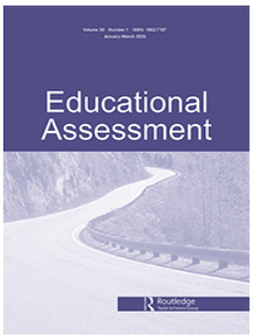

Educational Assessment

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/heda20

Impact of Displaying Grades Vs. Not Displaying Grades on Academic Performance and Emotional Outcomes While Delivering Feedback Comments: A Longitudinal Study

Ernesto Panadero & Iván Sánchez-Iglesias

To cite this article: Ernesto Panadero & Iván Sánchez-Iglesias (2025) Impact of Displaying Grades Vs. Not Displaying Grades on Academic Performance and Emotional Outcomes While Delivering Feedback Comments: A Longitudinal Study, Educational Assessment, 30:1, 21-40, DOI: 10.1080/10627197.2025.2455128

To link to this article: https://doi.org/10.1080/10627197.2025.2455128

Published online: 23 Jan 2025.

EDUCATIONAL ASSESSMENT
2025, VOL. 30, NO. 1, 21–40
https://doi.org/10.1080/10627197.2025.2455128

Routledge
Taylor & Francis Group

Impact of Displaying Grades Vs. Not Displaying Grades on Academic Performance and Emotional Outcomes While Delivering Feedback Comments: A Longitudinal Study

Ernesto Panadero [a,b,c] and Iván Sánchez-Iglesias [d]

[a]Centre for Assessment Research Policy and Practice in Education (CARPE), Institute of Education, Dublin City University, Dublin, Ireland; [b]Education Regulated Learning and Assessment group, Facultad de Educación y Deportes, Universidad de Deusto, Bilbao, Spain; [c]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain; [d]Universidad Complutense de Madrid, Spain

ABSTRACT

This study investigates the impact of displaying grades versus not displaying grades on student performance and emotions in secondary education, while keeping feedback comments constant. Using a longitudinal design, we examined intra-individual changes in performance and emotional responses among 99 students across four classroom groups as they experienced phases of grade display and non-display. Contrary to the notion of grades solely as detrimental, our findings reveal a complex relationship. Initially, grade display decreased performance and evoked negative emotions, but these effects dissipated over time, suggesting student adaptation. Our study highlights the significant impact of feedback comments, suggesting their importance independent of grades. These results call for a sophisticated understanding of grading practices, emphasizing longitudinal research to capture the evolving effects of grades and feedback. Practical implications for educators include maintaining stable grading practices and providing preparatory guidance to mitigate initial negative impacts, contributing to optimizing educational assessment strategies.

Study

Grading remains a cornerstone of educational assessment in a significant number of countries, serving as a critical mechanism to quantify student achievement and facilitate decisions of significant consequence, such as university admissions (Brookhart et al., 2016). Despite its ubiquity, the efficacy and impact of grades on student outcomes have ignited considerable debate among educational researchers and practitioners. At the heart of this discourse is a dual-edged narrative: on the one hand, grades are commonly used to estimate a level of achievement and provide a benchmark for learning progress (Brookhart & Guskey, 2019; Brookhart et al., 2016; Guskey, 2019). On the other hand, an emerging body of evidence suggests that grades may have unintended negative effects, particularly when considered in the broader spectrum of student motivation and emotional well-being (Koenka et al., 2019; Sadler, 2005). This paradox underscores the imperative to analyze deeper into the dynamics of grading practices, especially as they are used in classrooms around the world. This exploration is particularly crucial as we consider the nuanced interplay between grading and

CONTACT Ernesto Panadero ernesto.panadero@dcu.ie Centre for Assessment Research Policy and Practice in Education (CARPE), School of Policy and Practice, Institute of Education, Dublin City University, St. Patrick Campus (Drumcondra), C112 Main Building Dublin 9 D09 DY00, Ireland

This article has been republished with minor changes. These changes do not impact the academic content of the article.

feedback mechanisms, such as written comments, which have been shown to differently influence student outcomes (Koenka et al., 2019). Our study seeks to illuminate the complex relationships between grading practices (displaying vs. non-displaying) and feedback comments on academic performance and student emotions, contributing to the ongoing dialogue on optimizing educational assessment strategies.

Grades in education

"Grading refers to the symbols assigned to individual pieces of student work or to composite measures of student performance on student report cards" (Brookhart et al., 2016, p. 804). Grades have been a fundamental component of the educational system for over a century in many countries, serving as multidimensional indicators that reflect not only academic achievement but also non-cognitive factors such as effort and engagement (Brookhart et al., 2016). Grades are part of most formal educational systems around the world as ways to quantify students' educational achievement and to make high stakes decisions (e.g., university entry) (Guskey, 2019; Guskey & Brookhart, 2019). In this context, classroom activities contributing to grades are considered high stakes, as they cumulatively become a crucial indicator of educational attainment.

Importantly, there is ample research in the area of testing showing that activities with high stakes repercussions in comparison to low stakes enhance students' cognitive and academic performance (Finney, Perkins, et al., 2020; Wise & DeMars, 2005). One of the main arguments of that high stakes activities increases students' motivation as there is something in it for them, which further influences positively their performance (Attali, 2016; Wolf & Smith, 1995). Importantly, high stakes also influence students' emotions (Wolf & Smith, 1995), potentially amplifying the effects observed in low-stakes contexts on performance via affecting motivation (Satkus & Finney, 2021). In this line, the impact of emotions and perceived value on effort and performance has been studied in low-stakes contexts, with findings suggesting that managing emotions is more influential than the perceived importance of the test (Finney, Perkins, et al., 2020). However, further study is needed in contexts where the stakes are not as high (such as those without grades), given Wolf and Smith's (1995) findings on the association between importance and effort in both low and high-stakes contexts. In sum, it is crucial to better understand the impact of grades in classrooms, as grades constitute the most significant form of high stakes in everyday classrooms.

Building on the understanding of grades as significant indicators of educational achievement, it is crucial to examine their impact not just on performance, but also on student motivation and emotions (Goetz et al., 2018; Lipnevich et al., 2021a). Crucially, Koenka et al. (2019) in their comprehensive meta-analysis, shed some light on this complex relationship. Their findings reveal that while grades can enhance achievement, they often exert a negative influence on motivation when compared to the absence of feedback. More strikingly, students assessed solely with grades demonstrated lower performance and motivation than those who received feedback in the form of comments. Their study also revealed that the concurrent receipt of grades and feedback comments did not significantly differ in performance outcomes compared to grades alone, suggesting the potential for grades to mitigate the positive effects of feedback comments. This raises important questions, particularly as Koenka et al. (2019) highlighted the limited number of studies comparing the impact of combining grades with comments versus feedback comments alone. This prompts further investigation into the interplay between feedback comments and the display of grades. We have tasked ourselves with exploring this particular line of work in this study.

Given the enduring debate over grades' true educational value – as highlighted by Sadler (2005) – their prevalence in educational systems, especially where resources are limited, seems to be a fixed reality. This necessitates ongoing exploration into how grades are utilized, particularly in conjunction with another of the most common form of feedback: teachers' written comments. The insights from Koenka et al. (2019) serve as a critical reminder of the complex effects of grades on student motivation

and performance, emphasizing the need for continued research into finding the most effective strategies for using grades in educational assessment.

Summative and formative assessment: grades and feedback comments

Probably the two most common forms of feedback that students receive are grades and written comments, both being widely used in education in many countries. Grades and scores are the maximum exponent of summative assessment and are consequential for students' academic progression, such as selection for the next level educational system (Klapp, 2015), being therefore considered high stakes. In this study, we specifically explored whether informing students of their grades or not (display vs. non-display) influenced their academic performance and emotions.

Feedback comments are evaluative observations usually directed at a student's work, intended to assess its quality and to increase students' understanding of their mistakes and correct answers to help them move forward with their learning. Extensive research demonstrates the significant impact of formative feedback on student performance (Hattie & Timperley, 2007; Panadero & Lipnevich, 2022). In this study, we used teachers' feedback comments as a constant, while grades could be displayed (i.e., given to students) or not displayed (i.e., not given to students yet calculated by the teachers and the students being aware of such fact).

It is common to encounter three distinct scenarios in the context of grading and feedback: the provision of only comments, only grades, or a combination of both (Koenka et al., 2019; Lipnevich et al., 2021a). Despite this, research, as indicated by these studies, necessitates a deeper exploration of the nuanced effects of grading on students' cognitive, affective, and behavioral outcomes. Consequently, we focused on a specific aspect of grades: the impact of reporting them to students. The rationale behind this is that giving students both their grades and feedback comments might detract from their attention to the comments and influence them emotionally. Therefore, we aimed to isolate the impact of grades when coupled with feedback comments, allowing us to directly assess their influence on students' academic performance and emotional responses.

The impact of displaying grades on academic performance

Given that feedback comments will remain constant throughout our study, our investigation will specifically explore the effects of displaying grades. Grades, as indicators of a student's performance on a task, are assumed to influence future task performance (Bowers, 2019). The prevalent research question in this domain has been whether grades serve to enhance or diminish performance, with primary studies indicating both positive and negative outcomes (e.g., Baumert & Demmrich, 2001; Butler & Nisan, 1986). Notably, Lipnevich and Smith (2009a, 2009b) found in a comprehensive experimental study that students who received detailed written comments significantly outperformed those who received only grades or both comments and grades. While Koenka et al. (2019) provided a comprehensive meta-analytical overview supporting these findings, an area yet to be fully investigated is the specific effect of displaying or withholding grades. More specifically, there is a need to explore whether informing students of their grades (i.e., displaying vs. not displaying them) for each task influences their academic performance.

We chose this "less aggressive" approach, rather than fully introducing or removing grades, to examine whether simply displaying grades plays a beneficial or detrimental role in academic performance. In this study, students were always aware that their work was being graded; therefore, the primary question is whether the act of showing grades impacts their performance.

The impact of displaying grades on emotions

Academic environments are known to be emotionally charged (Pekrun et al., 2018), with students experiencing a broad spectrum of emotions, including happiness, anxiety, and relief, during

academic activities (Pekrun & Linnenbrink-Garcia, 2014). Certain academic tasks and events, such as test anxiety (e.g., Pekrun et al., 2011) and emotions associated with written tests and exams (Goetz et al., 2018; Vogl & Pekrun, 2016), have garnered significant attention. Notably, the research by Lipnevich and Smith (2009a, 2009b) demonstrated that students who received grades exhibited an increase in negative affect compared to those in non-graded conditions. Moreover, Van der Kleij and Lipnevich (2020) highlighted in their review a pronounced affective response. However, they noted a gap in linking students' perceptions and affective responses to academic performance or other significant outcomes, such as course completion or well-being. Our study aims to bridge this gap by examining the impact of grades on both academic performance and emotions.

Pekrun's Control-Value Theory (CVT) (Pekrun et al., 2006, 2018) serves as the theoretical backbone for our investigation, positing that students' academic emotions are profoundly influenced by their perceived control over learning tasks and the value they ascribe to these tasks. Following CVT, emotions are categorized based on their valence (positive or negative) and their activation level (activating or deactivating), suggesting a nuanced impact on students' learning outcomes, motivation, and self-regulation strategies. CVT is instrumental in examining the emotional landscape of academic settings, particularly in understanding how grades, as external indicators of academic success or failure, shape students' emotional experiences and, by extension, their engagement and performance in educational activities. Our study leverages this theory to explore the differential impact of the display or non-display of grades alongside feedback comments, hypothesizing that these grading practices significantly alter students' perceptions of control and task value, thereby eliciting distinct emotional responses, also influencing academic performance.

Building upon these foundational insights, CVT provides a comprehensive understanding of how emotional experiences in academic settings are shaped by the interplay of control and value appraisals. In line with CVT, students' emotional responses to grades are likely mediated by their beliefs about their ability to control academic outcomes (control appraisals) and the significance they place on these outcomes (value appraisals) (; Pekrun et al., 2002). When grades are perceived as highly valuable and students feel they have control over their performance, positive emotions such as pride or hope are more likely to be experienced (Pekrun & Bühner, 2014. Conversely, if students perceive grades as critical but feel a lack of control, negative emotions like anxiety or hopelessness may arise, potentially diminishing academic performance and motivation .

In our study, we hypothesize that the display or non-display of grades, combined with feedback comments, will influence students' emotional and academic responses through their effects on control and value perceptions. Displaying grades may initially provoke negative emotions due to perceived pressure and potential threats to self-worth, especially among students with low control beliefs (Lipnevich et al., 2016). However, over time, students might adjust to the grading practice, leading to stabilized emotions and performance. On the other hand, non-displaying grades, paired with constructive feedback, may enhance students' sense of control and intrinsic motivation, fostering positive emotions and improved academic outcomes (Hattie & Timperley, 2007).

Aim and research questions

Our study aims to investigate the impact of displaying versus non-displaying grades on two key variables: performance and emotions. We examined two research questions (RQs):

RQ1. How does displaying versus non-displaying grades affect students' performance while keeping feedback comments?

H1a. *Non-displaying grades but retaining feedback comments will enhance performance.*

H1b. *Displaying grades with feedback comments will lower performance.*

RQ2. How does displaying versus non-displaying grades impact students' emotions while keeping feedback comments?

H2a. *Non-displaying grades but retaining feedback comments will increase positive emotions.*

H2b. *Displaying grades with feedback comments will decrease positive emotions.*

H2c. *Non-displaying grades but retaining feedback comments will decrease negative emotions.*

H2d. *Displaying grades with feedback comments will increase negative emotions.*

To enhance our study's validity, we measured three relevant variables to assess comparability between the two conditions. First, we explored goal orientation as motivation can modulate the effect of feedback on students' emotions (Goetz et al., 2018). Second, we also checked personality traits using the Big Five questionnaire as especially conscientiousness has shown to influence learning (Zeidner & Matthews, 2012). And third, as students' receptivity to feedback could influence how students react to grades (Lipnevich et al., 2021b), we also investigated it. Additionally, a recent study has found that personality traits and receptivity to feedback are interlinked (Lipnevich et al., 2021b), thus the utility of assessing their equivalence.

Method

Research design

This study employs a quasi-experimental methodology with four classroom groups organized around two conditions that counterbalanced the presentation of grades. In the first condition, grades were displayed on two occasions and not displayed on two other occasions. In the second condition, grades were not displayed on two occasions and displayed on two others.

Both conditions received at all times feedback comments for two reasons. First, because it was an important research purpose to explore the combination of feedback comments & grades vs. feedback comments alone (Koenka et al., 2019). Second, because without feedback comments, the participants would not have received any type of feedback at the non-display phase.

Sample

A total of 99 students from a charter primary and secondary school in Madrid participated in the study, comprising 61 females (61.7%). Their ages ranged from 13 to 17 years, with an average age of 15.1 years ($SD = 0.6$). These participants were secondary education students enrolled in the 9th and 10th grades, corresponding to the 3rd and 4th ESO grades in the Spanish education system, spread across two classes (A and B) at each grade level. Participation in the study was voluntary and did not offer any form of reward. All students participated, as parental authorization was secured, and the activities conducted were integral to the compulsory curriculum. The research received ethical approval from the institution formerly affiliated with the first author. Additionally, instruction across the two grade levels was delivered by two teachers, with one teaching both 9th-grade classes and another teaching both 10th-grade classes. To create the conditions for our study, one class from the 9th grade and one from the 10th grade were randomly assigned to the condition Display/Non-display, while the remaining classes were assigned to Non-display/Display. This arrangement allowed us to partially control for any potential differences in instructional style between the teachers.

The population effect size that our study was able to detect was computed using a sensitivity analysis (Faul et al., 2007). For these analyses, we considered a mixed ANOVA for our 2 × 4 between-within factorial design, using $\alpha = .05$, $1 - \beta = .80$, and $N = 99$. This analysis yielded a minimum effect size of $f = .225$.

Instruments

Academic tasks and performance scores

These consisted of four social sciences – history – activities including open questions and text analyses. These tasks were graded in accordance with the Spanish scoring system, which ranges from 0 to 10 points. These grades were both our dependent variable for academic performance and the information the participants received when they were in the display grade phase. The tasks belonged to the official curriculum and, while different for the 9th and 10th groups, they were designed to present the same level of difficulty for both year levels. The tasks assigned were identical for the two classroom groups within each grade level, ensuring uniformity in the curriculum delivered. Additionally, before the experimental intervention started, we established a baseline by assessing performance on a task of the same type involved in the study. All participants were informed of their grades for these baseline tasks, making sure everyone started from a uniform starting point of receiving grades for their class work, as was customary for our participants in their regular classes. Furthermore, submission methods differed between grades due to the established practices of their respective teachers: 9th grade participants submitted their work electronically via online platform, while 10th grade participants submitted their assignments in paper form, handwritten. This differentiation was preserved for the teachers' convenience as they had been working that way throughout the academic year. Importantly, the classroom groups were distributed evenly among the two conditions to account for this confound.

Finally, to ensure uniformity across our measurement of academic performance, the first author independently graded the students' work without any familiarity with the students – i.e., blind grading. These grades showed an inter-rater reliability Cohen's kappa of .85 with Teacher A (9th grade), and .89 with Teacher B (10th grade). Agreement was defined as either an identical score or a discrepancy of less than 0.5 points on a 10-point scale. In cases of deviation of half point or above the first author score was selected as he had no contact with the participants and was not interpersonally biased.

Emotions scale

We employed an ad-hoc version of Pekrun's Achievement Emotions Questionnaire (Pekrun et al., 2005, 2011). Our scaling deviates from the original in five ways. First, our scale contained eight out of the nine original emotions: enjoyment, pride, hope, anger, boredom, anxiety, shame, and hopelessness. Relief was not included following the proposition by Pekrun et al. (2011, Appendix A; see also Bieleke et al., 2021) that when measuring learning-related emotion relief might not be a salient emotion experience by students. Second, instead of using the original items (e.g., "Thinking about class makes me feel uneasy"), we formulated a context and then list the emotions. The text used to formulate such context was: "Next, you will find different feelings and emotions. Read each item thinking about how you feel in relation to the comments your teacher has just provided you. Mark the appropriate response next to the word, using the following scale for your answers." Third, as previously mentioned, we formulated the emotions as a list and as adjective (e.g., proud instead of pride), the latter being a similar solution to the original items where emotions are formulated as adjectives (e.g., "Studying makes me irritated," "I feel ashamed that I can't absorb the simplest of details"). Fourth, while we also used a 5 points Likert-scale like the original, we slightly changed the meaning of the different points. Our scale was 1 = very slightly or not at all [experienced] – 5 = extremely [experienced], and the original being 1 = Strongly disagree – 5 = Strongly agree (Pekrun et al., 2005). Finally, we did not employ the activation vs deactivating dimension, and rather focus exclusively on the positive vs negative affect dimension for two reasons. First, following the recommendation presented in the first point above, leaving out relief meant that there was no representation

of a positive deactivating emotion. Second, focusing on the positive vs negative affect dimension streamlined our analysis, directly aligning with our study's objectives to explore the impact of emotions on academic performance and engagement. This binary classification simplified data processes and enhanced clarity in interpreting how emotions influence learning outcomes. Thus, we grouped the emotions just based on valence: positive and negative affect, also in line with how the PANAS, a widely used emotions questionnaire, groups its emotions (Watson et al., 1988). In addition to the emotions explored based in Pekrun's Achievement Emotions Questionnaire, we also analyzed the emotions measured by the PANAS. We identified two emotions that we added to our ad-hoc emotions scale as we believed there are usual for Secondary education students when receiving grades and feedback comments: Enthusiastic and Disappointed (i.e., "frustrado" in Spanish). Therefore, out positive emotions scale included enjoyment, pride, hope and enthusiasm (Cronbach's alpha of .81); while our negative emotions scale included: anger, boredom, anxiety, shame, hopelessness and disappointment, (Cronbach's alpha of .76); at T1, a floor effect was observed for Anger, Shame, and Hopelessness. The resulting distribution of 40% positive items and 60% negative items is intentional, reflecting the empirical and theoretical considerations that negative emotions in educational settings are more varied and complex, necessitating a more detailed examination (see for example the original distribution of the Achievement Emotion questionnaire, Pekrun et al., 2005). The summed emotions used to create the scales showed a positive correlation, both for positive emotions (ranging from .200 to .330) and negative emotions (ranging from .326 to .550). The correlations between the positive emotions and negative emotions subscales, across the four assessment time points, ranged from −.118 to −.339. Our ad-hoc instrument can be found in Appendix A.

Instruments to assess the comparability of conditions

Goal orientation questionnaire

Situated Goals Questionnaire (SGQ-U) (Alonso-Tapia et al., 2018). This questionnaire was used to assess student goal orientations and it is based in the trichotomous model of goal orientation (Elliot, 2005). It contains 30 items grouped in six first order scales each including five items: Desire to learn, Desire to be useful, Desire to succeed, Desire to pass, Desire to give up, and Desire to avoid failure. These scales are related to three second order factors that measure goal orientations: Learning orientation (Cronbach α = .86) example item: "When I am studying to prepare an exam, I try very hard because if I am competent, I will be able to help others;" Performance orientation (α = .87) example item: "If I have to study to prepare an exam, I think first of all on achieving a good grade;" and Avoidance orientation (α = .83) example item: "If I am preparing an exam, my main concern is whether I will do it worse than my peers and that they heard of it." The items are answered on a 5-points Likert scale, from totally disagree to totally agree.

Big five Questionnaire Children (BFQ-C) (Barbaranelli et al., 1998)

We used the Spanish validated version for underage participants (Carrasco Ortiz et al., 2005) named BFQ-NA (Big Five Questionnaire para niños y adolescentes). It contains 65 items to be answered in 5-point Likert scale ranging from 1 (Almost never) to 5 (Almost always) organized around five scales that we list next to their reliability values in our sample: openness (α = .79), extraversion (α = .66), conscientiousness (α = .77), agreeableness (α = .63), and neuroticism (α = .81).

Receptivity to feedback inventory (Lipnevich et al., 2021b)

This self-report tool measures students' acceptance of instructional feedback. We used a shorter version including only three scales and leaving out Behavioral Engagement as it was not relevant for our purposes. The version we employed contained 24 items to be answered in 5-point Likert scale (1 = strongly disagree and 5 = strongly agree), of which we eliminated a number of further items as they did not load in the theoretical factors after performing an exploratory factor analysis. We list next the three scales with their reliability values in our sample. First, experiential attitudes toward feedback (i.e.,

affect; e.g., I look forward to receiving the instructor's comments on my work) (α = .71); second, instrumental attitudes toward feedback (i.e., value for feedback; e.g., I find the comments I get on my assignment to be very helpful) (α = .74); and third, cognitive engagement with feedback (e.g., I know how to use feedback comments to improve my work) (α = .60). This scale demonstrated a low reliability index, markedly diverging from the original. A hypothesis for this discrepancy is that the original scale was validated with North American university students, whereas in our study, it was applied to Spanish secondary education students.

Instruments used in the intervention

Grades

As previously mentioned, participants engaged in five academic tasks (one baseline and four during the experimental phase), graded according to the Spanish scoring system, which ranges from 0 to 10 points. All participants were informed of their grades for the baseline task. For the subsequent four tasks, participants received their grades during the display phase, but not during the non-display phase. The participants' teachers were instructed to deliver the grades individually, and students were not shown a distribution of the classroom grades. Note that the grades serve the dual function of IV (whether or not they are shown to students) and DV (the numerical value used to operationalize performance).

Feedback comments

The feedback was provided to participants in writing, in two distinct forms. 9th graders received their feedback as comments embedded within the Word documents they had uploaded to a platform or sent via e-mail to the teachers. To access their feedback, they needed to download the document from the platform. 10th graders received their feedback as handwritten comments on the paper copies submitted to their teacher. For all participants, the feedback predominantly consisted of corrections to conceptual mistakes or typos, and positive reinforcement for correct responses or exceptional work. To categorize the feedback comments, we adopted the framework proposed by Panadero and Lipnevich (2022, Table 1 in the original source), which shows that most comments were aimed at verifying information, primarily correcting students with a neutral or negative valence and a low information load due to their brevity. The primary purpose of these comments was to provide information about learning and performance, with a minority also addressing motivation and affect to encourage students. Examples of the feedback are included in Appendix B.

Procedure

We approached the educational center to identify teachers interested in participating in our study. The two teachers who joined were ideal candidates, as they taught the same subject and each managed two classroom groups at consecutive year levels (i.e., Teacher A managed two 9th grade classroom groups and Teacher B managed two 10th grade groups). To initiate the process, we prepared a document outlining the study details for the center and parents, who then provided their parental consent via the previous agreement established with the educational center for research purposes. Subsequently,

Table 1. Organization of conditions regarding grades presence or removal.

	Composition	Baseline	Time 1	Time 2	Time 3	Time 4
Display/ Non-display	9th Group A $n = 27$ 10th Group A $n = 23$	Grade display	Grade display		Grade non-display	
Non-display/Display	9th Group B $n = 25$ 10th Group B $n = 24$	Grade display	Grade non-display		Grade display	

Note. Received grade = In phase. Grade withheld = Out phase.

participants completed the Big Five Personality Traits questionnaire for children, the Receptivity to Feedback Scale, and the Situated Goals Questionnaire during their regular class time.

A few days later, they engaged in the baseline task and received their grades. Subsequently, the four classroom groups were randomly assigned to one of two conditions (Display/Non-display vs Non-display/Display), continuing with their regular semester schedule in the social sciences courses. Over the course of two months and one week, students completed four tasks that were collected, evaluated, and returned with written feedback each time.

Although all tasks were graded by the teachers, students in each of the conditions were not informed of their grades (i.e., Non-display phase) on two occasions (Table 1). Importantly, students knew that their work was going to be graded in all occasions. The teachers were instructed to clearly indicate to their students before they perform the task whether the grades were going to be displayed or not. Teachers were also instructed that, when handing out their corrected work to the students, they should emphasize that the written comments represented the most valuable feedback for their learning improvement, regardless of the grading phase the students were in. Additionally, teachers were instructed to communicate to those students that did not receive grades that they should focus on leveraging this feedback to improve their learning. In all four times, right after receiving their work corrected, the participants filled out the emotions scale[1]

Therefore, all work was graded by the teachers, and students were made aware that all activities would be graded; the only variation was in the disclosure of these grades to the students. The reason why students were informed that all pieces of work would receive grades was to ensure transparency and fairness toward our participants. The method of informing students about their grades varied between year levels but was consistent within each level's two classroom groups. Specifically, 9th grade participants were orally informed of their grades immediately after they downloaded their work from the online platform and before completing the emotions questionnaires. Meanwhile, 10th grade participants received their graded and annotated essays in paper form before completing the emotions questionnaire.

As mentioned earlier, Teacher A was responsible for instructing both 9th grade groups, and Teacher B handled both 10th grade groups, ensuring each teacher had one classroom group in each experimental condition to minimize the potential impact of teacher variability on the study's outcomes. Additionally, to guarantee uniformity across the learning environments, both teachers adhered to the same curriculum within their respective year levels. They were also specifically instructed to maintain consistency in their instructional approach, particularly in providing identical types of feedback to both groups, irrespective of the experimental condition.

Data analyses

We explored the effects of grades display vs non-display on two dependent variables: academic performance and emotions (positive and negative). We calculated the means and standard deviations for the dependent variables over time, categorized by condition. We run mixed-design repeated measures ANOVAs using as within factor the four measurement times and as between factors the two conditions. For the variables used to assess comparability we performed t-tests employing the condition as the between factor variable. Also, we analyzed the intercorrelation between performance and emotions. In all the analyses we used .05 as the significance threshold. All analyses were run with SPSS 27.

[1]The participants completed the Perceptions of Value questionnaire immediately following the Emotions scale. Unfortunately, due to limitations in the design of the tool, the data obtained were deemed unreliable and, therefore, were not included in our analysis. The questionnaire is provided in Appendix A to offer readers comprehensive insight into the data collection process..

Results

We did not find previous, significant differences among the two conditions in goal orientation (learning orientation, $p = .880$; performance orientation, $p = .333$; avoidance orientation, $p = .901$), personality traits (openness, $p = .450$; extraversion, $p = .912$; conscientiousness, $p = .230$; agreeableness, $p = .948$; neuroticism, $p = .897$), or receptivity to feedback ($p = .426$). Even more importantly, we did not find differences in the participants' performance of the baseline task $t(97) = 0.217$, $p = .828$, $M_{display/non\text{-}display} = 5.93$, $SD = 1.83$; $M_{non\text{-}display/display} = 5.84$, $SD = 2.41$, thus the content knowledge expertise seemed the same between the two conditions. Considering all these results, the groups assigned to each condition were considered equivalent at least in these variables. The grades were positively correlated across the four time points, both for the Display/Non-display condition (r ranging from .450 to .675) and the Non-display/Display condition (r ranging from .432 to .552).

Table 2 shows the intercorrelations between performance and emotions (positive and negative). The grades were positively associated with positive emotions and negatively associated with negative emotions at each respective assessment time point. The pattern is repeated for both conditions. In the Display/Non-display condition, the explained variance ranged from 10.6% to 25.7% for positive emotions, and from 3.3% to 22.7% for negative emotions. In the Non-display/Display condition, the explained variance ranged from 1.8% to 31.6% for positive emotions, and from 3.9% to 21.6% for negative emotions. The correlations between the positive emotions and negative emotions subscales, across the four assessment time points, ranged from −.463 to .009 for the Display/Non-display condition, and from −.283 to −.114 for the Non-display/Display condition.

RQ1. How does displaying versus non-displaying grades affect students' performance while keeping feedback comments?

We used the grades as dependent variables, using condition and time as factors in a mixed-design ANOVA. Table 3 shows mean values and standard deviation. The main effect of time was significant $F(3, 270) = 13.48$, $p < .001$, $\eta^2 = .130$; but there was no effect of condition, $F(1, 90) = 0.54$, $p = .465$. Nevertheless, as there was a significant interaction effect between factors, this interaction should be interpreted instead of the main effects.

The interaction effect was significant $F(3, 270) = 13.72$, $p < .001$, $\eta^2 = .132$ (see Figure 1). We found a significant higher performance mean in the Display/Non-display condition at T3, $p < .001$. In line with that hypothesis H1a, we found significant increases in the Display/Non-

Table 2. Intercorrelations between performance and emotions, by condition.

			Positive emotions (PE)				Negative emotions (NE)			
Condition			T1	T2	T3	T4	T1	T2	T3	T4
Display/ Non-display ($n = 50$)	Display	T1	.325*	.431**	.164	.358*	−.181	−.219	−.188	−.160
		T2	.043	.368**	.246	.103	−.168	−.280*	−.151	−.065
	Non-display	T3	.260	.394**	.411**	.257	−.231	−.381**	−.444**	−.324*
		T4	.251	.223	.095	.507**	−.227	−.389**	−.405**	−.476**
Non-display/Display ($n = 49$)	Non-display	T1	.180	−.037	.035	.277	−.337*	−.157	−.137	−.435**
		T2	.051	.133	−.049	.213	−.271	−.197	−.284	−.173
	Display	T3	.200	−.028	.369*	.243	−.332*	−.233	−.465**	−.411**
		T4	.124	.097	.052	.562**	−.252	−.147	−.105	−.355*

Note. $N = 99$. * $p < .050$. ** $p < .010$.

Table 3. Mean and standard deviation for performance.

Condition	Time 1	Time 2	Time 3	Time 4
Display/Non-display	5.89 (1.33)	5.81 (1.36)	6.80 (1.37)	6.73 (1.63)
Non-display/Display	6.37 (1.60)	5.83 (1.51)	5.40 (1.73)	6.86 (1.86)

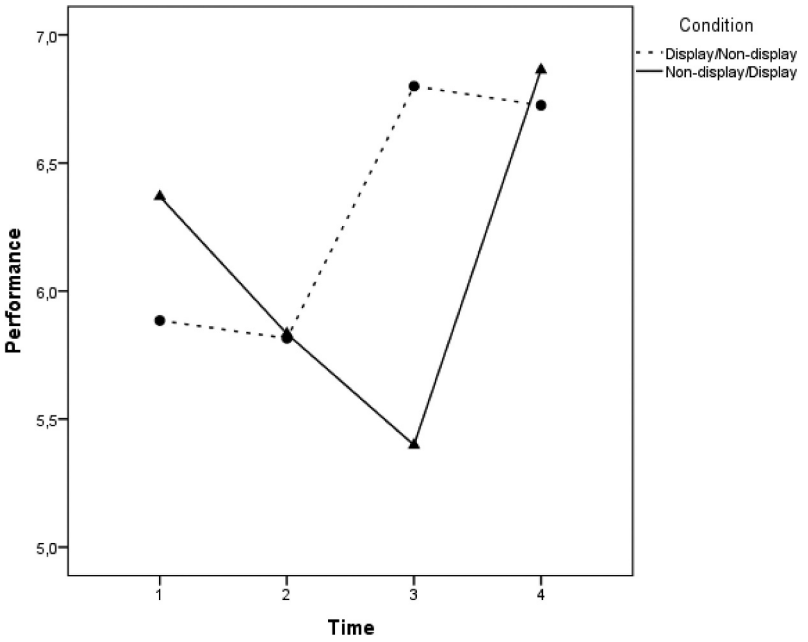

Figure 1. Comparison of the conditions for academic performance.

display condition between T1 and T3 ($p < .001$), T1 and T4 ($p = .002$), T2 and T3 ($p < .001$), and T2 and T4 ($p = .001$). That is, participants in the Display/Non-display condition had a lower performance in the graded times (T1 and T2) than in the non-graded times (T3 and T4). Regarding the Non-display/Display condition, we also found a result aligned with hypothesis H1b, a significant decrease in mean performance between T1 and T3 ($p < .001$). However, there was an abrupt increase in the performance of this condition in the fourth time of measurement that implied significant increases between T2 and T4 ($p = .001$) and T3 and T4 ($p < .001$), which does not align with that hypothesis.

In terms of the comparison of simple effects among conditions, at T3 the difference was significant, $p < .001$. This is relevant because the increase in performance in the Display/Non-display condition occurred between T2 and T3, when grades were removed. This result also aligns with hypotheses H1a and H1b.

RQ2. How does displaying versus non-displaying grades impact students' emotions while keeping feedback comments?

As mentioned earlier, positive emotions were the sum of enjoyment, pride, hope, and enthusiasm scores. Negative emotions were the sum of anger, boredom, anxiety, shame, hopelessness and disappointment scores. The descriptive statistics for each individual emotion can be found in Appendix C.

Table 4. Average and standard deviation for positive emotions.

Condition	Time 1	Time 2	Time 3	Time 4
Display/Non-display	1.79 (0.85)	1.52 (1.06)	1.75 (1.00)	1.92 (0.79)
Non-display/Display	1.78 (0.96)	1.71 (0.96)	1.29 (0.89)	1.96 (1.12)

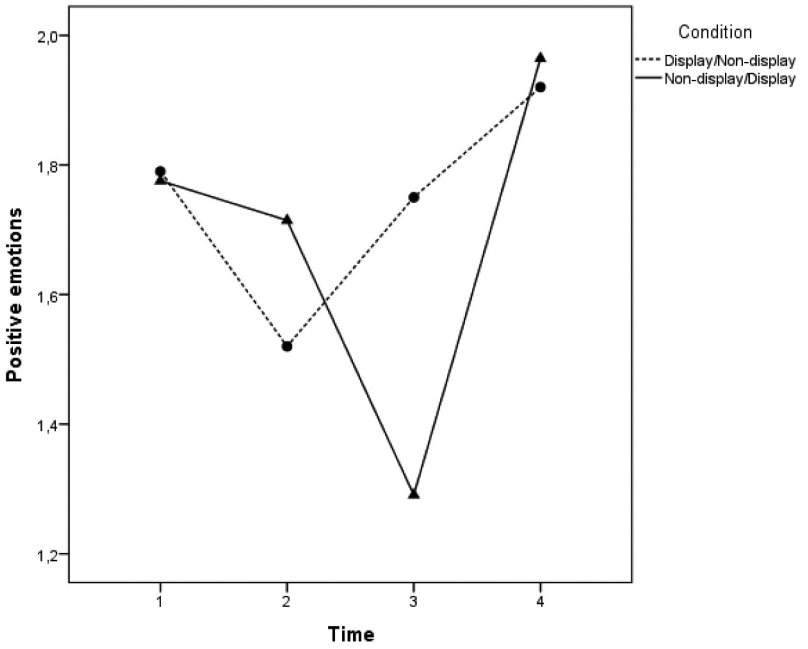

Figure 2. Comparison of the conditions for positive emotions.

Table 5. Average and standard deviation for negative emotions.

Condition	Time 1	Time 2	Time 3	Time 4
Display/Non-display	0.76 (0.72)	0.74 (0.87)	0.55 (0.73)	0.69 (0.84)
Non-display/Display	0.67 (0.53)	0.60 (0.56)	0.85 (0.69)	0.57 (0.47)

Positive emotions

We conducted a mixed-design ANOVA, using condition and time as factors and positive emotions score as dependent variable. Table 4 shows mean and standard deviations for positive emotions. Figure 2 shows the mean plot to help interpretation. We found a significant interaction effect between factors, $F(3, 291) = 2.92$, $p = .034$, $\eta^2 = .029$ (see Figure 2). Effect of time (irrelevant, as there is an interaction effect), $F(3, 291) = 5.11$, $p = .002$, $\eta^2 = .050$; and no effect of condition, $F(1, 97) = 0.20$, $p = .653$ $\eta^2 = .002$.

The interaction effect includes the significant difference between conditions at T3, $p = .018$. It also includes the significant decrease, for the Non-display/Display condition, in the mean scores between T1 and T3, $p = .012$; and an increase between T3 and T4, $p < .001$. Our interpretation is that H2a and H2b can be partially maintained, especially considering what happened in T3 and T2. We hypothesize that, at T1, the effect might have taken some time to show and at T4 the Non-display/Display condition might have bounced back as we will explain in the discussion.

Negative emotions

We conducted a mixed-design ANOVA, using condition and time as factors and negative emotions score as dependent variable, Table 5 shows average and standard deviations. There was an interaction effect between factors, $F(3, 291) = 4.48$, $p = .004$, $\eta^2 = .044$ (see Figure 3). No effect of time, $F(3, 291) = 0.55$, $p = .649$; and no effect of condition, $F(1, 97) = 0.01$, $p = .935$.

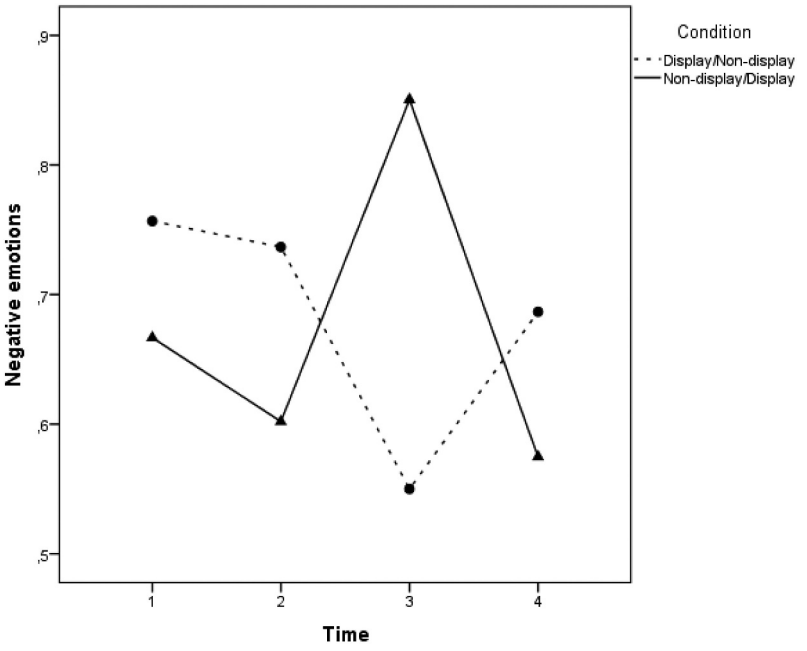

Figure 3. Comparison of the conditions for negative emotions.

The interaction effect includes the significant difference between conditions at T3, $p = .039$, consistent with the hypothesis H2c and H2d. Specifically, the Non-display/Display condition exhibited a higher average level of negative emotions compared to the Display/Non-display condition. However, there was also an unexpected significant decrease in the mean scores between T3 and T4 specifically for the Non-display/Display condition. Our interpretation is that H2c and H2d can be partially maintained, especially considering what happened in T3 and, to a much lesser extent, showed in the tendencies found at T1 and T2 but these were not significant. Regarding T4 both conditions seemed to bounce back but beyond the expected, as they returned to their T1 and T2 levels.

Discussion

Our aim was to explore the effects of displaying or not displaying grades on academic performance and emotions while maintaining written feedback comments. Throughout the study, students were aware that their work was being graded at all times. We explored two research questions with our 99 secondary education students sample. Crucially, we checked the absence of differences in four variables to assess the comparability of the two conditions: goal orientation, personality traits, receptivity to feedback, and academic performance in a prior task (i.e., baseline). We did not find significant differences between the conditions.

In RQ1, we explored the effects of displaying grades on students' academic performance. For the most part, our results showed that the display had an impact on students' performance as can be observed in the significant differences at Times 1 and 3, showing that when grades were non-displayed the students performed higher. This is what we hypothesized based on the effects found in previous research (e.g. Koenka et al., 2019; Lipnevich et al., 2009a). However, this effect dissipated in subsequent performances, as observed at Times 2 and 4, suggesting that the initial negative impact of displaying grades was transient, with performance returning to baseline levels once the display of grades became normalized. This study stands out for its unique intra-individual design, a distinction from previous research in the field, offering no direct comparisons. We propose that the display of

grades coming from a phase of non-display of grades may initially induce anxiety in students, hindering optimal performance and shifting their focus toward performance goals. However, this effect seems to diminish after the initial grade display experience. Our research introduces an innovative approach by examining intra-individual changes through a longitudinal design, addressing a significant gap highlighted by Goetz et al. (2018) and contrasting with the primarily inter-individual and cross-sectional studies included in the meta-analysis by Koenka et al. (2019).

In RQ2 we explored the effects of the display of grades on emotions. The results show that, in general, the display of grades decreased positive emotions and increased negative emotions, which aligns with previous research in which we based our hypotheses (Goetz et al., 2018; Lipnevich et al., 2009). Interestingly, the differences in the conditions were more remarkable at the "main moment of disruption" (Time 3) when both conditions experienced a change in the type of display they had been receiving up to that point. In Time 1 only one of the conditions changed (i.e., Non-display/Display) as the other remained the same from the baseline (i.e., grades were displayed). A surprising finding was the increase in positive emotions and decrease in negative emotions at Time 4 in the condition that was displayed grades at that time. We theorize this could be attributed to a "bounce back" effect related to performance in this condition: it is plausible that the changes at Time 4 resulted from students acclimatizing to receiving grades again, reverting to their typical baseline emotional state, which matched that of the students in the other condition. Regrettably, there is a lack of prior research on intra-individual changes through a longitudinal design for direct comparison (Goetz et al., 2018).

In light of the Control-Value Theory (CVT) of emotions, our findings offer substantive insights into the emotional dynamics of displaying grades as feedback. CVT posits that students' emotions in academic settings are significantly influenced by their perceptions of control over their learning outcomes and the value they assign to these outcomes (Pekrun et al., 2006). By integrating an intercorrelations table between performance and emotions into our results, we have directly linked students' emotional responses to their academic performance, thereby operationalizing CVT's premises in our empirical analysis.

Our analysis revealed that the display or non-display of grades affects students' emotions in a manner consistent with CVT (Pekrun et al., 2006). Initially, the display of grades was associated with increased negative emotions, which can be interpreted as a decrease in perceived control or value among students. This aligns with CVT's suggestion that negative emotions arise when students perceive a lack of control over their success or when they deem the task as having low value (Pekrun et al., 2002; Pekrun & Bühner, 2014).

However, as students adapted to the display of grades, the negative emotional impact diminished, suggesting a restoration of perceived control and value over their learning tasks. This dynamic reflects CVT's proposition that perceptions of control and value are malleable and can evolve in response to changes in the learning environment (Pekrun et al., 2007). The "bounce back" effect observed at Time 4 supports the idea that students' emotional responses stabilize as they acclimatize to the grading practice, potentially regaining a sense of control and reevaluating the value of their tasks (Hattie & Timperley, 2007).

The increased positive emotions and decreased negative emotions at Time May 4, also indicate a shift in students' appraisals. As students become accustomed to the feedback structure, they may develop enhanced coping mechanisms, leading to improved control perceptions and more adaptive emotional responses (Lipnevich et al., 2016). This suggests that, over time, students might learn to manage the emotional impact of grades better, aligning with CVT's emphasis on the adaptability of control and value appraisals in academic settings.

Our findings resonate with the conceptual framework proposed by Guskey (2019) and Guskey and Brookhart (2019), who underscored the multifaceted role of grades in shaping academic outcomes. Similar to our observation of the transient effects of the display of grades on student performance, their work suggests that the impact of grades may evolve over time as students adapt to assessment norms and expectations, aligning with our longitudinal insights. Furthermore, the nuanced improvement in student performance upon "grade removal" (i.e., non-display of grades) parallels the

discussions by Sadler (2005) and Finney, Perkins, et al. (2020, 2020), who argue for a more discerning application of grading practices to foster enhanced academic engagement and mitigate negative emotional responses. By extending these dialogs through our longitudinal analysis, our study underscores the dynamic interplay between grading, student motivation, and emotional well-being, highlighting the complexity of grading's impact on educational outcomes. The field of grading still requires the adoption of more sophisticated designs like the one used here, i.e., intraindividual longitudinal design-utilized in this study to fully grasp the nuanced effects of grades (Guskey, 2019).

Our study also underscored a critical distinction between the conventional dichotomy of high vs low stakes in educational settings and the nuanced dynamics inherent in grading practices. While grades undeniably introduce a high-stakes element (Guskey & Brookhart, 2019), our findings suggest that the absence of grades might not automatically equate to a low-stakes scenario. Indeed, feedback comments, even in the absence of grades, can introduce their own form of high stakes (e.g., when the feedback comments are perceived as harsh, have the potential to jeopardize a student's standing among peers). This observation is crucial for understanding the multifaceted nature of stakes in educational assessments, challenging the binary classification and inviting a more nuanced interpretation of what constitutes high stakes for students. By highlighting that both grades display vs non-display can carry significant weight, our study aligns with the broader discourse on assessment practices (Finney, Perkins, et al., 2020; Finney, Satkus, et al., 2020), suggesting that the emotional and motivational impacts of feedback are not solely contingent on the presence of grades but also on the content and delivery of feedback (Panadero & Lipnevich, 2022). This complexity underscores the need for educators to carefully consider how feedback is framed and delivered, recognizing that the stakes of educational assessments extend beyond the simple allocation of grades.

Limitations

This study has the following limitations. First, emotion are self-reported data. However, it made sense to evaluate emotions through questionnaires because we are aiming to understand the students' own perceptions of emotions and the academic field has a long and validated tradition of using self-report (Pekrun & Bühner, 2014). Second, our emotions scale was unbalanced with 4 items representing the positive valence scale, and 6 items representing the negative valence scale. As mentioned earlier, this was intentional as the negative emotions are more complex in educational settings (see original AEQ, Pekrun et al., 2005). To help readers understand this in more detailed we have provided the descriptive statistics for each emotion independently as an appendix (Appendix C). Third, we did not measure emotions before the intervention which impedes us from extracting stronger conclusions about the directionality in time 1. Fourth, while our sample size was somewhat constrained, it was sufficient to identify a realistic population effect size, as detailed in the Sample section. However, this limitation suggests the need for caution when generalizing our findings to broader populations. Fifth, while we instructed the teachers to ensure they treated both of their classroom groups equally except for grades manipulation, we did not control for it by assisting to their classrooms. Both teachers were well seasoned, over 15 years of experience, and we opted to keep their instructional context as natural for the teachers as possible. Sixth, our research design does not present the full spectrum as we did not have an all-time grade display condition and a total control condition without grades or feedback comments. Seventh, the submission methods differed between grades due to established teacher practices: 9th grade participants submitted work electronically, while 10th grade participants submitted handwritten assignments; however, classroom groups were evenly distributed across conditions to account for this potential confound.

And eight, a small percentage of the students' final work did not reach us: 9th grade group A first phase, 9th grade group B first phase, 10th grade group B second phase. Nevertheless, the teachers followed the same schema and performed the activities as requested. And we have the grades given by

the teachers that, as shown earlier, seemed to be quite reliable in comparison with an external rater, i.e., first author.

Implications

In terms of research implications, our study underscores the necessity for more longitudinal investigations to further explore intra-individual differences in educational assessment, particularly concerning the effects of grades. The prevailing body of literature predominantly comprises cross-sectional studies where experimental conditions either involve the allocation of grades or do not. Our design, echoing the insights of Brookhart et al. (2016) and Guskey (2019), reveals the complexities surrounding the effects of grading, suggesting that our research methodologies should reflect this complexity.

Regarding implications for teaching, our findings indicate that consistency in grading practices may be as critical as the practices themselves. The initial display of grades appears to exert negative effects; thus, educators who must implement grading are advised to mitigate these effects by thoroughly preparing students for the receipt of grades, such as by explicitly linking grades to learning outcomes (e.g., Guskey & Brookhart, 2019).

Conclusion

Our study explored the effects of displaying versus not displaying grades while delivering feedback comments using a novel longitudinal design, investigating both intra-individual and interpersonal effects. The main conclusion is that initially, displaying grades had negative effects on performance and emotions, but these effects stabilized in subsequent performances. This finding highlights the importance of considering how grades are displayed in educational settings and underscores the need for future research using more complex designs to disentangle the nuances of grading in students' learning.

Acknowledgments

We would like to thank our two participant teachers Paloma de Oñate and Antonio Roldán from Escuela IDEO (Madrid). Also thanks to Maria Valdivieso for helping with data collection. Additionally thanks to Anastasiya Lipnevich for participating in the planning of the research design and data collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the UAM - Banco Santander [2017/EEUU/12].

ORCID

Ernesto Panadero http://orcid.org/0000-0003-0859-3616
Iván Sánchez-Iglesias http://orcid.org/0000-0001-5419-7687

References

Alonso-Tapia, J., Nieto, C., Merino-Tejedor, E., Huertas, J. A., & Ruiz, M. (2018). Assessment of learning goals in university students from the perspective of 'person-situation interaction': The situated goals questionnaire (SGQ-U). *Estudios de Psicología*, *39*(1), 20–57. https://doi.org/10.1080/02109395.2017.1412707

Attali, Y. (2016). Effort in low-stakes assessments: What does it take to perform as well as in a high-stakes setting? *Educational and Psychological Measurement*, *76*(6), 1045–1058. https://doi.org/10.1177/0013164416634789

Barbaranelli, C., Caprara, G. V., & Rabasca, A. (1998). *Manuale del BFQC. Big Five Questionnaire Children*. O.S. Organizzaaioni Speciali-Firenze.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441–462. https://doi.org/10.1007/BF03173192

Bieleke, M., Gogol, K., Goetz, T., Daniels, L., & Pekrun, R. (2021). The AEQ-S: A short version of the achievement emotions questionnaire. *Contemporary Educational Psychology*, *65*, 101940. https://doi.org/10.1016/j.cedpsych.2020.101940

Bowers, A. J. (2019). Report card grades and educational outcomes. In T. R. Guskey & S. M. Brookhart (Eds.), *What we know about grading* (pp. 32–56). ASCD.

Brookhart, S. M., & Guskey, T. R. (2019). Reliability in grading and grading scales. In T. R. Guskey & S. M. Brookhart (Eds.), *What we know about grading*. ASCD. 13–31.

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. *Review of Educational Research*, *86*(4), 803–848. https://doi.org/10.3102/0034654316672069

Butler, R., & Nisan, M. (1986). Effects of no feedback, task-related comments and grades on intrinsic motivation and performance. *Journal of Educational Psychology*, *78*(3), 210–216. https://doi.org/10.1037/0022-0663.78.3.210

Carrasco Ortiz, M. A., Holgado Tello, F. P., & Del Barrio Gandara, M. V. (2005). Dimensionalidad del cuestionario de los cinco grandes (BFQ-N) en población infantil española. *Psicothema*, *17*(Número 2), 286–291. https://reunido.uniovi.es/index.php/PST/article/view/8270

Elliot, A. J. (2005). A conceptual history of achievement goal construct. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). Guilford.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Finney, S. J., Perkins, B. A., & Satkus, P. (2020). Examining the simultaneous change in emotions during a test: Relations with expended effort and test performance. *International Journal of Testing*, *20*(4), 274–298. https://doi.org/10.1080/15305058.2020.1786834

Finney, S. J., Satkus, P., & Perkins, B. A. (2020). The effect of perceived test importance and examinee emotions on expended effort during a low-stakes test: A longitudinal panel Model. *Educational Assessment*, *25*(2), 159–177. https://doi.org/10.1080/10627197.2020.1756254

Goetz, T., Lipnevich, A. A., Krannich, M., & Gogol, K. (2018). Performance feedback and emotions. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 554–574). Cambridge University Press.

Guskey, T. R. (2019). Grades versus comments: Research on student feedback. *Phi Delta Kappan*, *101*(3), 42–47. https://doi.org/10.1177/0031721719885920

Guskey, T. R., & Brookhart, S. M. (2019). *What we know about grading*. ASCD.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

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

Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2019). A meta-analysis on the impact of grades and comments on academic motivation and achievement: A case for written feedback. *Educational Psychology*, *41*(7), 1–22. https://doi.org/10.1080/01443410.2019.1659939

Lipnevich, A. A., Berg, D. A. G., & Smith, J. K. (2016). Toward a model of student response to feedback. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 169–185). Routledge.

Lipnevich, A. A., Gjicali, K., Asil, M., & Smith, J. K. (2021b). Development of a measure of receptivity to instructional feedback and examination of its links to personality. *Personality & Individual Differences*, *169*, 110086. https://doi.org/10.1016/j.paid.2020.110086

Lipnevich, A. A., Murano, D., Krannich, M., & Goetz, T. (2021a). Should I grade or should I comment: Links among feedback, emotions, and performance. *Learning & Individual Differences*, *89*, 102020. https://doi.org/10.1016/j.lindif.2021.102020

Lipnevich, A. A., & Smith, J. K. (2009a). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology Applied*, *15*(4), 319–333. https://doi.org/10.1037/a0017841

Lipnevich, A. A., & Smith, J. K. (2009b). "I really need feedback to learn: " students' perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability*, *21*(4), 347. https://doi.org/10.1007/s11092-009-9082-2

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback typologies and models: Towards an integrative model of feedback elements. *Educational Research Review*, *100416*, 100416. https://doi.org/10.1016/j.edurev.2021.100416

Pekrun, R., & Bühner, M. (2014). Self-report measures of academic emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 561–579). Routledge/Taylor & Francis Group.

Pekrun, R., Elliot, A. J., & Maier, M. A. (2006). Achievement goals and discrete achievement emotions: A theoretical model and prospective test. *Journal of Educational Psychology*, *98*(3), 583–597. https://doi.org/10.1037/0022-0663.98.3.583

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ). *Contemporary Educational Psychology*, *36*(1), 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002

Pekrun, R., Goetz, T., & Perry, R. P. (2005). Achievement Emotions Questionnaire (AEQ). *User Manual*.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, *37*(2), 91–105.

Pekrun, R., & Linnenbrink-Garcia, L. (Eds.). (2014). *International handbook of emotions in education*. Routledge.

Pekrun, R., Muis, K. R., Frenzel, A. C., & Goetz, T. (2018). *Emotions at school*. Routledge.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, *30*(2), 175–194. https://doi.org/10.1080/0260293042000264262

Satkus, P., & Finney, S. J. (2021). Antecedents of examinee motivation during low-stakes tests: Examining the variability in effects across different research designs. *Assessment & Evaluation in Higher Education*, *46*(7), 1065–1079. https://doi.org/10.1080/02602938.2020.1846680

Van der Kleij, F. M., & Lipnevich, A. A. (2020). Student perceptions of assessment feedback: A critical scoping review and call for research. *Educational Assessment, Evaluation and Accountability*, *33*(2), 345–373. https://doi.org/10.1007/s11092-020-09331-x

Vogl, E., & Pekrun, R. (2016). Emotions that matter to achievement: Student feelings about assessment. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 111–128). Routledge.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality & Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227–242. https://doi.org/10.1207/s15324818ame0803_3

Zeidner, M., & Matthews, G. (2012). Personality. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, & G. M. Sinatra (Eds.), *APA educational psychology handbook, vol, 2: Individual differences and cultural and contextual factors* (pp. 111–137). American Psychological Association. https://doi.org/10.1037/13274-005

Appendix A

Emotions questionnaire

Next, you will find different feelings and emotions. Read each item thinking about how you feel in relation to the (feedback) comments your teacher has just given you. Mark the appropriate response next to the word, using the following scale for your answers:

1	2	3	4	5	
Nothing or very little	A little	Moderately	Quite	Extremely	5
Angry	1	2	3	4	5
Hopeful	1	2	3	4	5
Proud	1	2	3	4	5
Anxious	1	2	3	4	5
Ashamed	1	2	3	4	5
Disappointed	1	2	3	4	5
Enthusiastic	1	2	3	4	5
Hopeless	1	2	3	4	5
Bored	1	2	3	4	5

Perceptions of value questionnaire

We are trying to understand what types of thoughts and emotions you experience when doing schoolwork. Therefore, we request that you indicate to what extent thoughts similar to the ones below occur to you, while you are engaged in academic assignments. Use the following scale:

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

1	Doing well on this assignment was important to me, no matter what grade I earn	1 2 3 4 5
2	It is important for me to get a good grade on this assignment	1 2 3 4 5
3	I usually get good grades on similar assignments	1 2 3 4 5
4	I have always done well on assignments like this one	1 2 3 4 5
5	The feedback I received is helpful	1 2 3 4 5
6	The feedback I received will help me to improve my assignment	1 2 3 4 5
7	The feedback I received is accurate	1 2 3 4 5

Appendix B

Samples of feedback comments given to the participants.

(1) You are improving, but there is still a need to delve deeper into the part about contextualization.
(2) The classification and the main ideas are adequate, but you haven't contextualized it.
(3) It wasn't asked for this.
(4) All kinds of data are missing.
(5) Some part of this reasoning is not entirely clear. For example, why does using materials from nature to make instruments make them more developed?
(6) In what? Why is that a sign of underdevelopment? You need to explain these kinds of things.

There are hundreds of examples of corrections of typos, grammar mistakes, etc., that we do not represent here due to their simplicity.

Appendix C

Emotions scale. Descriptive statistics by time and condition for each individual emotion

Condition	Emotion	Time 1	Time 2	Time 3	Time 4
In/Out	Enjoyment	2.26 (0.99)	1.86 (1.20)	2.18 (1.10)	2.44 (0.95)
n = 50	Pride	1.82 (1.08)	1.50 (1.40)	1.92 (1.18)	2.02 (1.15)
	Hope	1.72 (1.01)	1.60 (1.14)	1.52 (1.28)	1.74 (0.96)
	Enthusiasm	1.36 (1.12)	1.12 (1.22)	1.38 (1.18)	1.48 (1.20)
	Anger	0.64 (0.85)	0.76 (1.15)	0.48 (0.95)	0.60 (0.95)
	Boredom	0.68 (0.84)	0.92 (1.19)	0.86 (1.01)	0.84 (1.00)
	Anxiety	0.78 (0.91)	0.50 (0.81)	0.58 (0.99)	0.68 (1.00)
	Shame	0.76 (1.02)	0.60 (1.05)	0.38 (0.85)	0.56 (0.93)
	Hopelessness	0.64 (0.88)	0.74 (1.16)	0.44 (0.84)	0.74 (1.03)
	Disappointment	1.04 (1.07)	0.90 (1.22)	0.56 (0.97)	0.70 (1.09)
Out/In	Enjoyment	2.18 (1.11)	1.98 (1.11)	1.49 (1.21)	2.27 (1.30)
n = 49	Pride	1.67 (1.21)	1.69 (1.21)	1.10 (1.19)	2.04 (1.43)
	Hope	1.92 (1.29)	1.8 (1.15)	1.55 (1.17)	1.78 (1.23)
	Enthusiasm	1.33 (1.25)	1.39 (1.13)	1.02 (1.01)	1.78 (1.31)
	Anger	0.45 (0.79)	0.39 (0.86)	0.88 (1.07)	0.37 (0.73)
	Boredom	1.12 (1.25)	0.92 (1.30)	1.16 (1.34)	1.35 (1.28)
	Anxiety	0.82 (1.03)	0.80 (1.04)	0.71 (1.02)	0.65 (0.99)
	Shame	0.43 (0.76)	0.45 (0.87)	0.67 (1.07)	0.33 (0.75)
	Hopelessness	0.51 (0.79)	0.43 (0.87)	0.71 (0.96)	0.22 (0.47)
	Disappointment	0.67 (0.85)	0.63 (0.88)	0.96 (1.27)	0.53 (0.98)