward leveling of academic standards (Bischof, 2023). Furthermore, the importance of preserving the value of the Matura is emphasized (Mychalewicz, 2024) to ensure that it continues to serve its function as an entry qualification for tertiary education, a role that is increasingly being undermined by the growing number of entrance examinations conducted by universities and other tertiary education institutions for various fields of study (Kroisleitner & Wiesinger, 2025).

5.3.2 About the Third Dissertation Study

In my third study, I explored the third research question by examining teachers' roles as graders within the context of the 2020 grading reform, focusing on how the reform influenced their classroom grade decision-making. Recent research identifies the COVID-19 pandemic as a major catalyst for grade inflation (Berne, 2023; McManus et al., 2021; Sanchez, 2023; Tillinghast et al., 2023), showing that teacher-awarded grades rose substantially during and after the pandemic, while their correlation with standardized test scores declined (Goldhaber & Goodman, 2024).

In response to the disruptions caused by school closures and remote learning, many countries implemented more lenient grading policies in 2020 to prevent students from being disadvantaged (Castro et al., 2020; Chan, 2022; McManus et al., 2021). These measures often compelled teachers to pass students who would otherwise have failed, which, in turn, contributed to lower student motivation, increased absenteeism, and higher rates of academic dishonesty (Berne, 2023; Bowden et al., 2023). Moreover, teachers adapted their grading practices by emphasizing participation and task completion over mastery, resulting in inconsistencies in classroom grades, as teachers varied in the extent to which they required mastery, and it raised concerns about fairness and comparability (Doyle et al., 2021; Koller, 2021; McManus et al., 2021).

Based on McMillan's (2003) grade decision-making model, which conceptualizes grading as a process shaped by the interplay between teachers' beliefs, classroom realities, and external pressures, I particularly focused on changes in teachers' classroom grade decision-making.

Due to the findings from previous literature, I expected both the COVID-19 pandemic and the possibility of classroom grades compensating for low Matura grades to be catalysts for classroom grade inflation.

Using a sequential explanatory mixed-methods design (Ivankova et al., 2006), my third paper comprised two studies to examine changes in classroom grading following the 2020 grading reform. The first study conducted a quantitative longitudinal analysis of classroom grade developments over time, using chi-square tests and ordinal logistic regression. The second study employed qualitative methods (semi-structured interviews with 14 German and Mathematics teachers and a manifest qualitative content analysis following Graneheim & Lundman, 2004) to explore teachers' perceptions of these changes and potential shifts in their grading practices since 2020. Guided by McMillan's (2003) grade decision-making model, this second study also investigated teachers' views on the internal and external factors that have influenced their grading practices in the post-reform period.

Consistent with findings from other countries, the results point to a statistically significant increase in higher and a decline in lower classroom grades since 2020 in Mathematics and since 2021 in German, a trend that has persisted in subsequent years. The results suggest that the early reform years (2020–2021) represented a “transition period” shaped by both the disruptions of the pandemic and the introduction of the 2020 grading reform. Many teachers attributed the lasting leniency in grading to a broader “letting-everyone-pass” mentality, similar to California's “hold harmless” policy (Castro, 2021). While some welcomed the reform for reducing exam anxiety and promoting a more holistic assessment, others feared that it undermined the objectivity and credibility of the Matura.

One particularly striking metaphor raised by a teacher compared the current Matura to a “Potemkin village”: an assessment system that creates only an illusion of comparability and fairness while concealing profound inconsistencies in grading standards. This metaphor highlights the risk that the Matura, once conceived as a standardized measure of mastery, now serves a more symbolic function, signaling educational quality while shielding the state from accountability for actual learning outcomes (Benveniste, 2002).

Furthermore, the findings from my third study offer a deeper understanding of the mechanisms that shape teachers' grading decisions, particularly in the face of challenges introduced by the 2020 reform. At the systemic level, moral appeals to be lenient and the erosion of centralized standards were perceived to exert influence; at the interpersonal level,

pressures from students, parents, principals, and colleagues were described as affecting grading decisions in both subtle and explicit ways; and at the individual level, teachers reported to experience internal tensions as they navigate the balance between empathy, fairness, and their professional self-concept as supportive educators.

Over time, several of these mechanisms were perceived to have intensified, most notably the growing tendency toward grade leniency, teachers' increasing inner conflicts between maintaining positive relationships with students and fulfilling their role of a fair and consistent grader, and principals' greater "student-friendliness", which at times was seen to undermine teachers' professional judgment. Particularly following the 2020 grading reform, new student-teacher dynamics emerged, such as increased pressure and negotiation with students over final grades. While some teachers adapted their grading practices in response, others sought to preserve performance-based standards despite the changing atmosphere.

Ultimately, my third study underscores that grading is not merely a technical act, but a socially embedded process shaped by systemic pressures, interpersonal dynamics, and teachers' internal conflicts (Fjørtoft & Morud, 2021). Without addressing these contextual realities, reforms such as Austria's 2020 grading reform risk eroding the perceived legitimacy of assessment systems (Brown, 2024).

5.3.3 European Perspective: Evidence from Ireland

Not an identical, though similar, reform was implemented in Ireland. During the COVID-19 pandemic, responsibility for grading the Irish Leaving Certificate (LC) shifted from a team of specially trained external examiners (Schröter et al., 2022a) to the teachers who taught the students in class. Prior to this policy change, the LC was known for its "brutal fairness" (Schröter et al., 2022a); however, this reputation changed following the reform.

Doyle et al. (2021) provided important evidence showing that, as in Austria, the policy shift led to significantly higher LC grades compared to previous years, with students receiving, on average, one level higher grades.

Moreover, the study found that even within the same school and subject, teachers differed substantially in their grading, particularly in their use of achievement evidence to determine LC grades. Teachers also considered various non-academic achievement factors when assigning final grades. For example, 88% of teachers in Doyle's (2021) study reported using in-class formative assessments, such as observation, questioning, discussions, and self-assessments, to estimate final LC grades. Over 80% considered student engagement in class

important, and 56% also factored in engagement with learning outside of class (e.g., homework completion). Consequently, both the standardized nature and the fairness of LC grades after the reform were questioned by researchers and teachers alike.

Furthermore, allocation of study places continued to be based on the total points achieved on the final LC assessments. However, after grading responsibility shifted from the external examination board to the teachers, students' LC scores exceeded the number of available study places, leading to the need for the education system to provide thousands of additional new higher education study places (Doyle et al., 2021). This illustrates how changes within the upper secondary school system can have direct and far-reaching effects on the tertiary education system and, of course, on students' educational trajectories.

Besides offering insights into the effects of the "Irish 2020 grading reform" (CG2020, short for *Calculated Grades*; Doyle et al., 2025), Ireland provides a rich body of research examining various aspects and developmental stages of the national LC system (Lehane et al., 2025). A seminal study by Smyth et al. (2019) evaluated the strengths and limitations of upper secondary (*senior cycle*) education and assessment from the perspectives of students, teachers, and parents. The authors identified major challenges such as heavy workloads, exam-focused learning, and high pressure surrounding the LC, which limit extracurricular and social engagement. All three stakeholder groups strongly supported reform, emphasizing the need for more flexible, varied, and subject-sensitive assessment methods to be implemented throughout upper secondary education. However, Smyth et al. (2019) caution that reform must also consider the LC's continued role as a selection mechanism for tertiary education, as changes to assessment alone may not alleviate systemic pressures associated with upper secondary schooling.

Research on social justice and student well-being further challenges the perception of the LC as a fair and neutral system. Studies show that LC outcomes are influenced by students' social, economic, and cultural backgrounds, risking reproducing class advantages (Canny & Hamilton, 2018) and favoring those with access to private tuition (Moore, 2014). Other research highlights negative effects on student wellbeing: Burke and Minton (2019) found a steady decline in wellbeing throughout secondary school, and Banks and Smyth (2015) identified LC workload as a major stressor, particularly for high-achieving girls.

Despite these concerns, both teachers and parents appreciate certain strengths of the Irish assessment system, particularly the wide range of subject choices (students typically take six or more subjects for the LC) and the perceived objectivity ensured by external examination bodies (Lehane et al., 2025).

Considering both the positive and negative consequences of a system that embodies many of the goals pursued by the standardized Austrian Matura (ensuring high comparability of final results as a fair basis for university admission), it becomes evident that Ireland's model, which initially achieved these aims before shifting responsibility to teachers, is not without flaws. This highlights an important consideration for Austrian policymakers when contemplating future changes or reforms to the national assessment system.

The most relevant evidence from Irish research for my dissertation on teacher grading, however, comes from Doyle et al. (2025), who examined teachers' feelings, beliefs, struggles, and concerns regarding their new role as assessors within the CG2020 system. While Austrian teachers are already accustomed to grading their own students for high-stakes assessments, this responsibility was entirely new for Irish teachers. Consequently, Irish teachers' conscious reflections on their experience as graders provide valuable insights into teacher grading in Austria, where such responsibilities may be carried out more unconsciously. The Irish case offers a unique opportunity to compare professional identity before and after assuming this role.

In reflecting on this transition, teachers identified several positive aspects of their new responsibilities, including the heightened trust placed in their professional judgment, the reduction of pressures associated with external examinations, opportunities for reflection on their own professional identity, and increased collaboration with colleagues (Doyle et al., 2025). At the same time, teachers experienced significant tension and conflict. Many felt uncomfortable or guilty about judging and ranking their own students, as this challenged their established identity as advocates and supporters rather than assessors. This shift disrupted the student-teacher relationship, which teachers view as collaborative and grounded in trust and care. Adopting the dual role of mentor and examiner altered classroom dynamics and atmosphere, creating unease and prompting greater self-censorship in interactions with students.

Teachers' strong commitment to advocacy often led them to adopt more lenient grading practices, particularly for students from disadvantaged backgrounds. However, these same

impulses heightened concerns about bias and fairness. The shift also affected relationships among colleagues: teachers reported frustration with peers who did not adhere to official grading guidelines and inflated grades to support their students. Those who carefully followed the rules often felt their efforts were undermined and rendered meaningless, further adding to their stress.

As the findings of my third study align with these results, they underscore the need for further research on how assessment systems impact classroom climate, student–teacher relationships, and teacher–teacher relationships. Because these factors are important predictors of both student success (e.g., Kwong & Davis, 2015; Reyes et al., 2012) and the well-being of students (e.g., Van Petegem et al., 2008) and teachers (Farhah et al., 2021; Spilt et al., 2011), the interaction between grading practices and relational dynamics in schools warrants greater attention—particularly in systems, like Austria’s, that rely heavily on teachers as graders in high-stakes assessment contexts.

6. The Three Studies Forming my Dissertation

This section comprises the three studies that constitute my dissertation. To help the reader maintain an overall picture of the thesis as a whole, only an extended summary of each study is presented here, while the full-text versions are provided in the Appendix following my concluding thoughts in the subsequent section.

6.1 First Study – Extended Summary: Teacher Grades for University Entrance Qualification: Bias, But Not as We Expected

Vincent Schatz-Laliberté, Roman Zviagintsev, Nele Kampa, Gavin TL Brown

Deciding who should be granted access to higher education is a pivotal point in the transition from school to university. Ideally, final grades used for university admission should reflect valid, merit-based indicators of academic performance. However, countries differ widely in the methods they use to determine these grades (Hansen, 2011). The two main approaches are relying on teacher grades or standardized examination scores (Brookhart et al., 2016). While standardized exam scores offer comparability (Haertel, 2006; Heyneman, 1988), teacher-awarded grades often capture a broader spectrum of student abilities (Bowers, 2011; Johnson, 2013; Willingham et al., 2002) and can serve as strong predictors of later academic success (Cliffordson, 2008). Teacher-awarded grades tend to be higher than standardized scores (Hoffmann et al., 2022; Hübner et al., 2020), and numerous studies have shown that

the two are only moderately correlated (approximately $r = .5$; Brookhart et al., 2016). However, little is known about how teacher-awarded grades relate to standardized test scores when the same teacher assigns *both*, and they are not anonymized. Austria provides a unique context for addressing this research gap: unlike many other countries, the same teachers assign both students' final-year report card grades, reflecting their ongoing classroom performance (hereafter referred to as classroom grades), and the standardized exam scores on the Matura (hereafter referred to as Matura grades), the country's university entrance examination. This happens simultaneously at the end of upper secondary schooling. Moreover, in 2020, a grading reform permanently integrated final-year classroom grades into the calculation of final Matura grades, thus fundamentally altering teachers' stakes in classroom grading. Hence, this study examines how teachers' classroom grades relate to standardized Matura grades in general, and specifically before and after this reform. The analysis addresses four questions: (1) the overall alignment between classroom and Matura grades; (2) subject-specific differences in German, Mathematics, and English; (3) the effect of the 2020 grading reform on grade alignment; and (4) differences between academic and vocational upper secondary school types.

The study draws on a longitudinal dataset of 17,162 grades, consisting of 8,581 matched grade pairs of final-year classroom grades and corresponding Matura grades in German, Mathematics, and English. Data were collected from three academic and four vocational upper secondary schools in Vienna between 2017 and 2023. Grades were awarded by 154 teachers, most of whom graded students across multiple years. To quantify the alignment between classroom and Matura grades, Cohen's d , as a standardized grade discrepancy, was calculated for each teacher and academic year, comparing the ordinal grade distributions of classroom and Matura grades. Grades were inverted so that higher values corresponded to better performance. Positive effect sizes indicate higher (better) classroom grades compared to standardized Matura grades; negative values indicate lower classroom grading relative to Matura grades. These teacher-level annual effect sizes served as the dependent variable in an ANOVA examining the main effects of year, subject, school type, and teacher gender, including interaction terms. Post-hoc Tukey tests were used to explore significant effects further. To assess robustness, a sensitivity analysis was conducted with a subset of 39 teachers who provided grade pairs in at least four separate years.

Across all subjects and years, classroom and Matura grades showed a very high degree of alignment. In 56% of cases, the two grades were identical; differences of one level occurred

for about 41%, while discrepancies of two or more levels were rare. Effect sizes were generally small ($d < .20$), indicating limited grade differences when the same teacher grades both assessments. However, subject had a significant effect on grade discrepancy. Some fluctuations occurred in Mathematics, particularly between 2017 and 2018, where effect sizes shifted from negative (stricter classroom grading) to strongly positive (more lenient classroom grading). These swings likely reflect widely reported issues with variability in Mathematics exam difficulty and subsequent adjustments to grading procedures in 2019 (Neuhauser, 2019). In contrast, German and English displayed stable alignment patterns across all years. The high degree of alignment was not even strengthened by the 2020 grading reform, which integrated classroom grades into final Matura grades. The narrowing of grade differences over time was largely driven by the early fluctuations in Mathematics rather than by the reform itself. Moreover, the effects of gender and school type remained stable before and after 2020. Unexpectedly, teachers at academic schools significantly awarded higher classroom grades relative to Matura grades compared to teachers at vocational schools. This pattern persisted across all subjects and nearly all seven years investigated. The gap between school types narrowed somewhat after 2020, but without a statistically meaningful reform effect. Teachers' gender also influenced grading differences: male teachers tended to award statistically significantly higher classroom grades relative to Matura grades than female teachers. However, in the sensitivity analysis with teachers grading multiple years, the gender effect weakened ($p = .07$), suggesting caution in interpretation.

The high alignment between classroom and Matura grades suggests that when teachers assess both, they may consciously or unconsciously harmonize their judgments. This contrasts with research comparing classroom grades to externally marked tests, which typically reveals larger discrepancies (e.g., Arencibia Alemán, 2024). These findings raise important questions about teacher impartiality in high-stakes contexts (Doyle et al, 2021) and whether measures such as anonymization or external moderation could better support the Matura's original goal of standardization and comparability of results. The persistent differences between school types point to structural grading disparities between academic and vocational schools (Watermann et al., 2013) and underscore the need for further investigation into grading practices and cultures within each track.

In a system where the same teacher assigns both (non-standardized) classroom and standardized exam grades, the two grades are highly aligned—a pattern not even strengthened by a reform that merged the two grading systems. Nevertheless, the study identified

unexpected biases, including significant differences in grading between academic and vocational schools, as well as potential effects of teacher gender. Given the goal of standardized assessments at the end of upper secondary education to ensure comparable results and fair university admission, these findings merit careful consideration by teachers, researchers, and policymakers alike.

6.2 Second Study – Extended Summary: Self-Selection Bias in the Sensitive Topic of Teacher Grading: The Role of Subject and School Context

Vincent Schatz-Laliberté, Roman Zviagintsev, Nele Kampa

Self-selection bias in voluntary questionnaire surveys presents a well-documented threat to research validity (Bowden, 1986; Groves & Peytcheva, 2008), particularly when studies address sensitive topics. While nonresponse in general can be influenced by factors such as personality traits, survey length, and structural design (Korkeila et al., 2001; Saleh & Bista, 2017), self-selection bias arises when the likelihood of participation correlates with the topic itself (Tourangeau et al., 2010). In educational research, evidence of self-selection bias is scarce (Guillemot et al., 2025), despite high nonresponse rates among adult participants, such as teachers and university students (Meinck et al., 2017). The present study reports a case of self-selection bias discovered incidentally in research on teacher grading—a sensitive area due to potential intrusiveness, risk of personal harm, and social desirability concerns (Krumpal, 2013; Peytcheva & Yan, 2025). The study contributes methodologically by illustrating the mechanisms of self-selection bias in educational survey research, with a focus on teachers' school type as an institutional context.

The study employed a sequential explanatory mixed-methods design (Ivankova et al., 2006). The quantitative part of the study analyzed self-selection bias in a questionnaire on teachers' grading beliefs, which were linked to the same teachers' awarded classroom grades (i.e., final report card grades) and standardized exam grades on the Matura, Austria's national university entrance qualification. Grades were collected from 154 teachers across three academic and four vocational upper secondary schools in Vienna, covering 8,581 students and seven academic years (2017–2023) in German, Mathematics, and English. While grades were collected directly from schools, participation in the questionnaire was voluntary, yielding a 36% response rate (48% for academic schools and 30% for vocational schools). To test for self-selection bias, we created a binary response variable (missing vs. not missing) and

conducted an ANOVA with Cohen's d as the dependent variable and year, school type, subject, and the interaction between school type and response status as predictors. This analysis assessed whether respondents differed from nonrespondents in their grading (i.e., grade discrepancy between classroom and Matura exam grade) and whether school type moderated the association between response status and grade discrepancy (Cohen's d). The subsequent qualitative phase of the study employed semi-structured interviews with 14 teachers selected via purposive sampling. Seven participants were nonrespondents from the quantitative phase, while the other seven represented subgroups contributing most to the observed bias (Mathematics teachers in academic and technically oriented vocational schools). Interviews focused on both explanations for grade discrepancies and reasons for nonresponse. The qualitative data were analyzed using manifest qualitative content analysis (Graneheim & Lundman, 2004), with a focus on the explicit explanations provided by participants.

ANOVA revealed significant main effects of year and school type on Cohen's d , as well as a significant interaction between school type and response status, indicating the presence of self-selection bias. Responding teachers across school types exhibited similar grading patterns, with classroom and Matura grades being closely aligned (Cohen's $d < .20$). Nonrespondents, however, displayed significant discrepancies: nonrespondents from academic schools tended to award higher classroom grades than Matura exam grades, whereas nonrespondents from vocational schools, particularly those in technically oriented schools, awarded higher Matura exam grades than classroom grades. Mathematics teachers largely drove the effect, while German and English teachers showed minimal differences. Response rates were also lower in technical vocational schools compared to non-technical vocational schools and academic schools.

In academic schools, teachers attributed higher classroom grades to sustained student effort, participation, and opportunities to improve throughout the year. Teachers' primary explanation for nonresponse was to conceal strategically lenient grading and to avoid revealing their grading beliefs out of fear or shame. In technical vocational schools, higher exam grades were attributed to easier standardized exams relative to challenging, profession-specific classroom content. Nonresponse was explained by teachers' perception of low study relevance (since their students rarely had issues with the standardized Mathematics Matura) and a school-specific professional mentality.

Methodologically, these findings highlight the need to account for topic sensitivity as a potential source of self-selection bias and to consider contextual factors such as subject and school type when examining such bias in educational research: a simple comparison of respondents and nonrespondents using a binary response variable (missing vs. not missing) would not have revealed significant differences; self-selection bias only became apparent when moderated by school type, as opposing effects in academic and vocational schools would otherwise have cancelled each other out. The findings also underscore the value of mixed-methods approaches for uncovering underlying mechanisms of self-selection bias. Without teachers' explanations, the quantitative findings would have remained speculative and vague; the qualitative results clarified these patterns, offering both valuable insights into self-selection bias and into differences in teacher grading across school types. Future research should replicate and extend these findings in larger and more diverse samples to disentangle the effects of school type, subject, and school context on self-selection bias. Researchers should carefully select dependent variables appropriate to their educational system, as the discrepancy between classroom and standardized exam grades may only be meaningful in systems like Austria's. Experimental studies could further test how survey framing affects response rates, and cross-national comparisons could examine whether mechanisms such as strategic grading or shame are context-specific or generalizable. Finally, providing a transparent account of self-selection bias, including accidental findings like ours, is crucial to advancing this field, as it can raise researchers' awareness of potential biases and the risks of unrepresentative data, ultimately leading to more robust results, particularly in sensitive educational topics such as grading.

6.3 Third Study – Extended Summary: Grading After Reform: How Policy Change Affected Teachers' Grading for University Entrance Qualifications

Vincent Schatz-Laliberté & Marko Lüftenegger

The COVID-19 pandemic created unprecedented challenges for schools worldwide (Daumiller et al., 2023), including rapid shifts to remote learning (Fridkin et al., 2023), widening achievement gaps (Donnelly & Patrinos, 2022), and pressure to adapt grading and assessment practices (Panadero et al., 2022). One key concern was ensuring fair and valid assessment for school-leaving qualifications that determine access to higher education (Castro et al., 2020; Tarkar, 2020). Many countries, including the UK, Ireland, and France, responded by adjusting final assessment systems, often replacing standardized exams with

teacher-assigned grades (Daddi, 2020; Doyle et al., 2021; Isaacs & Murphy, 2022). Austria responded to the challenges of the pandemic with a grading reform introduced in 2020, which integrated teacher-awarded classroom grades into the standardized university entrance qualification examination, known as the *Matura*. Under this reform, which remains in effect today, classroom grades account for 50% of the final Matura grade and can automatically compensate for a failing standardized exam result if the student has earned at least a classroom grade 3 (“Satisfactory”) in the respective subject (BGBl, 2021). Research from other contexts suggests that policy changes during the COVID-19 pandemic catalyzed grade inflation and reduced alignment between grades and actual academic achievement (Goldhaber & Goodman, 2024; McManus et al., 2021; Sanchez, 2023). Literature indicates that teachers often respond with greater leniency, emphasizing participation and effort over mastery, and experience tensions between professional ideals and classroom realities (Berne, 2023; Koller, 2021). Pandemic-related policies such as California’s “hold harmless” approach, for example, show how external mandates can constrain teachers’ grading judgments, influencing final grades without corresponding changes in students’ academic performance (Castro et al., 2020). Based on these findings and drawing on McMillan’s (2003) classroom grade decision-making model, which highlights how internal beliefs and external pressures interact to shape grading decisions, five research questions addressed classroom grade changes since the reform (RQ1), teachers’ explanations for observed grade changes (RQ2), perceived shifts in classroom grading practices (RQ3), and perceived internal and external influences on grading decisions since 2020 (RQ4–RQ5).

This study used a sequential explanatory mixed-methods design, combining a longitudinal analysis of classroom grades from 6,798 students (2017–2023) with follow-up interviews. Phase 1 examined grade trends before and after the 2020 reform in German and Mathematics, using chi-square tests and ordinal logistic regression to assess distributional changes and control for school-type differences. Phase 2 involved semi-structured interviews with 14 purposefully selected teachers from diverse school types and regions who had taught Matura classes before and after the reform. Interviews were analyzed using manifest qualitative content analysis (Graneheim & Lundman, 2004), with systematic coding, category development, and intercoder discussions to ensure consistency and reliability.

Descriptive analyses showed that classroom grades in Mathematics shifted upward immediately after the 2020 reform, with significantly higher grades from 2020 onward. In German, a similar shift appeared later, first becoming significant in 2021. Chi-square tests

confirmed these patterns: Mathematics displayed immediate and stable post-reform changes, whereas German showed a delayed and less consistent pattern. Ordinal logistic regression analyses supported the chi-square results and additionally revealed that school type mattered: teachers at vocational schools awarded significantly higher classroom grades since 2020 than academic schools, and they largely drove the significant reform-related effects.

Interview analysis revealed 12 categories explaining teachers' perceptions of grade changes and influences on grading after the 2020 grading reform. For RQ2 and RQ3, teachers attributed early grade increases partly to pandemic-related leniency and disrupted assessment conditions due to the COVID-19 pandemic. They attributed the rise in grade 2s ("Good") and the decline in grade 4s ("Sufficient") to norm-referenced adjustments and strategic efforts to secure students' Matura outcomes. Teachers also described targeted practices, such as offering additional tests, conducting oral examinations, or rewarding effort, that contributed to an increase in grade 3s, widely viewed as the "safe grade" because it is the lowest grade capable of offsetting a failing exam result. At the same time, several teachers emphasized their commitment to professional integrity and their efforts to resist such grade inflation. Teachers further highlighted external pressures (RQ4) from students and parents, who increasingly pushed for a safe grade 3, as well as from principals and class heads favoring positive results. Systemic factors, such as pandemic norms of leniency and perceived erosion of centralized standards, were also mentioned as drivers of general leniency in grading. For RQ5, internal factors included empathy for students, emotional strain associated with grading high-stakes outcomes, and efforts to balance compassion with academic standards. Teachers noted tensions between wanting to be supportive and adhering to fair grading. Some linked their colleagues' student-friendly self-images to lenient practices. Across interviews, a widespread theme concerned strong criticism of the grading reform. Many teachers felt it devalued the Matura, weakened assessment standards, encouraged strategic minimal effort among students, and inflated grades. Others appreciated reduced exam pressure and the inclusion of broader indicators of student work.

This study shows that Austria's 2020 grading reform triggered a lasting upward shift in final-year classroom grades, especially in Mathematics. While some of this increase reflects pandemic-related leniency, much of it appears structurally linked to the reform, which weakened the standardized nature of the Matura by giving substantial weight to classroom grades. The results reveal two critical challenges: (1) the blurring of performance standards through sustained grade inflation and (2) emerging strategic grading practices due to both

external and internal pressures on teachers that undermine the credibility and comparability of Matura outcomes. An unanticipated further challenge (3) concerns the systematic differences between school types, with vocational schools driving most of the inflation. Teachers reported growing tensions between fairness, empathy, and external pressures, as well as concerns that students increasingly rely on the “safe grade” rather than engaging deeply with learning. Although some teachers noted benefits such as reduced exam anxiety and more holistic assessment, many feared that the reform fosters a symbolic rather than genuinely performance-based system. Overall, the findings highlight that grading reforms must account for the social, institutional, and value-laden nature of teachers’ grading decisions; otherwise, they risk unintended consequences that erode the legitimacy of the standardized assessment system itself.

The study suggests that, at the systemic level, policymakers should reconsider allowing teachers to adjust standardized exam results in order to preserve objectivity, prevent systematic grade inflation, and safeguard both student motivation and teachers’ sense of purpose and meaningfulness in their role as graders. At the institutional level, the study emphasizes the importance of common grading standards and consistent practices within schools, supported by principals, to reduce teacher role conflict and foster shared understanding and professional values among colleagues. At the individual level, teachers need to recognize that their grading affects both student outcomes and the perceived legitimacy of the entire assessment system, making responsible and transparent grading practices essential.

7. Concluding Thoughts

As outlined in Section 4, I began this dissertation project with the implicit assumption that university entrance examinations at the end of upper secondary education should yield comparable measures of students’ academic performance. Yet as my work on the topic progressed, I increasingly questioned the necessity and value of comparability in final Matura grades. Through continued research, engagement with teachers and school leaders, and many hours of reflection on possible alternative assessment systems in upper secondary school, four central arguments emerged that now consciously lead me to challenge my earlier, largely unconscious belief in the “golden” goal of ensuring comparable Matura grades.

First, the objective of comparable Matura grades may be unattainable within a system that simultaneously fosters heterogeneity across educational tracks. The coexistence of different school types and subtypes, each with distinct curricular emphases, undermines the coherence of a system that aims to produce comparable academic outcomes. Particularly, the existing structural differences, such as profession-specific tasks in certain vocational schools and varying opportunities for students to choose whether to take the standardized Mathematics or English exam while still earning the same Matura certificate (Bundesministerium für Bildung, 2025d) inherently challenges the notion of comparability of final grades. Consistent with this contradiction, both the evidence from Germany (Section 4.2.3) and findings from all three studies of this dissertation revealed systematic grading differences between school types, suggesting that true comparability of final results is unlikely to be achieved without sacrificing the educational diversity and specialization that are central to the system's strength.

Second, the National Education Report (Bundesministerium für Bildung, 2024), which traces educational pathways in Austria, shows that students entering different types of upper secondary schools begin their four-year (academic) or five-year (vocational) paths toward the Matura at very different levels of prior academic focus and achievement. School type plays a key role in the achievement-based selection of students. As described in detail in my second study, most students from academic lower secondary schools move on to academic upper secondary schools, whereas those from non-academic tracks typically transition into vocational upper secondary schools or into vocational training without Matura certification. Admission to an academic lower secondary school, however, already depends on performance in primary school, leading to systematic differences in achievement levels across school types from a very early stage. As a result, students enter their upper secondary trajectories from inherently non-comparable starting points, which raises questions about the validity and fairness of assessing students with such diverse educational backgrounds through the same standardized test. Moreover, evidence from Germany clearly points to reference group effects across school types (Neumann et al., 2009; Watermann et al., 2013). It is therefore highly plausible (an assumption also echoed by many teachers interviewed in Studies 2 and 3) that Austrian Matura graduates' actual achievement levels differ between school types, even though such differences may remain invisible in their final Matura grades (Watermann et al., 2013). This raises a fundamental question of fairness—or, more precisely, of what is *less* unfair: to subject students with different educational backgrounds and

professional orientations to the same standardized examination tasks (for a thoughtful discussion of this issue in a different yet more pronounced context, see Howie, 2015), or to award them the same certificate even though the competencies this certificate attests may differ substantially between school types?⁴

The third argument relates to Benveniste's (2002) discussion of the symbolic function of national assessment systems. He argues that such systems often serve primarily to signal alignment with international assessment norms and a commitment to quality education, rather than to provide meaningful diagnostic information on students' academic performance. In practice, they may function symbolically, projecting an appearance of accountability while shielding the state from responsibility for actual educational outcomes. My third study, however, revealed the practical effects of such symbolic assessment systems on teachers. Those who described themselves as committed to fair and achievement-oriented grading struggled with the fact that the politically claimed comparability of Matura results does not hold up in practice. Many reported growing dissatisfactions, feeling that the effort they invest in consistent and fair grading is increasingly meaningless. This sense of misalignment seems to have intensified following the 2020 grading reform, leaving teachers feeling that their standardized Matura grading no longer reflects their professional standards or educational values. From the perspective of Equity Theory (Adams, 2015), this frustration can be understood as a response to perceived inequity: teachers' inputs of time, effort, and expertise are not matched by meaningful outcomes, resulting in demotivation. This raises the question of whether it is worth having teachers experience such a loss of work satisfaction merely to maintain what one teacher described as a "Potemkin village" of achievement standards and comparability of final grades, particularly in the context of growing teacher shortages (Birch et al., 2018) and teacher burnout (Pishghadam et al., 2013).

Fourth, building and sustaining a standardized assessment system aimed at ensuring comparability requires enormous efforts and resources. According to a report by the Austrian Court of Audit (Rechnungshof Österreich, 2020), the standardized Matura in 2018 incurred personnel costs of approximately €5.23 million (including €3.80 million for employees of the Federal Ministry, €1.36 million for item writers, and the remainder categorized as "rewards"), in addition to €2.83 million in material expenses (such as service contracts, IT costs, printing

⁴ My response to this question would be to reorganize the upper secondary education and assessment structure around an achievement level based modular system. However, this proposal lies beyond the scope of this dissertation.

and transport, rental expenses, travel costs, and a substantial portion listed as “other material costs”). These expenditures have remained relatively stable over the years; even in 2023, the standardized Matura cost approximately €8.45 million (Wiederkehr, 2025). Given that the objective of ensuring comparability is fundamentally constrained by the system’s own internal regulations (i.e., differing Matura frameworks and examination tasks across school types, the inclusion of non-standardized classroom grades since 2020, systematic grading disparities between school types, and potential grader bias due to non-anonymized teacher grading), it seems reasonable to question whether such substantial financial and administrative investments are warranted. Moreover, the results of my first study revealed a remarkably high degree of alignment between standardized Matura grades and non-standardized classroom grades, particularly in language subjects. This suggests that the standardized examination might provide little additional information about students’ achievement beyond what teachers already assess in class, again raising doubts about the value of a costly assessment apparatus whose outcomes largely replicate teachers’ classroom grades.

To provide a more balanced perspective, however, additional insights from the teacher interviews in my third study deserve attention. Teachers’ reflections revealed clear subject-specific differences in how they perceived the added value of the standardized Matura and the 2020 grading reform for student learning. In Mathematics, many teachers emphasized that standardization encouraged greater adherence to the prescribed curriculum and ensured that students were taught the required content. Some even noted that the standardized exams compelled less diligent colleagues to fulfill their professional responsibility to teach and assess at a level appropriate for upper secondary education. Many Mathematics teachers also welcomed the 2020 grading reform, appreciating its potential to allow for more holistic assessment and to reduce students’ anxiety during the Matura examinations. In stark contrast, almost all German teachers reported that standardization of the Matura had led to a narrowing of the curriculum, substantially diminishing the quality of instruction, the natural breadth and diversity of the subject, and their own sense of professional autonomy and enjoyment of teaching their subject.⁵ Regarding the 2020 grading reform, German teachers almost unanimously called for the abolition of the 50% inclusion of classroom grades, arguing that it had transformed the final school year into a period dominated by negotiations over classroom

⁵ On the other hand, interesting suggestions for language teachers on how to prepare students for standardized tests without compromising good and joyful teaching practices is offered by Hollingworth (2007).

grades rather than meaningful teaching and learning. Hence, discussions about standardization and reform need to take subject-specific realities into account. While standardization in Mathematics may strengthen curricular coherence and accountability (Herman & Golan, 1990), in subjects like German, it risks constraining creativity, critical thinking, and authentic engagement with language and literature (Berliner, 2011; Mahmood et al., 2020). This underscores that educational policies cannot be evaluated in general terms but must be understood in light of each subject's nature and learning objectives.

I situated my three dissertation studies within the broader development of the Austrian Matura, tracing its trajectory from the initial phase of standardization in academic secondary schools to the effects of the 2020 grading reform through 2023. Accordingly, I would like to conclude by highlighting some of the reflections and wishes teachers shared regarding the possible future of the standardized Matura and their role as graders within it. These insights emerged from interviews conducted during the data collection for my second and third studies, but were not discussed in detail in order to maintain the specific research focus of those studies.

Most teachers, regardless of their subject, expressed a clear desire for greater recognition of their professional responsibility as graders, particularly in the context of final-year grading (both classroom and Matura assessments). They reported growing frustration that their grading decisions are frequently questioned, most notably by principals, and emphasized the need for greater trust in their professional judgment as well as stronger support from school leadership in their grading decisions. Teachers also highlighted the importance of shared grading standards within their schools to promote a sense of fairness and comparability in their immediate professional environment.

Regarding the 2020 grading reform, both German and Mathematics teachers suggested reversing the rule that students automatically receive a passing grade if their classroom grade is at least a 3. According to the interviewed teachers, such a change would not only reduce grade negotiations during the final school year but also restore the significance of the Matura exams and the perceived meaningfulness of their professional work as examiners.

Nearly all interviewed teachers further expressed a desire to reintroduce an external chairperson for oral Matura examinations, a role that was abolished in 2020. Many teachers emphasized that students' performance levels at the oral exams have noticeably declined since then. Partly because students can no longer fail the Matura in subjects where they already

hold good grades, and partly because the absence of an external chairperson allows for overly generous grading to avoid conflict and present the school favorably.

The findings of this dissertation should be viewed as initial insights into teacher grading within the context of Austria's university entrance qualifications, as well as into the sensitivity of this topic when studied empirically. The teachers' reflections and wishes described above are best understood as preliminary considerations and points of departure for future research in this field. My qualitative findings in particular warrant further quantitative investigation to obtain a more representative understanding of teachers' perspectives on grading. Hence, any definitive conclusion drawn from my dissertation must be made with caution. Furthermore, as highlighted at the very beginning of this dissertation, research is always shaped by the context in which it is conducted. The perspectives captured here reflect both teachers' views and my own interpretations of the results within the current educational, political, and societal landscape that frames our experiences. Thus, I regard my conclusions in this dissertation as provisional, open to ongoing reassessment and development through my continuing research, reflection, and engagement with assessment policies and practices.

8. References

- Abitur Berlin Brandenburg. (2024, January 30). *Berechnung der Abiturnote in Berlin*. <https://www.abitur-berlin.de/abiturnote>
- Adams, J. S. (2015). Equity theory. In *Organizational Behavior I* (pp. 134-158). Routledge.
- Austria Presse Agentur Science (2021, December 16). "Corona-Matura" brachte Maturantenrekord, 2021. <https://science.apa.at/power-search/8804223724611446300>
- Allensworth, E. M., Gwynne, J. A., Moore, P., & De la Torre, M. (2014). *Looking forward to high school and college: Middle grade indicators of readiness in chicago public schools*. University of Chicago Consortium on Chicago School Research. 1313 East 60th Street, Chicago, IL 60637.
- Arencibia Alemán, J. M., Sandsør, A. M. J., Zachrisson, H. D., & Blömeke, S. (2024). Teacher-assigned grades and external exams: sources of discrepancy. *Assessment in Education: Principles, Policy & Practice*, 31(2), 94–115. <https://doi.org/10.1080/0969594X.2024.2338764>
- Autonome Provinz Bozen Südtirol. (2025). *Die staatliche Abschlussprüfung der Oberschule*. <https://deutsche-bildung.provinz.bz.it/de/schule/staatliche-abschlusspruefung-oberschule>
- Banks, J., & Smyth, E. (2015). 'Your whole life depends on it': Academic stress and high-stakes testing in Ireland. *Journal of youth studies*, 18(5), 598-616. <https://doi.org/10.1080/13676261.2014.992317>
- Bayrhammer, B. (2024, September 7). "Im System geerdet": Wer die Matura-Aufgaben erstellt. *Die Presse*, 2024. <https://www.diepresse.com/4720444/im-system-geerdet-wer-die-matura-aufgaben-erstellt>

- Benveniste, L. (2002). The political structuration of assessment: Negotiating state power and legitimacy. *Comparative Education Review*, 46(1), 89-118. <https://doi.org/10.1086/324051>
- Berliner, D. (2011). Rational responses to high stakes testing: the case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education*, 41(3), 287-302. <https://doi.org/10.1080/0305764X.2011.607151>
- Berne, J. L. (2023). *Grading in Crisis: Examining the Impact of Covid-19 on Secondary Public School Teachers' Grading and Reporting in South Florida* [Dissertation]. University of Southern California.
- BGBI (2021). Bundesministers für Bildung, Wissenschaft und Forschung. Verordnung, mit der die Verordnung über die Vorbereitung und Durchführung abschließender Prüfungen für das Schuljahr 2020/21 geändert wird (2021). https://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2021_II_211/BGBLA_2021_II_211.html
- BIFIE. (2013). *Bericht zur Entwicklung der standardisierten kompetenzorientierten schriftlichen Reife- und Diplomprüfung*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.parlament.gv.at/dokument/XXV/III/120/imfname_370662.pdf
- Birch, P., Balcon, M. P., Bourgeois, A., Davydovskaia, O., & Tremosa, S. P. (2018). Teaching Careers in Europe: Access, Progression and Support. Eurydice Report. *Education, Audiovisual and Culture Executive Agency, European Commission*.
- Bischof, D. (2023, November 20). Matura abschaffen? Der zeitlose Traum der SPÖ. *Die Presse*, 2023. <https://www.diepresse.com/17844578/matura-abschaffen-der-zeitlose-traum-der-spo>
- Bowden, B., Rodriguez, V., & Weingarten, Z. (2023). *The Unintended Consequences of Academic Leniency*. https://static1.squarespace.com/static/5ec580a6be16d6545a82a17b/t/6563d0ab6290971bd1b88ed5/1701040300226/brw_leniency.pdf
- Bowden, R.J (1986). Self-selection biases in correlational studies based on questionnaires. *Psychometrika* 51, 313-325, <https://doi.org/10.1007/BF02293987>
- Bonner, S. M., & Chen, P. P. (2021). Development and Validation of the Survey of Unorthodox Grading Beliefs for Teachers and Teacher Candidates. *Journal of Psychoeducational Assessment*, 39(6), 746-760. <https://doi.org/10.1177/07342829211015462>
- 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>
- Brennan, R., Kim, J., Wenz-Gross, M., & Siperstein, G. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard educational review*, 71(2), 173-217. <https://doi.org/10.17763/haer.71.2.v51n6503372t4578>
- Brookhart, S. M. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice*, 10(1), 35-36. <https://doi.org/10.1111/j.1745-3992.1991.tb00182.x>
- Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H [James H.], Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. *Review of Educational Research*, 86(4), 803-348. <https://doi.org/10.3102/0034654316672069>

- Brown, G. T. (2024). An integrated approach to improve the credibility and quality of assessment in higher education. In *Research Handbook on Innovations in Assessment and Feedback in Higher Education* (pp. 372-388). Edward Elgar Publishing.
- Bundeskanzleramt Österreich. (2025). *Teilstandardisierte kompetenzorientierte Reifeprüfung an AHS (Zentralmatura)*.
https://www.oesterreich.gv.at/de/themen/bildung_und_ausbildung/schulen/2/1/Seite.110037
- Bundesministerium für Bildung. (2024). *Nationaler Bildungsbericht Österreich 2024*.
<https://www.bmb.gv.at/Themen/schule/bef/nbb.html>
- Bundesministerium für Bildung. (2025a). *Lehrpläne der Allgemeinbildenden höheren Schulen*.
https://www.bmb.gv.at/Themen/schule/schulpraxis/lp/lp_ahs.html
- Bundesministerium für Bildung. (2025b). *Standardisierte kompetenzorientierte Reifeprüfung an AHS (SRDP)*. https://www.bmb.gv.at/Themen/schule/schulpraxis/zentralmatura/srdp_ahs.html
- Bundesministerium für Bildung. (2025c). *Die standardisierte Reife- und -Diplomprüfung*.
<https://www.bmb.gv.at/Themen/schule/schulpraxis/zentralmatura/srdp.html>
- Bundesministerium für Bildung. (2025d). *Standardisierte Reife- und Diplomprüfung: Angewandte Mathematik*. <https://www.matura.gv.at/srdp/angewandte-mathematik>
- Bundesministerium für Bildung. (2025e). *Übersicht über die Teilprüfungen der BHS*.
https://www.bmb.gv.at/Themen/schule/schulpraxis/zentralmatura/srdp_bhs.html
- Burke, J., & Minton, S. J. (2019). Well-being in post-primary schools in Ireland: the assessment and contribution of character strengths. *Irish Educational Studies*, 38(2), 177-192.
<https://doi.org/10.1080/03323315.2018.1512887>
- Canny, A., & Hamilton, M. (2018). A state examination system and perpetuation of middle-class advantage: an Irish school context. *British Journal of Sociology of Education*, 39(5), 638-653.
<https://doi.org/10.1080/01425692.2017.1377599>
- Carter, R. S. (1952). How invalid are marks assigned by teachers? *Journal of Educational Psychology*, 43(4), 218–228. <https://doi.org/10.1037/h0061688>
- Castro, M., Choi, L., Knudson, J., & O'Day, J. (2020). Grading Policy in the Time of COVID-19: Considerations and Implications for Equity. *California Collaborative on District Reform*.
- Chan, C. K. Y. (2022). A review of the changes in higher education assessment and grading policy during covid-19. *Assessment & Evaluation in Higher Education*, 48(6), 874–887.
<https://doi.org/10.1080/02602938.2022.2140780>
- Cheung, T. K. (2008). An assessment blueprint in curriculum reform. *Journal of Quality School Education*, 5, 23–37.
- Cliffordson, C. (2004, January). Effects of Practice and Intellectual Growth on Performance on the Swedish Scholastic Aptitude Test (SweSAT). *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment*, 3(30), 192–204.
<https://doi.org/10.1027/1015-5759.20.3.192>
- Cliffordson, C. (2008). Differential Prediction of Study Success Across Academic Programs in the Swedish Context: The Validity of Grades and Tests as Selection Instruments for Higher Education. *Educational Assessment*, 13(1), 56–75. <https://doi.org/10.1080/10627190801968240>

- Coenen, J., Cornelisz, I., Groot, W., Maassen van den Brink, H., & Van Klaveren, C. (2018). Teacher characteristics and their effects on student test scores: A systematic review. *Journal of economic surveys*, 32(3), 848-877. <https://doi.org/10.1111/joes.12210>
- Cross, L. H., & Frary, R. B. (1999). Hodgepodge Grading: Endorsed by Students and Teachers Alike. *Applied Measurement in Education*, 12(1), 53–72. https://doi.org/10.1207/s15324818ame1201_4
- Daddi, S. (2020, April 30). Covid-19 France and Bac: Written part cancelled. *The Connexion*. <https://www.connexionfrance.com/news/covid-19-france-and-bac-written-part-cancelled/421397>
- Daumiller, M., Rinas, R., Schoon, I., & Lüftenegger, M. (2023). How Did COVID-19 Affect Education and What Can Be Learned Moving Forward? A Systematic Meta-Review of Systematic Reviews and Meta-Analyses. *Zeitschrift Für Psychologie*, 231(3), 177–191. <https://doi.org/10.1027/2151-2604/a000527>
- Donnelly, R., & Patrinos, H. A. (2022). Learning loss during Covid-19: An early systematic review. *Prospects*, 51(4), 601–609. <https://doi.org/10.1007/s11125-021-09582-6>
- Doyle, A., Lysaght, Z., & O'Leary, M. (2021). High stakes assessment policy implementation in the time of COVID-19: the case of calculated grades in Ireland. *Irish Educational Studies*, 40(2), 385–398. <https://doi.org/10.1080/03323315.2021.1916565>
- Doyle, A. (2025). Disturbing the teacher's role as assessor: The case of calculated grades 2020–2021 in Ireland. *Irish Educational Studies*, 44(1), 113-136. <https://doi.org/10.1080/03323315.2023.2236597>
- Doz, D. (2023). Factors influencing teachers' grading standards in mathematics. *Oxford Review of Education*, 49(6), 819–837. <https://doi.org/10.1080/03054985.2023.2185217>
- Duckworth, A. L., Shulman, E. P., Mastronarde, A. J., Patrick, S. D., Zhang, J., & Druckman, J. (2015). Will not want: Self-control rather than motivation explains the female advantage in report card grades. *Learning and Individual Differences*, 39, 13–23. <https://doi.org/10.1016/j.lindif.2015.02.006>
- Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What No Child Left Behind Leaves Behind: The Roles of IQ and Self-Control in Predicting Standardized Achievement Test Scores and Report Card Grades. *Journal of Educational Psychology*, 104(2), 439–451. <https://doi.org/10.1037/a0026280>
- Dufaux, S. (2012). Assessment for Qualification and Certification in Upper Secondary Education: A Review of Country Practices and Research Evidence. OECD Education Working Papers, No. 83. *OECD Publishing (NJ3)*.
- Erdost, I., & Cavallar, G. (2023, June 14). Weg mit dieser Matura? Zwei Positionen. *Der Standard*, 2023. <https://www.derstandard.at/story/3000000174469/weg-mit-dieser-matura-zwei-positionen>
- European Commission. (2024). *Eurydice*. European Union. <https://eurydice.eacea.ec.europa.eu/national-education-systems/germany/overview>
- Eurydice. (2025, March 28). *Sweden: Assessment in upper general and vocational secondary education*. <https://eurydice.eacea.ec.europa.eu/eurydice/sweden/assessment-upper-general-and-vocational-secondary-education>
- Farhah, I., Saleh, A. Y., & Safitri, S. (2021). The role of student-teacher relationship to teacher subjective well-being as moderated by teaching experience. *Journal of Education and Learning (edulearn)*, 15(2), 267-274.
- Federal Ministry of Education. (2025). *Centralised Matura*. https://www.bmb.gv.at/en/Topics/school/ev_sl/cm.htm

- Fjortoft, H., & Morud, E. B. (2021). Assessment decision-making in vocational education and training. *Studia Paedagogica*, 26(4). <https://doi.org/10.5817/SP2021-4-6>
- Fridkin, L., Fonts, N. B., Quy, K., & Zwiener-Collins, N. (2023). Understanding effects of COVID-19 on undergraduate academic stress, motivation and coping over time. *Higher Education Quarterly*, 77(4), 623–637. <https://doi.org/10.1111/hequ.12425>
- Fuller, B., & Izu, J. A. (1986). Explaining School Cohesion: What Shapes the Organizational Beliefs of Teachers? *American Journal of Education*, 94(4), 501–535. <https://doi.org/10.1086/443864>
- Gallagher, A. (1998). Gender and Antecedents of Performance in Mathematics Testing. *Teachers College Record*, 100(2), 297–314. <https://doi.org/10.1177/016146819810000204>
- Goldhaber, D., & Goodman Young, M. (2024). Course grades as a signal of student achievement: Evidence of grade inflation before and after COVID-19. *Journal of Policy Analysis and Management*, 43(4), 1270-1282. <https://doi.org/10.1002/pam.22618>
- Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse education today*, 24(2), 105-112.
- Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, 72(2), 167–189. <https://doi.org/10.1093/poq/nfn011>
- Gsenger, P. (2019). *Zentralmatura Mathematik*. <https://www.ahs-aktuell.at/zentralmatura-mathematik/>
- Guez, A., Peyre, H., & Ramus, F. (2020). Sex differences in academic achievement are modulated by evaluation type. *Learning and Individual Differences*, 83-84, 101935. <https://doi.org/10.1016/j.lindif.2020.101935>
- Guillemot, F., Lacroix, F., & Nocus, I. (2025). Teachers' non-response bias in questionnaires about their attitude towards inclusive education. *Journal of Research in Special Educational Needs*, 25(2), 312–322. <https://doi.org/10.1111/1471-3802.12726>
- Guskey, T. R., & Bailey, J.M. (2010). Developing standards-based report cards. Thousand Oaks, CA: Corwin Press.
- Gustafsson, J.-E., & Erickson, G. (2013). To trust or not to trust?—teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability*, 25(1), 69–87. <https://doi.org/10.1007/s11092-013-9158-x>
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Praeger.
- Hansen, H. (2011). Rethinking certification theory and the educational development of the United States and Germany. *Research in Social Stratification and Mobility*, 29(1), 31–55. <https://doi.org/10.1016/j.rssm.2011.01.003>
- Herman, J. L., & Golan, S. (1990). Effects of Standardized Testing on Teachers and Learning--Another Look.
- Heyneman, S. P. (1988). Improving university selection, educational research, and educational management in developing countries: The role of examinations and standardized testing. In S. P. Heyneman & I. Fagerlind (Eds.), *University Examinations and Standardized Testing: Principles, Experience, and Policy Options* (Vol. 78, pp. 179–216). World Bank.
- Hoffmann, L., Hübner, N., Neumann, M., & Schröter, P. (2022). 9: Und wenn man die Abiturprüfungen einfach ausfallen ließe? Empirische Befunde zu Unterschieden zwischen

Abiturprüfungsnoten und Kursnoten. In L. Hoffmann, P. Schröter, A. Groß, S. M. Schmid-Kühn, & P. Stanat (Eds.), *Das unvergleichliche Abitur: Entwicklungen-Herausforderungen-Empirische Analysen* (pp. 251–288). wbv Media. <https://doi.org/10.3278/9783763972494>

Hollingworth, L. (2007). Five ways to prepare for standardized tests without sacrificing best practice. *The Reading Teacher*, 61(4), 339–342. <https://doi.org/10.1598/RT.61.4.7>

Howie, S. (2015). High-stakes testing in South Africa: friend or foe?. In *High-Stakes Testing in Education* (pp. 81–98). Routledge. <https://doi.org/10.1080/0969594X.2011.613369>

Howley, A., Kusimo, P. S., & Parrott, L. (2000). Grading and the ethos of effort. *Learning Environments Research*, 3(3), 229–246. <https://doi.org/10.1023/A:1011469327430>

Hübner, N., Wagner, W., Hochweber, J., Neumann, M., Nagengast, B., Kendeou, P., & Graham, S. (2020). Comparing Apples and Oranges: Curricular Intensification Reforms Can Change the Meaning of Students' Grade. *Journal of Educational Psychology*, 112(1), 204–220. <https://doi.org/10.1037/edu0000351>

Isaacs, T., & Murphy, R. (2022). *The impact of COVID-19 on 2020 to 2021 assessment arrangements*. ofqual. <https://www.gov.uk/government/publications/the-impact-of-covid-19-on-2020-to-2021-assessment-arrangements/the-impact-of-covid-19-on-2020-to-2021-assessment-arrangements>

Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using Mixed-Methods Sequential Explanatory Design: From Theory to Practice. *Field Methods*, 18(1), 3–20. <https://doi.org/10.1177/1525822X05282260>

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education*, 28(1), 91–105. <https://doi.org/10.1080/02671522.2012.754229>

Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy & Practice*, 28(3), 212–227. <https://doi.org/10.1080/0969594X.2021.1884041>

Koller, R. (2021). *COVID-19 Impact on Middle School Teachers' Grading Practice* [Master Thesis]. California State University San Marcos, California.

Köller, O., & Baumert, J. (2002). Das Abitur – immer noch ein gültiger Indikator für Studierfähigkeit? *Aus Politik Und Zeitgeschichte*, 52(B26), 12–19.

Köller, O., Baumert, J., & Schnabel, K. U. (1999). Wege zur Hochschulreife: Offenheit des Systems und Sicherung vergleichbarer Standards. Analysen am Beispiel der Mathematikleistungen von Oberstufenschülern an integrierten Gesamtschulen und Gymnasien in Nordrhein-Westfalen. *Zeitschrift Für Erziehungswissenschaft*, 2, 385–422.

Korkeila, K., Suominen, S., Ahvenainen, J., Ojanlatva, A., Rautava, P., Helenius, H., & Koskenvuo, M. (2001). Non-Response and Related Factors in a Nation-Wide Health Survey. *European Journal of Epidemiology*, 17(11), 991–999. <https://doi.org/10.1023/A:1020016922473>

Kroisleitner, O., & Wiesinger, A. (2025, May 8). Ist die Zentralmatura noch zeitgemäß? *Der Standard*. <https://www.derstandard.at/story/3000000268456/ist-die-zentralmatura-noch-zeitgemaess>

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4), 2025–2047. <https://doi.org/10.1007/s11135-011-9640-9>

Kuhn, T. S. (1997). *The structure of scientific revolutions* (Vol. 962). Chicago: University of Chicago press.

- Kwong, D., & Davis, J. R. (2015). School Climate for Academic Success: A Multilevel Analysis of School Climate and Student Outcomes. *Journal of Research in Education*, 25(2), 68-81.
- Lehane, P., O'Connor, G., & O'Leary, M. (2025). Back to the future? Reflections from 25 years of empirical research on Leaving Certificate assessment in Ireland. *Irish Educational Studies*, 1–31. <https://doi.org/10.1080/03323315.2024.2441158>
- Lievore, I., & Triventi, M. (2023). Do teacher and classroom characteristics affect the way in which girls and boys are graded? A multilevel analysis of student–teacher matched data. *British Journal of Sociology of Education*, 44(1), 97-122. <https://doi.org/10.1080/01425692.2022.2122942>
- Link, L. J., & Guskey, T. R. (2022). Is standards-based grading effective? *Theory Into Practice*, 61(4), 406-417. <https://doi.org/10.1080/00405841.2022.2107338>
- Lundahl, C., Hultén, M., & Tveit, S. (2017). The power of teacher-assigned grades in outcome-based education. *Nordic Journal of Studies in Educational Policy*, 3(1), 56–66. <https://doi.org/10.1080/20020317.2017.1317229>
- Mahmood, M. I., Mobeen, M., & Abbas, S. (2020). Effect of high-stakes exams on creative writing skills of ESL learners in Pakistan. *Research Journal of Social Sciences and Economics Review*, 1(3), 169-179.
- McManus, I. C., Woolf, K., Harrison, D., Tiffin, P. A., Paton, L. W., Cheung, K. Y. F., & Smith, D. T. (2021). Predictive validity of A-level grades and teacher-predicted grades in UK medical school applicants: A retrospective analysis of administrative data in a time of COVID-19. *BMJ Open*, 11(12), e047354. <https://doi.org/10.1136/bmjopen-2020-047354>
- McMillan, J. H. (2003). Understanding and Improving Teachers' Classroom Assessment Decision Making: Implications for Theory and Practice. *Educational Measurement: Issues and Practice*, 22(4), 34–43. <https://doi.org/10.1111/j.1745-3992.2003.tb00142.x>
- Meinck, S., Cortes, D., & Tieck, S. (2017). Evaluating the risk of nonresponse bias in educational large-scale assessments with school nonresponse questionnaires: a theoretical study. *Large-Scale Assessments in Education*, 5(1), 1–21. <https://doi.org/10.1186/s40536-017-0038-6>
- Ministère de L'Education Nationale. (2025). *Déroulement des corrections aux examens du second degré à compter des épreuves 2024*. <https://www.education.gouv.fr/bo/2024/Hebdo5/MENE2335316N#:~:text=La%20commission%20d'entente%20nationale,-Une%20commission%20d>
- Mychalewicz, P. (2024, June 12). Ist die Matura noch zu retten? *Der Standard*. <https://www.derstandard.at/story/3000000223869/ist-die-matura-noch-zu-retten>
- Nagy, G. (2007). *Berufliche Interessen, kognitive und fachgebundene Kompetenzen: Ihre Bedeutung für die Studienfachwahl und die Bewährung im Studium. (Doctoral dissertation)*
- National Agency for Education. (2009). Bedömaröverensstämmelse vid bedömning av nationella prov [Interrater consistency in marking national tests]. Stockholm: National Agency for Education.
- National Agency for Education. (2011). Skillnaden mellan betygsresultat på nationella prov och ämnesbetyg i årskurs 9, läsåret 2010/11 [The difference between results on national tests and subject grades in Grade 9, 2010/11]. Stockholm: National Agency for Education.
- Neuhauser, J. (2019, January 28). Die Mathematik-Matura - eine Wellenbewegung. *Die Presse*. <https://www.diepresse.com/5568524/die-mathematik-matura-eine-wellenbewegung>

- Neumann, M., Nagy, G., Trautwein, U., & Lüdtke, O. (2009). Vergleichbarkeit von Abiturleistungen: Leistungs- und Bewertungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten und die Rolle zentraler Abiturprüfungen. *Zeitschrift Für Erziehungswissenschaft*, 12(4), 691–714. <https://doi.org/10.1007/s11618-009-0099-6>
- 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>
- Nimmervoll, L. (2024, May 2). Ist die Matura reif für die Abschaffung? *Der Standard*. <https://www.derstandard.at/story/3000000218092/ist-die-matura-reif-f252r-die-abschaffung>
- Nowruzi, M., Amerian, M., Yazdani, H., & Mohammadi, A. (2020). The Relationship between Graded and Tested Achievement: Do Gender and Proficiency Level Make a Difference? *Applied Research on English Language*, 9(4), 477–502. <https://doi.org/10.22108/are.2020.120578.1516>
- Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *International Journal of Public Opinion Quarterly*, 70(5), 737–758. <https://doi.org/10.1093/poq/nfl038>
- Ouazad, A. (2008). Assessed by a teacher like me: Race, gender and subjective evaluations.
- Panadero, E., Fraile, J., Pinedo, L., Rodríguez-Hernández, C., & Díez, F. (2022). Changes in classroom assessment practices during emergency remote teaching due to COVID-19. *Assessment in Education: Principles, Policy & Practice*, 29(3), 361–382. <https://doi.org/10.1080/0969594X.2022.2067123>
- Peytcheva, E., & Yan, T. (2025). Practical Guide to Survey Questionnaire Design and Evaluation. Routledge.
- Pilcher, J. K. (1994). The Value-Driven Meaning of Grades. *Educational Assessment*, 2(1), 69–88. https://doi.org/10.1207/s15326977ea0201_4
- Pattison, E., Grodsky, E., & Muller, C. (2013). Is the Sky Falling? Grade Inflation and the Signaling Power of Grades. *Educational Researcher*, 42(5), 259–265. <https://doi.org/10.3102/0013189X13481382>
- Pishghadam, R., Adamson, B., Sadafian, S. S., & Kan, F. L. F. (2013). Conceptions of assessment and teacher burnout. *Assessment in Education: Principles, Policy & Practice*, 21(1), 34–51. <https://doi.org/10.1080/0969594X.2013.817382>
- Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education*, 26(7), 1372–1380. <https://doi.org/10.1016/j.tate.2010.03.008>
- Rechnungshof Österreich. (2020). *Zentralmatura: Bericht des Rechnungshofes Reihe BUND 2020/22*. https://www.rechnungshof.gv.at/rh/home/home/004.743_Zentralmatura.pdf
- Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of educational psychology*, 104(3), 700. <https://doi.org/10.1037/a0027268>
- RIS. (2021). *II. Verordnung des Bundesministers für Bildung, Wissenschaft und Forschung über die Vorbereitung und Durchführung abschließender Prüfungen für das Schuljahr 2020/21*. RIS. <https://www.ris.bka.gv.at/eli/bgbl/II/2021/11/20210108>

- Ross, J. A., & Kostuch, L. (2011). Consistency of report card grades and external assessments in a Canadian province. *Educational Assessment, Evaluation and Accountability*, 23(2), 159-180. <https://doi.org/10.1007/s11092-011-9117-3>
- Saleh, A. & Bista, K. (2017). Examining Factors Impacting Online Survey Response Rates in Educational Research: Perceptions of Graduate Students. *Journal of Multidisciplinary Evaluation*, 13(29), 63–74. <https://doi.org/10.56645/jmde.v13i29.487>
- Sanchez, E. I. (August 2023). *Evidence of Grade Inflation Since 2010 in High School English, Mathematics, Social Studies, and Science Courses*. <https://www.act.org/content/act/en/research/pdfs/Evidence-of-Grade-Inflation-in-English-Math-Social-Studies-and-Science.html>
- Schröter, P., Hoffmann, L., & Schmid-Kühn, S. M. (2022a). Ein Blick in andere Länder: Abschlussprüfungen im Spannungsfeld zwischen Standardisierung und Autonomie. *Das unvergleichliche Abitur: Entwicklungen–Herausforderungen–Empirische Analysen*, 89-126.
- Schröter, P., Söldner, H., Hoffmann, L., Riemenschneider, A., Jost, J., Wieser, D. (2022b). Wie vergleichbar sind die Bewertungen von Abiturarbeiten im Fach Deutsch? Empirische Studien zu verschiedenen Bewertungsmodellen. *Das unvergleichliche Abitur: Entwicklungen–Herausforderungen–Empirische Analysen*, 213–250. <https://doi.org/10.3278/9783763972494>
- Smyth, E., McCoy, S., & Banks, J. (2019). Student, Teacher, and Parent Perspectives on Senior Cycle Education in Ireland. <https://doi.org/10.26504/rs94.pdf>
- Spilt, J. L., Koomen, H. M., & Thijs, J. T. (2011). Teacher wellbeing: The importance of teacher–student relationships. *Educational psychology review*, 23(4), 457-477. <https://doi.org/10.1007/s10648-011-9170-y>
- Staller, K. M. (2010). Qualitative Research. In N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 1156–1161). Thousand Oaks, CA: Sage.
- Stenlund, T., Eklöf, H., & Lyrén, P.-E. (2016). Group differences in test-taking behaviour: an example from a high-stakes testing program. *Assessment in Education: Principles, Policy & Practice*, 1(24), 4–20. <https://doi.org/10.1080/0969594X.2016.1142935>
- Sun, Y., & Cheng, L. (2014). Teachers’ grading practices: Meaning and values assigned. *Assessment in Education: Principles, Policy & Practice*, 21(3), 326-343. <https://doi.org/10.1080/0969594X.2013.768207>
- Swan, G. M., Guskey, T. R., & Jung, L. A. (2014). Parents’ and teachers’ perceptions of standards-based and traditional report cards. *Educational Assessment, Evaluation and Accountability*, 26(3), 289-299. <https://doi.org/10.1007/s11092-014-9191-4>
- Tarkar, P. (2020). Impact of COVID-19 pandemic on education system. *International Journal of Advanced Science and Technology*, 29(9), 3812–3814.
- Tierney, R., Simon, M., & Charland, J. (2011). Being Fair: Teachers' Interpretations of Principles for Standards-Based Grading. *The Educational Forum*, 75(3), 210–227. <https://doi.org/10.1080/00131725.2011.577669>
- Tillinghast, J. A., Mjelde, D. J. W., & Yeritsyan, A. (2023). COVID-19 and grade inflation: Analysis of undergraduate GPAs during the pandemic. *Sage Open*, 13(4), 21582440231209110. <https://doi.org/10.1177/21582440231209110>
- Tomlinson, C. A., & McTighe, J. (2006). *Integrating differentiated instruction & understanding by design: Connecting content and kids*. Association for Supervision and Curriculum Development.

Tourangeau, R., Groves, R. M., & Redline, C. D. (2010). Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error. *Public Opinion Quarterly*, 74(3), 413–432. <https://doi.org/10.1093/poq/nfq004>

Van Petegem, K., Aelterman, A., Van Keer, H., & Rosseel, Y. (2008). The influence of student characteristics and interpersonal teacher behaviour in the classroom on student's wellbeing. *Social indicators research*, 85(2), 279-291. <https://doi.org/10.1007/s11205-007-9093-7>

Watermann, R., Nagy, G [G.], & Köller, O. (2013). Mathematikleistungen in allgemein bildenden und beruflichen Gymnasien. 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). Springer-Verlag.

Welsh, M. E., D'Agostino, J. V., & Kaniskan, B. (2013). Grading as a Reform Effort: Do Standards-Based Grades Converge With Test Scores? *Educational Measurement, Issues and Practice*, 32(2), 26–36. <https://doi.org/10.1111/emip.12009>

Wiederkehr, C. (2025). *Dokument 395/AB*.
https://www.parlament.gv.at/dokument/XXVIII/AB/395/imfname_1675422.pdf

Wien Geschichte Wiki-Bearbeiter. (2014). *Matura*.
<https://www.geschichtewiki.wien.gv.at/index.php?title=Matura&oldid=94351>

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>

9. Appendix

9.1 First Study

Teacher Grades for University Entrance Qualification: Bias, But Not as We Expected

Schatz-Laliberté, V., Zviagintsev, R., Kampa, N., Brown GTL. (under review). Teacher Grades for University Entrance Qualification: Bias, But Not as We Expected. *Assessment in Education: Principles, Policy & Practice*

Schatz-Laliberté, Vincent^{1*} (ORCID 0009-0009-3686-4656)

Brown, Gavin TL² (ORCID 0000-0002-8352-2351)

Zviagintsev, Roman¹ (ORCID 0000-0002-4547-7748)

Kampa, Nele¹ (ORCID 0000-0001-5633-9335)

¹ Centre for Teacher Education, Department for Teacher Education, University of Vienna, Austria

² Faculty of Arts and Education, Learning, Development & Professional Practice,
University of Auckland, New Zealand

Status of the study

Submitted on 7 November 2024 and under review since 19 March 2025

Contribution of the individual authors

Vincent Schatz-Laliberté: Conceptualization, Data curation, Data analysis, Writing

Gavin TL Brown: Conceptualization, Data analysis, Supervision

Roman Zviagintsev: Critical Review

Nele Kampa: Critical Review

Teacher Grades for University Entrance Qualification: Bias, But Not as We Expected

Vincent Schatz-Laliberté, Roman Zviagintsev, Nele Kampa, Gavin TL Brown

Abstract

For university entrance qualifications, there is a tension between reliance on different assessment methods, including teacher grades and external examination results. In 2020, an Austrian grading reform allowed teacher grades to be averaged with examination results to form a university entrance grade, raising concerns about grading validity. With data from 142 teachers and 8.581 students, this study examines longitudinally (2017 to 2023) the grade discrepancy between the two systems. ANOVA analyses evaluated the correlation between subject, school type, and teachers' gender with grade discrepancy. Overall, grade discrepancies were small, especially in language subjects. The reform did not affect grade discrepancy; indeed, discrepancies in mathematics declined in 2019 following changes in exam design. Teachers at vocational schools awarded lower classroom grades than examination results, while academic teachers tended to grade classroom performance higher. Findings suggest that teachers' classroom grades align well with standardised exam expectations.

Keywords: teacher grading; university entrance qualification; standardised exam; reliability; bias

1 Introduction

The decision of who should be allowed to attend university is an individual needle eye for students as they proceed from school to the tertiary education system. This decision should be grounded in a valid merit-based representation of students' academic achievements and potential. To avoid biases due to variations in teacher grading, many countries rely on examination-based assessments. However, while test scores offer standardisation, teacher grades have been shown to provide a more holistic evaluation of student performance, often predicting first-year university success more accurately than standardised tests (Cliffordson, 2008).

Both teacher grades and exam scores have advantages and disadvantages (Bowers, 2011; Brookhart et al., 2016), as they assess distinct aspects of student achievement (Willingham et al., 2002). However, little is known about what happens to teacher grades when a policy

change combines teacher grades with standardised examination scores for qualification purposes, as it was the case in an Austrian grading reform in 2020. Our study adds to that by examining longitudinally how different teacher grades are compared to examination results before and after the policy change.

2 Theoretical Background

2.1 University Entrance Qualifications

University entrance qualification exams at the end of upper secondary school are critical milestones for students, marking not only the end of a long educational journey but also serving as gateways to a wide array of future career opportunities. The responsibility to ensure that university entrance certificates provide a trustworthy, valid representation of the students' knowledge and skills falls not only on schools and teachers but also on educational policymakers. The types of knowledge and skills deemed essential for a country's specific objectives at the end of upper secondary school shape how these examinations are designed and what factors of students' achievement are assessed to create the final grades. For instance, in the United States, assessment practices are primarily driven by academic interests, with many states employing centralised standardised testing systems to evaluate the students' academic proficiency (Hansen, 2011). In contrast, the German and Austrian systems are closely aligned with economic needs, viewing education as an important investment in skills that increase productivity (Hansen, 2011). Hence, students have multiple pathways to higher education, including university entrance examinations that assess both academic and vocational skills in fields like agriculture, nutrition, and social work (BMBWF, 2024a; Kultusministerkonferenz, 2024).

2.2 Examination Scores versus Teacher Grades

There are different systems for determining grades for university entrance qualifications. Two major attempts are to rely on teachers' classroom assessments and to rely on external examination scores. The two ways are based on different grading mechanisms and measure different types of achievement (Willingham et al., 2002). Examination scores primarily reflect students' academic knowledge, while classroom grades incorporate a range of additional factors, such as effort, participation, and behavior (Cross & Frary, 1999; McMillan, 2003; Pilcher, 1994). Literature differentiates between achievement factors and non-achievement factors (Brookhart et al., 2016), highlighting that classroom grades and examination scores differ significantly in both meaning and psychometric characteristics.

A key feature of classroom grades is their ability to provide information on a broad range of skills and abilities compared to standardised examinations (Johnson, 2013), which correlate with later success in higher education (Cliffordson, 2008). Teachers also have the advantage of drawing from a rich base of evidence gathered through their ongoing interactions with and observations of students (Johnson, 2013). Furthermore, using teacher grades for qualification purposes can minimize potential disruptive effects that external exams might have on coursework and learning (Johnson, 2013).

However, the reliability of teacher grades is questioned due to the challenge of maintaining objectivity when grading one's own students (Brookhart et al., 2016). Although teacher grades are very reliable in ranking students within the classroom, they lack comparability across classrooms, schools, and federal states (Hübner et al., 2020; Neumann et al., 2009). Teachers' grading practices vary based on students' abilities, with non-achievement factors being considered more for low-performing students to heighten their grades (McMillan, 2003; Pilcher, 1994; Tierney et al., 2011). Regarding gender differences, female students receive higher classroom grades relative to their performance on standardised examinations than males (Gallagher, 1998; Guez et al., 2020). This phenomenon may be partially explained by the fact that female students show more self-discipline and self-control in class (Duckworth et al., 2015), and teachers tend to adjust grades based on this (Ferman & Fontes, 2022).

In contrast to the unreliable characteristics of classroom grades, most external examinations are criterion-referenced, meaning that they assess students based on predefined performance criteria. In general, high-stakes examinations, especially when invigilated, are resistant to collusion, cheating, and nepotism in terms of subsequent decision-making (Cheung, 2008). By using objectively scored exam items, scoring tends to be reliable and consistent (Haertel, 2006), to diminish scorer biases, in part because the candidate is anonymous to any human marker assigned to the script (Heyneman, 1988). The standardised administration and marking of examinations mean that all candidates are compared with the same 'ruler', creating a sense of equal treatment (Heyneman, 1988). Several studies indicate that test scores and classroom grades further differ in achieved grade levels (Arencibia Alemán et al., 2024; Neumann et al., 2011; Willingham et al., 2002). Generally, test scores tend to be lower than classroom grades, particularly for students with low academic performance (Hoffmann et al., 2022; Hübner et al., 2020).

2.3 Teacher Grades for University Entrance Qualification: Two Examples

Deciding whether to rely on teacher grading or standardised performance examinations is a complex and political problem that is resolved uniquely within each nation (Heyneman & Fagerlind, 1988). Different societies take quite different approaches to the challenge of creating a fair, merit-based platform for ensuring that candidates meet university entrance standards. In Sweden, for example, students who have completed all required school courses and project work can be admitted to university based on the grades given by their teachers (European Commission, 2024). Swedish teachers' grading is loosely moderated by national test results, which raises concerns about the reliability of grades (Jönsson et al., 2021). As an alternate pathway, students may sit a series of objectively scored general ability and knowledge tests (European Commission, 2024; Stenlund et al., 2016): the Swedish Scholastic Aptitude Test (SweSAT; *Högskoleprovning*). Despite lower reliability, teacher grades have shown to be better predictors of success in higher education than SweSAT scores (Cliffordson, 2008).

Germany also uses teacher grades for university entrance by combining classroom grades from the last two years of school and scores from teacher-graded standardised national tests (*Abiturprüfungen*). The teacher grade both coursework and Abitur exams, but unlike Sweden, very strict guidelines are set by the Ministers of Education (*Kultusministerkonferenz*) for calculating final grades (Abitur Berlin Brandenburg, 2024).

Results show that Abitur results align more closely with external test scores than with classroom grades, indicating that adding a standardised exam to teacher assessments can provide a fairer, more objective evaluation of student performance (Neumann et al., 2009). The combination of Abitur exam and teacher coursework predicted university success better than either form of grading alone (Köller & Baumert, 2002; Nagy, 2007). Regarding differences between school types, Watermann et al. (2013) found significantly higher achievement levels from students in academic upper secondary schools compared to vocational tracks, but these differences were not reflected in teachers' classroom grades.

3 Austrian Context and Research Questions

The Austrian university entrance certificate, called *Matura* (signifying maturity), marks students' readiness for the next step in their educational journey. Austria's upper secondary education school system is built of two school types that lead to university entrance qualification: (a) general upper secondary schools (AHS; *Allgemeinbildende Höhere*

Schulen), and (b) vocational upper secondary schools (BHS; *Berufsbildende Höhere Schulen*). At BHS, students graduate with a university entrance certificate and a completed vocational qualification, such as in technical fields, tourism, or forestry (BMBWF, 2024a). BHS graduates can pursue any university program. To ensure vocational students achieve academic knowledge and skills at the same level as their AHS colleagues the BHS pathway takes five years, one more year than AHS.

In 2023, 61% of students attended an AHS and 39% a BHS (Statistik Austria, 2024b). AHS provides lower secondary education to prepare students for upper secondary AHS, requiring specific academic qualifications (BMBWF, 2024a). Students who do not meet these requirements attend middle school. Middle school students need higher grades to enter AHS, making entry into BHS easier for middle school students compared to entry into AHS. 62% of AHS graduates move on to upper AHS, while only 9% of middle school graduates do so (BMBWF, 2021b).

The Austrian grading system follows a 5-point scale: grade 1 (*very good*), grade 2 (*good*), grade 3 (*Satisfactory*), grade 4 (*Sufficient*, the minimum passing grade), and grade 5 (*Insufficient*, the failing grade). To qualify for the Matura examinations, students must complete the final school year by passing all subjects with at least grade 4.

Since 2015/2016, the Austrian Matura examinations are standards-based, aiming to ‘introduce greater fairness, the same conditions for all Matura students and easier comparability of examination qualifications for higher education institutions’ (BMBWF, 2024b, para. 1). Matura examinations are mandatory in the subjects German, mathematics, foreign language (foremost English), and up to two additional examination subjects. The Matura comprises written examinations, oral examinations, and the submission of a pre-scientific work. Annually, 180 specially trained teachers develop about 3,200 Matura tasks, which are reviewed by 50 experts and tested in schools (Bayrhammer, 2024). One day before the written Matura examinations, sealed envelopes with the tasks are delivered to schools by special couriers.

The grading responsibility of the Matura examinations lies with the students’ own teachers. Official assessment grids shall help to facilitate the standards-based assessment. However, the examination papers are not anonymised, and as teachers are grading both their students’ classroom performance and Matura examinations, teachers are put in a difficult dual grader role. A plethora of studies showed that teachers want their students to succeed, which is

especially true for low-achieving students (McMillan, 2003). Indeed, the failing rates at the Matura are exceedingly low. In 2023, for example, 6,7% of the students failed the Matura (Statistik Austria, 2024a), with the rates being similar in AHS (5,4%) and BHS (7,9%) (Statistik Austria, 2024c).

Besides these issues of reliability, the Matura has faced additional challenges. One challenge concerned the examination tasks for mathematics. They were described as a “wave motion” in the media (Neuhauser, 2019), as the difficulty level seems to have varied greatly from year to year. In 2015, 10.5% of students failed the written math exam which nearly doubled in 2016, with 21.8% receiving a failing grade. By 2017, the rate dropped back to 11.8%, only to rise once more in 2018, reaching 22.4% (Neuhauser, 2019). To diminish these fluctuations, the tasks (reduction of text-based tasks, more time for difficult tasks) and the grading key (possibility to award half-points, facilitation for achieving grade 4 in the grading key) were changed (Gsenger, 2019).

Another challenge of the Matura was the covid-19 pandemic. To address the loss of classroom time due to school closures several weeks before the Matura, the Ministry of Education introduced several measures for Matura candidates. The oral Matura exam became optional, while the written exams were limited to three subjects, and students were given six hours instead of four to complete the written exams (BMBWF, 2021a). Additionally, the final Matura grade was no longer based solely on exam results. Instead, 50% of the final Matura grade was drawn from the students’ final year classroom grade in the respective subject (BMBWF, 2021a). For example, if students had a classroom grade of 2 but received a 4 on the exam, their final grade was a 3. In case of a one-grade difference, the exam result took precedence. So, students with at least a classroom grade of 3 could no longer receive a failing grade (5).

While most of the measures implemented for Matura candidates in 2020 were withdrawn in subsequent years, the integration of final year classroom grades into the Matura grading remained. This raises the question of whether this change was truly intended as a temporary support for students during the covid-19 pandemic, as reported in the media (APA-Science, 2021), or if the challenging circumstances of 2020 were leveraged to introduce a lasting change to the Matura grading system.

The reasons for this shift can only be speculated upon, but a glance at the press coverage of education reveals that changing or even abolishing the Matura has long been a hotly debated

topic between various political parties and interest groups (Nimmervoll, 2024). Arguments range from viewing the Matura as a relic of a bygone era (Bischof, 2023) causing stressful teaching and learning focused on the Matura exam rather than preparing students for their future (Erdost & Cavallar, 2023) to concerns from conservative voices fearing a drastic downward leveling of educational standards (Bischof, 2023).

Considering the context of this policy reform, we derive the following research questions:

- (1) How do classroom and Matura grades relate to each other in a context where both grades are awarded by the same teacher?
- (2) What is the difference in grading among the subjects of mathematics, German and English in terms of alignment between classroom and Matura grades?
- (3) How has the 2020 grading reform affected the relationship between the two grades?
- (4) Does the alignment between the two grades differ between the school types AHS and BHS?

4 Methods

Three general upper secondary schools and four vocational upper secondary schools, varying in size, location, specialization, and student demographics participated in our study in Vienna, Austria. A total of 17,162 grades, with 8,581 final year classroom grades and corresponding Matura exam grades, matched for the individual students, were collected from 2017 to 2023. The grades were awarded by 154 teachers over the seven years, including the 2020 grading reform. On average, each teacher graded approximately 56 students with big differences between teachers ($SD = 38$). The dataset includes 3,505 grade pairs for German, 3,293 for mathematics, and 1,783 for English. 60% of the teachers are women, 24% are men, 16% provided no information regarding their gender. Table 1 depicts the number of students (respectively matched grade pairs) and teachers per school year, school type, and subject.

To explore the relationship between final-year classroom grades and Matura grades over the seven years, Cohen's d (1992) was calculated as an effect size measure to compare the grade distributions of the two groups (classroom and Matura grades) for each teacher and academic year. Grade levels 1-4 were included, while grade 5 was excluded due to its minimal representation of less than 1%.

We employed the online effect size calculator developed by Wilson (2023). It compares the frequency distribution in each grade category by assessing means (average ordinal values,

adjusted for category frequencies) and standard deviations (variability of ordinal data around the means) of both groups. Higher proportions of grades in higher grade levels in one condition compared to the other condition results in a positive effect, whereas a higher proportion in the lower grades results in a negative effect. Therefore, we inverted the grades so that higher values indicated better grades (the lowest grade 4 became 1, etc.). A positive effect indicates that higher classroom grades were awarded compared to Matura grades; a negative effect indicates the opposite. Therefore, the greater the positive value of Cohen's d , the stronger the tendency for teachers to give higher classroom grades compared to Matura grades.

The effect size was calculated for each teacher for each academic year separately and used as the dependent variable. An ANOVA was conducted using *jamovi* (The jamovi project, 2022) to analyze the main effects and potential interaction effects of the independent variables year, subject, school type, and teachers' gender on the standardised difference in grade distributions between classroom and Matura grades (Cohen's d). This analysis revealed (a) how each independent variable correlated with the grade difference while controlling for the other variables, and (b) if any interaction effects were present.

5 Results

Overall, the grade differences between Matura and classroom grades were small (research question 1). While 56% of students received the same grade in both assessments, 21% obtained a Matura grade one level lower than their classroom grade, the reverse was true for 20% of students. In 3% of comparisons, the grades differed by two levels, and in less than 1% by three grade levels.

Table 2 presents the grade distribution differences between classroom and Matura grades by school type, year, and subject, along with the average Cohen's d for each category. The table provides an overview of changes in effect sizes over time, allowing a first insight into whether the tendency to assign higher classroom grades compared to Matura grades changed over time, especially regarding the reform year 2020 (research question 3). From a statistical perspective, it indicates whether the effect sizes have moved closer to 0, and from a practical perspective, it shows whether the tendency to assign higher classroom grades compared to Matura grades changed over time.

The average effect sizes across the seven years tended to be small with no significant changes in 2020, especially for German and English. There was slightly more variability in the mathematics grades, with $SD = .18$ for AHS and $SD = .28$ in BHS.

Table 3 presents the ANOVA results (research questions 2 and 4), with four statistically significant predictors for the difference between classroom grades and Matura grades being identified: The main effects of year ($p = 0.031$), school type ($p = 0.003$), and teachers' gender ($p = 0.001$) as well as the interaction between year and subject ($p = 0.023$).

Table 1. Number of Students and Teachers per Year, School Type, and Subject

Year	AHS			BHS			Total
	German	English	Maths	German	English	Maths	
2017	154 [9]	42 [3]	97 [5]	263 [13]	208 [12]	273 [12]	1037 [54]
2018	179 [10]	15 [1]	182 [8]	268 [13]	204 [11]	230 [9]	1078 [52]
2019	140 [9]	71 [5]	125 [8]	359 [16]	21 [11]	318 [13]	1224 [62]
2020	162 [9]	60 [2]	181 [8]	354 [15]	194 [11]	365 [15]	1316 [60]
2021	194 [8]	80 [3]	214 [10]	360 [14]	225 [13]	324 [13]	1397 [61]
2022	187 [9]	40 [2]	185 [8]	358 [16]	98 [7]	370 [16]	1238 [58]
2023	187 [8]	109 [6]	203 [8]	340 [15]	226 [12]	226 [10]	1291 [59]
Total	1203 [62]	417 [22]	1187 [55]	2302 [102]	1366 [77]	2106 [88]	8581 [406]

Note. [] = number of teachers; AHS = academic upper secondary schools; BHS = vocational upper secondary schools. Since many teachers awarded grades across multiple years, the total number of teachers in the table exceeds their actual number.

Table 2. Standardised Differences (Cohen's *d*) in Grade Distribution Between Matura and Classroom Grades per School Type, Year, and Subject

Subject	Year	Classroom grades				AHS Matura grades				Cohen's <i>d</i>	BHS Classroom grades				BHS Matura grades				Cohen's <i>d</i>
		1	2	3	4	1	2	3	4		1	2	3	4	1	2	3	4	
Mathematics	2017	22	18	22	37	14	28	26	32	-0.02	7	11	28	55	15	26	26	33	-0.57
	2018	18	18	26	37	5	13	37	45	0.35	6	17	25	52	6	6	20	68	0.40
	2019	16	21	24	40	11	18	28	42	0.20	9	13	29	49	10	14	31	45	-0.11
	2020	13	20	26	41	4	15	29	52	0.41	13	20	33	35	13	25	33	30	-0.09
	2021	21	17	29	32	12	19	27	42	0.26	9	17	29	44	9	22	40	29	-0.21
	2022	25	24	29	22	19	23	32	26	0.13	10	15	26	49	16	18	37	29	-0.30
	2023	15	21	29	35	14	30	25	31	-0.12	14	20	30	36	15	17	35	33	0.11
	<i>M (SD)</i>									<i>0.17 (.18)</i>									<i>0.11 (.28)</i>
German	2017	25	24	30	21	23	26	32	19	-0.02	10	24	34	32	11	30	35	24	-0.16
	2018	27	23	26	24	16	31	31	22	0.13	15	22	41	22	14	28	35	23	0.02
	2019	24	24	31	22	17	23	36	21	0.21	16	18	35	32	12	26	35	26	-0.05
	2020	20	28	31	21	17	28	31	24	0.10	19	22	32	28	16	26	36	23	-0.05
	2021	23	25	32	21	17	29	28	26	0.11	22	30	32	17	17	26	38	19	0.18
	2022	30	27	26	17	22	27	33	18	0.18	14	23	32	31	13	23	36	28	0.05
	2023	23	25	33	20	19	24	30	26	0.15	17	22	38	23	15	30	31	24	-0.02
	<i>M (SD)</i>									<i>0.12 (.07)</i>									<i>0.00 (.10)</i>
English	2017	46	28	20	7	39	26	17	9	0.02	14	25	34	26	21	25	29	25	-0.19
	2018	25	31	37	6	19	19	56	0	0.17	26	27	24	23	29	26	21	24	-0.04
	2019	38	29	22	11	37	29	19	12	-0.02	20	28	28	24	28	22	18	31	-0.07
	2020	42	34	11	13	48	21	24	3	-0.11	24	30	23	24	26	27	22	25	0.00
	2021	49	34	12	5	47	31	18	4	0.04	24	33	25	19	27	32	20	20	-0.05
	2022	43	30	17	9	48	26	9	4	-0.39	14	28	37	21	13	28	27	33	0.15
	2023	30	26	25	18	28	31	18	19	-0.14	27	23	32	18	28	32	25	15	-0.22
	<i>M (SD)</i>									<i>0.06 (.16)</i>									<i>-0.06 (.11)</i>

Note. AHS = academic upper secondary schools; BHS = vocational upper secondary schools.

Table 3. ANOVA Predicting the Difference Between Classroom and Matura Grades (Cohen's d) by Year, School Type, Subject, and Teachers' Gender, Including Interactions

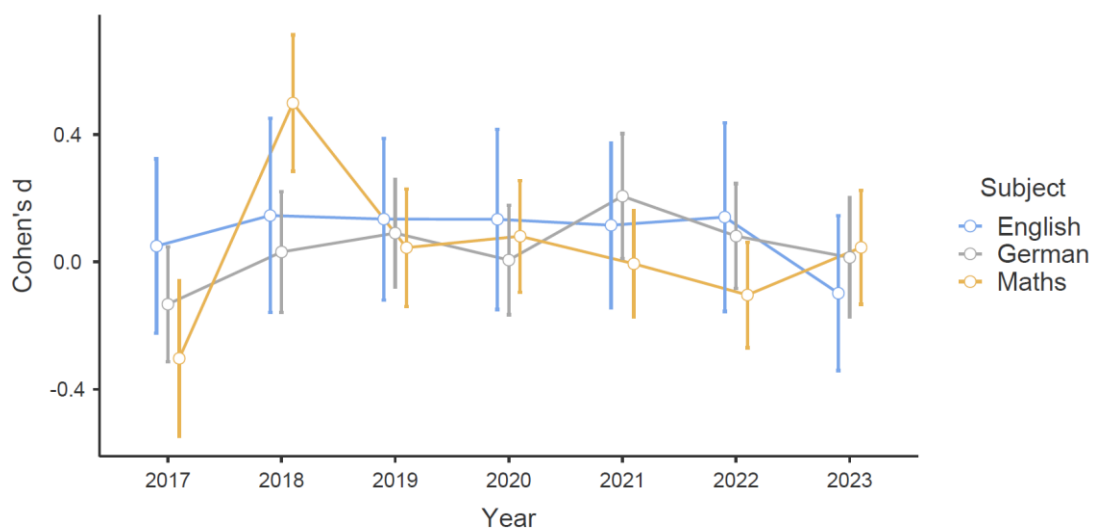
	df	Sum of Squares	Mean Square	F	p	η_p^2
Overall model	39	9.6348	0.2470	2.232	< .001	
Year	6	1.9529	0.3255	2.350	0.031	0.044
School type	1	1.2445	1.2445	8.985	0.003	0.029
Subject	2	0.0496	0.0248	0.179	0.836	0.001
Gender	1	1.4379	1.4379	10.380	0.001	0.033
Year*school type	6	0.2794	0.0453	0.327	0.923	0.006
Year*subject	12	3.3526	0.2794	2.017	0.023	0.074
School type*subject	2	0.2394	0.1197	0.864	0.422	0.006
Year*gender	6	0.3588	0.0598	0.432	0.858	0.008
School type*gender	1	0.0874	0.0874	0.631	0.428	0.002
Subject*gender	2	0.3200	0.3200	2.310	0.101	0.015
Residuals	305	0.1385	0.1385			

Note. df = degrees of freedom, F = F-statistic, p = p -Value, η_p^2 = partial Eta squared.

5.1 Year-by-Year Differences Between Matura and Classroom Grades

A Tukey post-hoc test revealed that the main effect of year and the interaction effect between year and school subject can be attributed to significant fluctuations in mathematics, particularly between 2017 and 2018. The only statistically significant year-to-year comparison was between 2017 and 2018 ($p_{\text{Tukey}} = 0.011$). An additional post-hoc test for year by subject showed that only mathematics had statistically significant year-to-year differences ($p_{\text{Tukey}2017-2018} < 0.001$, $p_{\text{Tukey}2018-2021} = 0.042$, and $p_{\text{Tukey}2018-2022} = 0.003$). Figure 1 illustrates subject-specific trends across years (research questions 2 and 3). In 2017, mathematics classroom grades were lower than Matura grades ($d = -0.30$). In 2018, this trend reversed, and classroom grades were significantly higher than Matura grades, ($d = 0.50$). From 2019 to 2023, these differences narrowed and showed little or no difference between classroom and Matura grades. In contrast, the two language subjects German and English showed a more stable trend across the seven years.

Figure 1. Subject-specific Differences Between Classroom and Matura Grades Across Years



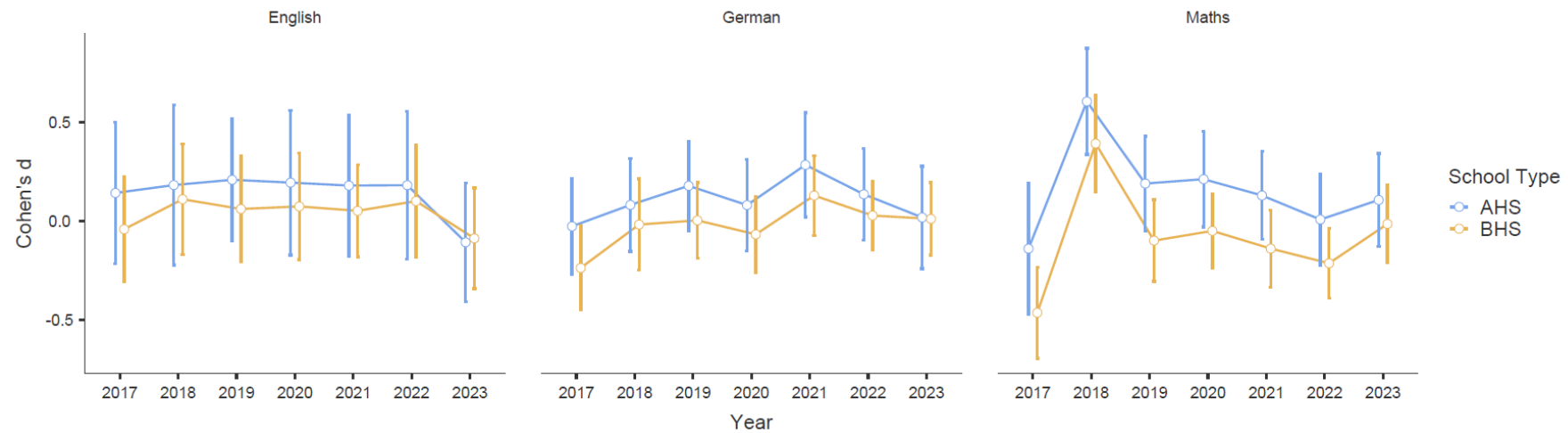
Note. A higher Cohen's d value indicates a higher tendency to award higher classroom grades compared to Matura grades.

It can be stated that the gap between classroom and Matura grades narrowed over the years (the values become closer to zero); however, this effect did not happen in the reform year 2020 but was largely driven by large fluctuations in mathematics between 2017 and 2018.

5.2 Differences by School Type and Teachers' Gender

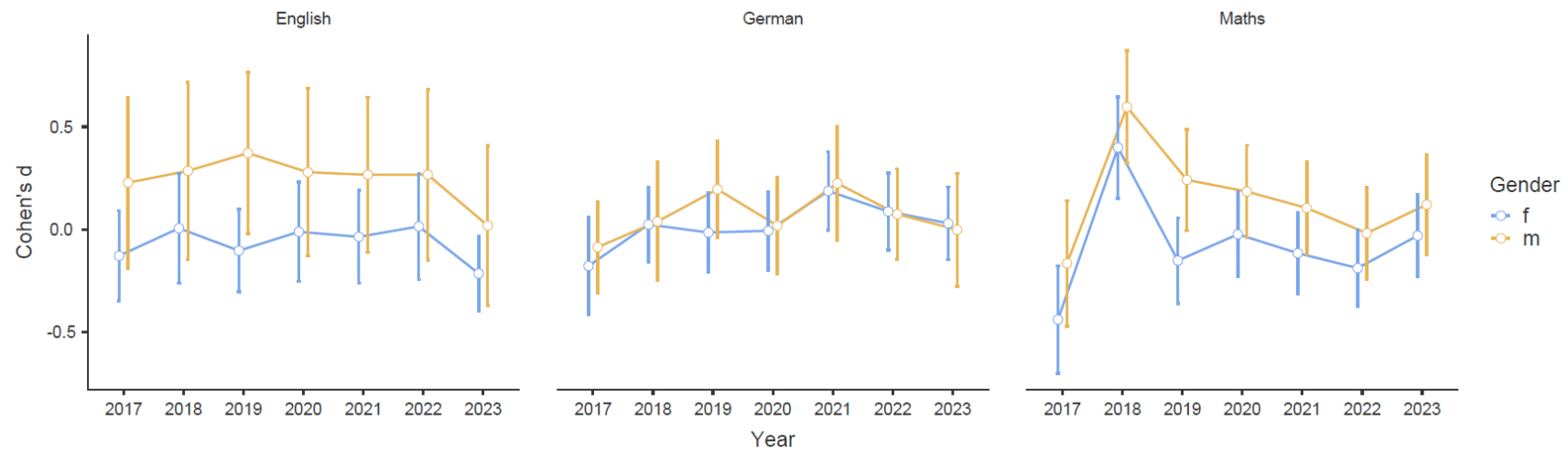
A post hoc test comparing school types (research question 4) showed that, overall, AHS teachers gave significantly higher classroom grades compared to Matura grades than BHS teachers ($p_{\text{Tukey}} = 0.003$). This trend is consistent across all three subjects, even though there are notable differences between them. AHS and BHS teachers awarded higher classroom than Matura grades in English (positive Cohen's d values), except for 2023, with AHS teachers showing a stronger tendency to give higher classroom grades than BHS teachers (see Figure 2). In German, BHS teachers awarded similar grades across the years, while AHS teachers, like in English, awarded higher classroom grades than Matura grades. In mathematics, AHS teachers also gave higher classroom grades compared to Matura grades, while BHS teachers assigned lower classroom grades. Comparing the differences between the two school types, the gap between them decreased after the 2020 grading reform (research question 3).

Figure 2. Effect of School Type on the Difference Between Classroom and Matura Grades Across Years for AHS and BHS



Note. AHS = academic upper secondary school, BHS = vocational upper secondary school. Higher Cohen's d value indicates higher classroom grades compared to Matura grades.

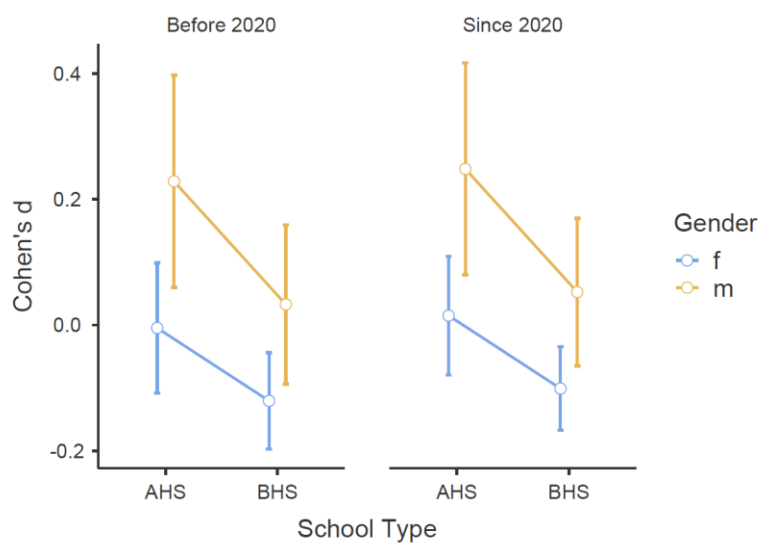
Figure 3. Effect of Teachers' Gender on the Difference Between Classroom and Matura Grades Across Years



Note. f = female teachers, m = male teachers. Higher Cohen's d value indicates higher classroom grades compared to Matura grades.

The second significant factor was teachers' gender (see Figure 3). While the difference between the two grades awarded by women was small and rather in favor of higher Matura grades (negative Cohen's d values), men awarded significantly ($p_{\text{Tukey}} = 0.001$) higher classroom grades compared to Matura grades. This was especially true for English and mathematics. Figure 4 compares these effects before and after the 2020 grading reform (research question 3).

Figure 4. Female and Male Teachers' Grading in AHS and BHS Before and After the Grading Reform



Note. f = female teachers, m = male teachers. Higher Cohen's d value indicates higher classroom grades compared to Matura grades.

As can be seen in Figure 4, the reform does not seem to have affected the grading differences between men and women. The pattern of male teachers awarding the highest grades in AHS compared to Matura grades remains consistent before and after the policy change.

Since not all teachers awarded grades every year in a row, a sensitivity analysis with the 39 teachers who provided data at least four times was subjected to the same analyses. These produced a statistically significant effect for school type ($p = 0.03$), but not for teachers' gender ($p = 0.07$).

6 Discussion

This study investigated longitudinally the alignment between teacher-awarded classroom

grades and standardised examination grades for university entrance qualifications in Austria over seven years (2017–2023). The focus lay on a 2020 grading reform. Before 2020, classroom grades were independent of Matura examination grades. From 2020, classroom grades were averaged with Matura examination grades to create a university entrance score for each student. The standardised difference (Cohen's d) in grade distribution between classroom and Matura grades was calculated and the impact of school subject, school type, and teachers' gender upon this difference was investigated. A statistically significant effect for subject and school type was found, but not for the grading reform.

Regarding the first research question, there was a high degree of alignment between classroom grades and standardised Matura grades, with most differences being quite low ($d < .20$). This contrasts with previous research findings highlighting differences between classroom grades and standardised tests scores (Arencibia Alemán et al., 2024; Neumann et al., 2011; Willingham et al., 2002). In our case, however, the same teachers graded both assessments. Since the teachers were familiar with the expectations of the Matura exam, they may have adjusted their classroom grading to align with these expectations. Alternatively, teachers may have graded the Matura exams in a way that validated their previous classroom grading decisions, leading to similar grades across both assessments. Future studies may explore whether teachers grade differently when they do not grade exams of their students or when the students are anonymous.

Regarding the second research question, our results showed high variability in mathematics grades before 2019, with differences flip-flopping between higher and lower Matura grades compared to classroom grades. As of 2019, the discrepancy decreased substantially when the Ministry of Education introduced measures aimed at increasing consistency in the difficulty of Matura examination tasks (Gsenger, 2019). In contrast, both language subjects showed very small discrepancies between Matura and classroom grades. This may be due to the use of the same assessment criteria in both formats or the relative ease of aligning grades in text-based assessments compared to mathematics. Additionally, potential differences between subjects could be attributed to differences in teacher training on the use of marking rubrics.

The 2020 grading reform had no statistically significant impact on the discrepancy between classroom and Matura grades (third research question). Although the difference between the two grades is narrower after 2020, this seems to be mainly due to the changes in 2019 which would be due to the new design of the mathematics Matura examination. This inference could

also be supported by the stable and small discrepancies for German and English where these changes did not occur. Given that teachers are familiar with Matura examination grading and probably have calibrated their classroom grading with the Matura standards (or vice versa), the policy of averaging the two grade types had no meaningful impact regarding the difference between the two grades. Existing literature emphasises that teachers consider a variety of non-achievement factors when assigning classroom grades, leading to higher grades than expected from standardised performance assessments (Cross & Frary, 1999; McMillan, 2003; Pilcher, 1994). Based on our results, it could also be suggested that this is not the case in a system where classroom grades are intended to prepare students for Matura examinations. However, this conclusion only holds if we assume that teachers did not consider any non-achievement factors when assessing the standardised Matura exams. Future comparative studies could explore how the grading practices in Austria might differ from those in other countries to further corroborate our inference.

Regarding school type differences (fourth research question), AHS teachers showed a significantly higher tendency to award higher classroom grades compared to Matura grades than BHS teachers. This contrasts findings from Germany (Watermann et al., 2013), which report a bias in favor of classroom grades in vocational school tracks. Given the limited research on grading differences between school types as well as specifics of each education system, further (international comparison) research is needed to understand the mechanisms behind this result. In Austria, especially in view of the 2020 grading reform, it is important to further investigate how classroom assessments are carried out in the final school year in both school types and what types of (non-)achievement factors are considered by teachers when grading. Nevertheless, it can be stated that without the merging of the two grading systems the lower grades in BHS would not be balanced out in the Matura.

A surprising result was that male teachers gave significantly higher classroom grades than female teachers across both school types and all subjects. However, this result did not stand a sensitivity analysis. To evaluate whether male teachers are more prone to grade inflation than women, a much larger sample is needed in future studies before any conclusions about the effect of teachers' gender can be drawn.

The policy implication of this study, given that the introduction of the grading reform policy did not affect grade discrepancy, is that the policy might have caught up with the reality of how teachers made grading decisions. Teachers are informed by the formal examination

standards and criteria embedded in the Matura examination, consequently their grades might seem to reflect such criteria. The policy may simply have given formal recognition to the reality of teacher grading. According to the existing literature, which shows that teacher grades (classroom as well as teacher-assigned standards-based grades) are valid predictors of future success at university (Cliffordson, 2008; Köller & Baumert, 2002; Nagy, 2007), allowing classroom grades to be part of the university entrance protocol seems to be justified.

Our findings are based on a sample in one urban area (Vienna), with schools being very different in terms of location, student background, size, and focus. However, our results should only cautiously be generalized to all students in Austria. To also take rural areas into account, other federal states from Austria need to be investigated to draw more general conclusions. Another limitation is that not all teachers award grades annually. Our sample also contained fewer male than female teachers. We addressed this issue with a sensitivity analysis which yielded robust results for school type, but not for teachers' gender.

Our results give rise to several questions that should be addressed in future studies. First, it is unclear what mechanisms led to the different grading behaviour of AHS and BHS teachers. Based on the theoretical background, we could assume that when assigning grades AHS teachers tend to give more weight to non-achievement factors, such as effort and participation (Brookhart et al., 2016). It is important to analyse how teachers in different school types vary in their pedagogical role as graders and the different educational goals they pursue through their grading practices. Second, the effect of teachers' gender on their grading behaviour is largely unexplored (Coenen et al., 2018), and the former studies showed significant but contradictory results (Nowruzi et al., 2020; Ouazad, 2008). Third, although our results revealed that the reform did not have a significant effect on grades, our study design did not allow us to investigate possible other effects of the grading reform. For example, there may have been changes in the way classroom tests are conducted in the final school year, changes in teachers' sense of responsibility in assigning classroom grades, or even changes in student-teacher-relationships due to possible negotiations over classroom grades because of their new weight in the university entrance protocol. Subsequent qualitative studies may explore these changes in depth.

In a system where the same teacher gives both classroom and standards-based exam grades, the two grades seem to be well aligned. This alignment wasn't even strengthened by a reform that merged the two grading systems. However, our research found biases that we didn't

expect, showing significant differences in teachers' grading behaviour between academic and vocational schools, as well as possible biases due to teachers' gender. Given the goal of standards-based assessments at the end of upper secondary education – ensuring comparable results as a fair basis for university entrance – our findings should be carefully considered by teachers, researchers, and policymakers alike.

Declaration of conflicting interests

The authors report there are no competing interests to declare.

Data availability statement

The data presented in this study are available on reasonable request from the corresponding author.

References

- Abitur Berlin Brandenburg. (2024, January 30). *Berechnung der Abiturnote in Berlin*. <https://www.abitur-berlin.de/abiturnote>
- APA-Science (2021, December 16). "Corona-Matura" brachte Maturantenrekord, 2021. <https://science.apa.at/power-search/8804223724611446300>
- Arencibia Alemán, J. M., Sandsør, A. M. J., Zachrisson, H. D., & Blömeke, S. (2024). Teacher assigned grades and external exams: sources of discrepancy. *Assessment in Education: Principles, Policy & Practice*, 31(2), 94–115. <https://doi.org/10.1080/0969594X.2024.2338764>
- Bayrhammer, B. (2024, September 7). "Im System geerdet": Wer die Matura-Aufgaben erstellt. *Die Presse*, 2024. <https://www.diepresse.com/4720444/im-system-geerdet-wer-die-matura-aufgaben-erstellt>
- Bischof, D. (2023, November 20). Matura abschaffen? Der zeitlose Traum der SPÖ. *Die Presse*, 2023. <https://www.diepresse.com/17844578/matura-abschaffen-der-zeitlose-traum-der-spo>
- BMBWF (Ed.). 11. *Verordnung: Vorbereitung und Durchführung abschließender Prüfungen für das Schuljahr 2020/21*.
- BMBWF. (2021b). Nationaler Bildungsbericht Österreich 2021. <https://www.bmbwf.gv.at/Themen/schule/bef/nbb.html>
- BMBWF. (2024a). Berufsbildende mittlere und höhere Schulen. <https://www.bmbwf.gv.at/Themen/schule/schulsystem/sa/bmhs.html>
- BMBWF. (2024b). Centralised Matura. https://www.bmbwf.gv.at/en/Topics/school/ev_sl/cm.html
- 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., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the

- Most Common Educational Measure. *Review of Educational Research*, 86(4), 803-348.
<https://doi.org/10.3102/0034654316672069>
- Cheung, T. K. (2008). An assessment blueprint in curriculum reform. *Journal of Quality School Education*, 5, 23–37.
- Cliffordson, C. (2008). Differential Prediction of Study Success Across Academic Programs in the Swedish Context: The Validity of Grades and Tests as Selection Instruments for Higher Education. *Educational Assessment*, 13(1), 56–75. <https://doi.org/10.1080/10627190801968240>
- Coenen, J., Cornelisz, I., Groot, W., van den Maassen Brink, H., & van Klaveren, C. (2018). Teacher Characteristics and their Effects on Student Test Scores: A systematic Review. *Journal of Economic Surveys*, 32(3), 848–877. <https://doi.org/10.1111/joes.12210>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.
<https://doi.org/10.1037//0033-2909.112.1.155>
- Cross, L. H., & Frary, R. B. (1999). Hodgepodge Grading: Endorsed by Students and Teachers Alike. *Applied Measurement in Education*, 12(1), 53–72. https://doi.org/10.1207/s15324818ame1201_4
- Duckworth, A. L., Shulman, E. P., Mastronarde, A. J., Patrick, S. D., Zhang, J., & Druckman, J. (2015). Will not want: Self-control rather than motivation explains the female advantage in report card grades. *Learning and Individual Differences*, 39, 13–23. <https://doi.org/10.1016/j.lindif.2015.02.006>
- Erdost, I., & Cavallar, G. (2023, June 14). Weg mit dieser Matura? Zwei Positionen. *Der Standard*, 2023. <https://www.derstandard.at/story/3000000174469/weg-mit-dieser-matura-zwei-positionen>
- European Commission. (2024). Eurydice. European Union.
<https://eurydice.eacea.ec.europa.eu/national-education-systems/germany/overview>
- Ferman, B., & Fontes, L. F. (2022). Assessing knowledge or classroom behavior? Evidence of teachers' grading bias. *Journal of Public Economics*, 216, 104773.
<https://doi.org/10.1016/j.jpubeco.2022.104773>
- Gallagher, A. (1998). Gender and Antecedents of Performance in Mathematics Testing. *Teachers College Record*, 100(2), 297–314. <https://doi.org/10.1177/016146819810000204>
- Gsenger, P. (2019). Zentralmatura Mathematik. <https://www.ahs-aktuell.at/zentralmatura-mathematik/>
- Guez, A., Peyre, H., & Ramus, F. (2020). Sex differences in academic achievement are modulated by evaluation type. *Learning and Individual Differences*, 83-84, 101935.
<https://doi.org/10.1016/j.lindif.2020.101935>
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Praeger.
- Hansen, H. (2011). Rethinking certification theory and the educational development of the United States and Germany. *Research in Social Stratification and Mobility*, 29(1), 31–55.
<https://doi.org/10.1016/j.rssm.2011.01.003>
- Heyneman, S. P. (1988). Improving university selection, educational research, and educational management in developing countries: The role of examinations and standardized testing. In S. P. Heyneman & I. Fagerlind (Eds.), *University Examinations and Standardized Testing: Principles, Experience, and Policy Options* (Vol. 78, pp. 179–216). World Bank.

- Heyneman, S. P., & Fagerlind, I. (Eds.). (1988). *University Examinations and Standardized Testing: Principles, Experience, and Policy Options*. World Bank.
<https://files.eric.ed.gov/fulltext/ED328559.pdf>
- Hoffmann, L., Schröter, P., Groß, A., Schmid-Kühn, S. M., & Stanat, P. (Eds.). (2022). *Das unvergleichliche Abitur: Entwicklungen - Herausforderungen - Empirische Analysen // Das unvergleichliche Abitur: Entwicklungen - Herausforderungen - Empirische Analysen*. wbv Media; wbv. <https://doi.org/10.3278/9783763972494>
- Hübner, N., Wagner, W., Hochweber, J., Neumann, M., Nagengast, B., Kendeou, P., & Graham, S. (2020). Comparing Apples and Oranges: Curricular Intensification Reforms Can Change the Meaning of Students' Grade. *Journal of Educational Psychology*, 112(1), 204–220.
<https://doi.org/10.1037/edu0000351>
- The jamovi project. (2022). jamovi (Version 2.3) [Computer software]. <https://www.jamovi.org>
- Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education*, 28(1), 91–105. <https://doi.org/10.1080/02671522.2012.754229>
- Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy & Practice*, 28(3), 212–227. <https://doi.org/10.1080/0969594X.2021.1884041>
- Köller, O., & Baumert, J. (2002). Das Abitur – immer noch ein gültiger Indikator für Studierfähigkeit? *Aus Politik Und Zeitgeschichte*, 52(B26), 12–19.
- Kultusministerkonferenz. (2024). Hochschulreife über berufliche Bildungsgänge.
<https://www.kmk.org/themen/berufliche-schulen/hochschulzugang-ueber-berufliche-bildung.html>
- McMillan, J. H. (2003). Understanding and Improving Teachers' Classroom Assessment Decision Making: Implications for Theory and Practice. *Educational Measurement: Issues and Practice*, 22(4), 34–43. <https://doi.org/10.1111/j.1745-3992.2003.tb00142.x>
- Nagy, G [G.]. (2007). Berufliche Interessen, kognitive und fachgebundene Kompetenzen: Ihre Bedeutung für die Studienfachwahl und die Bewährung im Studium.
- Neuhauser, J. (2019, January 28). Die Mathematik-Matura - eine Wellenbewegung. *Die Presse*.
<https://www.diepresse.com/5568524/die-mathematik-matura-eine-wellenbewegung>
- Neumann, M., Nagy, G [Gabriel], Trautwein, U [Ulrich], & Lüdtke, O [Oliver] (2009). Vergleichbarkeit von Abiturleistungen: Leistungs- und Bewertungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten und die Rolle zentraler Abiturprüfungen. *Zeitschrift Für Erziehungswissenschaft*, 12(4), 691–714. <https://doi.org/10.1007/s11618-009-0099-6>
- Neumann, M., Trautwein, U [Ulrich], & Nagy, G [Gabriel] (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>
- Nimmervoll, L. (2024, May 2). Ist die Matura reif für die Abschaffung? *Der Standard*.
<https://www.derstandard.at/story/3000000218092/ist-die-matura-reif-f252r-die-abschaffung>
- Nowruzi, M., Amerian, M., Yazdani, H., & Mohammadi, A. (2020). The Relationship between Graded and Tested Achievement: Do Gender and Proficiency Level Make a Difference? *Applied Research on English Language*, 9(4), 477–502. <https://doi.org/10.22108/are.2020.120578.1516>
- Ouazad, A. (2008). Assessed by a Teacher Like Me: Race, Gender, and Subjective Evaluations.

- Pilcher, J. K. (1994). The Value-Driven Meaning of Grades. *Educational Assessment*, 2(1), 69–88. https://doi.org/10.1207/s15326977ea0201_4
- Statistik Austria. (2024a). Reife- und Diplomprüfungen. <https://www.statistik.at/statistiken/bevoelkerung-und-soziales/bildung/reife-und-diplompruefungen>
- Statistik Austria. (2024b). Schüler:innen. <https://www.statistik.at/ueber-uns/aufgaben-und-grundsaeetze/rechtsgrundlagen/impressum>
- Statistik Austria. (2024c). sRDP AHS und BHS Gesamtbeurteilung nach Bundesland Haupttermin 2023. <https://www.statistik.at/statistiken/bevoelkerung-und-soziales/bildung/reife-und-diplompruefungen>
- Stenlund, T., Eklöf, H., & Lyrén, P.-E. (2016). Group differences in test-taking behaviour: an example from a high-stakes testing program. *Assessment in Education: Principles, Policy & Practice*, 1(24), 4–20. <https://doi.org/10.1080/0969594X.2016.1142935>
- Tierney, R., Simon, M., & Charland, J. (2011). Being Fair: Teachers' Interpretations of Principles for Standards-Based Grading. *The Educational Forum*, 75(3), 210–227. <https://doi.org/10.1080/00131725.2011.577669>
- Watermann, R., Nagy, G [G.], & Köller, O. (2013). Mathematikleistungen in allgemein bildenden und beruflichen Gymnasien. 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). Springer-Verlag.
- 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>
- Wilson, D. B. (2023). Practical meta-analysis effect size calculator (Version 2023.11.27). <https://www.campbellcollaboration.org/escalc/equations.php#ordinal-distribution-by-group-using-cell-frequencies>

9.2 Second Study

Self-Selection Bias in the Sensitive Topic of Teacher Grading: The Role of Subject and School Context

Schatz-Laliberté, V., Zviagintsev, R., Kampa, N. (2026). Self-Selection Bias in the Sensitive Topic of Teacher Grading: The Role of Subject and School Context. *Educational Assessment, Evaluation and Accountability*.

Schatz-Laliberté, Vincent^{1*} (ORCID 0009-0009-3686-4656)

Zviagintsev, Roman¹ (ORCID 0000-0002-4547-7748)

Kampa, Nele¹ (ORCID 0000-0001-5633-9335)

¹ Centre for Teacher Education, Department for Teacher Education, University of Vienna, Austria

Status of the study

Accepted for publication in January 2026

Contribution of the individual authors

Vincent Schatz-Laliberté: Conceptualization, Data curation, Quantitative and qualitative data analysis, Writing – original draft, review & editing

Roman Zviagintsev: Quantitative data analysis, Critical Review

Nele Kampa: Critical Review

Self-Selection Bias in the Sensitive Topic of Teacher Grading: The Role of Subject and School Context

Vincent Schatz-Laliberté, Roman Zviagintsev, Nele Kampa, Gavin TL Brown

Abstract

Self-selection bias in voluntary questionnaire surveys poses a threat to the validity of research findings, particularly when studies address sensitive topics. This paper reports on an unexpected case of self-selection bias encountered in research on teachers' grading beliefs and practices in Austria. Originally designed to examine the relationship between teachers' grading beliefs and their actual grading behavior, the study revealed systematic differences between respondents and nonrespondents. Drawing on grade data from 154 teachers, linked with teacher questionnaire responses ($n = 56$) and nonresponses ($n = 98$), we found that survey respondents showed significantly different grading patterns compared to nonrespondents. Furthermore, school type (academic vs. vocational upper secondary schools) emerged as a key moderating factor. A subsequent qualitative study with 14 teachers from the nonresponding subgroups (mathematics teachers from both academic and technically orientated vocational schools) provided further insights into the mechanisms underlying this bias, including perceived low value of research topic for nonrespondents, school type specific teacher mentality, reluctance to disclose potentially questionable grading practices, and professional reputation concerns. By documenting this finding, the paper highlights the role of institutional factors (i.e., school type) and subject area in self-selection bias and emphasizes the need for more explicit consideration of such bias in research on teachers, especially when addressing sensitive topics like grading.

Keywords: self-selection bias, nonresponse, teacher grading, school type, sensitive educational topics

1 Introduction

Self-selection bias in voluntary questionnaire surveys poses a threat to research validity (Bowden, 1986) because the results reflect only the responses of a specific subgroup of participants and are therefore not representative (Groves et al., 1992). Besides reasons for nonresponse in questionnaire-based surveys in general (e.g., personality traits; Korkeila et al., 2001; structure and length of the survey; Saleh & Bista, 2017), self-selection bias is a particular risk when research concerns sensitive topics (Tourangeau et al., 2010). In

educational research, literature on self-selection bias is scarce (Guillemot et al., 2025), although high nonresponse rates in questionnaire surveys among adult participants (i.e., teachers and university students) are reported (Meinck et al., 2017). The present study contributes to this area by reporting the case of self-selection bias in the sensitive topic of teacher grading. It did not result from a deliberate investigation of self-selection bias; rather, the findings were discovered accidentally during a robustness check. Accordingly, this paper provides a transparent account of how research originally designed to examine teachers' grading practices and beliefs developed into a methodological contribution to research on self-selection bias in educational research.

2 Theoretical Background

2.1 Self-Selection Bias in (Educational) Research

The phenomenon of self-selection bias, also referred to as nonresponse bias (Berg, 2005), has been identified as a particularly salient issue in the context of questionnaire-based research (Bowden, 1986). It commonly occurs when respondents' likelihood of participating in the survey is linked to the topic being investigated. In such cases, the validity of data may be threatened because those who choose to respond differ systematically from those who do not, resulting in a sample that does not accurately represent the target population (Olsen, 2008).

As primary reasons for nonresponse in general, underlying sociodemographic and behavioral characteristics, the length and potentially sensitive content of the questionnaire, concerns about providing written consent, and the possibility of being personally linked to the data (Korkeila et al., 2001). In the context of educational research, literature describes that the response rate to surveys is strongly affected by participants' interests, the design and format of the survey, the ways in which communication is carried out, and (again) the level of trust in the protection of privacy and confidentiality (Saleh & Bista, 2017).

Self-selection bias has been widely studied in fields such as medicine (e.g., Pahwa et al., 2012; Robinson et al., 2023) and psychology (e.g., Haynes & Robinson, 2019; Dawson et al., 2019). However, research on self-selection bias in education is still scarce (Guillemot et al., 2025). Furthermore, Meinck et al. (2017) found that adults, such as teachers and university students, tend to have particularly low response rates in surveys, compared to school students who generally demonstrate a greater willingness to participate or may be required to do so under school regulations, making their participation less voluntary. The authors examined nonresponse patterns in the international educational studies TIMSS (Trends in International

Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study). They also found that schools (as distinct, autonomous entities) are at a higher risk of nonresponse than individual participants within schools. This highlights the importance of considering not only the individual characteristics of participants, but also the systemic levels when accounting for the possibility of self-selection bias in questionnaire-based educational research.

2.2 Self-Selection Bias in Research Addressing Sensitive Topics

The risk of self-selection bias particularly exists when research concerns sensitive topics, such as sexual attitudes and behavior (Dunne et al., 1997) or politics (Tourangeau et al., 2010).

Three types of sensitive questions in questionnaire-based research are outlined by the literature (Peytcheva & Yan, 2025): The first type of question is perceived as intrusive, involving topics that invade personal privacy and are seen as inappropriate in everyday conversation. The second type involves potential personal harm if the response is shared. The third type is very similar to second type, but relates more specifically to social desirability, where individuals may feel compelled to give answers that align with social norms rather than their true opinions or behaviors (Krumpal, 2013).

In the field of education, Guillemot et al. (2025) provide an interesting example of self-selection bias concerning social desirability. In their study on teachers' attitudes toward inclusive education, they found that responding teachers were more likely to work in schools where parents reported their children to feel happier, more socially included, and experienced less bullying. In contrast, teachers from schools where inclusive education posed greater challenges were less inclined to participate. Based on these findings, the authors did not only raise concerns about a potential positivity bias in previous research on the topic, but they also point out the limited number of studies that explicitly address self-selection bias among teachers in surveys on sensitive educational issues. Our study responds to these concerns as it contributes to the research on self-selection bias in educational contexts by examining its potential impact on studies of teachers' attitudes toward grading.

Based on our theoretical framework, teachers' attitudes toward grading are considered a sensitive topic because they represent all three types of sensitive questions. First, questions about grading can feel intrusive, as they touch on personal beliefs, values and professional judgment (McMillan, 2003; Pilcher, 1994). Second, teachers may fear potential harm if their responses—especially those contradicting with official grading guidelines—are shared with

authorities. Additionally, grading often involves navigating moral dilemmas and situations in which teachers may feel compelled to act against their personal ethical standards (Galevska, 2019). Third, grading is shaped by strong norms about fairness and objectivity; while Sætre (2025) offers a useful illustration of these dynamics in higher education, similar pressures may also lead upper secondary teachers to respond in socially desirable ways rather than describe their actual practices.

To assess nonresponse bias, Olsen (2006) summarizes four main approaches described in previous literature, three of which are applicable for questionnaire-based research. First, researchers compare the demographic characteristics of survey respondents with those from a reliable external source to identify any under- or overrepresented groups. Second, they use frame information available for both respondents and nonrespondents to detect systematic differences between the two groups. Third, they design specific experiments to investigate response rates across groups known to differ in survey outcomes, helping to reveal how nonresponse might distort results. As the results section will show, this paper reveals self-selection bias in a questionnaire survey on the sensitive topic of grading in a way similar to the third approach described above (Olsen, 2006). However, this was not the result of a well-thought-out experiment, but rather an unintentional discovery.

2.3 The Unexpected Finding of Self-Selection Bias in the Sensitive Topic of Grading

To ensure transparency, a brief description of the initially planned research that accidentally revealed self-selection bias including initial theoretical background, aims and study background is provided here.

The initial aim of our research was to investigate whether teachers' beliefs regarding non-academic achievement factors in grading can predict their tendency to award higher classroom grades compared to standardized test scores⁶. Research has shown that teachers' classroom grades and standardized test scores assess different aspects of student achievement (Willingham et al., 2002). Teacher grades often incorporate both academic and non-academic factors, such as effort, attendance, and behavior, offering a broad perspective on student performance (Brookhart et al., 2016; Cross & Frary, 1999; Bowers, 2011). However, including certain non-academic factors reported in the literature is considered an unorthodox, yet common, grading practice (Bonner & Chen, 2021), as they may conflict with official

⁶ An analysis of variance (ANOVA) was planned using teachers' grading beliefs as independent variables and their tendency to award higher classroom grades than standardized Matura grades – measured by Cohen's *d* as standardized discrepancy between the two grades – as dependent variable.

grading guidelines. In Austria, where this study is situated, official guidelines explicitly state that student behavior must not be considered in any aspect of performance evaluation (see §11(5) of RIS, 2025).

While standardized tests assess student performance based on predetermined criteria, teachers' classroom grading also reflects their personal beliefs about what grades should measure (Annerstedt & Larsson, 2010; McMillan, 2003). Teachers strive to ensure that their grading decisions are consistent with these beliefs (McMillan & Nash, 2000; Pilcher, 1994), sometimes resulting in challenging tension between their ideal grading practices and the realities imposed on them by external demands, such as grading policies or negotiations with parents (McMillan, 2003).

Further, some researchers posit that teachers' beliefs are also tied to their broader institutional context (Fuller & Izu, 1986; Goddard, 2003; Rubie-Davies et al., 2012). It has been argued that teachers tend to adopt informal, implicit school-specific norms—referred to as *cultural beliefs* (Fuller & Izu, 1986)—that shape their attitudes and behaviors. This process of socialization is described as the subtle alignment of teachers with their colleagues' beliefs, resulting in unified practices within a given school (Fuller & Izu, 1986).

With respect to our focus on grading, school type has been shown to be a predictor of grade differences between teacher grades and standardized test scores on an institutional level (Doz, 2023; Köller et al., 1999; Watermann et al., 2013). In the Austrian context, school type (academic versus vocational upper secondary schools) similarly functions as an important institutional predictor of differences in grading practices (Authors, in review). The underlying mechanisms for this differential grading, however, have not been sufficiently explored (Doz, 2023).

In this context, our initial research addressed this lack of evidence by examining the extent to which teachers' grading beliefs are linked with their actual grading behavior, both at the individual and school type levels. We particularly focused on teachers' grading beliefs regarding the “unorthodox” non-academic achievement factors described by Bonner and Chen (2021; see the ‘Methods’ section and supplementary material for details: Appendix A describes the adaptation and validation of the scale, and Appendix B its German adaptation) as potential mechanisms behind the differences between teacher grades and standardized test scores.

During the data analysis, however, a self-selection bias became evident: respondents displayed significantly different grading patterns than nonrespondents. In addition, participation patterns varied significantly between teachers from different school types.

2.4 Novelty of the Present Study

To our knowledge, no prior study has examined self-selection bias, nor its relation to institutional factors such as school type, in educational research on teacher grading. By documenting this unexpected finding in teachers' response behavior, the present study aims to provide insights into the mechanisms that may lead to self-selection bias among different subgroups of teachers, with particular focus on school type and subject area. This knowledge can inform future research on teacher grading to better anticipate and account for potential biases threatening sample representativeness. As a basis for interpreting our results, the following section outlines the Austrian context of teacher grading.

3 The Austrian Context

Austrian upper secondary education has a distinctive grading system. In our study, we focus on teachers' grading in the final year of upper secondary education, during which teachers play a dual role. On the one hand, teachers grade students on their classroom performance in their final school year. On the other hand, teachers also grade the Matura, the Austrian standardized school-leaving examination, which serves as a gateway to higher education.

The Matura in Austria is a standardized exam that assesses German, mathematics, and English as a foreign language. The Matura exam is developed centrally at the federal level so that all students, regardless of school, school type, or federal state, are given standardized exam tasks. Teachers are provided with standardized assessment criteria by the Ministry of Education to increase the comparability of grades. However, the grading of the standardized tasks is not anonymous and is the responsibility of the students' own teachers without moderation by colleagues or external bodies, ultimately weakening the degree of objectivity and comparability of the Matura grades.

In addition, a grading reform in 2020 fundamentally changed the composition of the final Matura grades. Prior to 2020, only the results of the standardized Matura exams were considered for the Matura certificate (the national university entrance qualification). Since 2020, students' classroom grades from the final school year contribute 50% to the Matura certificate, with the remaining 50% based on the standardized Matura exams. This new

grading system again called into question the intended standardized nature of the Matura (Schatz-Laliberté & Lüftenegger, 2025).

In the Austrian secondary school system, students can choose between attending an academic secondary school (Allgemeinbildende Höhere Schule, AHS) and a vocational secondary school (Berufsbildende Höhere Schule, BHS) to qualify for university entrance. AHS prepares students for university studies and ends with Matura graduation. BHS focuses on vocational training, but students also graduate with the Matura, thereby obtaining the same university entrance qualifications as AHS students. BHS schooling lasts one year longer (5 years) than AHS schooling does (4 years).

The AHS and BHS school contexts differ substantially. While AHS schooling primarily focuses on academic interests and prepares students for later university studies, BHS schooling offers special vocational training that prepares students for the labor market (in addition to the possibility of pursuing university studies). Austria is home to a variety of BHS types, each with a different focus, including for example, technical schools, schools for economic professions, fashion, agriculture, federal sports, or music academies (Bundesministerium für Bildung, 2025a). Another difference between AHS and BHS lies in the students' educational backgrounds. While many AHS students already pursued an academic track in lower secondary school (with entry requirements in core subjects), many BHS students attended a non-academic secondary school. Sixty-two percent of the students from lower academic tracks enter AHS, whereas only 33% enter BHS. In contrast, only 9% of non-academic lower secondary graduates enter AHS, but 35% of non-academic lower secondary graduates enter BHS (BMBWF, 2021).

Further, AHS and BHS differ in their Matura examinations in mathematics and English, and the various types of BHS also show distinctions in this regard. Unlike the AHS Matura in mathematics, which focuses more strongly on basic competencies, the BHS Matura—referred to as Applied Mathematics—consists of two components: one assessing general knowledge, identical for all students, and another assessing profession-specific knowledge, which varies across tracks (SRDP, 2025). Also, in some BHS tracks, students may choose to take mathematics or English orally rather than in written form. However, these oral examinations are designed by the students' own teachers and are therefore not standardized, even though they still count as equivalent standardized university entrance exams (Bundesministerium für Bildung, 2025b).

Unlike several European countries, such as Ireland (Lehane et al., 2025), Italy (Autonomous Province of Bolzano South Tyrol, 2025), or France (Ministère de L'Education Nationale, 2025), where external bodies are involved in or are responsible for assessing final grades, Austria has a different model. In the Austrian system, students' own teachers are responsible for grading both the final classroom grades and the standardized Matura exams. While variation in classroom grades is well documented (e.g., Brookhart et al., 2016), variation in the grading of standardized exams should be minimal if the Ministry's goal of ensuring comparable Matura grades is to be achieved. Although the Ministry of Education provides a uniform assessment framework, the grading process is neither anonymized nor subject to internal or external moderation.

4 Methods and Results

This paper employs a sequential explanatory mixed-methods design (Ivankova et al., 2006), consisting of a quantitative analysis of self-selection bias in questionnaire-based research on teachers' grading (Study 1), followed by qualitative interviews with teachers (Study 2). In Study 1, we document the presence of self-selection bias in teachers' responses, showing that teachers who completed the questionnaire on grading beliefs differed significantly in their grading behavior from those who did not. In Study 2, teachers representing the target group of nonrespondents are interviewed to explore their perspectives on the patterns identified in the quantitative findings of Study 1.

4.1 Study 1

4.1.1 Sample

We collected matched grade pairs (classroom and Matura grades) from a total of 8,581 students in three AHS and four BHS schools in Vienna, Austria. These grades were awarded by 154 teachers over seven years from 2017 to 2023 in three subjects: German, mathematics, and English as a foreign language. The full dataset includes 3,505 grade pairs for German, 3,293 for mathematics, and 1,783 for English. On average, each teacher graded approximately 56 students, with large differences between teachers ($SD = 38$). Sixty percent of the teachers were women, 24% were men, and 16% provided no information on their gender. The average age of the teachers was $M = 47$ ($SD = 10$).

In 2023, the same teachers were asked to respond to a questionnaire about the Matura in general and on their beliefs about non-academic achievement factors in grading (Bonner &

Chen, 2021). While we received the grade data directly from the schools, participation in the questionnaire was voluntary. Of the 154 teachers who were invited to participate, 36% (56 teachers) completed the questionnaire. Response rates differed slightly by school type, with 48% (25 out of 52) of AHS teachers and 30% (31 out of 102) of BHS teachers completing the questionnaire. Among the respondents, 43% teach in AHS schools, whereas 57% teach in BHS schools.

4.1.2 Variables

Dependent variable

As the dependent variable, we used Cohen's d as a standardized effect size measure of the discrepancy between teacher-awarded classroom and Matura grades at the teacher level. Cohen's d was calculated for each teacher and each academic year separately. A positive Cohen's d indicates that, on average, a teacher awarded higher classroom grades than Matura grades, whereas a negative value indicates the opposite (higher Matura grades compared to classroom grades). Thus, this measure describes the tendency of teachers to be biased toward either classroom grading or standardized Matura exam grading. The higher the positive or negative value of Cohen's d , the greater this tendency.

Independent variables

School type served as an independent variable. We also included school subject (German, mathematics, and English), year (2017 to 2023), and teachers' gender as additional control variables.⁷

4.1.3 Questionnaire

Besides general questions regarding the teachers' views about the importance of the Matura⁸, the value and validity of final grades, we employed the "Survey of Unorthodox Grading Beliefs for Teachers and Teacher Candidates" by Bonner and Chen (2021) to assess the teachers' grading beliefs related to non-academic achievement dimensions of grading: *effort* (a belief referring to account for students' effort and hard work when assigning grades, associated with a constructivist view of learning as an experiential process, rather than solely an outcome; Bonner & Chen, 2009), *manage* (drawing upon the behaviorist logic of

⁷ In our initially planned research, teachers' grading beliefs (described in the subsequent section *Questionnaire*) were also included as independent variables.

⁸ E.g., "In my opinion, the grades for the written high school graduation exams in my subject are generally accurate", or "In my opinion, the grades for the written high school graduation exams in my subject are generally fair".

reinforcing students' appropriate behavior by using grades as a tool for classroom management and discipline), *pull* (a success-oriented grading belief which contrasts with behaviorist views of learning), and *lenient* (a belief reflecting support for more lenient grading practices, without being tied to a specific educational objective). Since the original questionnaire was in English, we adapted the scale into German language following the five-step process recommended by Borsa et al. (2012). The complete adaptation and validation of the German version of the questionnaire are provided in Appendix A in the supplementary material.

4.1.4 Robustness Check: Test on Self-Selection Bias

To test for self-selection bias among the respondents, we analyzed whether the 56 teachers who completed the questionnaire constituted a representative sample of the full dataset of 154 teachers. First, we added a binary variable (*missing/not missing*), indicating each teacher's response status. We then ran an ANOVA to test whether this new variable was connected to our dependent variable or served as a moderator⁹. Cohen's *d* served as the dependent variable, and year, school type, subject, and the interaction term of school type and response status (*missing/not missing*) served as the independent variables. This approach determined (a) whether respondents differed from nonrespondents regarding their grading behavior and (b) whether there was an interaction between school type and response status.

4.1.5 Results

Besides the significant main effects of year and school type, the results revealed a statistically significant effect of the interaction between school type and response status (Table 1), indicating the presence of selection bias. The effect of this interaction was rather large and significant, suggesting that response status was systematically related to the differences in teachers' grading behavior across school types.

⁹ This procedure was necessary because we originally aimed to investigate the difference in grading beliefs between school types, which was in question after the results from our previous study (Authors, in review).

Table 1. Test for selection bias: Prediction of the difference between classroom and Matura grades (Cohen's d) by year, school type, subject, and the interaction term of school type, and response status

	df	Sum of Squares	Mean Square	F	p	η_p^2
Overall model	11	8.731	0.974	7.90	< .001	
Year	6	3.563	0.594	4.56	< .001	0.065
School type	1	1.263	1.263	9.70	0.002	0.024
Subject	2	0.789	0.394	3.03	0.049	0.015
NOT missing	1	1.154	0.154	1.18	0.278	0.003
School type*NOT missing	1	2.963	2.963	22.76	< .001	0.055
Residuals	391	50.897	0.130			

Note. df = degrees of freedom, F = F statistic, p = p value, η_p^2 = partial Eta squared.

Hence, teachers who responded to the questionnaire on grading beliefs exhibited significantly different grading behavior than the nonresponding teachers. To further explore this phenomenon, a post hoc test was performed (Table 2).

Table 2. Post hoc comparisons between the two school types and their interactions with missing questionnaire responses

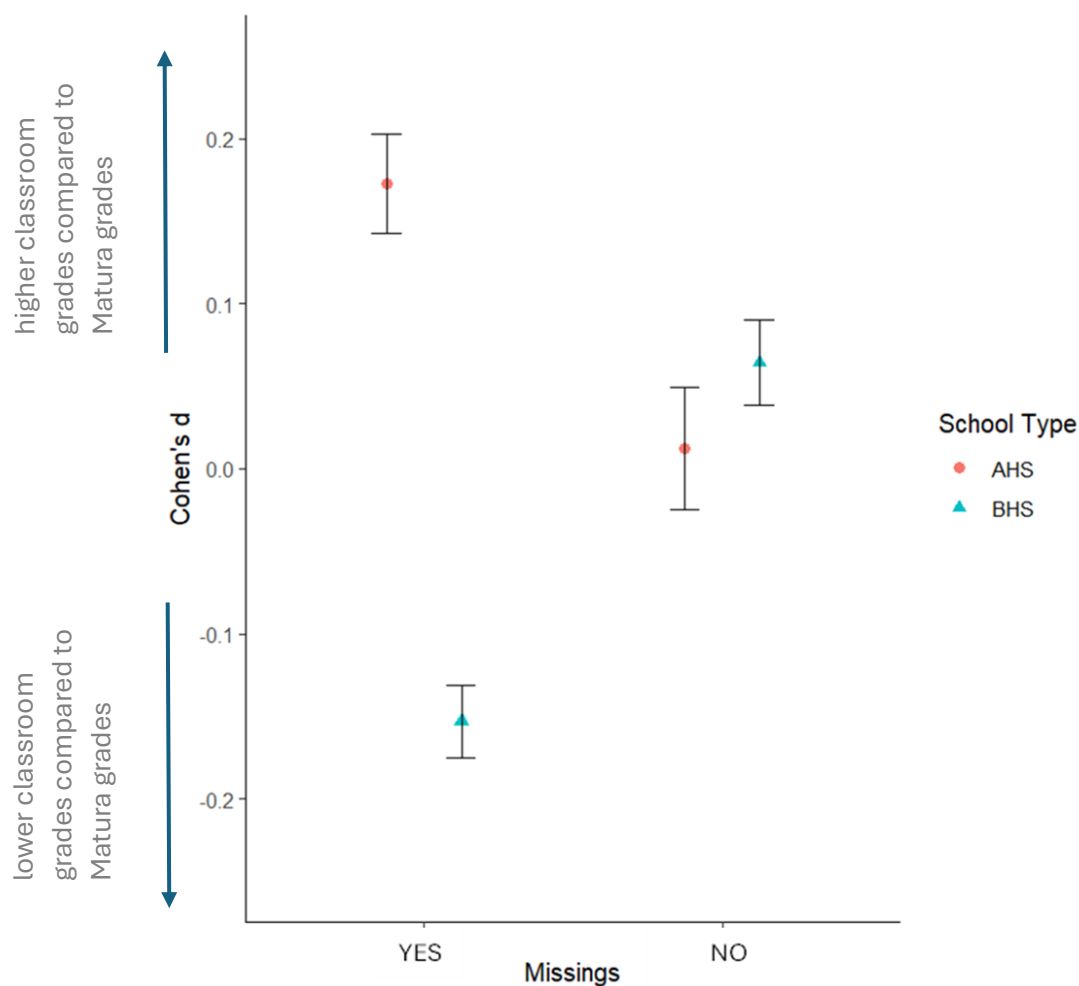
Comparison				Mean difference	SE	df	p_{tukey}
School type	Missing	School type	Missing				
AHS	yes	AHS	no	0.1463	0.0654	391	0.115
	yes	BHS	yes	0.3147	0.0475	391	<.001
	yes	BHS	no	0.0812	0.0529	391	0.417
	no	BHS	yes	0.1684	0.0603	391	0.028
	no	BHS	no	-0.0652	0.0642	391	0.741
BHS	yes	BHS	no	-0.2335	0.0459	391	<.001

Note. SE = standard error of the mean difference; df = degrees of freedom; p_{tukey} = Tukey's p value.

Table 2 provides insights into the interaction between school type and response status. For AHS teachers, no statistically significant difference was found in grading behavior between respondents and nonrespondents. However, AHS teachers who did not respond exhibited significantly different grading behavior than BHS teachers who did not respond. Furthermore, AHS teachers who responded differed significantly from BHS teachers who did not respond. For BHS teachers, a statistically significant difference emerged between those who responded

and those who did not. In other words, teachers who decided to respond to the questionnaire showed similar grading behavior, irrespective of their school type. Conversely, teachers who opted not to answer demonstrated significantly different grading behavior, with notable discrepancies observed between AHS and BHS schools. Figure 1 (created in R with ggplot2; Wickham, 2016) illustrates this observed discrepancy (model-predicted scores averaged at the school type level, full dataset).

Figure 1. Self-Selection bias in questionnaire responses



Note. Cohen's d = standardized difference in grade distribution between Matura and classroom grades; AHS = academic upper secondary schools; BHS = vocational upper secondary schools.

Figure 1 illustrates that respondents (*Missing* = NO) awarded Matura grades that were highly aligned with classroom grades. The Cohen's d values were close to zero. In contrast, nonrespondents (*Missing* = YES) exhibited a greater discrepancy between the two grades, with AHS and BHS teachers showing opposite grading behavior to each other: AHS teachers

who did not respond tended to be biased toward higher classroom grading, whereas BHS nonrespondents awarded significantly lower classroom grades than Matura grades.

Examining the data revealed that the significant effect of BHS nonrespondents was driven solely by mathematics teachers at technically oriented BHS schools. The average Cohen's d values showed high negative values for both BHS respondents ($d = -0.35$) and nonrespondents at technically oriented BHS schools ($d = -0.30$) in mathematics. At non-technical BHS schools, the Cohen's d values of both respondents ($d = 0.04$) and nonrespondents ($d = 0.11$) in mathematics were positive and small. Thus, within both technical and non-technical BHS, responding mathematics teachers did not differ from nonrespondents in their grading behavior. German and English teachers in BHS showed a high degree of alignment between classroom and Matura grades, with no significant differences between respondents and nonrespondents.

Although not statistically significant, the effect of AHS nonrespondents was also largely driven by mathematics teachers. In AHS, nonrespondents showed greater grade discrepancies ($d = 0.29$) than responding mathematics teachers ($d = 0.14$). Like in BHS, AHS language teachers showed only marginal differences between classroom and Matura grades ($d_{\text{German}} = 0.12$, and for English, $d_{\text{English}} = -0.02$), with no differences between respondents and nonrespondents regarding the dependent variable.

Regarding response rate, while only 20% of the teachers from technically oriented BHS (independent of subject area) completed the questionnaire, 44% of the teachers at non-technical BHS and 42% of the teachers at AHS did so. A table summarizing Cohen's d values and response rates for school subtypes and subject area is provided in Appendix D in the supplementary material.

4.2 Study 2

4.2.1 Methods

Due to the lack of theoretical evidence that could explain the unexpected interaction effect observed in Study 1, we conducted a follow-up qualitative study (Study 2) to explore potential underlying mechanisms. Study 2 consisted of semi-structured interviews with 14 participants recruited through purposive sampling from the researchers' professional networks. Seven participants were nonrespondents of Study 1, and the other seven were selected because they belonged to the subgroups that largely drove the observed self-selection bias: mathematics teachers from AHS and technically oriented BHS. In the interviews we

asked them to explain 1) why technically oriented BHS mathematics teachers tend to give significantly higher Matura grades than classroom grades, while in nontechnical BHS, students generally receive similar Matura and classroom grades, 2) why technically oriented BHS teachers are much more likely than teachers in other BHS types not to complete a questionnaire on grading, and 3) why AHS mathematics teachers are more likely not to complete a questionnaire when they give higher classroom grades than Matura grades. Although the questions were school-type specific, we asked all participants the same set of questions to not only maintain methodological consistency, but also to capture alternative perspectives from outside of a teacher's current institutional context.

Besides the basic inclusion criteria for participants (i.e., relevant school types and the subject of mathematics), participants had to have graded at least four final-year Matura classes within the past 10 years. Although not a mandatory criterion, we aimed to include teachers who also had prior teaching experience in school types other than their current one. Table 3 provides an overview of participating teachers with a short description of their teaching experience. The number of participants was not predetermined; instead, we followed the principle of data saturation. Consequently, interviews continued until no new insights emerged, resulting in a total of 14 interviews, with seven teachers from technically oriented BHS and seven from AHS.

We analyzed the interview data using manifest qualitative content analysis (Graneheim & Lundman, 2004). Since we asked the participants for explicit explanations for grading and response patterns, we concentrated on coding what was directly expressed in the data (Downe-Wamboldt, 1992). Our analysis followed the steps outlined by Kleinheksel et al. (2020). First, we identified meaning units, usually sentences or parts of sentences, which we condensed when needed by removing fillers, repeated words, or simplifying complex phrasing. Second, we assigned codes to equivalent units, and third, we organized related codes into categories to structure the material. Although Kleinheksel et al. (2020) propose a fourth step (integrating categories into overarching themes), they note that this stage is more applicable to latent analyses. Since our focus was on manifest content, we limited our procedure to the first three steps.

Table 3. Participating teachers in qualitative Study 2

ID	Gender	School Type	Teaching Experience	Non-Respondent from Study 1
P01	Male	Technical BHS	16 years at two different technical BHS	Yes
P02	Male	Technical BHS	7 years at technical BHS; before that, 3 years at AHS	Yes
P03	Female	Technical BHS	11 years at technical BHS; before that, 2 years at AHS	No
P04	Male	Technical BHS	8 years at technical BHS; before that, 2 years at AHS	No
P05	Female	Technical BHS	7 years at technical BHS; before that, 3 years at AHS	Yes
P06	Female	Technical BHS	2 years at technical BHS; before that, 25 years at non-technical BHS	No
P07	Male	Technical BHS	Both 6 years at technical and non-technical BHS	No
P08	Female	AHS	7 years at AHS; before that, 12 years at non-technical BHS	Yes
P09	Female	AHS	23 years at AHS	No
P10	Female	AHS	12 years at AHS; before that, 5 years at non-technical BHS	No
P11	Female	AHS	17 years at AHS, with 4 years in technical BHS at between	Yes
P12	Male	AHS	8 years at AHS; intense exchange with colleagues from technical BHS through professional working groups	No
P13	Female	AHS	23 years at AHS	Yes
P14	Female	AHS	13 years at AHS; before that, 7 years at technical BHS	Yes

Note. AHS = academic upper secondary schools; BHS = vocational upper secondary schools.

4.2.2 Results

The results are presented separately for school type. Part A reports findings from AHS teachers, covering (1) their explanations for awarding higher classroom than Matura grades

and (2) their explanations regarding questionnaire (non)response. Part B presents the perspectives of technical BHS teachers, including (3) their explanations for higher Matura than classroom grades and (4) also their views on questionnaire (non)response. While categories (1) and (3) are not directly linked to the topic of self-selection bias, they provide important contextual insights into teachers' grading practices that help to situate and interpret their explanations for (non)response. Appendix C provides a table with all four categories, their codes, code descriptions, and example quotes.

4.2.2.A Categories of Grading and Questionnaire (Non)Response Among Mathematics Teachers in AHS

Category 1: Teachers' Explanations for Higher Classroom than Matura Grades in AHS Mathematics

As a central explanation for higher classroom than Matura grades in AHS mathematics, teachers emphasized the role of students' effort and participation in AHS classroom grades. They highlighted that AHS students are offered both a greater number and a wider variety of opportunities to improve their grades throughout the year, for example through homework, oral contributions, or persistence in class. One teacher explained that "AHS teachers [...] place a great deal of value on other things, such as how much effort a student shows and how much commitment they display during the year" (P02). Another added that "especially in the AHS, there are these girls who are always very diligent, hand in the same homework multiple times, regularly participate, and in this way manage to move up to the Matura. [...] In the HTL [technical oriented BHS], they don't even do homework, but when it counts, they deliver" (P09). Such practices were described as rewarding diligence and engagement rather than reflecting only achievement, which, as one teacher noted, can leave students struggling at the final exam: "These little pats on the back that we [in the AHS] give to every student don't exist in the HTL. [...] In the AHS, the effort made for the weaker students is very pronounced [...] and then, of course, they face the final exam and struggle accordingly" (P13).

Some teachers connected these differences in grading between school types to structural conditions. As one teacher put it, "At the AHS, we have many weekly lessons [...] and have much smaller groups in the core subjects. I can do many oral repetitions, where students can improve their classroom grade. [...] In the BHS, that's not possible; few hours, large classes" (P14). In contrast, teachers explained that in BHS, large class sizes and fewer teaching hours result in classroom grading relying mainly on written assessments, leaving little room to consider participation or individual progress.

Category 2: Teachers' Explanations for Nonresponse in AHS

Nearly all participants highlighted the potential concealment of grading practices that contribute to grade inflation. Several teachers described that colleagues who strategically awarded higher classroom grades to compensate for weaker Matura results might avoid participating to prevent exposure. As one teacher explained, “I don’t want to accuse anyone, but I do think some simply protect themselves. You are actually revealing your teaching style, because if the student already has a good classroom grade, not much can go wrong anymore” (P03). Others noted that final classroom grades often rose sharply in the last months prior to the Matura exams, “with little connection to actual performance” (P09), or that “teachers are well aware they need to balance out with the classroom grade what they may have neglected during lessons [...] this is a loophole in the system” (P04). Such practices were also linked to “fear, so teachers sometimes give the better [classroom] grades to make sure students pass and so they don’t embarrass themselves” (P11).

Some participants also connected conscientiousness in grading with conscientiousness in responding to a questionnaire. They described that teachers whose classroom and Matura grades match closely usually experience a sense of validation and therefore may feel more comfortable in completing a questionnaire on grading. One participant explained, “When classroom grade and Matura grade fit well, then you probably judged the student quite accurately. [...] And these teachers are generally more conscientious, and then they are also more likely to fill out a questionnaire” (P05). Similarly, another noted that those who responded often “had the feeling they had done everything right” (P08). At the same time, some admitted that others might hand out better grades “more thoughtlessly [...] so they have less work” (P11).

Another explanation revolved around teachers’ concerns about their teaching quality being scrutinized. Some respondents believed that colleagues might avoid the questionnaire because participation would make their teaching and grading practices too transparent. One teacher suggested that nonresponse could stem from the realization “that this [classroom grading] actually has little to do with actual subject performance, but rather with diligence and obedience, and then the students perform worse in the Matura, and teachers are aware they may not have prepared their students sufficiently” (P01). Others emphasized the discomfort of having to confront one’s own errors, noting that “this could be a reason not to fill it in, because then I would have to face my mistakes” (P04).

In a similar vein, some teachers addressed the role of shame as a barrier to participation. Nonresponse, they suggested, could result from feelings of embarrassment when Matura results frequently do not match teachers' classroom grading. As one teacher phrased it, "That is shame. They thought: 'Oh dear, the student I gave a top grade to now got a four [lowest passing grade] on the exam; everyone will think I cannot assess properly.' They don't want to show their hand" (P10). Others described the issue as "shame-laden" (P09) or noted that some simply felt "uncomfortable putting their grading beliefs out into the open" (P13).

In contrast, some teachers argued that grading itself was a secondary concern for colleagues, leading to little motivation to participate in a questionnaire on the topic. For these teachers, other aspects of teaching, such as relationships with students, their well-being, and fostering a positive learning environment, were considered more meaningful. As one teacher explained, they were "fed up with constantly having to justify that my grades are too good. The quality of my teaching is measured by how strict my grades are. For me, grades are by far not as important as other things [...] for example my relationship with my students, how they are doing, and that they enjoy the lessons. For such a topic, I would be much more willing to complete a questionnaire" (P10).

4.2.2.B Categories of Grading and Questionnaire (Non)Response Among Mathematics Teachers in Technical Oriented BHS

Category 3: Teachers' Explanations for Higher Matura than Classroom Grades in Mathematics in Technical BHS

As main explanation for higher Matura than classroom grades, teachers emphasized a gap between Matura and classroom demands, describing the standardized exam as relatively easy compared with the higher expectations of technical BHS classrooms. One teacher explained that classroom tests "definitely exceed what is asked at the Matura" and are designed to challenge students with profession-specific content (P01), while another noted that "during the year we cover topics that often go well beyond what is required in the Matura, and this means students sometimes receive lower classroom grades" (P02). Others stressed that "the tasks in our HTL textbooks are simply much more demanding than what is then required at the Matura" (P04) and that "the Matura only covers part of the curriculum" (P11), leading one teacher to conclude that "HTL students breeze through the Matura" (P13).

In this context, teachers highlighted the cross-subject reinforcement of mathematical skills as a factor contributing to strong Matura results. Because mathematics is embedded in other, profession-specific subjects, students practiced and consolidated their skills beyond the

mathematics classroom. One teacher explained that “they are simply good at math. They manage [the Matura] without having to make much effort [...] and they already bring a lot of mathematical competence from other subjects” (P03).

Category 4: Teachers’ Explanations for Nonresponse in Technical BHS

Technical BHS teachers offered three main explanations for nonresponse. A reason pointed out by many teachers was low interest in participating in a study about the Matura and teacher grading beliefs due to the perceived ease of the Matura in mathematics. Teachers noted that because the standardized exam was rarely challenging in their school context, the topic seemed less relevant. One explained, “If I am satisfied with something, I am less interested. And now there are simply no difficulties in math here. [...] You are less interested in studies if you think everything is going well” (P04). Another highlighted that in other types of BHS, teachers face more challenges in mathematics and are therefore more engaged, “but for us the Matura is not a major event in that sense” (P02). Similarly, one participant explained, “Teachers only feel the need to say something if they recognize a problem in a topic. HTL [technical oriented BHS] teachers have nothing to say because the Matura works fine for them.” (P13).

A second explanation focused on the lower perceived value of grades in academic subjects (i.e., mathematics, German and English) compared to profession-specific subjects, such as construction or structural engineering. One teacher stated, “General education subjects also have less importance here than specific subjects, which are important for students’ future professions. The topic of grading in math and a questionnaire is simply secondary [...]; what matters is what they can do in, for example, construction technology or other areas” (P05). Another emphasized that for many HTL students, “the Matura is not so highly valued because they already have a clear vocational path, and the Matura is not the main goal” (P08). A third teacher framed it in terms of priorities: “They [teachers at technical BHS] do not need to worry so much about pedagogy. At the end of the day, they want good engineers. [...] They have a different background, a different goal” (P11).

Finally, some teachers pointed to the mentality of technical BHS teachers as driving nonresponse. One participant noted, “I can well imagine that they wouldn’t fill out a questionnaire. It is clearly a matter of mentality” (P09). Another explained that long-term exposure to a particular school type changes one’s approach: “Principally, HTL teachers have a different mentality than AHS teachers [...]. I taught at an AHS before, and I can say that

when you teach longer at a particular school type, you adapt to its mentality” (P05). A third summarized the pragmatic orientation: “We HTL teachers are very pragmatic and application-oriented [...] In the priority list, a questionnaire comes rather far down” (P07).

5 Discussion

Our study uncovered systematic self-selection bias in research on teachers’ grading. Respondents and nonrespondents differed significantly in their grading patterns, with school type emerging as a crucial moderating factor. Complementary qualitative insights into teachers’ perceptions of differences in grading between school types, as well as their explanations for the discovered effect provide an initial understanding of the potential mechanisms underlying this bias. Our findings point to several methodological considerations for the detection and interpretation of self-selection bias in educational research. First, institutional factors must be taken into account (Guillemot et al., 2025). However, our results showed that this needs to be done in a differentiated way. A simple comparison of respondents and nonrespondents using a binary variable of response status (missing vs. not missing) did not reveal significant differences. Only when moderated by school type did self-selection bias become apparent, as the opposing effects across school types (higher classroom grades relative to Matura results in AHS and lower classroom grades in BHS) would have cancelled each other out. Future research on sensitive topics such as teachers’ grading practices and beliefs should therefore gather as much information as possible of both respondents and nonrespondents, as otherwise, possible self-selection bias will lead to erroneous results.

This study also emphasizes the importance of thoroughly examining the data when attempting to explain the presence of self-selection bias. Our data analysis showed that the effect was driven solely by a specific group of the sample: teachers in technically oriented BHS, but not by teachers from other BHS sub types. Additionally, our research showed the importance of also taking subject-specific response patterns into account: while language teachers (i.e., German and English as a foreign language) showed no differences between respondents and nonrespondents, only math teachers contributed to the self-selection bias effect. These findings emphasize the need for researchers to examine nonresponse in questionnaire-based research beyond broad institutional categories, also accounting for subject-specific and school-type subgroup dynamics.

Underlying mechanisms for self-selection bias could be revealed by Study 2, showing that these mechanisms differed between school types. The mechanism most frequently declared by

respondents as an explanation for nonresponse in the technically oriented BHS is what Saleh and Bista (2017) described as one of the most important issues for nonresponse in educational research: interest. Mathematics teachers in these schools considered the Matura to be a low-stakes assessment, particularly since its standardization, and therefore attached little value to participating in a study on grading in this context. The mathematics exam was perceived either as too easy or as a settled issue with no need for reform, reducing teachers' motivation to engage in the research. This suggests that perceived value and relevance of a research topic (shaped by the stakes of the Matura exam for students and its perceived difficulty) are important drivers of survey participation (Saleh & Bista, 2017).

Second, several teachers pointed to the pragmatic and practice-oriented mentality of technical BHS teachers as a factor connected to nonresponse. While previous studies have linked sociodemographic characteristics to nonresponse patterns (Korkeila et al., 2001), little evidence exists in the educational context regarding self-selection bias shaped by teachers' professional attitudes and mindsets. Our findings therefore point to the need to explore such professional mentalities as potential determinants of participation behavior in survey-based studies.

Although not statistically significant, the pattern of nonresponding AHS math teachers awarding higher classroom grades than standardized Matura grades was not surprising to participants in our qualitative interviews. Nearly all participants connected this result to teachers strategically awarding higher classroom grades to compensate for students' expected low or failing Matura grades in mathematics. From a methodological perspective, this highlights the importance of considering the sensitivity of a topic in questionnaire-based research (Guillemot et al., 2025). Our results indicate that grading is a sensitive topic in educational research. Teachers argued that nonrespondents did not want to reveal information on their grading practices, described as associated with feelings of shame. This addresses the first type of sensitive questions described by Peytcheva and Yan (2025), which are perceived as intrusive and invading personal privacy (Tourangeau & Yan, 2007; Yan, 2021).

Teachers also suggested that nonresponding might serve to conceal low teaching quality when higher classroom grades are awarded compared to Matura grades. This addresses the second type of sensitive questions involving potential personal harm if the response is shared (Peytcheva & Yan, 2025). Regarding the third type of sensitive questions addressing social desirability, teachers reported that awarding classroom grades substantially higher than

Matura grades is seen—by themselves or by colleagues—as unprofessional grading practices. Inflated classroom grades were associated with reduced conflicts with students and principals, and with less workload (e.g., fewer repeat exams), making the practice attractive but simultaneously stigmatized within the profession.

Large-scale teacher surveys typically apply weighting procedures that treat nonresponse as missing at random (OECD, 2018). However, these approaches do not account for the mechanisms of nonresponse identified in our study. This limitation is particularly relevant when surveys address sensitive issues such as school climate or job satisfaction (OECD, 2018), where the likelihood of participation may be systematically related to teachers' views. In such cases, standard weighting procedures risk overlooking underlying self-selection bias.

Our findings also underscore the value of qualitative and mixed method approaches for uncovering the mechanisms underlying self-selection bias (Groves & Peytcheva, 2008). In our case, it was important to interview teachers who, according to our qualitative data, seemed to endorse different grading approaches. For example, with respect to leniency in grading, it was essential to capture perspectives from both sides: while some teachers critically portrayed colleagues' low teaching quality and strategic grading behaviors, others emphasized a view of grades as secondary to broader educational goals, such as fostering positive student-teacher relationships and supporting students' well-being in the classroom and learning process. At the same time, these teachers reported feeling pressured and stigmatized by their colleagues as unprofessional graders. Without such diversity in respondents, it would not have been possible to draw this broad understanding of reluctance in responding to our questionnaire.

Besides the methodological implications for research on sensitive topics such as grading, our findings may also prompt a broader debate about the value of a standardized Matura assessment system that aims to ensure comparability of grades while simultaneously addressing educational heterogeneity by providing partially different mathematics tasks which, according to our results, appear to differ in their level of difficulty across school types. Statements from technical vocational school teachers, in particular, portrayed the standardized Matura exams in their school type as very easy and low stakes, regarding them as little more than a formality. These teachers did not criticize this situation and did not appear to seek a more demanding Mathematics Matura examination. The findings nonetheless raise the political question as to whether the annual cost of €8.45 million (Austrian Court of Audit, 2020) for developing and administering standardized Matura exams is justified when, at least

in this school context, it functions primarily as a symbolic act rather than as a serious assessment that qualifies students for tertiary education.

6 Limitations

Three major limitations must be noted. First, the study involved only a small number of schools (3 AHS, 2 technically oriented BHS, and 2 non-technical BHS). Although the interview respondents were very consistent in their explanations, the findings cannot be generalized to all schools of these types (Infante-Rivard & Cusson, 2018). While interview participants clearly addressed school type issues, it is nevertheless possible that school-specific effects might be influential, too. Second, of the teachers who participated in the qualitative study, only four AHS teachers and three BHS teachers were also participating nonrespondents of Study 1. The other teachers were from the same subgroup (mathematics teachers at technically oriented BHS or AHS), but from different schools and federal states. While their uninvolvement and neutral perspective may lend credibility to their statements (e.g., because they may speak freely without feeling being questioned in their grading practices and beliefs), it is also possible that any specific mechanisms within participating schools of Study 1 are not captured. Third, our qualitative data focus exclusively on mathematics teachers' perspectives, so potential mechanisms underlying response and grading patterns among language teachers remain unexplored. Finally, a separate replication study could focus exclusively on our finding and verifying the identified effects.

7 Directions for Future Research

The quantitative and qualitative results of our study provide empirical evidence of the mechanisms underlying self-selection bias in research on the teaching profession. However, we would argue that a follow-up larger-scale study may more precisely disentangle the effects of school type, subject, and local schools on self-selection bias when investigating sensitive topics. It is important to note that researchers from other countries and/or assessment systems should carefully consider which dependent variable is suitable for their specific research context. Since the variable should reflect grading sensitivity to yield a comparable effect, researchers could base their choice on grading guidelines or implicit grading norms in their respective country. Teachers' grading patterns during transition decisions, for example, could be an interesting choice. Using the discrepancy between classroom and Matura grades as the dependent variable is presumably only suitable for systems like Austria's, where teachers

grade both assessments. The Austrian "loophole" of allowing teacher classroom grades to compensate for low or failing Matura grades provided an unplanned but excellent opportunity to detect self-selection bias related to teacher grading.

Future experimental studies could test how different framings of survey invitations (e.g., focus on grading vs. teaching practices more broadly) impact response rates in sensitive areas such as grading. In addition to such overall effects, it would be valuable to investigate whether framing impact differs by, for example, teacher characteristics (e.g., age, gender, professional attitudes), subject area or school type. Further, comparative perspectives (e.g., cross-national) could shed light on whether the mechanisms identified here, like strategic grading or the role of shame, are context-specific or generalizable across different education systems.

Research on self-selection bias in the educational context is still scarce (Guillemot et al., 2025), presumably because it is difficult to detect, requires relatively complex designs (Olsen, 2006), and is often not the primary focus of researchers' interests. However, providing a transparent account of self-selection bias, including accidental findings, is important to advance this field, as it can raise researchers' awareness of potential biases and risks of unrepresentative data, ultimately leading to more robust and generalizable results.

Statements and declarations

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

The data that support the findings of this study are available from the corresponding author upon reasonable request.

All the authors declare that they have no conflicts of interest.

References

- Annerstedt, C., & Larsson, S. (2010). 'I have my own picture of what the demands are...': Grading in Swedish PEH — problems of validity, comparability and fairness. *European Physical Education Review*, 16(2), 97-115. <https://doi.org/10.1177/1356336X10381299> (Original work published 2010)
- Austrian Court of Audit. (2020). Zentralmatura: Bericht des Rechnungshofes Reihe BUND 2020/22 [Centralized Matura: Report by the Court of Auditors, BUND series 2020/22]. https://www.rechnungshof.gv.at/rh/home/home/004.743_Zentralmatura.pdf

- Autonomous Province of Bolzano South Tyrol. (2025). Die staatliche Abschlussprüfung der Oberschule [The state final examination for secondary school]. <https://deutsche-bildung.provinz.bz.it/de/schule/staatliche-abschlusspruefung-oberschule>
- Berg, N. (2005). Non-Response Bias. *Encyclopedia of Social Measurement*, Vol. 2, pp. 865-873, Kempf-Leonard, K., ed., London, Academic Press, Available at SSRN: <https://ssrn.com/abstract=1691967>
- BMBWF. (2021). Nationaler Bildungsbericht Österreich 2021 [National Education Report Austria 2021]. <https://www.bmbwf.gv.at/Themen/schule/bef/nbb.html>
- Bonner, S. M., & Chen, P. P. (2009). Teacher candidates' perceptions about grading and constructivist teaching. *Educational Assessment*, 14, 57-77. doi:10.1080/10627190903039411
- Bonner, S. M., & Chen, P. P. (2021). Development and Validation of the Survey of Unorthodox Grading Beliefs for Teachers and Teacher Candidates. *Journal of Psychoeducational Assessment*, 39(6), 746–760. <https://doi.org/10.1177/07342829211015462>
- Borsa, J. C., Damásio, B. F., & Bandeira, D. R. (2012). Cross-Cultural Adaptation and Validation of Psychological Instruments: Some Considerations. *Paidéia* (Ribeirão Preto), 22, 423–432.
- Bowden, R.J (1986). Self-selection biases in correlational studies based on questionnaires. *Psychometrika* 51, 313–325, <https://doi.org/10.1007/BF02293987>
- 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., Guskey, T. R., Bowers, A. J., McMillan, J. H [James H.], Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. *Review of Educational Research*, 86(4), 803-348. <https://doi.org/10.3102/0034654316672069>
- Bundesministerium für Bildung. (2025a). Berufsbildende mittlere und höhere Schulen [Vocational middle and high schools]. <https://www.bmb.gv.at/Themen/schule/schulsystem/sa/bmhs.html>
- Bundesministerium für Bildung. (2025b). Die Zentralmatura [The Centralized Matura]. <https://www.bmb.gv.at/Themen/schule/schulpraxis/zentralmatura.html>
- Cross, L. H., & Frary, R. B. (1999). Hodgepodge Grading: Endorsed by Students and Teachers Alike. *Applied Measurement in Education*, 12(1), 53–72. https://doi.org/10.1207/s15324818ame1201_4
- Dawson, S. J., Huberman, J. S., Bouchard, K. N., McInnis, M. K., Pukall, C. F., & Chivers, M. L. (2019). Effects of Individual Difference Variables, Gender, and Exclusivity of Sexual Attraction on Volunteer Bias in Sexuality Research. *Archives of Sexual Behavior*, 48(8), 2403–2417. <https://doi.org/10.1007/s10508-019-1451-4>
- Downe-Wamboldt, B. (1992). Content analysis: Method, applications, and issues. *Health Care for Women International*, 13(3), 313–321. <https://doi.org/10.1080/07399339209516006>
- Doz, D. (2023). Factors influencing teachers' grading standards in mathematics. *Oxford Review of Education*, 49(6), 819–837. <https://doi.org/10.1080/03054985.2023.2185217>
- Dunne, M. P., Martin, N. G., Bailey, J. M., Heath, A. C., Bucholz, K. K., Madden, P. A., & Statham, D. J. (1997). Participation bias in a sexuality survey: psychological and behavioural characteristics of responders and non-responders. *International journal of epidemiology*, 26(4), 844-854.

- Fuller, B., & Izu, J. A. (1986). Explaining School Cohesion: What Shapes the Organizational Beliefs of Teachers? *American Journal of Education*, 94(4), 501–535. <https://www.jstor.org/stable/1085339>
- Galevska, N. A. (2019). Challenges of Teachers in the Process of Evaluation and Grading. *The Eurasia Proceedings of Educational and Social Sciences*, 15, 94-99.
- Goddard, R. D. (2003). The impact of schools on teacher beliefs, influence, and student achievement: the role of collective efficacy beliefs. In J. Raths & A. C. McAninch (Eds.), *Teacher beliefs and classroom performance: The impact of teacher education*. (pp. 183–202). IAP.
- Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*, 24(2), 105–112. <https://doi.org/10.1016/j.nedt.2003.10.001>
- Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, 72(2), 167–189. <https://doi.org/10.1093/poq/nfn011>
- Groves, R. M., Presser, S., & Dipko, S. (2004). The Role of Topic Interest in Survey Participation Decisions. *Public Opinion Quarterly*, 68(1), 2–31. <https://doi.org/10.1093/poq/nfh002>
- Guillemot, F., Lacroix, F., & Nocus, I. (2025). Teachers' non-response bias in questionnaires about their attitude towards inclusive education. *Journal of Research in Special Educational Needs*, 25(2), 312–322. <https://doi.org/10.1111/1471-3802.12726>
- Haynes, A., & Robinson, E. (2019). Who are we testing? Self-selection bias in laboratory-based eating behaviour studies. *Appetite*, 141, 104330. <https://doi.org/10.1016/j.appet.2019.104330>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Infante-Rivard, C., & Cusson, A. (2018). Reflection on modern methods: selection bias—a review of recent developments. *International Journal of Epidemiology*, 47(5), 1714–1722. <https://doi.org/10.1093/ije/dyy138>
- Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using Mixed-Methods Sequential Explanatory Design: From Theory to Practice. *Field Methods*, 18(1), 3–20. <https://doi.org/10.1177/1525822X05282260>
- The jamovi project. (2022). jamovi (Version 2.3) [Computer software]. <https://www.jamovi.org>
- Kleinheksel, A. J., Rockich-Winston, N., Tawfik, H., & Wyatt, T. R. (2020). Demystifying Content Analysis. *American Journal of Pharmaceutical Education*, 84(1), 7113–7137. <https://doi.org/10.5688/ajpe7113>
- Korkeila, K., Suominen, S., Ahvenainen, J., Ojanlatva, A., Rautava, P., Helenius, H., & Koskenvuo, M. (2001). Non-Response and Related Factors in a Nation-Wide Health Survey. *European Journal of Epidemiology*, 17(11), 991–999. <https://doi.org/10.1023/A:1020016922473>
- Köller, O., Baumert, J., & Schnabel, K. U. (1999). Wege zur Hochschulreife: Offenheit des Systems und Sicherung vergleichbarer Standards. Analysen am Beispiel der Mathematikleistungen von Oberstufenschülern an integrierten Gesamtschulen und Gymnasien in Nordrhein-Westfalen [Paths to university entrance qualifications: openness of the system and ensuring comparable standards. Analyses based on the example of mathematics performance by upper secondary school students at integrated comprehensive schools and grammar schools in North Rhine-Westphalia]. *Zeitschrift Für Erziehungswissenschaft*, 2, 385–422.

- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4), 2025-2047.
- Kutner, M. H., Nachtsheim, C. J., & Neter, J., & Li, W. (2005). Applied linear statistical models. McGraw-hill.
- Lehane, P., O'Connor, G., & O'Leary, M. (2025). Back to the future? Reflections from 25 years of empirical research on Leaving Certificate assessment in Ireland. *Irish Educational Studies*, 1–31. <https://doi.org/10.1080/03323315.2024.2441158>
- McMillan, J. H. (2003). Understanding and Improving Teachers' Classroom Assessment Decision Making: Implications for Theory and Practice. *Educational Measurement, Issues and Practice*, 22(4), 34–43. <https://doi.org/10.1111/j.1745-3992.2003.tb00142.x>
- McMillan, J. H., & Nash, S. (2000). Teacher classroom assessment and grading practices decision making: Paper presented at the 2000 annual meeting of the National Council on Measurement in Education., New Orleans.
- Meinck, S., Cortes, D., & Tieck, S. (2017). Evaluating the risk of nonresponse bias in educational large-scale assessments with school nonresponse questionnaires: a theoretical study. *Large-Scale Assessments in Education*, 5(1), 1–21. <https://doi.org/10.1186/s40536-017-0038-6>
- Menard, S. (2001). Applied logistic regression analysis. SAGE Publications.
- Ministère de L'Education Nationale. (2025). Déroulement des corrections aux examens du second degré à compter des épreuves 2024. <https://www.education.gouv.fr/bo/2024/Hebdo5/MENE2335316N#:~:text=La%20commission%20d'entente%20nationale,-Une%20commission%20d>
- Muller, A., De La Rochebrochard, E., Labbé-Declèves, C., Jouannet, P., Bujan, L., Mieusset, R., Le Lannou, D., Guerin, J. F., Benchaib, M., Slama, R., & Spira, A. (2004). Selection bias in semen studies due to self-selection of volunteers. *Human Reproduction (Oxford)*, 19(12), 2838–2844. <https://doi.org/10.1093/humrep/deh521>
- Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *International Journal of Public Opinion Quarterly*, 70(5), 737-758. <https://doi.org/10.1093/poq/nfl038>
- Olsen, R. (2008). Self-selection bias. In Encyclopedia of survey research methods (Vol. 0, pp. 809-809). Sage Publications, Inc., <https://doi.org/10.4135/9781412963947.n526>
- Organisation for Economic Co-operation and Development (OECD). (2018). Teaching and Learning International Survey (TALIS) 2018 teacher questionnaire: Main survey version. OECD.
- Pahwa, P., Karunanayake, C., Hagel, L., Janzen, B., Rennie, D., Lawson, J., Pickett, W., & Dosman, J. A. (2012). Self-Selection Bias in an Epidemiological Study of Respiratory Health of a Rural Population. *Journal of Agromedicine*, 17(3), 316–325. <https://doi.org/10.1080/1059924X.2012.686381>
- Peytcheva, E., & Yan, T. (2025). Practical Guide to Survey Questionnaire Design and Evaluation. Routledge.
- Pilcher, J. K. (1994). The Value-Driven Meaning of Grades. *Educational Assessment*, 2(1), 69–88. https://doi.org/10.1207/s15326977ea0201_4
- RIS, Rechtsinformation des Bundes (2025, August 28). Verordnung des Bundesministers für Unterricht und Kunst vom 24. Juni 1974 über die Leistungsbeurteilung in Pflichtschulen sowie mittleren und höheren Schulen (Leistungsbeurteilungsverordnung) [Regulation of the Federal Minister of Education and Art of June 24, 1974, on performance assessment in compulsory schools and middle

and high schools (Performance Assessment Regulation)].

<https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10009375>

Robinson, K., Dayer, K. F., Mirichlis, S., Hasking, P. A., & Wilson, M. S. (2023). Who are we missing? Self-selection bias in nonsuicidal self-injury research. *Suicide & life-threatening behavior*, 53(5), 843–852. <https://doi.org/10.1111/sltb.12987>

Rubie-Davies, C. M., Flint, A., & McDonald, L. G. (2012). Teacher beliefs, teacher characteristics, and school contextual factors: What are the relationships? *The British Journal of Educational Psychology*, 82(Pt 2), 270–288. <https://doi.org/10.1111/j.2044-8279.2011.02025.x>

Sætre, C. (2025). Behind the grades: co-constructing fairness to reach agreement in evaluative judgement. *Assessment & Evaluation in Higher Education*, 50(2), 236–249.

Saleh, A. & Bista, K. (2017). Examining Factors Impacting Online Survey Response Rates in Educational Research: Perceptions of Graduate Students. *Journal of Multidisciplinary Evaluation*, 13(29), 63–74. <https://doi.org/10.56645/jmde.v13i29.487>

Schatz-Laliberté, V., & Lüftenegger, M. (2025). Grading after reform: How policy change affected teachers' grading for university entrance qualifications. *Studies in Educational Evaluation*, 87, 101540. <https://doi.org/10.1016/j.stueduc.2025.101540>

SRDP. (2025). Angewandte Mathematik [Applied Mathematics]. <https://www.matura.gv.at/srdp/angewandte-mathematik>

Statistik Austria. (2024). Schüler:innen [Students]. <https://www.statistik.at/ueber-uns/aufgaben-und-grundsaeetze/rechtsgrundlagen/impressum>

TALIS 2018 Technical Report. (2019). OECD Publishing

Tourangeau, R., Groves, R. M., & Redline, C. D. (2010). Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error. *Public Opinion Quarterly*, 74(3), 413–432. <https://doi.org/10.1093/poq/nfq004>

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). The psychology of survey response. Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859–883. <https://doi.org/10.1037/0033-2909.133.5.859>

Watermann, R., Nagy, G., & Köller, O. (2013). Mathematikleistungen in allgemein bildenden und beruflichen Gymnasien [Mathematics performance in general education and vocational high schools]. 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). Springer-Verlag.

Wickham, H. (2016). Data analysis. In *ggplot2: elegant graphics for data analysis* (pp. 189–201). Cham: Springer international publishing.

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>

Yan, T. (2021). Consequences of asking sensitive questions in surveys. *Annual Review of Statistics and Its Application*, 8(1), 109–127.

Supplementary Material

Appendix A

Adaptation and Validation of the German Version of the “Survey of Unorthodox Grading Beliefs” (Bonner and Chen, 2021)

Since the original questionnaire was in English, we translated the vignettes into German following the five-step process recommended by Borsa et al. (2012). First, two independent translations from English to German were created and compared to reach a consensus with the help of a third person. Second, three independent researchers from the authors’ institute reviewed the translation, and their feedback was incorporated. Third, a think-aloud study was conducted with four upper secondary school teachers from different subject areas (German, mathematics, and English) who verbalized their thoughts while responding to the questionnaire. Fourth, the questionnaire was revised based on their feedback, followed by a prestudy with 16 teachers as the final step of the translation process.

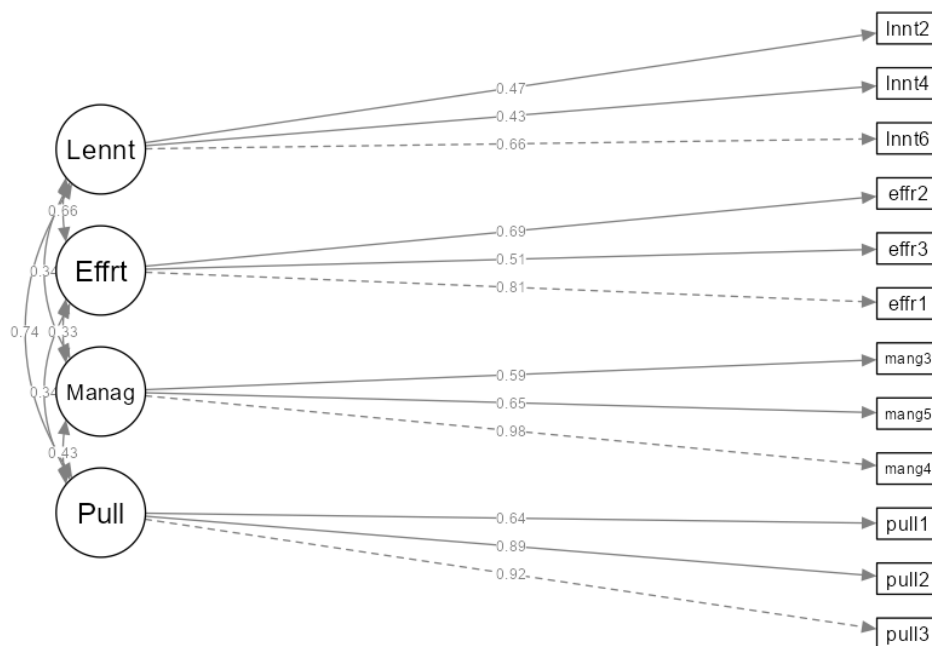
To examine the four-factor structure of the German version of the "Survey of Unorthodox Grading Beliefs", we conducted a confirmatory factor analysis (CFA) with maximum likelihood estimation using jamovi (jamovi project, 2022). The four-factor structure reported by Bonner and Chen (2021) was tested to ensure that the translated version maintained the same latent dimensions as the original scale. Missing values were handled with full information maximum likelihood estimation (FIML).

The final version was slightly changed on the basis of the initial CFA results and a low case-to-item ratio of 3.30 (56 cases, 17 items). Hence, we trimmed the model by removing five items that showed cross-loadings and violated the original structure. Specifically, three items of the *lenient* factor showed cross-loadings with either items of the *pull* (modification indices: 6.69 and 6.32) or the *effort* factor (modification index: 5.10). Additionally, one *manage* item cross-loaded with an item of the *pull* factor (modification index: 5.70). As a final step, we removed one additional *manage* item due to its low factor loading of less than 0.3. The final questionnaire in German can be found in the supplementary material, too.

These steps resulted in a four-factor solution with three items per factor and an acceptable to good model fit ($\chi^2 = 53.3$ ($df = 48$), $\chi^2/df = 1.11$, CFI = 0.975, TLI = 0.965, RMSEA = 0.044, SRMR = 0.072; Hu & Bentler, 1999). The reliability indices for the four dimensions are $\alpha_{pull} = 0.848$, $\alpha_{manage} = 0.758$, $\alpha_{effort} = 0.693$, and $\alpha_{lenient} = 0.583$. Figure 1

shows the pathway diagram of the trimmed version of the model. Full analysis reports for the original model and the trimmed version are provided in the supplementary material.

Fig. 1. Pathway diagram (trimmed version) of the German version of the survey of unorthodox grading beliefs questionnaire (Bonner & Chen, 2021)



As shown in Figure 1, correlations were observed between the four grading belief dimensions, especially between *pull* and *lenient* ($r = 0.74$). To assess multicollinearity among the four beliefs, variance inflation factors (VIFs) and tolerance values were calculated. While the values for *effort* (VIF = 4.48, Tolerance = 0.2231) and *manage* (VIF = 1.71, Tolerance = 0.5864) were acceptable, the values for *pull* (VIF = 13.63, Tolerance = 0.0743) and *lenient* (VIF = 17.91, Tolerance = 0.0558) indicated multicollinearity (Kutner et al., 2005; Menard, 2001). The *pull* and *lenient* dimensions seem to be rather similar for teachers, as they might share the common feature of overlooking inconsistencies in student performance across the academic year. In contrast to *lenient* belief (which refers more to overall “sloppy”/“lazy” grading), however, *pull* deliberately disregards

instances of poor performance, thereby emphasizing the potential of the students in the grading process rather than their deficits. Hence, we removed the *lenient* dimension from our model and included only the *pull* dimension in our analyses.

References

- Bonner, S. M., & Chen, P. P. (2021). Development and Validation of the Survey of Unorthodox Grading Beliefs for Teachers and Teacher Candidates. *Journal of Psychoeducational Assessment*, 39(6), 746–760. <https://doi.org/10.1177/07342829211015462>
- Borsa, J. C., Damásio, B. F., & Bandeira, D. R. (2012). Cross-Cultural Adaptation and Validation of Psychological Instruments: Some Considerations. *Paidéia* (Ribeirão Preto), 22, 423–432.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- The jamovi project. (2022). jamovi (Version 2.3) [Computer software]. <https://www.jamovi.org>
- Kutner, M. H., Nachtsheim, C. J., & Neter, J., & Li, W. (2005). Applied linear statistical models. McGraw-hill.
- Menard, S. (2001). Applied logistic regression analysis. SAGE Publications.

Appendix B

German Adaptation of the Survey of Unorthodox Grading Beliefs

Hinweis: Die englische Originalversion des „Survey of Unorthodox Grading Beliefs“ (Bonner & Chen, 2021) wurde mithilfe des fünf-stufigen Adaptionsprozesses nach Borsa et al. (2012) ins Deutsche übersetzt.

Item-Zuordnung

Die Namen der in Summe 17 Items richten sich nach den Namen der Lehrer:innen in den jeweiligen Vignetten.

Pull: Klinger, Stein, Maurer

Manage: Gant, Petrovic, Müller, Danner, Szabo

Effort: Funk, Pacher, Leitner

Lenient: Braun, Kobler, Hartl, Schmidt, Kraus, Grün

Skala

Vollkommen einverstanden

Einverstanden

Eher schon einverstanden

Eher nicht einverstanden

Nicht einverstanden

Überhaupt nicht einverstanden

Keine Antwort

Fragebogen

Bitte geben Sie bei den folgenden Situationen an, inwieweit Sie mit der Beurteilungspraxis der jeweiligen Lehrperson einverstanden sind oder nicht.

Alle Situationen beziehen sich auf die Sekundarstufe II (Oberstufe).

Herr Danner rechnet das Verhalten der Schüler:innen im Unterricht in deren Noten mit ein. Dementsprechend können die Noten von Schüler:innen sinken, wenn sie den Unterricht stören, sich respektlos verhalten oder anderweitig die Unterrichtsregeln missachten. Inwieweit sind Sie mit Herrn Danners Beurteilungspraxis einverstanden?

Wenn es darum geht, die Abschlussnoten ihrer Schüler:innen zu erfassen, stützt sich Frau Grün hauptsächlich auf die Ergebnisse von Tests und Projekten. Manchmal jedoch hat sie das Gefühl, dass ein:e Schüler:in mit niedrigen Punktezahlen bei Tests und Schularbeiten den Inhalt in Wirklichkeit besser verstanden hat, als die Zahlen vermuten lassen. Dann passt sie die Note der Schülerin:des Schülers an. Inwieweit sind Sie mit der Beurteilungspraxis von Frau Grün einverstanden?

Frau Funk hat Schüler:innen mit unterschiedlichen Hintergründen und Vorkenntnissen. Daher beschließt sie, Anstrengung und individuellen Fortschritt in die Benotung miteinfließen zu lassen. Am Ende des Schuljahres hebt sie die Noten derjenigen Schüler:innen an, die sich bemüht und verbessert haben. Inwieweit sind Sie mit der Benotungspraxis von Frau Funk einverstanden?

Frau Szabo beobachtet, dass sich zwei ihrer Schüler:innen unterhalten, während sie einen Test durchführen möchte. Sie ermahnt sie ein Mal, als sie kurze Zeit darauf jedoch weiterflüstern, nimmt sie ihnen die Testbögen ab und erklärt ihnen, dass jede:r von ihnen ein Nicht genügend für diese Prüfung erhalten werde. Inwieweit sind Sie (unabhängig von der rechtlichen Lage) mit Frau Szabos Vorgehen einverstanden?

Herr Kraus unterhält sich mit seinen Kolleg:innen über seine Klassen und sagt: „Ich weiß, dass meine Schüler:innen diese Dinge wissen, aber die Tests spiegeln das einfach nicht wider.“ Bei der Vergabe von Jahresnoten stützt sich Herr Kraus daher sowohl auf die Testergebnisse als auch auf seine Intuition über das tatsächliche Können seiner Schüler:innen. Inwieweit sind Sie mit der Benotungspraxis von Herrn Kraus einverstanden?

Herr Maurer erstellt Tests für seine Klassen so, dass diese aus vier Fragen bestehen, die sowohl inhaltliches Wissen als auch analytisches Denken erfordern. Er legt die Punktzahl jeder Schülerin:jedes Schülers auf die besten drei von vier Antworten fest. Auf diese Weise hat er das Gefühl, dass die Testergebnisse der Schüler:innen ihre Stärken widerspiegeln. Inwieweit sind Sie mit der Benotungspraxis von Herrn Maurer einverstanden?

In diesem Schuljahr haben die Schüler:innen von Frau Braun drei große Tests. Jeder Test geht zu gleichen Teilen in die Benotung mit ein. Maria, eine Schülerin der Klasse, hat Schwierigkeiten beim ersten der drei Tests und erhält darauf eine schlechte Note. Frau Braun beschließt, Marias ersten Test weniger zu gewichten als die anderen beiden Tests (Frau Braun würde dies auch für alle anderen ihrer Schüler:innen tun). Inwieweit sind Sie mit der Benotungspraxis von Frau Braun einverstanden?

Herr Müller ist der Meinung, dass Schüler:innen für gutes Verhalten belohnt und für schlechtes Verhalten bestraft werden sollten. Jede Woche werden Verhaltenspunkte zu ihrer Gesamtnote hinzugezählt, wenn sie keine diszipliniären Probleme verursacht haben. Hingegen werden Punkte abgezogen, wenn sie sich nicht ordentlich verhalten haben. Inwieweit sind Sie mit der Vorgehensweise von Herrn Müller einverstanden?

Während eines Elterngesprächs wird Frau Pacher von einem Elternteil gefragt, wie sie die Noten der Schüler:innen in ihren Klassen verteile. Frau Pacher antwortet, dass sie, wenn die Schüler:innen Einsatz zeigen, dies bei der Vergabe der Jahresnoten berücksichtigt und die Noten entsprechend anpasst. Inwieweit sind Sie mit der Beurteilungspraxis von Frau Pacher einverstanden?

Frau Hartl ist der Meinung, dass es für alle Schüler:innen wichtig ist, in der Schule erfolgreich zu sein. Sie entwirft ihre Tests so, dass sie hauptsächlich einfache und grundlegende Fragen enthalten, damit alle Schüler:innen die Möglichkeit haben, gute Noten erzielen zu können. Inwieweit sind Sie mit Frau Hartls Benotungspraxis einverstanden?

In den Klassen von Herrn Klinger schreiben die Schüler:innen vier Tests pro Schuljahr. Bei der Jahresnotenvergabe beschließt Herr Klinger, die niedrigste Testpunktezahl für jede Schülerin:jeden Schüler zu streichen und nur vier von fünf Tests zu zählen. Herr Klinger erklärt seinen Schüler:innen, dass sie dadurch in der Lage wären, das Beste aus ihren Jahresnoten rauszuholen. Inwieweit sind Sie mit Herrn Klingers Benotungspraxis einverstanden?

Frau Petrovic hat ein System entwickelt, mit dem sie positives Verhalten ihrer Schüler:innen bei der Notenvergabe berücksichtigen kann. Demgemäß werden die Schüler:innen für gutes Verhalten und die Einhaltung der Unterrichtsregeln mit Punkten belohnt, die der Note für den jeweiligen Beurteilungszeitraum angerechnet werden. Inwieweit sind Sie mit der Benotungspraxis von Frau Petrovic einverstanden?

Bei der Jahresnotenvergabe berücksichtigt Herr Kobler, dass die Schüler:innen auch mal schlechte Tage haben können. Daher erhöht er die Noten der Schüler:innen automatisch um ein paar Punkte. Inwieweit sind Sie mit der Benotungspraxis von Herrn Kobler einverstanden?

Herr Leitner legt Wert auf eine positive Unterrichtskultur. Er vergibt bessere Noten an Schüler:innen, die aktives Engagement im Unterricht zeigen und sich bemühen. Inwieweit sind Sie mit Herrn Leitners Benotungspraxis einverstanden?

Die Schüler:innen in Herrn Gants Klassen wissen, dass sie sich an seine Unterrichtsregeln halten müssen, da ein Verstoß zu einer Verschlechterung der Abschlussnote führen kann. Inwieweit sind Sie mit der Beurteilungspraxis von Herrn Gant einverstanden?

Frau Schmidt hat eine:n Schüler:in, der:die in ihrem Fach durchfallen würde; sie ist jedoch der Überzeugung, dass der:die Schüler:in die Fähigkeit hat, sich zu verbessern. Frau Schmidt beschließt, dem:r Schüler:in einen Vertrauensvorschuss zu geben und ihn:sie im Zweifelsfall als bestanden zu bewerten. Inwieweit sind Sie mit der Benotungspraxis von Frau Schmidt einverstanden?

Die Noten in Frau Steins Klasse basieren auf vier Tests. Frau Stein berechnet die Gesamtnoten der Schüler:innen jedoch nur auf der Grundlage der drei besten Testergebnisse jeder Schülerin:jeden Schülers und lässt das niedrigste Ergebnis weg, um auf die Bedürfnisse der Schüler:innen einzugehen, die möglicherweise einen schlechten Testtag hatten. Inwieweit sind Sie mit der Beurteilungspraxis von Frau Stein einverstanden?

Literatur

Bonner, S. M., & Chen, P. P. (2021). Development and Validation of the Survey of Unorthodox Grading Beliefs for Teachers and Teacher Candidates. *Journal of Psychoeducational Assessment*, 39(6), 746–760. <https://doi.org/10.1177/07342829211015462>

Borsa, J. C., Damásio, B. F., & Bandeira, D. R. (2012). Cross-Cultural Adaptation and Validation of Psychological Instruments: Some Considerations. *Paidéia* (Ribeirão Preto), 22, 423–43.

Appendix C

Categories and Codes from Qualitative Study 2

A) Categories of Grading and Questionnaire (Non-)Response Among Mathematics Teachers in AHS

Category 1: Teachers' Explanations for Higher Classroom than Matura Grades in AHS Mathematics

Code	Coding Description	Examples
Emphasis on Effort and Participation in AHS	When teachers describe that in AHS, students are provided with more diverse opportunities to improve their classroom grades, primarily through effort and participation in class.	„AHS-Lehrer [...] legen auch sehr viel Wert auf andere Dinge, wie, wie sehr bemüht sich der Schüler und wieviel Einsatz zeigt er unterm Jahr.“ (P02) „Vor allem in der AHS gibt's halt diese Mädels, die immer ganz brav sind, fünf Mal dieselben Hausübungen abgeben, regelmäßig aufzeigen und sich so ganz gut hinaufbewegen bis zur Matura. [...] In der HTL machen sie schon gar keine Hausübung, aber wenn's drauf ankommt, liefern die.“ (P09) „Diese Streicheleinheiten, die wir [in der AHS] jedem Schüler geben, gibt es in der HTL nicht [...] In der AHS ist das Bemühen um den Schwachen sehr ausgeprägt [...] und dann stehen die Schüler vorm Abschlusstest und tun sich natürlich dementsprechend schwer.“ (P13)

		„[I]n der AHS haben die Schüler unterm Jahr viel mehr und [...] diversere Möglichkeiten, ihre Noten aufzubessern. Bei mir [in der BHS] gibt's Tests und Schularbeiten und das war's.“ (P02)
Structural Conditions Affecting Classroom Grading	When teachers describe how structural conditions, such as the number of lessons and class size, affect focus, frequency, and types of achievement factors captured by classroom assessments.	„Wir [an der BHS] haben halt relativ wenig Stunden und wenig Schularbeiten und durch das kommt halt die Leistung oft nicht so raus, als wenn man mehr Zeit und mehrere Schularbeiten hat.“ (P04) „Bei mir [in der BHS] zählt, was die Schüler wirklich können und schriftlich zeigen. Bei 36 in der Klasse weiß ich nicht, wer da mitarbeitet und wer nur drinsitzt [...] da würd ich mich massiv überschätzen, wenn ich da Mitarbeitsnoten machen würde.“ (P07) „An der AHS haben wir halt auch viele Wochenstunden, [...] kennen die Schüler zum Teil seit acht Jahren und haben viel kleinere Gruppen in den Hauptfächern. Da kann ich viele mündliche Wiederholungen machen, wo sich die Schüler ihre Jahresnote verbessern können. [...] In der BHS geht das nicht; wenig Stunden, große Klassen.“ (P14)

Category 2: Teachers' Explanations for Non-Response in AHS

Concealing Grade Inflation Practices	When teachers describe that colleagues who strategically award higher classroom grades to compensate for potential low Matura results may be reluctant to complete questionnaires on grading to avoid revealing their grading practices.	„Ich will das jetzt keinem unterstellen oder so, aber ich glaube schon, dass man sich da einfach selber absichert. [D]u deckst da ja auch eigentlich deinen Unterrichtsstil, denn wenn der gute Neujahresnoten hat, kann nicht mehr viel passieren.“ (P03)
--------------------------------------	--	--

		„Da werden dann am Ende noch fünf Prüfungen durchgeführt, die Jahresnoten klettern steil bergauf, mit Leistung hat das oft nicht wirklich etwas zu tun.“ (P09) „Ja, gut, die sind sich dem Problem natürlich sehr wohl bewusst, dass sie das mit der Jahresnote ausgleichen müssen, was sie während dem Unterricht ein bisschen versäumt haben [...] Wenn ich weiß, dass meine Schüler nicht so viel können, muss man jetzt halt im Vorhinein schon ein bisschen vorbauen [...] das ist jetzt ein Schlupfloch im System [...]“(P04) „Da spielt auch bei vielen Lehrern auch die Angst mit. Und drum geben die auch die besseren [Jahres]Noten manchmal [...], damit die Schüler durchkommen, damit sie sich nicht blamieren.“ (P11)
Conscientious Teachers as Conscientious Respondents	When teachers link conscientiousness in responding to a questionnaire to conscientiousness in grading, highlighting a possible connection between careful grading and a sense of professional responsibility.	„Das kann auch mit einer gewissen Gewissenhaftigkeit tun haben. Also ich hatte das Gefühl, wenn Jahresnote und Maturanote gut zusammenpassen, dann hat man den Schüler wahrscheinlich relativ gut eingeschätzt. [...] Und diese Leute sind dann insgesamt gewissenhafter und dann füllen sie auch eher einen Fragebogen aus, aus Gewissenhaftigkeit.“ (P05) „Weil die [, die den Fragebogen ausfüllen] eher das Gefühl haben, alles richtig gemacht zu haben. Die haben ja schon vor der Matura so bewertet, wies dann bei der Matura auch verlangt wird. Und es gibt auch ein gutes Gefühl zu sehen, dass man sozusagen in der Jahresnote die Leistung der Schüler schon realistisch vorhergesagt hat.“ (P8) „[I]ch kann mir gut vorstellen, dass Lehrer oft die bessere Note leicht, also unüberlegter [...] geben, damit sie weniger Arbeit haben.“ (P11)

(No) Fear of Exposure of Teaching Shortcomings	When teachers discuss nonresponse as driven by fear of revealing professional weaknesses, highlighting that those confident in their instruction and grading accuracy may showing less hesitation to participate.	„[W]eil sie wissen, dass das eigentlich nichts mit tatsächlicher Leistung in einem Fach zu tun hat, sondern eben eher mit Fleiß und brav und so. Und dann schneiden die Schüler bei der Matura schlechter ab und den Lehrern ist bewusst, dass sie ihre Schüler vielleicht nicht ausreichend vorbereitet haben.“ (P01) Ja, logisch, weil jeder einfach Angst hat, gell, dass du da draufkommst, wie manche ihre Arbeit erledigen [...], du machst dich da halt schon transparent.“ (P03) „Und das könnte halt ein Grund sein, wieso ich das dann nicht ausfüllen will, weil da müsste ich mich meinen Fehlern mal stellen.“ (P04) „Das betrifft Lehrpersonen, die, ich sag's plakativ, die vielleicht oberflächlicher unterrichten, schneller eine bessere Note hergeben und die dann bei der Matura aber draufkommen, dass der Unterricht doch nicht ganz so passend war.“ (P10) „Ich für mich habe das Gefühl, dass es im Großen und Ganzen relativ gut passt bei mir. Insofern hab ich die Situation nicht, dass ich mir denke, das ist aber komisch, und zögere nicht, was auszufüllen.“ (P12)
Grading Topic as a Secondary Concern	When teachers describe grading as less important than other aspects of teaching, resulting in only little interest in responding to questionnaires on the topic.	„Für die ist auch die Notengebung vielleicht nicht das große Thema. Es gibt sicher auch Lehrpersonen, die sich bei den Noten vielleicht auch ein bisschen weniger überlegen [...] und füllen dann einen Fragebogen zum Thema eben nicht aus.“ (P11)

		„Also ich hab die Nase voll davon, mich immer rechtfertigen zu müssen, dass meine Noten zu gut sind. Die Qualität meines Unterrichts wird daran gemessen, wie streng meine Noten sind. Für mich sind Noten aber bei weitem nicht so wichtig wie andere Dinge. [...] zum Beispiel meine Beziehung zu meinen Schülern, wie es ihnen geht und dass sie Freude am Unterricht haben [...] Zu so etwas würde ich deutlich eher einen Fragebogen ausfüllen, ja.“ (P10)
		„[M]anche geraten da einfach unter Druck, weil sie halt immer die besseren Noten vergeben und die Schüler dann alle durchkommen. Ist sicher ein Druck, der auf einigen lastet.“ (P11)
Non-Response Driven by Shame	When teachers describe that avoiding participation in a questionnaire may stem from feelings of shame related to classroom grading practices, particularly when awarded classroom grades may not accurately reflect students' Matura performance.	„Das ist Scham. Die haben sich gedacht ‚Oh je, der Schüler, dem habe ich einen 1er gegeben und jetzt schreibt der was weiß ich einen Vierer, die werden alle glauben, ich kann nicht verbessern‘. Die wollen sich nicht in die Karten schauen lassen.“ (P10) „Das ist natürlich schambesetzt. Man will sich nicht mit eigenen möglichen Unzulänglichkeiten konfrontieren. Da übt man sich lieber in Zurückhaltung.“ (P09) „Weil sie sich genieren dafür [...], das ist ihnen unangenehm, ihre Überzeugungen zur Notengebung nach außen zu tragen.“ (P13)

B) Categories of Grading and Questionnaire (Non-)Response Among Mathematics Teachers in Technical BHS

Category 3: Teachers' Explanations for Higher Matura than Classroom Grades in Technical BHS

Gap Between Matura and Classroom Demands

When teachers explain that, due to standardization, the mathematics Matura has become relatively easy for technical BHS students, leading to higher Matura grades compared with the more demanding standards in classroom assessments.

„Also das, was ich in meinem Unterricht verlange, übersteigt definitiv das, was dann bei der Matura gefragt ist. Wir nehmen sehr viel Fachspezifisches durch, also was die Schüler halt auch später in ihrem Beruf brauchen, und da gestalte ich die Schularbeiten unterm Jahr schon so, dass die Notenskala ausgeschöpft ist, halt einfach auch, damit die Schüler gefordert sind.“ (P01)

„Die HTL-Schüler spielen sich mit der Matura.“ (P13)

„Unterm Jahr behandeln wir durchaus Themen, die oft deutlich über das hinausgehen, was in der Matura verlangt wird, und dadurch kriegen die Schüler halt auch mal schlechtere Jahresnoten, weil unterm Jahr prüfen wir schon [...] auf höherem Niveau“ (P02)

„Die Aufgaben in unseren HTL-Schulbüchern sind einfach deutlich anspruchsvoller als das, was dann bei der Matura verlangt wird.“ (P04)

„Das muss man sich immer bewusst sein, dass die Matura dort [in der technisch orientierten BHS] nicht den Lehrplan abdeckt. Der Matura ist eigentlich nur ein Teil des Lehrplans.“ (P11)

Cross-Subject Reinforcement of Math Skills	When teachers explain students' high math ability by noting that students also engage with mathematics in other, profession-specific subjects.	„Die sind einfach gut in Mathe, gell. Die schaffen das, ohne dass sie sich jetzt angestrengt hätten. [...] Und die bringen ja auch schon viel Können in Mathematik von anderen Fächern mit.“ (P03) „In vielen anderen Fächern, wie zum Beispiel Statik, machen sie ja auch nichts anderes als wie Rechnen.“ (P05)
--	--	---

Category 4: Teachers' Explanations for Non-Response in Technical BHS

Low Interest Due to Lack of Challenge in Easy Matura	When teachers describe that they are less likely to participate in a questionnaire on grading because the mathematics Matura is perceived as easy or unchallenging in their school context, making the topic seem less relevant or pressing.	„Wenn ich mit etwas zufrieden bin, interessiert mich das eher weniger. Und jetzt gibt es halt da bei uns keine Schwierigkeiten in Mathe. [...] Du bist an Studien weniger interessiert, wenn du denkst, es läuft eh gut.“ (P04) “[D]adurch, dass sich Schüler in anderen BHS deutlich schwerer tun in Mathe, [sind] die Lehrer [...] mehr mit dem Thema konfrontiert [...], aber bei uns ist die Matura jetzt kein Großevent in dem Sinne.“ (P02) „Lehrer haben nur dann das Bedürfnis, etwas zu sagen, wenn sie das Problem in einem Thema erkennen. HTL-Lehrer haben nichts zu sagen, weil die Matura passt für sie ja eh. Die haben nicht das Bedürfnis, der Welt zu dem Thema was mitzugeben, [...] laut um Hilfe zu schreien.“ (P13)
Lower Value of General Education Grades	When teachers describe grading in a general education subject, such as mathematics, as being of lower value or	„Allgemeinbildende Fächer haben bei uns auch weniger Stellenwert, sag ich jetzt mal, als spezifische Fächer, die dann auch wichtig sind für den späteren Beruf der Schüler. Und da lernen sie ja auch, Mathematik

	importance compared with the students' achievement in profession-specific subjects	richtig anzuwenden. Das Thema Notengebung in Mathe und eben auch so ein Fragebogen ist da einfach zweitrangig [...], wichtig ist, was sie jetzt zum Beispiel in Bautechnik oder eben in anderen Bereichen so draufhaben.“ (P05) „Und dann kommt dazu, dass für viele in der HTL die Matura nicht so einen hohen Stellenwert hat, weil die Schüler ja mit der Ausbildung schon eine klare Berufsperspektive haben und die Matura für viele gar nicht das Hauptziel ist.“ (P8) „Die [Lehrer an technisch orientierten BHS] brauchen sich nicht so um das Pädagogische zu kümmern. Die wollen am Ende des Tages gute Ingenieure haben. [...] Die haben einen anderen Hintergrund, die haben ein anderes Ziel.“ (P11)
Mentality of Technical BHS Teachers	When teachers refer to the mentality of technical BHS colleagues, describing a pragmatic, application-oriented attitude in which completing questionnaires ranks low in priority.	“Ja das kann ich mir ganz gut vorstellen, dass die keinen Fragebogen ausfüllen. Das ist ganz klar Mentalitätssache.“ (P09) „Also prinzipiell haben wir HTL-Lehrer schon eine andere Mentalität als AHS-Lehrer [...]. Ich habe früher an einer AHS unterrichtet und ich kann sagen, wenn man länger an einer bestimmten Schulform unterrichtet, verändert man sich bzw. passt sich an die Mentalität der Schule an.“ (P05) „Wir HTLer sind sehr pragmatisch und anwendungsorientiert [...] In der Prioritätenliste steht da so ein Fragebogen dann eher weit unten.“ (P07)

Appendix D

Cohen's d Values and Response Rates for School Subtypes and Subject Area

	Mathematics		German		English	
	Not missing	Missing	Not missing	Missing	Not missing	Missing
AHS	$d = 0.14$ (n = 11)	$d = 0.29$ (n = 11)	$d = 0.12$ (n = 12)	$d = 0.12$ (n = 16)	$d = 0.06$ (n = 2)	$d = -0.07$ (n = 7)
Technical BHS	$d = -0.30$ (n = 4)	$d = -0.35$ (n = 14)	$d = 0.12$ (n = 3)	$d = -0.03$ (n = 13)	$d = -0.08$ (n = 3)	$d = -0.08$ (n = 13)
Non-technical BHS	$d = 0.11$ (n = 6)	$d = 0.04$ (n = 2)	$d = 0.02$ (n = 7)	$d = -0.01$ (n = 13)	$d = -0.06$ (n = 8)	$d = -0.06$ (n = 9)

Note. Cohen's d = standardized difference in grade distribution between Matura and classroom grades; AHS = academic upper secondary schools; BHS = vocational upper secondary schools; the higher the positive value of Cohen's d , the higher the classroom grade compared to Matura grade; the higher the negative value of Cohen's d , the lower the classroom grade compared to Matura grade.

9.3 Third Study

Grading after reform: How policy change affected teachers' grading for university entrance qualifications

Schatz-Laliberté, V., & Lüftenegger, M. (2025). Grading after reform: How policy change affected teachers' grading for university entrance qualifications. *Studies in Educational Evaluation*, 87. <https://doi.org/10.1016/j.stueduc.2025.101540>

Schatz-Laliberté, Vincent^{1*} (ORCID 0009-0009-3686-4656)

Lüftenegger, Marko^{1,2} (ORCID 0000-0001-8112-976X)

¹ Centre for Teacher Education, Department for Teacher Education, University of Vienna, Austria

² Faculty of Psychology, Department of Developmental and Educational Psychology, University of Vienna, Austria

Status of the study

Published in December 2025

Contribution of the individual authors:

Vincent Schatz-Laliberté: Conceptualization, Data curation, Quantitative and qualitative data analysis, Writing – original draft, review & editing

Marko Lüftenegger: Conceptualization, Supervision, Critical Review



Grading after reform: How policy change affected teachers' grading for university entrance qualifications

Vincent Schatz-Laliberté^{a,*}, Marko Lüftenegger^b

^a Centre for Teacher Education, University of Vienna, Porzellangasse 4, Vienna 1090, Austria

^b Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna, Porzellangasse 4, Vienna 1090, Austria

ARTICLE INFO

Keywords:

Reform
Grading
Grade inflation
Classroom assessment
University entrance qualification

ABSTRACT

This paper investigates the impact of a 2020 Austrian grading reform on teachers' grading practices for university entrance qualifications. The reform introduced a combined model integrating teacher-awarded classroom grades with standardized exam scores. Employing a sequential explanatory mixed-methods design, we analyze longitudinal data on classroom grades in German and Mathematics from 2017 to 2023, alongside 14 in-depth teacher interviews. Quantitative findings reveal significant grade inflation following the reform, particularly in Mathematics. Qualitative results offer insight into the reform's influence on teachers' grading behavior, highlighting a shift toward greater leniency and strategic decision-making to secure students' passing of final exams. Teachers reported a decline in grading standards, along with increased internal conflict and external pressure, raising concerns about the credibility of final outcomes. These studies contribute to educational research by highlighting how policy changes can affect final grades and teacher decision-making, revealing far-reaching, unintended consequences of structural shifts in high-stakes contexts.

1. Introduction

The COVID-19 pandemic posed unprecedented challenges not only for public health and the economy (Greer & King, 2021) but also for education systems worldwide (Daumiller et al., 2023). Schools faced rapid transitions to remote learning (Fridkin et al., 2023), growing learning gaps (Donnelly & Patrinos, 2022), and increased pressure to adapt grading and assessment practices (Panadero et al., 2022). A key concern was ensuring fair and valid assessment for school-leaving qualifications that determine access to higher education (Castro et al., 2020; Tarkar, 2020).

In response, many countries restructured their final assessment systems in secondary school. For example, the UK and Ireland replaced standardized exams with teacher-assessed grades (Doyle et al., 2021; Isaacs & Murphy, 2022), and France replaced its baccalauréat with continuous teacher assessment from the final school year (Daddi, 2020). Austria similarly introduced a lasting grading reform during the pandemic, integrating classroom grades (end-of-year report card grades) with standardized test results for the *Matura*, the national school-leaving exam (Bundesministers für Bildung, 2021). This reform aimed to reduce reliance on high-stakes testing and align with international assessment

practices (Kurier, 2020). Unlike temporary pandemic measures, this reform remains in place (BMBWF, 2023). While studies from the UK suggest that reduced standardization may affect future academic outcomes (McManus et al., 2021), the effects of Austria's grading reform have not yet been scientifically examined.

Using a sequential explanatory mixed-methods design, this paper investigates the impact of Austria's 2020 grading reform from teachers' perspectives. A quantitative study examines grade trends from 2017 to 2023, followed by a qualitative interview study that explores teachers' grading practices and perceived influences, guided by McMillan's (2003) classroom assessment decision-making model. The studies contribute to educational policy and assessment research by using Austria's grading reform as a case to explore how changes in grading policy influence final qualification grades, as well as the underlying processes behind teachers' decisions when assigning grades.

2. Theoretical Framework

A large body of research points to ongoing grade inflation at all levels of education, from primary school to university (e.g., Cheng & Kong, 2023; Chowdhury, 2018; Tyner & Gershenson, 2020), with evidence

* Corresponding author.

E-mail addresses: vincent.schatz@univie.ac.at (V. Schatz-Laliberté), marko.lueftenegger@univie.ac.at (M. Lüftenegger).

<https://doi.org/10.1016/j.stueduc.2025.101540>

Received 8 August 2025; Received in revised form 20 October 2025; Accepted 24 November 2025

Available online 27 November 2025

0191-491X/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

going back some 70 years (Miller, 2014; Walsh, 1979). Literature differentiates between three distinct forms of grade inflation (Tyner & Gershenson, 2020). The first is *static grade inflation*, which refers to a situation where grades exaggerate students' actual achievement at a specific point in time. The second is *dynamic grade inflation*, which highlights changes over time in the relationship between grades and student performance, indicating a shifting standard. Finally, *differential grade inflation* occurs when the alignment between grades and achievement varies across different types of schools or student groups.

As a recent catalyst for grade inflation, both static and dynamic, the Covid-19 pandemic is discussed in the literature (Berne, 2023; Doyle et al., 2021; Goldhaber & Goodman Young, 2024; McManus et al., 2021; Sanchez, 2023; Tillinghast et al., 2023). For instance, a comparison between students' high school grade point averages (HSGPA) and their results on standardized tests in the U.S. demonstrated persistent grade inflation over the last three decades, with a significant increase in inflation during the years 2020 and 2021 (Sanchez, 2023). A notable finding in this study was a ceiling effect, indicating a decline in the number of students receiving the second-highest grade and an increase in the number of students attaining the highest grade. While the researchers are cautious in directly attributing this sudden increase in grade inflation to the Covid-19 pandemic (the disparity between rising HSGPAs and declining standardized test scores already grew more pronounced starting in 2018), they highlight the potential risk associated with nonachievement factors (i.e., effort, participation, behavior, and the teachers' personal impressions) in grades, which create challenges in comparing students' academic achievement nationwide (Sanchez, 2023).

Similar findings are reported by Goldhaber and Goodman (2024), who compared middle and high school grades with standardized test scores in Washington State before and after the COVID-19 pandemic. Their findings, too, revealed a significant grade inflation over time, particularly following the onset of the pandemic. The grades declined in the year following the pandemic, yet remained higher than pre-pandemic levels (particularly in Mathematics). In addition, findings indicate a decline in the correlation between grades and standardized test scores. For instance, in the school year 2021/22, only 19 % of students who did not meet the grade level expectations in Mathematics based on the results of their standardized test scores received a failing grade from their teacher, while 40 % received the highest (15 %) or second highest (25 %) grade.

One of the main reasons given in the literature for the sudden increase in high grades and grade inflation is that many countries changed their grading policies due to the COVID-19 pandemic (Doyle et al., 2021; Goldhaber & Goodman Young, 2024; McManus et al., 2021; O'Leary & Lysaght, 2024; Berne, 2023). In the UK, the standardized higher education entrance exams were canceled in response to the school closures in March 2020. Instead, the teachers had to assess their students by estimating the grades they would likely have achieved on the standardized exams (McManus et al., 2021). McManus et al. (2021) explored possible implications of this change by comparing the predictive validity of standardized exam scores and teacher grades for study outcomes in higher education (i.e., medical studies). Their findings revealed that standardized exam scores correlated significantly better with study outcomes than teacher grades (teacher grades were accurate in 48.8 % of cases, but 44.7 % were overpredicted and only 6.5 % underpredicted), suggesting that replacing standardized exam scores with teacher grades may lead to lower academic performance for 2020 higher education (i.e., medical school) entrants compared with previous years.

A similar policy change in grading took place in Ireland. Before the COVID-19 pandemic, the final secondary school leaving examinations were known for their "brutal fairness" (Schröter et al., 2022, p. 92), due to the central grading by a specially trained external examination board. In 2020, however, this grading responsibility was transferred to the students' own teachers (Schröter et al., 2022). Allocation of study places was (and is still) based on the total number of points achieved on these

assessments. On average, students received one level higher grades than in previous years, which led to the need for the education system to provide thousands of additional new higher education study places (Doyle et al., 2021) – an example of the far-reaching consequences of a seemingly minor change to the assessment mechanism.

3. Literature review

The observation of higher teacher-awarded grades since the pandemic appears to be just the tip of the iceberg, underneath which research reveals huge challenges for teachers in grading (Doyle et al., 2021; Koller, 2021; Vahle et al., 2023). The year 2020 has been accompanied by changes in some countries' grading policies to ensure that students' educational success isn't hindered by the challenges posed by the pandemic (e.g., school closures, distance learning, health issues) (Castro et al., 2020; Chan, 2022; McManus et al., 2021). For example, California's official grading policy adopted a "hold harmless" (Castro et al., 2020) policy that ensured that a student's final grade for the second semester could not be lower than their grade on the day schools were closed (but higher grades could be awarded).

A qualitative study in Florida found that teachers felt compelled by official grading policies to pass students who would have failed under normal circumstances (Berne, 2023). However, this leniency in grading seriously impacted students' motivation to attend, participate, and submit work, which in turn challenged teachers as they reported an increase in student cheating, academic dishonesty, and observed grade inflation (Berne, 2023). Indeed, a study investigating unintended consequences of lenient grading policies found that greater grading leniency resulted in higher grades without a corresponding increase in student achievement; at the same time, however, it led to a significant rise in student absenteeism (Bowden et al., 2023).

Another study (Koller, 2021) revealed that teachers adapted their grading by showing greater leniency, reducing penalties for tardiness, emphasizing attendance, participation, and task completion over mastery, and allowing multiple attempts or additional opportunities for success. Moreover, there were major differences between individual teachers in adapting changes in grading guidelines, which was a challenge for the teachers themselves (Berne, 2023) as well as for policy-makers and researchers regarding fairness and comparability of final grades (Doyle et al., 2021; McManus et al., 2021).

These changes in teachers' grading practices clearly illustrate what McMillan (2003) described in his grade decision-making model about 20 years earlier. The model provides a valuable framework for understanding the observed shifts in grading practices, as it highlights that teachers' grading decisions are connected (among others) to their internal beliefs and values as well as external factors, such as classroom realities and external pressures from parents or policy guidelines. First, the model describes a teacher belief called "pulling for students" (McMillan, 2003, p. 36), meaning that teachers want to support their students' success by raising their grades, especially for borderline cases. According to Koller (2021), increased leniency, reduced penalties for tardiness, and emphasis on participation and task completion rather than mastery might reflect teachers' commitment to ensuring that students were not disadvantaged by the challenging circumstances. Second, the pandemic intensified classroom realities, such as absenteeism, disengagement, and differences in student ability (Berne, 2023). Teachers reported frustration with declining student motivation and increased academic dishonesty (Berne, 2023), further complicating grading. These challenges demonstrate McMillan's earlier findings that teachers struggle to balance their ideal grading approaches with the practical realities they face in class. Third, external factors had an even stronger influence on teachers' grading decisions during the pandemic (Tarkar, 2020). McMillan found that teachers often follow official grading policies only loosely or sometimes ignore them. However, during the pandemic, policies like California's "hold harmless" approach (Castro et al., 2020) required teachers to be more lenient, restricting

their professional judgment in grading. McMillan's model predicts such tensions as teachers try to balance institutional rules like grading policies with their own beliefs about fair and meaningful grading.

4. The current study

In Austria, the pandemic opened the opportunity to implement a grading reform in 2020 rapidly. This reform led to the merging of the grades given in the final year of school (non-standardized, traditional grading; from now on referred to as classroom grades) with the standardized grades from the university entrance certificate examinations (*Matura*). Both grades are given by the students' own teachers, without anonymization or moderation by colleagues, teachers from other schools, or central examination boards. Hence, the 2020 grading reform meant that students' classroom grades from their final school year now account for 50 % of their final *Matura* grade.

The Austrian grading system consists of 5 grades, with grade 1 being the highest, grade 5 being the lowest, and the only failing grade. This means that since the grading reform, it is no longer possible for students to fail the *Matura* examinations if they have received a classroom grade of at least 3 from their teachers in the subjects relevant to the *Matura* (primarily German, Mathematics, and English).¹ This paper will explore how this fact might have altered teachers' grading decisions and which mechanisms underlie them.

Therefore, by drawing on the theoretical background, this study uses two phases to explore changes in classroom grading resulting from the 2020 grading reform. The first phase involves a quantitative longitudinal investigation examining classroom grade changes. The second phase involves a qualitative investigation of teachers' perceptions of these changes, as well as possible changes in their grading practices since 2020. Based on McMillan's (2003) grading decision-making model, this second investigation also examines teachers' perceptions of internal and external factors that may have influenced their grading practices since 2020. Internally, teachers may feel personally responsible for supporting students by awarding higher grades, particularly in difficult circumstances. Externally, they may face pressure from students, parents, or school administrators to award grades that will improve students' university entrance prospects.

The following research questions are addressed in this paper:

Phase 1

RQ1. How have final classroom grades changed since the 2020 grading reform in Austria?

Phase 2

RQ2. How do teachers explain these changes?

RQ3. What are teachers' perceptions of possible changes in grading practices since the 2020 grading reform?

RQ4. What external factors did teachers perceive to impact their grading, and to what extent have these factors changed since 2020?

RQ5. What internal factors did teachers perceive to impact their grading, and to what extent have these factors changed since 2020?

¹ An exception to this was introduced one year after the grading reform. Namely, students must have completed at least 30 % of the tasks in the *Matura* (applying to both the written and oral *Matura* exams) so that the classroom grade can compensate for a *Matura* failing grade.

5. Methods

5.1. Research design

The present paper employs a sequential explanatory mixed methods design (Ivankova et al., 2006), consisting of a quantitative longitudinal analysis (Phase 1) followed by qualitative interviews (Phase 2). In the first, quantitative phase (RQ1), changes in classroom grades over time are examined using a longitudinal data set, focusing on trends before and after the 2020 grading reform in the *Matura*-relevant subjects German and Mathematics. In the subsequent qualitative phase (RQ2–5), German and Mathematics teachers are interviewed to investigate their perceptions of observed classroom grade changes identified in the quantitative analysis, potential changes in grading practices since the 2020 grading reform, and external and internal factors that may have influenced their grading practices since 2020 (see Appendix A for the full interview guideline).

5.2. Data

For the quantitative phase, we collected classroom grades from 6798 students in seven schools in Vienna, Austria. The schools were very different in terms of school type, size, district, and socioeconomic background of the students. Specifically, the data included three upper academic schools and four upper vocational schools, indicating that the students in our dataset differed in their educational backgrounds. Students from academic schools are primarily prepared for university studies, whereas students from vocational schools obtain a diploma that qualifies them for entry into the labour market in their chosen field, in addition to the *Matura*. The vocational schools in our sample were mainly oriented toward technical and economic fields.

The collected classroom grades were awarded by 122 teachers over seven years, from 2017 to 2023, in the *Matura*-relevant subjects German and Mathematics. The dataset includes 3505 grade pairs for German and 3293 for Mathematics. On average, each teacher graded approximately 56 students, with large differences between teachers ($SD = 38$). The average age of the teachers was 47 ($SD = 10$).

According to official grading guidelines, classroom grades reflect a combination of academic achievement, assessed through written tests and oral examinations, as well as students' effort and participation in class (RIS, Rechtsinformation des Bundes, 2025). Because the grades in our sample were awarded in the final school year, we assume that the assessments were already aligned with *Matura* standards, as the Austrian Ministry provides a centralized task pool to help teachers prepare their students for the *Matura* (Bundesministerium für Bildung, 2025). The Austrian grading scale ranges from 1 to 5, with 1 as the highest (best) grade and 5 as the lowest (worst). A grade of 5 is the only failing grade and disqualifies students from taking the *Matura* examination in the respective subject.

For the qualitative phase, semi-structured interviews were conducted with participants recruited via purposive sampling through the researchers' professional networks. This procedure enabled the targeted selection of teachers whose profiles aligned with the study's objectives. To ensure the inclusion of information-rich cases with a diverse representation of perspectives and experiences within the participant group, participants had to meet the following criteria: (a) be active secondary school teachers at an academic or vocational upper secondary school, (b) teach either German or Mathematics, and (c) have graded at least two final-year *Matura* classes both before and after the 2020 grading reform, specifically during the period from 2017 to 2023. To circumvent the possible issue of self-selection bias, the participants were only provided with general information about the aim of the study (i.e., it was explained to them that the survey was about grading in general without mentioning its focus on the 2020 grading reform). In addition, some participants were asked to suggest teaching colleagues with whom they strongly disagree regarding their opinions, values, and behaviors when

grading to maximize the diversity of the views in subsequent interviews.

The interview participants were drawn from four Austrian federal states (Vienna, Tyrol, Upper Austria, and Lower Austria). They represented diverse school contexts, including rural and urban settings and different school types. A table with brief teacher profiles—covering gender, subject, current school context (school type, location, and federal state), and years of teaching experience—is provided in Appendix C. Interview durations ranged from 13 to 54 min, with a mean duration of approximately 36. The number of participants in this phase was not predetermined; the principle of data saturation guided it. Consequently, the interview process continued until no new insights or thematic developments were observed, resulting in 14 interviews (seven with German teachers and seven with Mathematics teachers). This approach allowed for an in-depth investigation of the research questions and ensured the collection of comprehensive and meaningful qualitative data.

5.3. Data analysis

To address Research Question 1 (examining the extent to which final classroom grades have changed since the 2020 grading reform), we analyzed descriptive statistics and conducted year-to-year comparisons of grade distributions in Mathematics and German from 2017 to 2023 via chi-square tests using RStudio (RStudio Team, 2020). The year-to-year comparisons were performed between consecutive years and between 2019, the final year before the reform, and each subsequent year up to 2023. To enhance the robustness of the analysis, the chi-square tests were complemented by ordinal logistic regression analyses for mathematics and German, conducted in jamovi (The jamovi project, 2022). This combination leveraged the respective strengths of both methods: chi-square tests offered detailed insights into distributional changes between individual years—the focus of our quantitative research question—while ordinal logistic regressions served as robustness checks, verifying whether these observed patterns held when controlling for school type and allowing for subgroup comparisons across different school types.

The semi-structured interviews (RQ 2–5) were analyzed by applying manifest qualitative content analysis (Graneheim & Lundman, 2004). Since Research Questions 2–5 address explicit aspects (i.e., teachers' explanations for grade changes, their perceptions of possible shifts in grading practices, and factors influencing their grading behavior following the 2020 grading reform), we focused on coding manifest content that was directly observable in the data (Downe-Wamboldt, 1992). We followed the analysis process outlined by Kleinheksel et al. (2020). First, we identified units of meaning, typically sentences or sentence fragments, which we condensed when necessary by removing verbal fillers, duplicate words, or simplifying compound sentences. Second, we labelled equivalent units with a code, and third, we grouped similar codes into categories to structure the data. Kleinheksel et al. (2020) describe a fourth step involving the combination of related categories into themes, but emphasize that this step is primarily suited for latent analyses. Since our study focused on manifest content, we restricted the analysis to the first three steps.

To ensure consistency, a subset of four transcripts from each subject was independently coded by the first author and a teacher collaborator familiar with both the school context and qualitative data analysis methods. Any discrepancies were discussed and resolved through consensus. The remaining interviews were then coded using the agreed-upon coding scheme. Categories were developed by the first author, incorporating feedback from both the teacher and the second author.

6. Results

6.1. Classroom grade changes

6.1.1. Descriptive statistics

We calculated descriptive statistics as percentages of grades 1–4 for each subject (Mathematics and German) and year (2017–2023). These were used to gain first insights into possible shifts in classroom grade distributions, with particular attention to changes following the 2020 grading reform. In Mathematics (Fig. 1), a noticeable shift appeared immediately in 2020: the proportion of grade 4 dropped, while higher grades (1–3) became more frequent compared to the pre-reform years 2017–2019. This trend continued in subsequent years, indicating an overall shift towards higher classroom grading in Mathematics.

In contrast, grade distributions in German (Fig. 2) remained relatively stable through 2020, with a more noticeable shift emerging in 2021. Like in Mathematics, the proportion of grade 4 declined, accompanied by an increase in grade 1 and particularly in grade 2. Grade 3 remained relatively stable.

In both subjects, grade 5 is rarely awarded and remains consistently low across all years, suggesting that failing grades play a minimal role in final-year classroom grading. Regarding subject differences, the grading patterns indicate that the grading reform may have affected the two subjects differently, with changes in Mathematics grading emerging one year earlier than in German.

6.1.2. Differences in grade distributions across the years

Chi-square tests were performed using R to test for statistically significant differences in classroom grade distributions between consecutive years (RStudio Team, 2020). Additionally, pre-reform grading was compared with that in subsequent years following the 2020 grading reform. As no significant differences in grade distributions were observed between 2017 and 2019, 2019 was defined as a reference point for pre-reform grading, as it was the last year before the 2020 grading reform. Table 1 summarizes the results of these tests for both the subjects of German and Mathematics.

In the subject German, no statistically significant differences in grade distributions were observed between 2017 and 2019, confirming the stability of grading practices in the pre-reform period. Comparisons between 2019 and the post-reform years revealed no significant change in 2020, the first year after the reform. However, a statistically significant shift happened in 2021 ($p < .001$), showing the largest effect size (Cramér's $V = .158$) among all comparisons and highlighting a notable change in grading distribution one year after the reform. Statistically significant differences were also detected for 2019 vs. 2023 and 2020 vs. 2021, with small effect sizes (Cramér's $V \approx .10$). Other comparisons, including those between consecutive post-reform years (e.g., 2022 vs. 2023), did not yield statistically significant results.

In contrast to German, Mathematics showed a statistically significant shift in grade distributions immediately in the first year following the 2020 grading reform: compared to 2019, a statistically significant change was already evident in 2020. This changed grade distribution persisted in subsequent years, as shown by additional significant differences in 2019 vs. 2021 and 2019 vs. 2023. In Mathematics, effect sizes were slightly larger than in German, averaging around Cramér's $V \approx .12$. Grade distributions among the post-reform years themselves, such as 2020 vs. 2021, 2021 vs. 2022, and 2022 vs. 2023, did not differ significantly, suggesting a stable post-reform grading pattern in Mathematics.

Comparing German and Mathematics reveals distinct patterns in the timing and stability of grade changes following the 2020 reform. While grades in German showed a delayed response, with the first significant shift only occurring in 2021, grades in Mathematics showed an immediate and significant change in 2020 itself. Furthermore, Mathematics grades appeared to stabilize after the initial shift, with no further significant differences observed between post-reform years. In contrast,

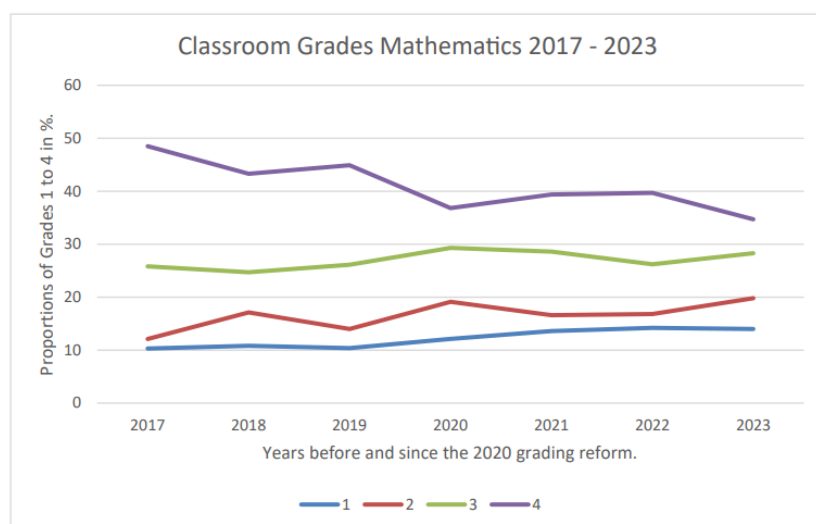


Fig. 1. Development of final classroom grades in mathematics 1–4 from 2017 to 2023, including the reform year 2020.

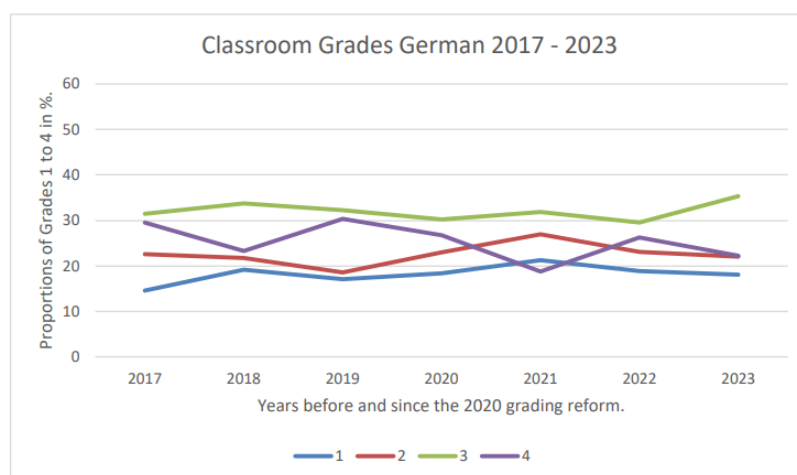


Fig. 2. Development of final classroom grades in German 1–4 from 2017 to 2023, including the reform year 2020.

German grades showed a less consistent pattern, with differences between subsequent years.

To enhance the robustness of our results, we supplemented the chi-square tests with ordinal logistic regression analyses using jamovi (The jamovi project, 2022). These analyses were performed separately for German and mathematics. Classroom grades, treated as ordinal values, served as the dependent variable. School type, distinguishing between academic and vocational schools, and year, spanning 2017–2023, with 2019 defined as the reference year, were included as independent variables. Classroom grades, treated as ordinal values, served as the dependent variable. School type, distinguishing between academic and vocational schools, and year, spanning 2017–2023, with 2019 defined as the reference year (in line with the chi-square tests), were included as independent variables.

The ordinal logistic regression models' results supported the chi-square test findings. In German (Table 2) and Mathematics (Table 3), the same years that were significant in the chi-square analyses also

showed significant effects in the regression models. The size of the effects was also very similar. In German, the strongest effect was again found for 2021, with an odds ratio of 1.64. In Mathematics, the effects were again a bit stronger than in German, with odds ratios averaging around 1.4 across the significant years.

In addition, school type had a significant effect in both subjects, with students in vocational schools receiving more favourable grades than those in academic schools. We ran separate ordinal logistic regression models for each school type to explore this further. The results showed that vocational schools drove the significant effects almost entirely. Vocational schools account for about two-thirds of our dataset, yet the divergent pattern between school types deserves closer attention in future research. Although these differences were not the focus of this study, we provide results for academic and vocational schools in German and Mathematics in Appendix D (Table 4: German, academic schools; Table 5: German, vocational schools; Table 6: Mathematics, academic schools; Table 7: Mathematics, vocational schools).

Table 1
Differences in classroom grade distributions (chi-square tests) across 2017 – 2023.

German					
Years Compared	χ^2	df	N	p-value	Cramér's v
2017 / 2018	7.929	4	442, 468	.094	.094
2018 / 2019	8.053	4	468, 533	.090	.090
2019 / 2020	4.476	4	533, 538	.345	.065
2019 / 2021	27.732	4	533, 574	< .001	.158
2019 / 2022	5.617	4	533, 577	.230	.071
2019 / 2023	9.763	4	533, 557	.045	.095
2020 / 2021	11.738	4	538, 574	.019	.103
2021 / 2022	13.134	4	574, 577	.011	.107
2022 / 2023	5.022	4	577, 557	.285	.067
Mathematics					
Years Compared	χ^2	df	N	p-value	Cramér's v
2017 / 2018	4.729	4	388, 434	.316	.076
2018 / 2019	2.880	4	434, 472	.578	.057
2019 / 2020	13.497	4	472, 570	.009	.114
2019 / 2021	13.457	4	472, 553	.009	.115
2019 / 2022	7.387	4	472, 585	.117	.084
2019 / 2023	14.961	4	472, 449	.005	.128
2020 / 2021	2.916	4	570, 553	.572	.051
2021 / 2022	3.727	4	553, 585	.444	.057
2022 / 2023	3.402	4	585, 449	.493	.057

Note. Dark blue: comparison of the first year of grading reform (2020) with the baseline year (2019); light blue: comparison of subsequent years since grading reform (2021–2023) with the baseline year (2019).

Table 2
Ordinal logistic regression predicting classroom grades in German (2017–2023).

Model Fit / Predictors							
Model Fit Measures	Estimate	SE	Z	p	Odds ratio	95 % CI Lower	95 % CI Upper
Deviance	10435						
AIC	10457						
R ² _{MCF}	0.00540						
Predictors							
Schooltype: BHS – AHS	-0.324	0.063	-5.146	<.001	0.724	0.640	0.818
year:							
2017 – 2019	-0.011	0.116	-0.095	0.924	0.989	0.787	1.242
2018 – 2019	0.258	0.114	2.251	0.024	1.294	1.034	1.620
2020 – 2019	0.184	0.111	1.659	0.097	1.202	0.967	1.493
2021 – 2019	0.497	0.108	4.585	<.001	1.643	1.329	2.032
2022 – 2019	0.190	0.109	1.743	0.081	1.209	0.977	1.497
2023 – 2019	0.217	0.109	1.987	0.047	1.242	1.003	1.538

Note. The dependent variable 'classroomgrade' has the following order: 5 | 4 | 3 | 2 | 1; Models estimated using sample size of N = 3682; BHS = vocational upper secondary schools, AHS = academic upper secondary schools.

6.2. Teacher perceptions of classroom grade changes

Overall, the analysis of the interviews revealed 12 main categories (see Appendix B). The first four categories are related to research questions 2 and 3, while categories 5–11 are related to research questions 4 (categories 5–7) and 5 (categories 8–11). Although not directly related to our initial research questions, almost every participant addressed one additional category (12). Given its prominence across interviews, we included it in the results section of our qualitative study.

The first two research questions of the qualitative study addressed

teachers' explanations for the grade changes over time (RQ2) and their perceptions of shifts in grading practices since the 2020 reform (RQ3). Because the categories emerging from the interviews for both questions largely overlapped, they are discussed together here.

6.2.1. Category 1: extraordinary circumstances in pandemic years

Many teachers attributed the early rise in grades not only to the 2020 grading reform but also to the exceptional circumstances of the pandemic. They described a general climate of leniency, as one teacher noted: "During the COVID years, we teachers were much more lenient in

Table 3
Ordinal logistic regression predicting classroom grades in Mathematics (2017–2023).

Model Fit / Predictors							
Model Fit Measures	Estimate	SE	Z	p	Odds ratio	95 % CI Lower	95 % CI Upper
Deviance	9448						
AIC	9470						
R ² _{McF}	0.00813						
Predictors							
Schooltype: BHS – AHS	-0.442	0.066	-6.715	< .001	0.643	0.565	0.731
year:							
2017 – 2019	-0.038	0.127	-0.303	0.762	0.962	0.750	1.234
2018 – 2019	0.076	0.124	0.617	0.537	1.079	0.847	1.375
2020 – 2019	0.363	0.114	3.179	0.001	1.438	1.150	1.799
2021 – 2019	0.309	0.115	2.678	0.007	1.362	1.086	1.707
2022 – 2019	0.289	0.114	2.531	0.011	1.335	1.068	1.671
2023 – 2019	0.378	0.122	3.107	0.002	1.460	1.150	1.853

Note. The dependent variable 'classroomgrade' has the following order: 5 | 4 | 3 | 2 | 1; Models estimated using sample size of N = 3682; BHS = vocational upper secondary schools, AHS = academic upper secondary schools.

our grading" (P02). Also, the disrupted learning environment and limited assessment opportunities were frequently cited. Many teachers felt they couldn't apply their usual standards with classes being moved online and distance learning. One participant stated, "They didn't really learn anything, and teaching was only online" (P01). Others emphasized the difficulty of objective assessment: "There were fewer opportunities to assess objectively because fewer students were even present" (P02). This led to more generous grading decisions, such as "At the beginning, I definitely gave more 3 s than I would have given 4 s" (P13).

6.2.2. Category 2: explanations for grade curves over the years

In contrast to the immediate effect of the 2020 reform on Mathematics grades, the effect on grading in German was not immediate. Some German teachers noted that the decision to include classroom grades into final Matura grades came too late in the school year to affect grading practices in that first year. "The measure in 2020 came just before the Matura – the classroom grades were already done" (P06). Another teacher reflected, "A lot had already been completed; there wasn't much time left to really change anything" (P05). Teachers also reported a delayed recognition of the reform's strategic consequences. One reflected, "Maybe it took teachers a year to understand what kind of impact this would have" (P07). Another admitted, "In 2020, the pandemic only started in March, and classroom grades were already mostly finalized. Then, suddenly, it was announced that classroom grades would count towards the Matura" (P01).

Many teachers believed that the number of classroom grade 3 s had increased since the grading reform – while classroom grade 4 s had become less common – because a classroom grade 3 is the "safe grade" needed to pass the Matura even if the exam is failed (unlike a classroom grade 4, which doesn't compensate for a failed Matura exam). However, our quantitative data tell a different story: It was mainly the number of classroom grade 2 s (and even grade 1 s) that increased, while classroom grade 4 s declined. Teachers used two mechanisms to explain this observed increase in classroom grade 2. First, some described a norm-referenced grading logic, in which relative performance within the class led to upward adjustments: "If I'm giving borderline 3 s, then grades that were originally 3 s but leaning toward a 2 are no longer comparable in the same way, so those also get bumped up to a 2" (P10). Second, others referred to strategic compensation, deliberately awarding higher classroom grades to offset anticipated weaker Matura results: "If someone writes a 4 on the Matura and has a 2 in the classroom grade, they end up with a 3 [in the Matura] overall. But if they have a 3 [classroom grade] and then a 4 [Matura exam grade], it's a 4 [in the Matura] overall" (P13).

6.2.3. Category 3: strategies to secure positive matura outcomes

Teachers described how the reform incentivized various strategies to help students achieve a minimum passing grade on the Matura. A commonly mentioned development was an increase in oral exams between classroom grades 4 and 3: "We never used to have so many tests between 3 and 4" (P09); "I've noticed a rise in exams, especially between 3 and 4, hoping to use the classroom grade as a backup" (P10). Several teachers also described an increased willingness among teaching colleagues to reward effort or participation. "Everything is offered, everything is accepted. Students ask, 'How can I improve my grade?' – and teachers do everything to grade them positively" (P05). One teacher summed it up by saying, "They show effort, they participate – then they get the 3" (P11). The result seems to be a systemic inflation of the classroom grade 3, with teachers describing it as a kind of default. "There are definitely more 3 s and fewer 4 s now" (P06). This was a tactical move for some teachers to prevent failure: "If I give her a 3, then she can write a 5 on the Matura and still pass" (P06). Another explained, "We build a buffer in the classroom grade, because we don't want to be too lenient in the Matura, which is still controlled. So, what used to be a [classroom grade] 4 is now a 3" (P07).

6.2.4. Category 4: professional integrity and achievement orientation

Teachers also expressed a strong commitment to grading integrity and fairness. They reported trying to resist grade inflation and emphasized the importance of performance-based assessment. Several explicitly stated that their own grading behavior had remained consistent, in contrast to what they perceived as growing leniency among colleagues: "I always tried to resist that [...] I don't think I gave more 3 s than before" (P05). Others reported rejecting non-academic forms of compensation, such as extra assignments or effort-based bonuses: "You [to students] can give presentations or use ChatGPT every day to work on something, but that still won't turn a 4 into a 3" (P01). Some teachers even described a countermovement after an initial period of leniency in grading, motivated by concerns about declining student motivation and the perceived devaluation of the Matura. As one teacher reflected: "After the classroom grade started counting, I initially tended to give the better grade. But over the past two years, I've gone back to being stricter with students who were really more like a 4. I noticed that once they got a 3, they stopped putting in effort, thinking, 'That's enough, I'll pass the Matura with the 30 %.' I don't think that's right – if the Matura still exists and is supposed to have any value, then students should at least put in some effort" (P13).

Research questions 4 and 5 addressed external (RQ4) and internal (RQ5) factors perceived by teachers to impact their grading, and the extent to which these factors have changed since 2020. Concerning external factors (RQ4), the interviews revealed three main categories

(5–7).

6.2.5. Category 5: student and parental influences on grading

The most commonly cited pressure teachers experienced was increased pressure from students and negotiations for better classroom grades since the grading reform. In Matura-relevant subjects, students increasingly pushed to receive at least a classroom grade of 3: “The hunt for the 3 is massive. What annoys us throughout the whole Matura year is that they just want their 3 to feel safe” (P01). Teachers described persistent discussions and student attempts to negotiate: “They ask, ‘What else can I do, can I give a presentation, take another test, how can I still get a 3?’” (P05). According to several, this dynamic has intensified since classroom grades began contributing to Matura outcomes. At the same time, teachers voiced frustration about unrealistic student expectations, especially from those with minimal effort: “They do nothing for five years and then want a 3” (P01). Some expressed that students expected favorable grades without proportional achievement: “They wrote a 5 during the semester and then came to me saying, ‘What can I do to get a 3?’” (P13).

Parental pressure was described as a factor that existed before the reform. Some parents approached teachers even before the Matura to advocate for leniency: “They wanted to make sure we understood their child was applying to England or the US and that we should be generous with the grades” (P03). In more extreme cases, teachers reported confrontations, such as: “I also had a conversation with a father who shouted at me: ‘You are not what my daughter needs right now!’” (P04).

However, teachers in vocational upper secondary schools or rural areas described less parental interference: “Out here in the countryside, we really have a very decent clientele. We don’t have the problems that schools in larger cities do” (P02).

6.2.6. Category 6: pressures from principals and head teachers

Several teachers described how principals tend to side with students: “Our principal is very student-friendly. He almost always takes the students’ side, and you as a teacher have to justify yourself” (P02). Some reported that principals reviewed grade distributions and questioned teachers when negative outliers (e.g., a 5 in an otherwise good report card) appeared. There was also pressure to avoid failing students, often driven by administrative convenience: “The school leadership wants students to pass because compensation exams are a huge effort. Anything that means more work is not appreciated” (P05). This pressure was often unspoken but perceived: “It’s something you feel, even if no one ever says it” (P06).

While this ‘student-friendliness’ of principals seems to be a general trend, the following dynamic is enabled directly by the 2020 grading reform: Teachers reported being pressured to fail students strategically, so a final grade of 4 wouldn’t appear on the Matura report card – since a 5 could still be raised via compensation: “Parents were so persistent – they wanted a 5 so that she’d have a chance to earn a 3 later. The principal was under pressure too” (P03).

Lastly, class head teachers were described as another source of ongoing pressure. According to interviewed teachers, some class heads emotionally identified with their class and pushed for better outcomes: “Some see the class head role like being a mom or dad – ‘these are my children, and they all need to get through the Matura!’” (P01). Others sought to preserve their class’s reputation: “He always had the best class in school and asked me not to give a 2 but a 1.” (P02).

6.2.7. Category 7: systemic drivers of grade leniency

Teachers described informal norms and moral appeals during the pandemic as external pressure to be lenient. As one put it: “They were poor kids, we were all struggling – so we were told to be kind” (P01). Another teacher referred to moral responsibility: “We were given a kind of aftercare mission as teachers” (P03).

Some saw the grading reform itself as an external factor and just one piece in a series of systemic measures that undermined grading

standards. One teacher noted: “When there were too many fails [in the Matura], they shortened the Math word problems. Then they said classroom grades would count. It has nothing to do with central exams anymore” (P05).

Finally, some teachers viewed the reform as a form of protection for underperforming teachers, whose weak preparation would no longer be as visible due to the 50 % weight of classroom grades: “Some teachers just hand out 3 s to protect themselves. If a student gets a 5 on the Matura, that turns into a 4 overall – and no one ever looks into it again” (P01).

In addition to these external factors, the analysis of the interviews also revealed four key categories regarding internal factors (RQ5) affecting grading decisions.

6.2.8. Category 8: empathy and compassion for students

Teachers’ empathy for students was a strong internal factor shaping grading decisions, particularly during and shortly after the pandemic. Many teachers described more lenient grading in response to the emotional burden on students: “There were students with tough situations [...] when I heard those stories, I was more lenient. I didn’t want to put 10,000 obstacles in their way to passing the Matura” (P13). Others reflected more generally: “Didn’t we all feel deep pity for the COVID cohorts?” (P06). Teachers also reported being particularly affected by students’ psychological challenges or difficult family circumstances, which made it harder for them to grade objectively: “If I see that a student is really doing badly psychologically [...] then I’m more likely to think, ‘Okay, I’ll give her that one extra point’” (P08). As one teacher put it: “You know the background, the home situation [...] and sometimes you’re just glad that you can give a classroom grade ‘3’ or even a ‘2’ with a clear conscience” (P12). A special case of this was compassionate grading for hardworking non-native speakers, especially girls: “We really have very diligent, hardworking girls who don’t have German as a first language [...] so sometimes I say, ‘Fine!’” (P01). Teachers described balancing formal performance with effort and context: “She just can’t write a passing text, but she works so hard all year [...] so I say, okay, she gets her classroom grade 3 and she’s fine for the Matura” (P02).

6.2.9. Category 9: torn nature of the teacher role

Teachers frequently described being emotionally torn between their (preferred) supportive role and grading responsibility. The emotional burden of grading students they had known for years was intense: “You talk to students about their personal problems, and then suddenly you’re their German teacher again and have to give them a failing grade. That’s really hard. [...] You can’t sleep, you panic before correcting, thinking: ‘Please let it be positive!’” (P05). Some found themselves or their colleagues emotionally blocked: “She can’t manage to write a grade under the paper because she’s so emotionally attached to her students” (P02). Many teachers emphasized their self-image as supporters, not gatekeepers: “Very few teachers become teachers just to do the job – they generally see themselves as supporters, not examiners” (P10), leading some to go to great lengths to provide additional opportunities: “She felt it was important to offer the students the chance and the time to take these extra exams” (P10).

Some saw the grading reform as a relief from the emotional conflict of failing students: “Some colleagues were probably glad the new system existed so they didn’t have to give a failing grade on the Matura and know the student didn’t pass” (P05). One teacher admitted: “I don’t really like doing it [grading], especially in the final year. I’d prefer someone else would come in and be the ‘bad guy’ instead of me” (P12).

6.2.10. Category 10: teachers’ achievement values in grading

Despite these emotional pressures, several teachers described a renewed focus on performance and academic standards in response to perceived grade inflation. “Last year, I was more generous – if they participated and handed in homework, I gave a 3 even with a 4 on the test. I’m not doing that anymore” (P11). Others spoke of balancing

compassion with standards: “Sometimes I want them to succeed, but then I think: I also don’t want it to feel like none of this matters anymore” (P04). Some expressed frustration with colleagues who gave higher classroom grades simply to avoid extra work: “There have always been teachers who avoid September exams – you never see them there. Their grades range from 1 to 3, maybe a 4 at worst. They just don’t want the extra hassle” (P06).

Finally, teachers reflected on the subjectivity built into classroom grading, which allows for intentional grade manipulation since grading reform: “The classroom grades that get included [into the final Matura grades] are actually quite arbitrary. Every teacher can decide what kind of test they give – whether it’s on Matura level or just ‘re-chewed’ stuff” (P05). As one put it: “I can make a test hard or easy, depending on what I want the outcome to be” (P06).

6.2.11. Category 11: student-friendliness as self-image

A commonly mentioned internal factor influencing grading – though not directly tied to the reform – was the desire to be perceived as student-friendly. Several teachers saw this as a barrier to fair and objective grading. Some noted that colleagues seemed reluctant to assign failing grades out of a need for student approval: “We have a colleague who simply can’t give a 5. [...] She always wants to be liked. And I think that need to be loved is a bit misplaced for a teacher, because you’re supposed to grade fairly” (P05).

Others observed this tendency, especially among younger teachers striving for popularity: “I’ve seen a trend among younger teachers who do everything to be popular – TikTok channels, showing their private life, going to parties with students. [...] That just doesn’t fit with giving bad grades” (P06).

Teachers also described pressure to avoid being seen as strict or unfriendly when applying higher standards. One said: “If you insist on strict grading, you get portrayed as frustrated or student-unfriendly [...] and then you think, ‘I want to be the one who gives good grades too’” (P04). This created emotional tension between maintaining relationships and upholding expectations.

Some criticized what they viewed as excessive student-friendliness, including overly generous interpretations of student work: “This ‘benefit-of-the-doubt’ correcting – ‘he probably meant that,’ ‘he just can’t express himself’ [...] I doubt it’s always neutral and objective” (P05). Another summed it up more broadly: “And the other side is just a bit of a misunderstood kind of student-friendliness, you know?” (P01).

6.2.12. Category 12: teachers’ attitudes and criticism of the grading reform

While a few participants acknowledged the benefits of the 2020 grading reform, most expressed strong criticism. Many questioned its impact on assessment standards, fairness, student motivation, and the integrity of the Matura.

A key concern was the perceived devaluation of the oral exam. Teachers reported cases where students passed the Matura despite poor oral performance, solely due to high classroom grades. One teacher remarked: “Since the reform, the oral Matura is just an excuse to dress nicely. [...] They get a 3 because 1 and 5 is 3” (P01). Others described the oral as a “farce” (P09), with students passing despite lacking basic knowledge (P02). Several teachers criticized the reform for allowing students to pass despite failing the standardized exam. As one put it: “If the Matura is negative, I think the classroom grade should not be included. [...] Otherwise, you can abolish the Matura” (P02). Participants also raised concerns about declining centralized standards. Some argued the Central Matura had lost credibility: “What annoys me is that we still pretend as if there were objective grades, an objective central assessment, an objective exam that is comparable in Europe or internationally. That’s not the case – it’s just window dressing. It’s a Potemkin village” (P05). Another recurring category was the rise of a “checklist mentality.” Students were seen aiming for the minimum (a classroom grade of 3) to pass, rather than engaging with content. One teacher noted: “They just tick off requirements on a superficial level.

That blocks real learning or forming their own opinions” (P04). Most teachers described this trend as highly frustrating, with one noting that “as long as we get a classroom grade 3, we don’t need to study for the Matura anymore” had become a typical student attitude (P13). Some teachers observed more students failing the Matura since the reform, attributing it to lower motivation or strategic failing to access easier compensatory exams. One noted: “There were students who just submitted a blank sheet – I don’t think that happened before” (P03). A few questioned whether the reform was designed to inflate success rates: “Let’s be honest: it was introduced so that grades improve on paper” (P02). Lastly, the reform’s 30 % rule for passing the Matura came under sharp criticism. Teachers described it as “outrageous” and “misleading,” with some noting that minimal participation was enough to pass. “You just need to appear like you’ve thought about the question, and you’re through. That’s pretty outrageous,” one teacher stated (P13).

Despite widespread concerns, some teachers supported the reform. They appreciated its attempt to reduce the pressure of high-stakes exams and include broader measures of student effort and performance: “I actually see it as an advantage that effort and homework count toward the Matura grade” (P11). One added: “Students go into the Matura with less pressure and probably perform better” (P10). A few mathematics teachers felt the reform fit well with their subject, as students were typically prepared by spring: “In Maths, they are top-fit by March. [...] They have already learned most of the material to get their year grade” (P08).

7. Discussion

The present paper provides a longitudinal perspective on the development of final-year classroom grades in Austria over seven years in German and mathematics. It captures not only the disruptions caused by the pandemic but also the effects of a grading reform, introduced in the first pandemic year of 2020, that fundamentally changed how university entrance qualifications are determined. As many education systems grapple with questions of fairness, reliability, and the role of teacher judgment in high-stakes grading (Jönsson et al., 2021), Austria’s grading reform provided a valuable case for examining how such policy changes affect final qualification grades and the underlying mechanisms of teachers’ grade decision-making.

Our study results support international research on grade inflation (e.g., McManus et al., 2021). As in other countries, there was a marked increase in higher (and a decrease in lower) classroom grades in Austria in 2020, which has remained the case ever since, particularly in the case of Mathematics. Within this overall trend, our findings suggest indications of all three forms of grade inflation (Tyner & Gershenson, 2020) within the Austrian context. First, static grade inflation was most evident in 2020 and 2021, when teachers reported awarding grades that overstated students’ actual achievement. Similar to previous studies (e.g., Berne, 2023), teacher largely attributed this to the extraordinary circumstances of the COVID-19 pandemic, including disrupted learning environments, limited assessment opportunities, and teachers’ heightened empathy and compassion toward students during this period. Second, dynamic grade inflation appears to be linked to the 2020 reform that integrated classroom grades, which are typically higher than standardized exam scores (Hübner et al., 2020), into the final Matura result. Teachers observed a noticeable reduction in final failing Matura grades since the reform, indicating a shift in grading standards over time. This observation is consistent with official statistics, showing that the proportion of students failing the Matura dropped sharply from 12.3 to 13.5 % in 2017–2019–3.4–4.3 % in 2020–2021, and then gradually increased to 6.4–6.9 % in 2022–2024, remaining below pre-2020 levels (Statistik Austria, 2023). Finally, differential grade inflation emerged in our analyses, as the effect of higher classroom grades was driven almost entirely by vocational schools. This pattern points to systematic differences across school types, highlighting the need for further research on the mechanisms underlying school type specific grading practices.

It is difficult to disentangle the effects of the pandemic from those of the 2020 grading reform. Especially for the first two years after the COVID-19 pandemic, the results seem to reflect a 'transition period', shaped by both the 2020 grading reform and the extraordinary pandemic conditions. Teachers described the years of 2020 and 2021 as exceptional years due to pandemic-related disruptions, which led them both to adjust their grading practices in response to these challenging circumstances and to leverage the 2020 grading reform to help students achieve favourable Matura outcomes. In the years following 2021, however, teachers consistently attributed the ongoing upward shift in grades (i.e., the decline in grade 4 and rise in grade 2) to the structural change of the reform. Furthermore, several participants perceived the reform as part of a broader trend towards a "letting-everyone-pass" mentality, which might be an unofficial, yet analogous trend to that described by [Castro et al. \(2020\)](#) for California's official "hold harmless" grading policy.

An important point to consider here is the effect of lenient grading on student motivation. Previous research ([Berne, 2023](#); [Bowden et al., 2023](#)) indicates that teachers' lenient grading does not lead to the hoped-for increase in student motivation. Instead, it has been shown to lead to a significant rise in student absenteeism. Our qualitative results support this; many teachers reported appallingly poor Matura results when students received their compensatory 'safe grade' of classroom grade 3. This is particularly apparent in the oral Matura, despite students being required to demonstrate their knowledge and skills to a committee and in a relatively public setting. However, the seemingly logical conclusion that students need extrinsic motivation to learn does not fully capture this challenge. Teachers primarily reported on students' 'checklist mentality', which hinders their in-depth engagement with the learning material. Therefore, teaching should focus more on the joy of the learning process itself, fostering students' intrinsic learning motivation ([Ryan & Deci, 2000](#)), rather than on achieving certificates. This is certainly not easy to implement given the existence of national educational certificates such as the Matura, on which many other educational pathways are based. Nevertheless, research has shown that high-stakes tests can positively affect teaching and learning. These effects include, among others, higher performance outcomes ([Bishop, 2000](#)) and increased teacher understanding of the nature and structure of their own disciplines ([Cizek, 2001](#)).

However, our results have shown that the intended criterion-referenced nature of the Austrian Matura has largely been lost since the 2020 grading reform due to the high degree of arbitrariness of the classroom grades included since then. According to teachers, Mathematics appears to be less affected by this due to the subject's greater "tangibility" when it comes to performance assessment ([Brookhart et al., 2016](#); [Koretz, 2008](#)). Nevertheless, the grading reform significantly impacted oral examinations. Teachers reported that students are increasingly choosing subjects in which they can typically expect to receive good classroom grades due to lower assessment standards – often without the corresponding level of performance – safe in the knowledge that their classroom grades will compensate for poor results in the oral exam. While this certainly affects only some students, the system should encourage and support students to graduate in subjects where they have discovered their interests and talents, rather than subjects in which they encounter the least resistance.

Consequently, many teachers expressed concerns about reducing centralized, mastery-oriented standards. One teacher likened the existing assessment system to a "Potemkin village" that creates only the illusion of objectivity and comparability of final Matura grades. This resonates with [Benveniste's \(2002\)](#) argument that assessment systems often serve a symbolic rather than a diagnostic function. From his institutional sociology perspective, states adopt such systems mainly to demonstrate alignment with international policy norms and to signal a commitment to quality education. In practice, however, these systems would often operate in a ceremonial way: to uphold the appearance of accountability while shielding the state from responsibility for actual

educational outcomes.

[Brown \(2024\)](#) conceptualizes assessment as a process exposed to legitimacy risks across three domains: the psychometric limitations of score accuracy, the sociocultural influences of policy and cultural norms, and the biases introduced by human actors. Our findings highlight particular challenges for the Austrian Matura in the latter two domains. Many teachers reported that both the policy change and prevailing norms of leniency during the pandemic years weakened their perception of the Matura as a valid reflection of students' academic performance. Moreover, the biases of teachers (e.g., strategic grading, empathy, compassionate grading) as well as pressures from other stakeholders (e.g., student negotiations, parental influence, or demands from principals) further undermined their trust in the credibility and quality of Matura outcomes.

While the need for comparability in grading may be open to debate ([Berman et al., 2020](#); [Coe, 2010](#)), any system claiming to uphold objective, performance-based standards must deliver on those promises. Otherwise, as the interviews have shown, it fosters deep frustration among teachers. This particularly applies to teachers who are committed to fair, achievement-oriented assessment. However, even teachers who said they did not care too much about performance-based assessments expressed frustration due to the perceived reduction in student engagement with Matura-relevant content since the grading reform. Some teachers even expressed a growing dissatisfaction with their work, feeling that the effort they invest in consistent and fair grading is increasingly meaningless.

However, some arguments highlighted certain aspects of the reform as positive. These emphasized that integrating classroom grades into Matura grades provides a more holistic assessment of student performance ([Bowers, 2011](#)) and helps balance the pressure of a one-day high-stakes exam ([Kruger et al., 2018](#)). Researchers may argue that this diminishes the comparability of students' academic achievement across the country ([Sanchez, 2023](#)) or lowers the predictive validity of Matura exam scores for later study outcomes ([McManus et al., 2021](#)). However, there are also conflicting findings in the literature. In the case of Sweden, for example, research shows that teacher assessments can be even better predictors of university success than standardized test scores ([Cliffordson, 2008](#)).

Teachers' support for the reform was particularly notable in Mathematics, where they emphasized benefits such as increased student motivation and classroom engagement to obtain higher classroom grades, earlier and more thorough exam preparation, and reduced anxiety during the Matura exam—factors that, according to some teachers, may even contribute to higher exam achievement.

Our research results revealed a deeper understanding of the possible mechanisms underlying teachers' grading decisions in general, and within the context of reform-induced challenges in particular. These challenges include perceived influential factors at the systemic level (e.g., moral appeals to be lenient and the undermining of centralized standards), interpersonal level (pressures from students, parents, principals and colleagues influencing grading practices in subtle and direct ways) and individual level where internal conflicts and values play a key role in balancing empathy, fairness and self-image as a supportive educator.

Some of these mechanisms were perceived to have intensified over the years, such as the trend towards greater grade leniency, teachers' increasing inner role conflict in their relationships with students, and the increased "student-friendliness" of principals, which sometimes undermines teachers' professional assessments. Other mechanisms were perceived to have explicitly shifted following the introduction of the 2020 grading reform. This pertains foremost to increased pressures from and negotiations with students about final classroom grades, leading some to adapt their grading practices while others resisted to uphold performance-based standards.

Indeed, as earlier studies have shown ([Berne, 2023](#)), teachers react differently to externally imposed systematic changes. Many teachers

expressed concerns about their colleagues' strategic grading behavior, which was described by McMillan (2003) as 'pulling for students', to prevent their students from failing. Others reported that they had realized their adaptation to more lenient grading had backfired, prompting them to become stricter again to maintain performance standards and ensure that students still take the Matura exams seriously. These qualitative findings are also reflected in the results of our quantitative study, which illustrate a wave-motion trend of German classroom grades over time.

In any case, our findings underscore that grading is not merely a technical act of evaluation but a deeply contextualized and value-laden process (Pilcher, 1994). Any grading reform must account for the complex social dynamics and professional norms that govern teachers' decision-making. Otherwise, as our results show, reforms risk producing unintended consequences, like strategic grading behaviors and, with that, a loss of assessment credibility, ultimately eroding the legitimacy of the grading system itself.

8. Implications, limitations, and directions for future research

From our findings, cautious implications can be drawn at different levels. At the macro (systemic) level, policymakers should reconsider allowing students' own teachers to assess standardized test results, not only to ensure objectivity and performance-based grading, but also to uphold student learning and motivation. However, external high-stakes testing also carries risks, such as narrowing the curriculum or encouraging teaching to the test (Farvis & Hay, 2020; Jacob, 2005; Muth et al., 2025; Muth & Lüftenegger, 2025). To identify the best approach for their country, policymakers must carefully weigh the benefits and potential drawbacks of various assessment methods. In the Austrian context, policymakers should, at a minimum, reverse the ability of teacher-assigned grades to compensate for failing standardized test scores to prevent unintended grade inflation through internal adjustments and preserve the integrity of Matura grades. At the meso (institutional) level, our findings underscore the importance of shared values and coherent grading practices at the school level (Fuller & Izu, 1986). Without a common ethos, committed teachers striving for performance-based grading may feel undermined, while others experience a role conflict between their supportive teacher identity and the demands of objectively scoring students. Principals need to cultivate this shared ethos to ensure consistency and fairness in grading, as well as a collective trust in the school-wide grading process itself (Engels et al., 2008). At the micro (individual) level, teachers should be aware that their grading practices not only determine their students' academic outcomes but also affect the perceived legitimacy of the entire assessment system. While the role of grades as performance indicators can be debated (Butler, 2025; McClam & Sevier, 2010; McMorran & Ragupathi, 2019), they must be used responsibly as long as they remain in place. Otherwise, teachers have shown themselves to uphold a symbolic "Potemkin village" of assessment that consumes energy without contributing meaningfully to teaching and learning.

This paper has several limitations. First, it draws solely on teachers' perspectives, while voices from students, parents, and principals are missing. Hence, the single-stakeholder view presented in this paper offers only a partial understanding of grading reform dynamics. Both students and parents likely perceive the 2020 grading reform as a valuable opportunity for students' effort and other non-academic achievement factors to be more strongly recognized in the overall outcome. From their perspective, the reform may increase the value placed on continuous classroom work rather than relying solely on exam performance. They might also welcome the reform as an opportunity to secure more favourable final grades while experiencing reduced exam pressure. Future studies, therefore, should include the views of students and parents to receive a balanced picture of the systemic effects of this policy change. This also includes school leaders' perspectives, who, according to the interviewed teachers, especially play an important role

in teachers' grading behavior, particularly under reform conditions (Kunnath, 2017; Weaver, 2018). In this study, school principals are portrayed somewhat negatively (e.g., exerting pressure on teachers, siding with students and parents, undermining teachers' professional judgment). However, their perspectives and challenges as school leaders (Tamadoni et al., 2024), which could help explain such behaviours, are not considered. Second, our analysis lacks data beyond 2023. Some teachers already reported signs of a reactive shift in grading, particularly in German, which is also reflected in the quantitative data. Similar trends may emerge in Mathematics with a delay. Longer-term data will be needed to assess the full scope of the reform's long-term effects across subjects. A third limitation concerns possible selection bias. Self-selection bias is particularly concerning when investigating sensitive educational topics such as grading (Authors, in review). It appears that teachers with a strong sense of grading reliability were more likely to participate in our study, which may also be linked to a greater interest in taking part and sharing their views (Saleh & Bista, 2017). Consequently, these teachers could be overrepresented in our sample, while those supportive of the 2020 grading reform may be underrepresented. Despite our efforts to counter this through sampling strategies, our qualitative sample still appeared to include many teachers who emphasized performance-based grading and personal grading integrity. Many participants spoke critically of their colleagues' grading practices while highlighting their own commitment to fairness and transparency. In order to address this potential imbalance, we expanded on the few arguments in favour of the reform raised by participants in the discussion section. Future research should strive to include a broader range of perspectives, especially from teachers who favor alternative or gradeless assessment models. Notably, many participants expressed discomfort with traditional grading, indicating a need to explore their preferred methods for addressing performance differences without relying on conventional grades.

CRedit authorship contribution statement

Vincent Schatz-Laliberté: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. Marko Lüftenegger: Supervision, Conceptualization.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.stueduc.2025.101540.

References

- Benveniste, L. (2002). The political structuration of assessment: Negotiating state power and legitimacy. *Comparative Education Review*, 46(1), 89–118.
- Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (2020). *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*. National Academy of Education.
- Berne, J.L. (2023). *Grading in Crisis: Examining the Impact of Covid-19 on Secondary Public School Teachers' Grading and Reporting in South Florida* [Dissertation]. University of Southern California.
- Bishop, J. H. (2000). Curriculum-based external exit exam systems: Do students learn more? *How? Psychology, Public Policy, and Law*, 6(1), 199–215. <https://doi.org/10.1037/1076-8971.6.1.199>
- BMBWF. (2023). Informationen zur Beurteilung der abschließenden Prüfungen. *Gesamthafte Beurteilung von Prüfungsgebieten*. (<https://www.bmhs-wien.at/wp-content/uploads/2023-0-215.251-RInformationenzurBeurteilungderabschliessendenPruefungen.pdf>).
- Bowden, B., Rodriguez, V., & Weingarten, Z. (2023). *The Unintended Consequences of Academic Leniency*. (https://static1.squarespace.com/static/5ec580a6be16d6545a82a17b/t/6563d0ab6290971bd1b88ed5/1701040300226/bwv_leniency.pdf).

- 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., Guskey, T. R., Bowers, A. J., McMillan, J. H., [James H.], Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, 86(4). <https://doi.org/10.3102/0034654316672069>, 803–848.
- Brown, G. T. (2024). An integrated approach to improve the credibility and quality of assessment in higher education. *Research Handbook on Innovations in Assessment and Feedback in Higher Education* (pp. 372–388). Edward Elgar Publishing.
- Bundesministerium für Bildung (2025). *Aufgabenpool*. (<https://aufgabenpool.at/srdp/sta/rseite/index.php>).
- Bundesministers für Bildung. (2021). *Verordnung, mit der die Verordnung über die Vorbereitung und Durchführung abschließender Prüfungen für das Schuljahr 2020/21 geändert wird*. Wissenschaft und Forschung. (https://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2021_II_211/BGBLA_2021_II_211.html).
- Butler, M. (2025). Alternative grading systems and student outcomes: A comparative analysis of motivation, enjoyment, engagement, stress, and perceptions of final grades. *Assessment Evaluation in Higher Education*, 50(5), 807–817. <https://doi.org/10.1080/02602938.2025.2475068>
- Castro, M., Choi, L., Knudson, J., & O'Day, J. (2020). Grading policy in the time of COVID-19: Considerations and implications for equity. *California Collaborative on District Reform*.
- Chan, C. K. Y. (2022). A review of the changes in higher education assessment and grading policy during covid-19. *Assessment Evaluation in Higher Education*, 48(6), 874–887. <https://doi.org/10.1080/02602938.2022.2140780>
- Cheng, Y., & Kong, D. (2023). Are female teachers more likely to practice grade inflation? Evidence from China. *China Economic Review*, 80, 101987.
- Chowdhury, F. (2018). Grade inflation: Causes, consequences and cure. *Journal of Education and Learning*, 7(6), 86–92.
- Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4), 19–27. <https://doi.org/10.1111/j.1745-3992.2001.tb00072.x>
- Cliffordson, C. (2008). Differential prediction of study success across academic programs in the Swedish context: The validity of grades and tests as selection instruments for higher education. *Educational Assessment*, 13(1), 56–75.
- Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education*, 25(3), 271–284. <https://doi.org/10.1080/02671522.2010.498143>
- Daddi, S. (2020). Covid-19 France and Bac: Written part cancelled. *The Connexion*. (<https://www.connexionfrance.com/news/covid-19-france-and-bac-written-part-cancelled/421397>).
- Daumiller, M., Rinas, R., Schoon, I., & Lüftenegger, M. (2023). How did COVID-19 affect education and what can be learned moving forward? A systematic meta-review of systematic reviews and meta-analyses. *Zeitschrift Für Psychologie*, 231(3), 177–191. <https://doi.org/10.1027/2151-2604/a000527>
- Donnelly, R., & Patrinos, H. A. (2022). Learning loss during Covid-19: An early systematic review. *Prospects*, 51(4), 601–609.
- Downe-Wamboldt, B. (1992). Content analysis: Method, applications, and issues. *Health Care for Women International*, 13(3), 313–321. <https://doi.org/10.1080/07399339209516006>
- Doyle, A., Lysaght, Z., & O'Leary, M. (2021). High stakes assessment policy implementation in the time of COVID-19: The case of calculated grades in Ireland. *Irish Educational Studies*, 40(2), 385–398. <https://doi.org/10.1080/03323315.2021.1916565>
- Engels, N., Hutton, G., Devos, G., Bouckennooghe, D., & Aelterman, A. (2008). Principals in schools with a positive school culture. *Educational Studies*, 34(3), 159–174. <https://doi.org/10.1080/0305569070181126>
- Farvis, J., & Hay, S. (2020). Undermining teaching: How education consultants view the impact of high-stakes test preparation on teaching. *Policy Futures in Education*, 18(8), 1058–1074. <https://doi.org/10.1177/1478210320919541>
- Fridkin, L., Font, N. B., Qu, K., & Zwiener-Collins, N. (2023). Understanding effects of COVID-19 on undergraduate academic stress, motivation and coping over time. *Higher Education Quarterly*, 77(4), 623–637.
- Fuller, B., & Izu, J. A. (1986). Explaining school cohesion: What shapes the organizational beliefs of teachers? *American Journal of Education*, 94(4), 501–535. (<https://www.jstor.org/stable/108533>).
- Goldhaber, D., & Goodman Young, M. (2024). Course grades as a signal of student achievement: Evidence of grade inflation before and after COVID-19. *Journal of Policy Analysis and Management*, 43(4), 1270–1282.
- Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*, 24(2), 105–112. <https://doi.org/10.1016/j.nedt.2003.10.001>
- Greer, S. L., King, E., Da Massard Fonseca, E., & Peralta-Santos, A. (Eds.). (2021). *Coronavirus politics: The comparative politics and policy of COVID-19*. University of Michigan Press.
- Hübner, N., Wagner, W., Hochweber, J., Neumann, M., Nagengast, B., Kendeou, P., & Graham, S. (2020). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grade. *Journal of Educational Psychology*, 112(1), 204–220. <https://doi.org/10.1037/edu0000351>
- Isaacs, T., & Murphy, R. (2022). The impact of COVID-19 on 2020–2021 assessment arrangements. *ofqual*. (<https://www.gov.uk/government/publications/the-impact-of-covid-19-on-2020-to-2021-assessment-arrangements/the-impact-of-covid-19-on-2020-to-2021-assessment-arrangements>).
- Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods*, 18(1), 3–20. <https://doi.org/10.1177/1525822X05282260>
- Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5–6), 761–796. <https://doi.org/10.1016/j.jpubeco.2004.08.004>
- Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy Practice*, 28(3), 212–227. <https://doi.org/10.1080/0969594X.2021.1884041>
- Kleinheksel, A. J., Rockich-Winston, N., Tawfik, H., & Wyatt, T. R. (2020). Demystifying Content Analysis. *American Journal of Pharmaceutical Education*, 84(1), 7113–7137. <https://doi.org/10.5688/ajpe7113>
- Koller, R. (2021). *COVID-19 Impact on Middle School Teachers' Grading Practice* [Master's Thesis]. California State University San Marcos, California.
- Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Harvard University Press.
- Kruger, L. J., Wandle, C., & Struzziero, J. (2018). Coping with the stress of high stakes testing. *High Stakes Testing*, 109–128.
- Kunnath, J. P. (2017). Teacher grading decisions: Influences, rationale, and practices. *American Secondary Education*, 45(3), 68–88.
- Kurier. (2020). Matura trotz Fünfer: Neue Benotung soll auch nach Corona bleiben. *Kurier*. (<https://kurier.at/politik/inland/matura-trotz-fuenfer-neue-benotung-soll-auch-nach-corona-bleiben/400850858>).
- McClam, S., & Sevier, B. (2010). Troubles with grades, grading, and change: Learning from adventures in alternative assessment practices in teacher education. *Teaching and Teacher Education*, 26(7), 1460–1470.
- McManus, I. C., Woolf, K., Harrison, D., Tiffin, P. A., Paton, L. W., Cheung, K. Y. F., & Smith, D. T. (2021). Predictive validity of A-level grades and teacher-predicted grades in UK medical school applicants: A retrospective analysis of administrative data in a time of COVID-19. *BMJ Open*, 11(12), Article e047354. <https://doi.org/10.1136/bmjopen-2020-047354>
- McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, 22(4), 34–43. <https://doi.org/10.1111/j.1745-3992.2003.tb00142.x>
- McMorran, C., & Ragupathi, K. (2019). The promise and pitfalls of gradeless learning: Responses to an alternative approach to grading. *Journal of Further and Higher Education*, 44(7), 925–938. <https://doi.org/10.1080/0309877X.2019.1619073>
- Miller, G. (2014). Grade inflation, gatekeeping, and social work education: Ethics and perils. *Journal of Social Work Values and Ethics*, 11(1), 12–22.
- Muth, J., Grützmacher, L., & Lüftenegger, M. (2025). Teaching to the test in the English language classroom: Development and validation of a measurement instrument. *Language Teaching Research*. <https://doi.org/10.1177/13621688251351250>
- Muth, J., & Lüftenegger, M. (2025). Teaching to the test: Unraveling the consequences for student motivation. *Learning and Individual Differences*, 121, Article 102707. <https://doi.org/10.1016/j.lindif.2025.102707>
- O'Leary, M., & Lysaght, Z. (2024). Considering reform at senior cycle: Teachers' feelings and beliefs about high-stakes assessment following the onset of COVID-19. *The Irish Journal of Education*, 48(1), 1–29.
- Panadero, E., Fraile, J., Pinedo, L., Rodríguez-Hernández, C., & Díez, F. (2022). Changes in classroom assessment practices during emergency remote teaching due to COVID-19. *Assessment in Education: Principles, Policy Practice*, 29(3), 361–382. <https://doi.org/10.1080/0969594X.2022.2067123>
- Pilcher, J. K. (1994). The value-driven meaning of grades. *Educational Assessment*, 2(1), 69–88. https://doi.org/10.1207/s15326977ea0201_4
- RIS, Rechtsinformation des Bundes (2025, August 28). *Verordnung des Bundesministers für Unterricht und Kunst vom 24. Juni 1974 über die Leistungsbeurteilung in Pflichtschulen sowie mittleren und höheren Schulen (Leistungsbeurteilungsverordnung)*. (<https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10009375>).
- RStudio Team (2020). *Integrated Development for R* [Computer software].
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–87. <https://doi.org/10.1037/0003-066X.55.1.68>
- Saleh, A., & Bista, K. (2017). Examining factors impacting online survey response rates in educational research: Perceptions of graduate students. *Journal of Multidisciplinary Evaluation*, 13(29), 63–74. <https://doi.org/10.56645/jmde.v13i29.487>
- Sanchez, E. I. (2023). *Evidence of Grade Inflation Since 2010 in High School English, Mathematics, Social Studies, and Science Courses*. (<https://www.act.org/content/act/en/research/pdfs/Evidence-of-Grade-Inflation-in-English-Math-Social-Studies-and-Science.html>).
- Schröter, P., Hoffmann, L., & Schmid-Kühn, S. M. (2022). Ein Blick in andere Länder: Abschlussprüfungen im Spannungsfeld zwischen Standardisierung und Autonomie. *Das unvergleichliche Abitur. Entwicklungen-Herausforderungen-Empirische Analysen*, 89–126.
- Statistik Austria (2023). *Statistik Austria: Reife- und Diplomprüfungen*. (<https://www.statistik.at/statistiken/bevoelkerung-und-soziales/bildung/reife-und-diplompruefung>).
- Tamadoni, A., Hosseingholizadeh, R., & Bellibas, M.S. (2024). A systematic review of key contextual challenges facing school principals: Research-informed coping solutions. *Educational Management Administration & Leadership*, 52(1), 116–150.
- Tarkar, P. (2020). Impact of COVID-19 pandemic on education system. *International Journal of Advanced Science and Technology*, 29(9), 3812–3814.
- The jamovi project. (2022). *jamovi* (Version 2.3) [Computer software]. (<https://www.jamovi.org>).
- Tillinghast, J. A., Mjelde, D. J. W., & Yeritsyan, A. (2023). COVID-19 and grade inflation: Analysis of undergraduate GPAs during the pandemic. *Sage Open*, 13(4), Article 21582440231209110.

Tyner, A., & Gershenson, S. (2020). Conceptualizing grade inflation. *Economics of Education Review*, 78, 102037.

Vahle, C., Araujo, Z. de, Han, J., & Otten, S. (2023). Teachers' instructional responses to the COVID-19 pandemic. *Teaching and Teacher Education*, 124, Article 104040. <https://doi.org/10.1016/j.tate.2023.104040>

Walsh, J. (1979). Does high school grade inflation mask a more alarming trend? *Science (American Association for the Advancement of Science)*, 203(4384), 982. <https://doi.org/10.1126/science.203.4384.982>

Weaver, D.J. (2018). High School Principals' Leadership Activities in Policy Development and Implementation of Standards-Based Grading [Dissertation]. University of Illinois Urbana-Champaign, Urbana, Illinois. (<https://www.ideals.illinois.edu/items/109479>).

Supplementary Material

Appendix A

Interview Guide

1.) Have you noticed any changes in the awarding of final grades in the final Matura classes since 2020? If so, what kind of changes?

1a.) Can you recall a specific situation in which this became apparent or clearly manifested?

2.) I conducted a study at seven schools on changes in final grades in mathematics and German over several years. I'd like to show you this graph and briefly explain it. [Graph with results from the preceding quantitative study]

Please describe how these grade developments correspond with or differ from your personal perception. How would you explain the development of individual grades (1–4) over the years in German/mathematics? According to your personal opinion and experience, what could explain this trend?

2a.) Can you describe a specific situation in which this trend became visible or particularly evident?

3.) Do you believe that this phenomenon (or: this development, or: what you have described) is subject-specific, or does it also occur in other subjects? How would you explain the reason for this?

I'm particularly interested in which different influencing factors may have contributed to these changes.

4.) Have you noticed any external influences (e.g., from students, colleagues, school administration, parents, the education system itself, etc.) that have directly or perhaps indirectly affected your grading?

4a.) Have you noticed any changes in this regard since 2020 compared to previous years?

4b.) Can you describe a specific situation in which this became visible or clearly noticeable?

5.) Now let's turn to possible internal influences. This includes, for example, your attitude toward the Matura exam or grading in general, as well as your "philosophy" regarding education or grading specifically – your beliefs, values, views, etc. To what extent do you feel these factors have influenced your final grade decisions since 2020?

5a.) Have you noticed any changes in this regard since 2020 compared to previous years?

5b.) Can you describe a specific situation in which this influence became evident?

6.) Is there anything else you would like to share regarding the reform?

7.) Is there anything else you would like to share regarding grading practices more generally?

8.) Is there anything else you would like to talk about?

Appendix B

Categories and Codes from Qualitative Study 2

Categories with Referring Codes for Research Questions 2 (*How do teachers explain grade changes since the 2020 grading reform in Austria?*) and 3 (*What are teachers' perceptions of possible changes in grading practices since the 2020 grading reform?*)

Category 1: Extraordinary Circumstances in Pandemic Years

Code	Code Description	Example
General climate of leniency during pandemic years	When teachers speak of a broad tendency to grade more generously.	„Also ich glaube schon prinzipiell, dass wir Lehrer während der Coronajahre wesentlich sanfter in der Beurteilung waren.“ (P02)
Disrupted Learning Environment and Limited Assessment Opportunities in Pandemic Years	When leniency is explained with reference to non-ideal teaching conditions and fewer chances to assess students during pandemic years.	„[...]sie haben ja nichts gelernt und der Unterricht war nur online [...]“ (P01)

Category 2: Teachers' Explanations for Grade Curves Over the Years

Code	Code Description	Examples
Too Late for Grading Adjustment in 2020	When teachers explain that the 2020 grading reform did not immediately influence classroom grades because it was introduced too late in the school year. As a result, teachers had already finalized most grades without considering the	„Da war vielleicht auch schon Vieles einfach erledigt, also das Jahr schon fortgeschritten und sozusagen auch nicht mehr wahnsinnig viel Zeit, um da jetzt tatsächlich die Noten zu verändern oder zu beeinflussen.“ (P05)

new regulation, and grading practices only changed in the following year once the implications of the reform were fully understood.

Stronger Disruptions in the 2021 Cohort

When teachers describe how the consequences of the pandemic on grading only became fully apparent in 2021, not 2020, due to more school closures in the later phase (2021).

„Also die [der 2020er Maturajahrgang] haben noch eine relativ normale Schulzeit vorher gehabt, würde ich sagen, also die haben quasi nur die Matura in der Corona-Zeit gemacht, aber die [der 2021er Maturajahrgang] haben das letzte Jahr schon mit den Beschränkungen, Zugangsbeschränkungen und so weiter halt da, ja, leben müssen.“ (P03)

Delayed Recognition of Reform Benefits

When teachers explain that they may have recognized the strategic advantage of assigning a higher year grade only one year later, especially in the subject German.

„[V]ielleicht haben die Lehrer*innen ein Jahr gebraucht, um zu checken, was es eigentlich für Auswirkungen hat [...], während sie im darauffolgenden Schuljahr gleich von Beginn an vom Anfang des Schuljahres alle Maßnahmen ergriffen haben, dass die Noten wirklich gut werden.“ (P07)

Norm-Referencing in Grading as an Explanation for the Increase in Grade 2

Code statements in which teachers explain that the increase in grade 2 ('Good') results from internal comparisons within the class. Specifically, when some students receive grade 3s ('Satisfactory'), others who were originally seen as good grade 3 are reclassified as grade 2 to maintain coherence in the relative grading logic.

„[W]eil es dann vielleicht nicht mehr fair ist. Jenen, denen man früher einen Dreier gegeben hätte, gibt man jetzt vielleicht ein Gut, [...] weil man eben sagt, die, die ein Genügend gehabt haben und dann mit einer Prüfung noch ein Befriedigend bekommen, dass man dann den klassischen Befriedigend vielleicht eher ein Gut gibt.“ (P12)

Compensatory Classroom Grading Based on Matura Forecast as Explanation for the Increase in Grade 2	Teachers explaining the increase in classroom grade of 2 by anticipating that students will perform worse on the Matura, aiming to balance the overall final Matura grade outcome to at least grade 3.	“[W]enn jemand halt einen Vierer auf die Matura schreibt und er hat einen Zweier im Jahreszeugnis, dann kriegt er halt gesamt einen Dreier, während, wenn er einen Dreier in der Jahresnote hat und dann einen Vierer schreibt auf die Matura, hat er gesamt einen Vierer.“ (P13)
--	--	---

Category 3: Strategies to Secure Positive Matura Outcomes

Code	Code Description	Example
Increase in Oral Exams to Reach Grade 3	Teachers observing that more oral exams or additional tasks are now being used to help students raise their grade from 4 to 3 post-reform.	„Das, was mir bewusst aufgefallen ist, das ist eine Zunahme von Prüfungen. Vor allem zwischen den beiden Notengraden Genügend und Befriedigend, sodass dann, in der Hoffnung der Schüler, eben eine Absicherung durch die Jahresnote auch gegeben ist.“ (P10)
(Collegial) Willingness to Offer Non-Academic Compensation	Teacher describes how colleagues offer and accept non-performance-based opportunities to help students improve their grades	„[W]eil sie sich ja dann doch beteiligen und was tun, also ihren Willen zeigen, kriegen die halt den Dreier.“ (P11)
Systemic Inflation of Classroom Grade 3	Teachers reflecting on (colleagues’) tendency to assign a grade 3 instead of 4.	„Also bei mir ist es so, dass die Viererschüler halt einen Dreier kriegen, aber die Dreierschüler [...] auf einem Dreier bleiben.“ (P08)
Tactical Awarding of Save Classroom Grade (3) to Shield Students from Failing the Matura	When teachers intentionally assign a classroom grade of 3 to compensate for an expected grade	„Ich glaube, dass Mathematik immer anders tickt, dass es Kolleg:innen gibt, die wirklich so laut denken und sagen, ,wenn ich der einen Dreier

of 5 on the Matura, ensuring the student passes overall.

gebe, dann kann sie auf die Matura einen Fünfer schreiben, weil die schreibt ja eh immer Fünfer und dann kriegt sie einen Vierer“.“ (P06)

Category 4: Professional Integrity and Achievement Orientation in Grading

Code	Code Description	Examples
Perceived Stability in Own Grading Practice vs. Change in Others	Teachers state that their own grading behavior remained unchanged, but they observed changes in colleagues' grading practices.	„Na ja, so generell, also bei mir jetzt eigentlich nicht, weil ich mich dagegen verwehrt habe, da irgendwie jetzt automatisch mehr Dreier zu geben. [...] Ich hab immer versucht, mich dagegen zu wehren und ich denke nicht, dass ich mehr Dreier jetzt vergeben habe als früher.“ (P05)
Teacher Resistance to Grade Inflation via Non-Academic or Non-Performance-Based Grading Criteria	When teachers state that they resist assigning better grades based on non-performance-based activities (e.g., effort, extra tasks)	„Ihr könnt's Referate machen oder [...] mit ChatGPT jeden Tag irgendwas ausarbeiten und dann gibt es einfach keinen Dreier, sondern [...] einen Vierer.“ (P01)
Countermovement After Perceived Grade Inflation and Decline in Student Engagement	When teachers describe a reaction to initially inflated classroom grades by later adjusting their grading behavior to restore achievement standards, particularly to ensure that students still take the Matura seriously.	„Ich hatte [...] diese beiden Bewegungen mir. Manchmal hat man so dieses Gefühl, man will ja eh, dass sie es schaffen [...] und dann [stellt sich] die Gegenbewegung wieder ein von 'Ich will auch nicht, dass es bei uns um gar nichts geht'.“ (P04)

Grading Integrity to Preserve Fairness and Matura Significance	When teachers describe carefully considering grade assignments to maintain fairness in grading and value of final Matura exam	„Man [denkt] dann schon nach, was man ihnen gibt. Weil wenn man ihnen einen Dreier gibt, sind sie safe, und manchmal will man das nicht, weil man weiß, dass die Leistung das ganze Jahr über eher bei 4 war. [...] Wenn es nicht eindeutig ist, überlege ich es mir zweimal [, eine bessere Note zu geben], weil ich es unfair fände, dass die Matura in ihrer Wichtigkeit, wie sie früher einmal war, so geschmälert wird, dass eine negative Note dann keine Rolle spielt für diese Schüler.“ (P06)
--	---	--

Categories with Referring Codes for Research Question 4 (*What external factors did teachers perceive to have an impact on their grading and to what extent have these factors changed since 2020?*)

Category 5: Student and Parental Influences on Grading

Code	Code Description	Examples
Increased Student Pressure and Negotiation for Better Grades	When teachers describe that students increasingly exert pressure and negotiate for better grades, especially to receive at least a grade 3 in Matura-relevant subjects since grading reform.	„Die Jagd auf den Dreier ist massiv, [das ist] jetzt noch festzustellen. Was uns wahnsinnig ärgert, das ganze Maturajahr hindurch, ist, dass sie einfach an ihren Dreier wollen, um safe zu sein. Und da sind wir einfach in dem Dilemma, in allen Fächern oder in allen Maturafächern, dass wir damit einfach gut umgehen müssen.“ (P01)
Unrealistic Student Expectation in Receiving a Safe Grade	When teachers express frustration that students who have shown minimal effort or engagement	„[E]s ist halt teilweise auch ein bisschen mühsam, wo ich mir dann denke, ‚bitte sei froh, dass du

	over their entire school career in high school still expect to receive a safe classroom grade without demonstrating sufficient achievement.	einen Vierer hast, und jetzt willst du Dreier haben‘.” (P12)
Increased Parental Pressure for Better Grades	When teachers describe how parental expectations or pressures influence grading decisions, particularly when parents implicitly or explicitly advocate for better grades regardless of student performance.	„Ich meine aber, was nicht so gut ist, ist, dass der Druck für die Lehrperson natürlich auch massiv steigt [...], dass du dann halt im März Eltern von achtzehnjährigen Schülern da hast, die sich halt beschweren, warum das Kind keinen Dreier hat, sondern einen Vierer.“ (P08)
Strategic Grade Inflation as a Shield Against Parental Confrontation	When teachers describe feeling pressured to award higher grades to avoid conflict with influential parents, even when the performance does not justify a positive assessment.	„Und andererseits könnte schon eine Rolle spielen, dass man weiß, dass manche Eltern vielleicht oder Schüler*innen dann Einspruch erheben bei einer negativen Note. Das möchte man sich vielleicht auch ersparen.“ (P07)
Perceived Absence of Parental Pressure in BHS or Rural Areas	When teachers report that they do not experience parental pressure to raise grades due to the school’s type (BHS) or rural location.	„Wir spüren das nicht so sehr, denke ich, weil wir eine BHS sind und nur Oberstufenklassen haben. Ich nehme an, in einer AHS schaut das ein bisschen anders aus, weil da die Kinder mehr behütet sind und da die Eltern eine größere Rolle spielen.“ (P06)

Category 6: Pressures from Principal and Head of Class

Code	Code Description	Examples
Bias of Principals Toward Students	When teachers report that principles tend to side with students in grading conflicts, creating a situation where teachers must defend their grading decisions, while student perspectives are more readily accepted.	„Also unser Direktor ist sehr schülerfreundlich. Er stellt sich eigentlich immer auf die Seite der Schüler und du als Lehrer musst dich rechtfertigen. Umgekehrt ist es selten.“ (P02)
Pressure from Principles to Avoid Failing Students to Minimize Administrative Workload	When teachers describe explicit or implicit pressure from school leadership to pass students in order to avoid administrative burdens.	„Weil es einfach gewünscht wird von der Direktion, dass die Schüler bestehen [...] und weil es umständlich ist, also Kompensationsprüfungen – das ist ein Riesenaufwand [...]. Alles, was mehr Arbeit ist, wird an unserer Schule nicht goutiert.“ (P05)
Principal Compliance with Parental Pressure Undermining Teacher Authority	When teachers report that principals, under pressure from parents, tries to persuade teachers to award a failing grade (5) instead of a 4 in the written Matura so that the student can take a compensation exam (strategy used to avoid a final grade of 4 on the certificate, as due to the integration of the classroom grade it is possible for students to receive a higher final outcome after oral compensation).	„Weil ich einfach sehr sauer bin, wie sehr die Elternvertreterin die Lehrpersonen unterminiert [...] bis sie erreicht, was sie will. Und der Direktor hat sich halt da irgendwie auch schon einkaufen lassen.“ (P04)

Desire to Avoid Disapproval from School Leadership	When teachers adjust their grading practices to avoid negative attention or implicit criticism from school leadership.	„Um sich nicht dem Missmut der Führung unserer Schule auszusetzen, dass der und der Lehrer jetzt wieder mal Prüfungen hat.“ (P06)
Pressure from Head of Class (Klassenvorstand) to Award Higher Grades	When teachers report that head teachers of classes (Klassenvorstände) emotionally identify with their class or want to be seen as “successful class teacher” and pressure colleagues to award passing or better grades, sometimes leading to conflicts over grade decisions.	„Ja, dass manche den Job als Klassenvorstand einfach wirklich wie die Mama oder den Papa verstehen, ,das sind meine Kinder, die müssen alle zur Matura‘. [...] Eine junge Kollegin von mir hat das letztes Jahr gehabt: massiven Kampf mit dem KV um eine Note, zuerst um den Dreier in der Jahresnote, dann um den Vierer in der Matura.“ (P01)

Category 7: Systemic Drivers of Grade Leniency

Code	Code Description	Examples
Informal Norms and Moral Pressure to be Lenient in Pandemic Years	When teachers refer to moral appeals to grade more leniently due to the high burden on students during the pandemic.	„Naja, da in dieser Zeit ist ja Corona gewesen, und da hat es eine Art Moratorium gegeben. Das war einfach so quasi „die armen Schüler, die haben nicht in die Schule gehen können, die haben nichts lernen können, die sind so allein gewesen, zum Teil haben sie niemanden gehabt, der sie in der Früh aus dem Bett geschmissen hat“ und wir haben als Lehrer da schon, wie soll ich sagen, einen Nachsorgeauftrag mitgeteilt gekriegt“ (P03)

The Grading Reform Itself as External Factor Undermining Grading Standards

Captures teacher concerns about centrally imposed modifications to assessment practices – like the grading reform – that, while aimed at improving pass rates, are viewed as undermining grading standards.

„Weil das so gesteuert wurde. [...] Wenn man gemerkt hat, ui, da gibt es zu viele Nicht genügend, dann müssen wir in Mathe die Textbeispiele kürzen [...]. Dann eben hier, das letzte war, die Jahresnoten miteinzuberechnen. Und das hat ja dann mit zentral eigentlich gar nichts mehr zu tun.“ (P05)

The Grading Reform as Protection for Underperforming Teachers

Applies when teachers describe the integration of the classroom grade into the Matura as a form of protection for teachers who do not adequately prepare students or meet professional standards.

„Aus meiner Sicht ist es wirklich ein Schutz für schlechte Lehrer. Die Matura war ja immer auch ein Stress für Lehrer, dass man einfach da nach seiner Leistung beurteilt wird, das ist eine öffentliche Prüfung und jeder kriegt mit, was du machst oder was du nicht machst.“ (P01)

Long-Term Increase in Pressure to Avoid Failing Students

When teachers describe a noticeable increase over time in the pressure or scrutiny directed at those who give failing marks, reflecting broader shifts in school culture or societal expectations.

„Ich finde, der Druck ist überhaupt auf Lehrpersonen – oder sagen wir so, auf Lehrpersonen, die Nicht genügend geben – enorm gestiegen. [...] Das kann ich über meine lange Unterrichtszeit schon sehr gut beurteilen.“ (P08)

Categories with Referring Codes for Research Question 5 (*What internal factors did teachers perceive to have an impact on their grading and to what extent have these factors changed since 2020?*)

Category 8: Empathy and Compassion for Students

Code	Code Description	Examples
Empathy for student burden in pandemic years	When teachers explain lenient grading by referring to students' general emotional or psychological stress during the pandemic.	„Es gab dann teilweise schon Schicksale von Schülern [...], wenn ich solche Geschichten bisschen mitbekommen hab, dann war ich schon kulanter, also da hab ich dann schon eher dazu geneigt, ihnen da jetzt nicht 10.000 Türen in den Weg zu legen, dass sie eine Matura schaffen.“ (P13)
Empathy for students with psychological problems or difficult background	When teachers describe how knowing about their students' psychological problems or challenging home circumstances leads them to be more generous or lenient when grading.	„Wenn ich seh, dass es der Schülerin psychisch wirklich sehr schlecht geht [...] oder der Vater gestorben ist, kurz vor der Matura – da ist es am ehesten, dass von mir aus kommt ‚Jetzt geb ich ihr halt den einen Punkt noch‘.“ (P08)
Compassionate Grading for Hardworking Non-Native Speakers	When teachers report making lenient or accommodating grading decisions out of empathy for students (often girls) who work hard but face disadvantages due to not having German as their first language.	„Wir haben in der Schule wirklich ganz, ganz, ganz bemühte, ganz fleißige, meistens Mädels, muss man sagen, die einfach nicht Deutsch als Muttersprache haben, also da bin ich dann manchmal so, dass ich sage, passt.“ (P01)

Category 9: Torn nature of the teacher role

Code	Code Description	Examples
Emotional Burden of Grading due to Relationship with Students	When teachers describe how relationships with students can create emotional conflict during the grading process.	„Viele bauen eine eher freundschaftliche Beziehung zu Schülern auf und schaffen es dann nicht zu sagen, ‘diese Leistung ist Nicht genügend’.“ (P05)
Tension Between Supportive Intentions and Grading Responsibility	When teachers describe a self-concept focused on student support rather than gatekeeping, highlighting intrinsic motivation to offer students opportunities (e.g., for additional exams for grade improvement).	„Ich hab noch eine Aussage einer Kollegin im Kopf, die gemeint hat, sie muss jetzt eben diese Befriedigend-Prüfungen alle noch machen. Das heißt, es war ihr offensichtlich selber ein Anliegen, diese Prüfungen zu machen und [...] den Schülern und Schülerinnen die Möglichkeit und Gelegenheit bieten zu wollen, auch zeitlich diese Prüfungen zu machen.“ (P10)
Emotional Conflict in Awarding Failing Grades	When teachers see the grading reform as a way to avoid the emotional or ethical conflict of failing students in the Matura.	„Und dass vielleicht auch manche Kollegen froh waren, dass es diese Möglichkeit gibt, damit sie nicht in diesen Konflikt geraten, dann jemanden bei der Matura ein Nicht genügend geben zu müssen und zu wissen, der hat jetzt nicht bestanden.“ (P05)

Category 10: Teachers' Achievement Values in Grading

Code	Code Description	Examples
Securing Achievement Orientation Through Stricter Classroom Grading	Describes a shift in grading practice from a previously generous approach to a stricter, performance-based standard to maintain academic standards and foster student learning, motivated by perceived misuse of leniency.	„Naja, nachdem relativ viel schriftlich angetreten sind in Mathematik und dann gerade mal eben auf diese 30%-Schwelle hingearbeitet haben, habe ich mir heuer gedacht, nein, ich schau mir das heuer ganz genau an, wem ich einen Dreier geb. Und bin heuer mit dem sehr gut gefahren. Und letztes Jahr habe ich wirklich halt gesagt, ‚Na ja, ok, Mitarbeit, Hausübungen, wenn sie einen Vierer schreiben auf die Schularbeit, kriegen sie einen Dreier, wenn sie was tun in der Zeit‘. So großzügig bin ich nicht mehr.“ (P11)
Critique in Colleagues' Avoidance of Extra Work Following Failing Grades	When teachers are perceived that they are awarding higher grades not because of student performance, but to avoid extra work or discomfort with repeat examinations.	„Also ich glaube, dass es immer schon diese Lehrer gab, die zum Beispiel, um eine September-Prüfung zu vermeiden, die siehst du auch nie im September bei den Wiederholungsprüfungen, deren Noten sind 1 bis 3, 1 bis 4 im allerschlimmsten Fall. Die wollen sich einfach weniger Arbeit antun.“ (P06)
Subjectivity in Classroom Grading	When teachers reflect on how the flexible design of (colleagues') classroom assessments allows them to influence classroom grade outcomes intentionally.	„Denn jede Schule und jeder Lehrer kann ja für sich entscheiden, ‘was gebe ich zur Schularbeit [...]’. Ob das immer auf Maturaniveau ist oder nur Reproduktion von Wiedergekäutem, das weiß ja niemand. Das heißt, diese Noten, die da

eingerechnet werden, die sind eigentlich ziemlich willkürlich.“ (P05)

Category 11: Student-Friendliness as Self-Image

Code	Code Description	Examples
Need for Student Love and Approval as Obstacle to Fair Grading	When teachers describe how a strong desire to be liked by students can interfere with fair and objective grading, including avoiding the assignment of low or failing grades, leading to overly generous interpretations of student work or the blurring of professional boundaries.	„Wir haben zum Beispiel eine Kollegin, die kann keine Fünfer geben. [...] Sie will immer nur gemocht werden, und ich glaube, dieses Liebesbedürfnis ist ein bisschen fehl am Platz als Lehrer, weil man eigentlich fair benoten sollte.“ (P05)
Internal Pressure from not Wanting to be Seen as Hostile Towards Students	When teachers describe experiencing internal pressure or adapting their grading behavior because they fear being perceived as hostile towards students.	„Dass man halt sonst, wenn man irgendwie auf eine strenge Notengebung beharrt, ja, als frustriert oder schülerfeindlich irgendwie dargestellt wird. [...] Es wird dann vielleicht eher zum inneren Druck oder dass man sich denkt, ‚ich will ja auch gute Noten haben‘, oder ‚mir ist auch alles wurscht‘.“ (P04)
Adaptation to Perceived Leniency in Colleagues' Grading	Teacher reports adjusting their own grading practices upward in response to perceived overly generous grading by colleagues	„Ich muss sagen, ich hab dann auch damals ein bisschen nachgezogen, hab mir gedacht, OK, wenn die 10 Sehr gut vergibt, dann ich auch [...] weil offensichtlich nimmts hier eh keiner so ernst.“ (P04)

Critique of Misguided Student-Friendliness	When teachers criticize their colleagues for being overly accommodating or kind to students in ways that compromise fair grading.	„Dieses Wohlwollend-pro-Schüler-Korrigieren und ‘der hat doch sicher das gemeint’ und ‘ich kenne ja den Schüler und der kann nicht formulieren’. [...] Es wird immer geschaut, dass man das Beste herausholt. Und ob das immer so neutral und objektiv ist, das wage ich zu bezweifeln.“ (P05)
--	---	--

Codes Belonging to the Additional Category 12

Category 12: Teachers’ Attitudes and Criticism of the Grading Reform

Code	Code Description	Examples
Cons		
Perceived Devaluation of the Oral Matura Through Grade Averaging	When teachers express concern that the grading reform has undermined the credibility and standards of the oral Matura, describing situations where students receive passing or even good final grades despite poor oral exam performance, due to averaging with strong classroom grades.	„Ich sag das auch ganz laut, das kann gern in die Presse: Die mündliche Matura ist nur mehr ein Grund, sich ein schönes Gewand anzuziehen, weil die Matura ist quasi eine Farce. Die [Schüler] treten an in Fächern, wo sie nur Einser und Zweier haben. Und dann haben wir es teilweise wirklich so, die nehmen den Zettel, sagen nichts dazu, geben ab und dann haben sie einen Dreier, weil 1 und 5 ist 3.“ (P01)
Rejection of Classroom Grade Inclusion When Matura Is Failing	When teachers express disapproval of including the classroom grade in the final result if the standardized Matura exam is failed. They view	“Also prinzipiell finde ich die Einrechnung der Jahresnote einen Schwachsinn. Also ich finde nicht okay. Vor allen Dingen im Hinblick, im Hinblick auf, dass man dadurch positiv werden kann. Also

this as undermining the purpose and value of the Matura.

ich finde es prinzipiell OK, wenn man eine Jahresnote miteinberechnet, ja, aber wenn die Matura negativ ist, dann finde ich, sollte keine Jahresnote mehr einberechnet werden können. Also ein Dreier, der dann quasi die Matura zunichtemacht, dann kann man die Matura abschaffen.“ (P02)

Perceived Decline of Centralized Standards

When teachers express the view that the standardized nature of the Central Matura has been lost, resulting in a perceived lowering of assessment quality.

„Ich finde es einen Rückschritt zur Zentralmatura. Ich finde, dass die Zentralmatura den zentralen Charakter verloren hat.“ (P14)

Critique of Reform for Promoting Checklist Mentality

When teachers criticize the Grading Reform for giving the students the chance to focus on meeting minimal requirements (classroom grade 3) rather than engaging deeply with learning and Matura relevant content, leading to frustration among teachers.

„Ich glaub generell, dass es ist zwar wichtig für Schüler, auch zu wissen, was muss ich tun, um durchzukommen oder was sind die Anforderungen, die soll man auch transparent kommunizieren. Aber irgendwie der Zugang, dass man Dinge einfach nur abhakt und sie so auf einem sehr teilweise sehr oberflächlichen Level damit nur auseinandersetzt, glaube ich, das verbaut die Chance, ja, tatsächlich was zu lernen, oder zu irgendwelchen eigenen Erkenntnissen zu kommen oder zu, ja, einfach eigenständig zu denken oder einfach Meinungen zu entwickeln.“ (P04)

Increase in Failing Matura Grades since Grading Reform as Negative Side Effect

When teachers report a noticeable rise in failing Matura grades (grade 5) since Grading Reform.

„Also in unserer Schule ist das enorm, wie viele Schüler Fünfer haben bei der Matura, seitdem die Jahresnote eingerechnet wird, viel viel mehr.“ (P02)

Skepticism About Reform's Purpose: Inflating Success Rates	When teachers question the true intent behind the grading reform, suspecting that it was introduced primarily to increase the number of students passing the Matura.	„Das können wir uns alle vorstellen, was der Grund [für die Notenreform] ist: Damit die Noten besser werden. Man rühmt sich ja dann damit, wieviel nicht 99,9% die Deutschmatura auf Anhieb geschafft haben.“ (P01)
Criticism of the 30% Rule	When teachers criticize the 30% threshold for passing the Matura as too low, expressing concerns that it allows students to pass with minimal effort, drastically undermining the exam's standards (especially in oral Matura exams).	„Diese 30% hat man, wenn man irgendwie aktiv an der Prüfung teilnimmt. Das fand ich ganz schlecht. Also wenn jemand da so faktisch nichts mehr tun muss, außer so zu tun, als hätte er sich ein bisschen was zu den Fragen überlegt, und dann ist er durch. Also das find ich ziemlich daneben.“ (P13)
Pros		
Valuing Holistic Assessment to Balance One-Time High-Stakes Testing	When teachers support the integration of classroom grades, emphasizing that it compensates for the pressure of a one-day high-stakes exam and that a holistic view better represents students' overall performance.	„[Ich] finde diese Richtung eigentlich schon gut. Ich finde es auch in Ordnung, dass man sagt, es zählt nicht nur diese Tagesverfassung bei der Matura selber.“ (P03)
Agreement with the Grade Reform in Mathematics	When teachers express agreement for the grading reform in mathematics, highlighting that it might increase student motivation, encourages early preparation, or just because both assessments (year grade and Matura performance) reflect the same content.	„Weil es die Motivation der Schülerinnen steigern würde, also jetzt in Mathematik, [...] aber in allen anderen Nebenfächern finde ich es eigentlich schlecht. [...] Der Unterschied ist glaube ich, dass die in Mathe im März wirklich alle topfit schon drauf sind. Das heißt, die müssen schon alles halbwegs gelernt haben, damit sie die positive Jahresnote kriegen.“ (P08)

Appendix C

Interview Participants

Teacher ID	Gender	Subject	School Context	Years of Teaching Experience
P01	female	German	Academic School in urban area of Tyrol	25
P02	female	German	Vocational School in rural area of Tyrol	13
P03	male	German	Elite Academic School in urban area of Vienna	12
P04	female	German	Academic School in rural area of Upper Austria	8
P05	female	German	Vocational School in rural area of Lower Austria	23
P06	female	German	Vocational School in urban area of Lower Austria	26
P07	male	German	Vocational School in urban area of Vienna	17
P08	female	Mathematics	Academic School in urban area of Tyrol	27
P09	female	Mathematics	Vocational School in rural area of Lower Austria	23
P10	male	Mathematics	Academic School in rural area of Lower Austria	15
P11	female	Mathematics	Vocational School in urban area of Vienna	20
P12	female	Mathematics	Vocational School in urban area of Vienna	11
P13	female	Mathematics	Vocational school in urban area of Lower Austria	12
P14	female	Mathematics	Elite Academic School in urban area of Vienna	22

Appendix D

Table 4. Ordinal logistic regression predicting classroom grades in German at academic secondary schools (2017–2023)

Model Fit / Predictor							
Model Fit Measures	Estimate	SE	Z	p	Odds ratio	95% CI Lower	95% CI Upper
Deviance	3733						
AIC	3735						
R ² _{McF}	0.00176						
Predictor							
year:							
2017 – 2019	0.017	0.201	0.084	0.933	1.017	0.686	1.51
2018 – 2019	-0.128	0.196	-0.653	0.514	0.880	0.599	1.29
2020 – 2019	0.047	0.197	0.238	0.812	1.048	0.712	1.54
2021 – 2019	-0.058	0.191	-0.301	0.763	0.944	0.649	1.37
2022 – 2019	-0.360	0.194	-1.859	0.063	0.698	0.477	1.02
2023 – 2019	-0.079	0.192	-0.413	0.680	0.924	0.635	1.35

Note. The dependent variable 'classroomgrade' has the following order: 1 | 2 | 3 | 4 | 5; Models estimated using sample size of N = 1281.

Table 5. Ordinal logistic regression predicting classroom grades in German at vocational secondary schools (2017–2023)

Model Fit / Predictor							
Model Fit Measures	Estimate	SE	Z	<i>p</i>	Odds ratio	95% CI Lower	95% CI Upper
Deviance	6656						
AIC	6676						
R ² _{McF}	0.00677						
Predictor							
year:							
2017 – 2019	0.0250	0.144	0.174	0.862	1.025	0.774	1.359
2018 – 2019	-0.3245	0.142	-2.284	0.022	0.723	0.547	0.955
2020 – 2019	-0.3003	0.135	-2.230	0.026	0.741	0.569	0.964
2021 – 2019	-0.7396	0.132	-5.595	<.001	0.477	0.368	0.618
2022 – 2019	-0.0869	0.133	-0.655	0.513	0.917	0.707	1.189
2023 – 2019	-0.2854	0.134	-2.137	0.033	0.752	0.578	0.977

Note. The dependent variable 'classroomgrade' has the following order: 1 | 2 | 3 | 4 | 5; Models estimated using sample size of N = 2401.

Table 6. Ordinal logistic regression predicting classroom grades in Mathematics at academic secondary schools (2017–2023)

Model Fit / Predictor							
Model Fit Measures	Estimate	SE	Z	<i>p</i>	Odds ratio	95% CI Lower	95% CI Upper
Deviance	3635						
AIC	3655						
R ² _{McF}	0.00545						
Predictor							
year:							
2017 – 2019	-0.2277	0.239	-0.951	0.342	0.796	0.498	1.274
2018 – 2019	-0.1493	0.205	-0.730	0.465	0.861	0.576	1.286
2020 – 2019	0.0637	0.204	0.312	0.755	1.066	0.714	1.589
2021 – 2019	-0.3165	0.198	-1.598	0.110	0.729	0.494	1.073
2022 – 2019	-0.6659	0.202	-3.289	0.001	0.514	0.345	0.763
2023 – 2019	-0.2005	0.198	1.010	0.312	0.818	0.554	1.207

Note. The dependent variable 'classroomgrade' has the following order: 1 | 2 | 3 | 4 | 5; Models estimated using sample size of N = 1256.

Table 7. Ordinal logistic regression predicting classroom grades in Mathematics at vocational secondary schools (2017–2023)

Model Fit / Predictor							
Model Fit Measures	Estimate	SE	Z	<i>p</i>	Odds ratio	95% CI Lower	95% CI Upper
Deviance	5751						
AIC	5771						
R ² _{McF}	0.00798						
Predictor							
year:							
2017 – 2019	0.1449	0.151	0.957	0.338	1.156	0.859	1.556
2018 – 2019	-0.0178	0.159	-0.112	0.910	0.982	0.720	1.341
2020 – 2019	-0.6028	0.138	-4.354	<.001	0.547	0.417	0.718
2021 – 2019	-0.3079	0.143	-2.149	0.032	0.735	0.555	0.973
2022 – 2019	-0.0746	0.140	-0.533	0.594	0.928	0.705	1.221
2023 – 2019	-0.5632	0.158	-3.570	<.001	0.569	0.418	0.776

Note. The dependent variable 'classroomgrade' has the following order: 1 | 2 | 3 | 4 | 5; Models estimated using sample size of N = 2188