frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 15 April 2024
DOI 10.3389/feduc.2024.1386016

OPEN ACCESS

EDITED BY
Kwok Kuen Tsang,
The Education University of Hong Kong, Hong Kong SAR, China

REVIEWED BY
Nicolas Hübner,
University of Tübingen, Germany
Grzegorz Szumski,
University of Warsaw, Poland

*CORRESPONDENCE
Meike Bonefeld
✉ meike.bonefeld@ezw.uni-freiburg.de

RECEIVED 14 February 2024
ACCEPTED 02 April 2024
PUBLISHED 15 April 2024

CITATION
Peter S, Karst K and Bonefeld M (2024) Objective assessment criteria reduce the influence of judgmental bias on grading. *Front. Educ.* 9:1386016.
doi: 10.3389/feduc.2024.1386016

Objective assessment criteria reduce the influence of judgmental bias on grading

Sophia Peter[1], Karina Karst[2] and Meike Bonefeld[3]*

[1]Institute for Psychosocial Prevention, Heidelberg University Hospital, Heidelberg, Germany, [2]School of Social Sciences, University of Mannheim, Mannheim, Germany, [3]Department of Educational Science, University of Freiburg, Freiburg, Germany

Past research has indicated that students with a migration background are graded worse than those without a migration background, despite them showing the same level of academic performance. Negative implicit associations of teachers associated with these student characteristics could explain these findings. Objective assessment criteria, such as error tables, provide user-independent rules for the interpretation of results and could therefore help to ensure that the influence of student characteristics on assessment is reduced. To test this hypothesis, 157 pre-service teachers assessed a dictation. Two aspects were varied: the presentation of an error table for assessment and the name of the student who had written the dictation (with vs. without a supposed Turkish migration background). An implicit association test measured implicit associations of the pre-service teachers toward the performance of Turkish and German people. When no error table was used and the pre-service teachers had negative implicit associations toward the performance of people with a Turkish migration background, they graded students with a migration background worse than students without a migration background. No grading disparities were found when the error table was used. To reduce judgmental bias, the use of objective assessment criteria can therefore be recommended.

KEYWORDS

objective assessment criteria, judgmental bias, implicit associations, migration background, teacher bias

1 Introduction

Personal features such as social origin and migration background play a decisive role in educational success (OECD, 2019). Prior research showed that educational careers as well as the performance in school differs between students with and without a migration background (Kristen et al., 2008; Henschel et al., 2022). For example, students with a migration background are more likely to drop out of school than their fellow students without a migration background (González-Rodríguez et al., 2019). Further, Turkish people form the largest minority group in Germany (DESTATIS, 2019; Henschel et al., 2022). Especially in Germany, where this investigation takes place, Turkish students represent the majority of students in lower school tracks and the minority in higher school tracks (Henschel et al., 2022). Turkish students perform worse in standardized tests compared to students without a migration background and face negative stereotyping, especially concerning performance (Gebhardt et al., 2013; Froehlich et al., 2016). Given these findings, it is particularly important to pay special attention to Turkish students when examining educational disparities in Germany.

For years now, research and policymakers have been discussing the causes of differences in performance depending on students' backgrounds (Bundesregierung, 2023). Even though many factors contribute to some student groups being disadvantaged, the role of teachers has also been discussed (Jussim and Harber, 2005; McKown and Weinstein, 2008). Within this debate, it is of crucial importance that reported differences between students with a migration background and those without a migration background are observable, even after taking differences in standardized tests into account (Bonefeld et al., 2017; Triventi, 2020). Especially judgments (e.g., grading) that are not guided by a set of clear rules seem to be influenced not only by performance but also by student characteristics such as migration background. The current work investigates potential reasons for differences in grading dependent on the migration background of students beyond performance, on the one hand, and examines first approaches to prevent these differences, on the other hand.

1.1 Teachers' judgments and their formation

Teachers' judgments, for example in the form of grades can in both positive and negative ways, to shape students' educational paths. For example, grades continue to be decisive for admission to universities or for obtaining a job and can therefore fulfill a gatekeeping function. In addition to deciding on the exercises that students complete and the support they receive, teachers also assign performance evaluations and grades. The awarding of grades faces various challenges that can influence both the assessment process and the outcomes (Trapmann et al., 2007). A fundamental issue is that grades are often considered uniform standards for complex performances, which can lead to simplification and distortion of actual performance. Additionally, there are often no clear guidelines on which level of performance corresponds to which grade. This leads to significant variance among teachers in their grading practices (Südkamp et al., 2012; Drüke-Noe, 2014). School grades therefore often show a lack of comparability, e.g., between individual schools and their individual grading practices (Hübner et al., 2024). Moreover, personal biases, subjective judgments, and unconscious biases on the part of teachers can influence the assessment process, resulting in unfair grading and unjust outcomes. In this performance assessment in particular, other factors besides performance also play a role. These are, for example, student characteristics like personality traits, migration background or gender (Glock, 2016; Bonefeld et al., 2017; Holder and Kessels, 2017; Bonefeld and Dickhäuser, 2018; Hübner et al., 2024).

Previous research has revealed a connection between migration background and the judgments of teachers. For instance, data from experimental studies has shown that pre-and in-service teachers graded students with a migration background worse compared to students without a migration background, even when their performance was comparable or exactly the same (Glock et al., 2015). This bias holds for language subjects as well as mathematics.

1.1.1 Formation of judgments and role of associations

One angle to explain the influence of student characteristics on teachers' judgments is the formation of judgments. Dual process models propose that individuals have the capacity to employ two distinct strategies when forming evaluations of others (e.g., Fiske and Neuberg, 1990). One strategy is known as the category-based or heuristic approach, which is assumed to operate in a mostly automatic manner. This involves initiating information processing by focusing on the social categories to which the subject being evaluated belongs (e.g., Fiske and Neuberg, 1990). These social categories are triggered by relevant information and are utilized during the judgment formation process (Macrae and Bodenhausen, 2000). Various types of information can serve as cues in this process, including easily accessible categories like perceived gender (Swim et al., 1989; Hoffman and Hurst, 1990) or ethnicity (Devine, 1989; Hewstone et al., 1991). By employing social categories to process personal information, individuals can quickly form judgments while conserving cognitive resources (Macrae et al., 1994). However, this heuristic approach often leads to biased judgments. The heuristic process is particularly evident when making judgments about individuals who exhibit information that aligns with stereotypes and therefore the activated categories. For instance, teachers tend to exhibit bias when evaluating students who possess stereotype-consistent characteristics, such as a migration background and lower academic performance (Pit-ten Cate and Glock, 2018). Another approach involves a more deliberate strategy that is integrative in nature (e.g., Fiske and Neuberg, 1990) or rule-based (Smith and DeCoster, 1999; Ferreira et al., 2006). In this scenario, judges do not rely solely on a limited set of category-related details; instead, they examine all pertinent information to inform their judgments. Integrating all relevant information facilitates more accurate judgments that are less influenced by social categories like immigrant background (Fiske et al., 1999).

1.2 Implicit associations and judgmental biases

The influence of social categories on judgments in the judgment process can be explained, among other things, by stereotypes and implicit associations of the person (Greenwald and Banaji, 1995). These implicit associations of objects (e.g., members of a group) correspond to positive or negative evaluations (Fazio, 2007). Past research has shown that people with a migration background in Germany and people with a migration background in general are associated less strongly with performance and success-related attributes by other Germans compared to Germans without a migration background (Kahraman and Knoblich, 2000; Froehlich et al., 2016;). This finding holds as well for the associations of teachers, who show negative associations concerning the performance of students with migration background in Germany (Bonefeld and Dickhäuser, 2018; Bonefeld et al., 2020). The impact of these associations is evident in the school context: Past research suggests that especially biases against students with a migration background, depend on the type of association a teacher holds (van den Bergh et al., 2010; Peterson et al., 2016; Bonefeld and Dickhäuser, 2018). Furthermore, it has been shown that teachers rate students with a migration background as less capable than students without a migration background (Glock and Krolak-Schwerdt, 2013). Such perceptions and associations can influence information processing and guide attention (Roskos-Ewoldsen and Fazio, 1992; Bonefeld et al., 2020) and in turn affect teachers' judgments (Dee, 2005; Parks and Kennedy, 2007; Wiggan, 2007). Judgmental biases could for example result from specific

perceptions of teachers about the future performance level of students from a specific group (Jussim and Harber, 2005; Lorenz, 2018). This is particularly important for the assessment of students with a migration background and could be one explanation for the above mentioned biases in the grading process (Bonefeld et al., 2017; Bonefeld and Dickhäuser, 2018): If teachers hold negative associations about the performance potential of students with a migration background in general, they could make judgments based on additional information about the student related to characteristics that do not exist or, at least, that cannot be observed and this in turn leads to biases in grading depending on student characteristics like the migration background (Taylor, 1981; Walton and Spencer, 2009; van Ewijk, 2011).

Whereas explicit associations are conscious evaluations of an object that require cognitive effort and control, implicit associations are often unconscious evaluations of an object that are triggered by the mere presence of the object (Fazio, 2001; Gawronski and Bodenhausen, 2006). Given this nature in the context of biased grading, implicit associations are particularly important because they lead to automatic, heuristic information processing (Gawronski and Bodenhausen, 2006; van den Bergh et al., 2010) and a person is often not even consciously aware of them which means that these associations cannot be controlled or corrected. Thereby they can impact teachers' perceptions of their students (Olson and Fazio, 2009) without teachers even knowing. Past research provides findings on these theoretical assumptions and it was demonstrated that implicit associations toward students from minorities predict teachers' non-verbal behavior and the general impression they form (Nosek et al., 2002; van den Bergh et al., 2010). Furthermore, previous research could show the influence of implicit associations on teachers' judgments like for example grading (Bonefeld and Dickhäuser, 2018). Thus, they may provide an explanation for how teacher judgments are influenced by student characteristics.

1.3 Reduction of the influence of judgmental bias

From the assumptions made so far, it can be concluded that (1) student characteristics can influence teachers' judgments (Glock et al., 2015), (2) this can be explained, e.g., by implicit associations related to these students' characteristics (e.g., Bonefeld and Dickhäuser, 2018) and (3) implicit associations are automatic, often unconscious and therefore harder to control (Gawronski and Bodenhausen, 2006). Therefore, it is crucial to determine how the influence of student characteristics, moderated by implicit associations, can be reduced. As implicit associations in particular are difficult to regulate due to their automatic nature, they must be countered by other means. The basic prerequisite for heuristic or systematic information processing is that the available material can be viewed and evaluated objectively (Petty and Cacioppo, 1983; Chaiken et al., 1989). It is argued that one should reduce the discretion in judgments by developing evidence-based guidelines (Spencer et al., 2016). In school, these are then, for example, evaluation criteria that reduce the room for discretion. This can be achieved by a clear standardization of requirements to prevent subjective components from being included in grading (Birkel and Tarnai, 2018). As mentioned above, previous research has shown that pre-service and in-service teachers' grading can be biased based on student characteristics such as migration background (Glock and Krolak-Schwerdt, 2013; Bonefeld et al., 2017; Bonefeld and Dickhäuser, 2018), while a simple true/false judgment, such as the detection of errors is not biased by these features (Bonefeld and Dickhäuser, 2018). While judgments regarding errors detected based on universally known rules (e.g., spelling errors) are comparably easy to form, the absence of clear assessment norms for grading might lead to biases. This could be because without clear standards and, consequently, higher ambiguity in judging, it is more difficult to judge systematically and heuristic components can take up more space in judgments, since one has to fill in missing information (Petty and Cacioppo, 1986; Chaiken et al., 1989). This finding from many social psychological experiments is transferable to performance assessment without clear standards or evaluation criteria.

In the case of unclear or incomplete information, which includes grading without a clear evaluation scheme, this information is usually categorized and interpreted using existing knowledge (Smith and DeCoster, 1999). This process can be compared with heuristic information processing (Chaiken et al., 1989). Non-observed details can be replaced by recourse to memory, and existing, automatically activated knowledge based on associations can influence the assessment (Smith and DeCoster, 1999). The use of objective assessment criteria can be helpful in this respect in three ways. One aspect is that their use reduces the influence of associations on behavior. According to dual process models, the context and social situation play a role in whether an association influences behavior (Fazio, 1986). So-called standards, i.e., the knowledge of what is appropriate in a particular situation or not, are decisive (Fazio, 1986). Objective assessment criteria, such as error tables, which are used to assign grades based on a specific number of errors, form an external standard as orientation. As a result, this external standard determines the grade rather than associations. Objective assessment criteria therefore constitute a guideline, so that teachers no longer need to draw on associations as a source of information in cases of uncertainty. When applied, error tables based on clear criteria provide the basis for an objective performance assessment. In addition, these objective assessment criteria provide user-independent rules for the interpretation of results and contribute to an independent assessment.

Uhlmann and Cohen (2005) also argued that biases in judgments take place, above all when there is ambiguity regarding the criteria to be applied. This ambiguity is reduced by the use of error tables. As an error table assigns a grade depending on the number of errors, grading is less ambiguous and should be performed using an error table. Associations can obviously still exist, but their influence on grading should be reduced and participants are less likely to draw on associations in the assessment process since they have clear criteria and rules for grading. Closely related to this is another advantage of such objective assessment criteria: the judges themselves do not have to be aware of their implicit associations toward different student groups. This is a very important aspect, because it is in the nature of implicit associations that they are activated automatically, are often unconscious, and are difficult to control (Gawronski and Bodenhausen, 2006).

To date, research on the application and use of objective assessment criteria to avoid biases in teachers' judgments dependent on student characteristics is still rare. One example in this field is the ASSET project (Assessing Students' English Texts), where an assessment tool based on didactic criteria was provided

to pre-service teachers to assess students' performance. The results suggested that irrelevant characteristics such as migration background did not influence performance assessment when using the tool (Jansen et al., 2019). Another study has shown that the objectivity of oral examinations could also be increased by using assessment criteria, which precisely define what performance and skills must be shown to what grade (Westhoff et al., 2002). The disadvantage of these studies is the lack of a control group, where no evaluation criteria are used. In an experimental study, teachers showed racial bias only when they used a vague grade-level evaluation scale to assess a writing sample. When teachers used a rubric, which defined the scoring domain of interest and provided more specific scale points - therefore providing clearer evaluation criteria - no racial bias was evident (Quinn, 2020).

To summarize: Social information processing models can be used to explain how errors of judgment occur and the role that person characteristics might play in this (e.g., Fiske and Neuberg, 1990). A way of objectifying judgment processes and reducing errors of judgment based on student characteristics and related subjective associations could be the specification of objective assessment criteria. These provide user-independent rules for the assessment of performance. Further, they might help the judge to focus on performance-relevant characteristics and reinforce more objective assessments (Jansen et al., 2019). In this work we aim to examine whether objective assessment criteria can reduce migration-related judgmental bias by limiting the influence of negative implicit associations.

1.4 The present study

In this experimental study we investigated judgmental bias regarding grading and how the influence of student characteristics on it can be reduced using objective assessment criteria. Therefore, we focused on judgments of pre-service teachers who assessed students' performance in a dictation. The judgment involved counting errors and giving a grade for the performance. We were interested to see whether grading depended on the students' supposed migration background and the underlying implicit associations of pre-service teachers about the performance of people with a migration background. Most importantly, we looked at the role of objective assessment criteria and their effects on grading. Therefore, all participants received the same dictation with an identical number of errors. We experimentally varied the perceived migration background of the student (with vs. without migration background) and the use of objective assessment criteria (no error table vs. error table). The following two hypotheses were formulated.

> *H1*: Negative implicit associations toward people with a migration background were expected to lead to worse grading of the student with a migration background if no error table was provided.
>
> *H2*: If an error table was provided, the influence of implicit associations on grading the student with migration background would be reduced.

Error counting is a rule-based process which should not be influenced by associations. Therefore, the hypotheses do not apply to the counted errors.

2 Materials and methods

2.1 Sample

Our sample consisted of 188 pre-service teachers (72.34% female, 1.06% diverse, 26.60% male). The participants mean age was 21.59 years ($SD = 2.94$). Out of them 16 majored in primary school education, 168 in secondary school education and 4 in special needs education. They had completed an average of 3.90 semesters ($SD = 2.94$) of their teacher training program and 65.43% of our sample completed a school teaching internship. In terms of prior experience 17.55% stated they had experiences in grading dictations and 60.63% studied the subject German.

The participants were recruited via personal contacts, flyers at schools and universities, e-mails, and social media. The participants either received study participation credits which they must collect during their studies at some universities in Germany, had the chance to win a 10 Euro Amazon voucher or choose to receive no reward.

2.2 Materials

2.2.1 Dictation

The participants assessed a dictation by a third grader used in a previous study by Bonefeld and Dickhäuser (2018). Dictations are a widely used method to detect spelling skills. Teachers read texts out loud, and students are instructed to write down the text they have heard correctly. The dictation had 30 errors, which corresponds to a low performance. In a pre-test, pre-service teachers counted an average of 26.28 ($SD = 5.51$) errors and awarded an average grade of 4.85 ($SD = 1.16$) (Bonefeld and Dickhäuser, 2018).

2.2.2 Objective assessment criteria

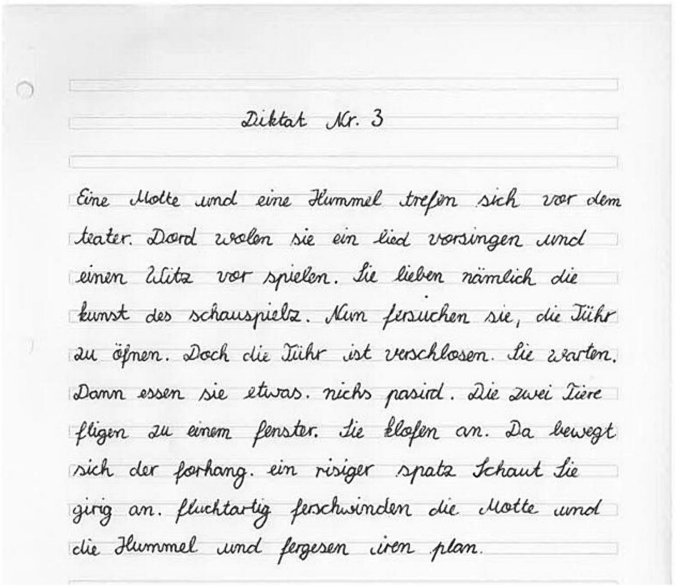
An error table to assess dictations in elementary schools operationalized objective assessment criteria. An error table assigns a specific grade to each number of errors, ranging from 1 (very good) to 6 (insufficient) (see Figure 1). The error table was created by school experts.

2.2.3 Migration background

We manipulated the migration background by changing students' names. The participants were briefly informed that the dictation was written either by Emre or Lukas: "The following dictation was written by Emre [Lukas]. He is a third-grade student and 8 years old." In a pre-test, the name Emre was strongly associated with a Turkish background and the name Lukas with a German background. In addition, both names were associated with a male name, and the differences regarding perceived intelligence and attractiveness were also relatively low (Bonefeld and Dickhäuser, 2018).

2.2.4 Implicit association test

An online-based implicit association test (IAT) using word and visual stimuli was performed to measure performance-related implicit associations toward Turkish and German people (see Figure 2; Bonefeld, 2019). Aim of an IAT is to test how strongly people associate certain attribute concepts with certain target concepts. During the test, participants are asked to rapidly sort words and/or images into categories according to predetermined criteria. In case of strong



Fehlertabelle

Note	Fehler
1	0
1-	1
2+	2
2	3
2-	4
3+	5
3	6
3-	7
4+	8
4	9-10
4-	11
5+	12
5	13-14
5-	15
6+	16
6	>16

FIGURE 1
Split-screen view of instruction, dictation, and provided error table. Imagine your school implemented a standardized error table in the teachers' conference. The table assigns a specific grade to each error. Please evaluate the dictation! German – English translation: Fehlertabelle – error table, Note – grade, Fehler – error, Diktat – dictation.

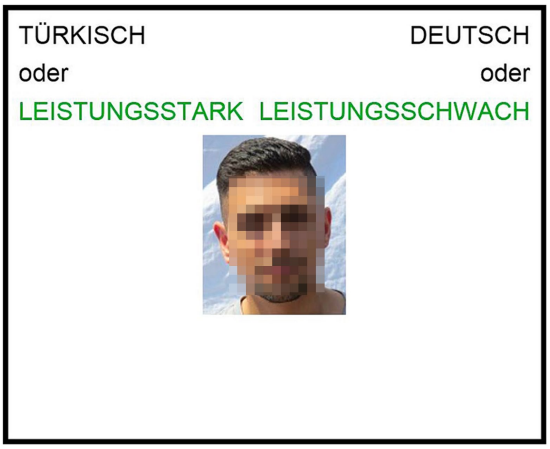


FIGURE 2
Split-screen view of the IAT (incompatible block). German – English translation: Türkisch – Turkish, Deutsch –German, leistungsstark – high performance, leistungsschwach – low performance.

associations between certain attribute concepts and target concepts, the reaction time should be faster if the same reaction (=pressing the same key) is necessary for the categorization of the corresponding attributes and targets. Based on the response time researchers can determine the strength of association between the target and attribute concepts (Greenwald et al., 1998; Nosek et al., 2005).

In our IAT the participants completed 12 practice trials (for each category), the participants assigned visual material (photographs of German- and Turkish-looking faces) to the right target concepts (German vs. Turkish) and assigned performance-related attributes (e.g., gifted, untalented) to the right attribute concepts (high performance vs. low performance). After that, they assigned the faces and attributes at the same time in another 24 practice trials and in 48 critical trials (measurement trials).

In the next step, the categorization was changed. The target concepts changed their position, whereas the attribute concepts stayed in the same position. Again, the participants reported their associations in 12 practice trials for each concept and sorted them afterwards in 24 trials together. This was followed by 48 critical trials. Therefore, we had a compatible block (German-looking faces paired with high performance adjectives) and an incompatible block (Turkish-looking faces paired with high performance adjectives). To reduce sequential effects, the order of the blocks varied randomly between the participants.

Participants with negative implicit associations toward people with a migration background showed faster response times when Turkish faces were linked with weak adjectives rather than with strong ones. A *D*-score indicates the strength of an implicit association. In our study, this was calculated based on the improved scoring algorithm by Nosek et al. (2005). In this study, a *D*-score below zero indicated more positive associations toward people with a migration background compared to people without a migration background. A *D*-score over zero indicated more negative implicit associations toward people with a migration background compared to people without a migration background.

2.3 Procedure and design

This study was based on a 2 (migration background: yes vs. no) × 2 (objective assessment criteria: yes vs. no) between-subjects experimental design. The participants were therefore randomly assigned to four experimental conditions [student without migration background and error table ($n_{total} = 43$), student without migration background and no error table ($n_{total} = 44$), student with migration

background and error table ($n_{total} = 45$), student with migration background and no error table ($n_{total} = 44$)]. Implicit associations were assessed as a continuous variable using the *D-score* (Nosek et al., 2005).

The participants were informed about the contents of the study and provided an informed consent in accordance with the guidelines of the German Research Foundation (DFG). They were informed that they were participating in a study investigating the effectiveness of dictation in assessing student performance and testing different rating scales and rater characteristics that may play a role in this process. Further, we informed them, that they therefore will assess a dictation and participate in a reaction time task. First, the participants filled in demographic information. Then, they were provided with a short overview of the following steps and information on the name, age, and class level of the student. Finally, they were given the dictation and, if applicable, the error table (Figure 1). Participants who received the error table were instructed as follows: "Imagine your school implemented a standardized error table in the teachers' conference. The table assigns a specific grade to each error number. Please evaluate the dictation." Otherwise, the instruction was as follows: "Please evaluate the dictation." Afterwards all participants were asked to count the errors and to grade the dictation ("How many errors did the dictation have?" and "What grade would you award the student for this dictation?"). To grade the dictation, the German grading system was used (ranging from 1 to 6 with 1 indicating the best performance and 6 the worst performance). This means high values of the variable grade represented a low performance assessment. The performance assessment was followed by a manipulation check. Participants were asked to specify the name and class level of the student and whether they thought the student had a migration background or not. As a last step, the IAT was completed. Participants were given the opportunity to provide their email address in order to learn more about the research project, the study background (with debriefing) and the results of the study in the follow-up.

2.4 Data analysis

To check whether the manipulation of the migration background worked, we looked at the names and the background assigned to the students by the participants. A total of 87.77% of the participants mentioned the correct or a suitable German or Turkish name or identified the students with a supposed migration background correctly. However, 9.04% of the participants did not remember a name and mentioned no migration background. We excluded six participants (3.19%) from the analysis because they either specified a name associated with another ethnic group (e.g., Tim instead of Emre), associated Emre with no migration background, or specified a name used in the study published by Bonefeld and Dickhäuser (2018). Second, we analyzed whether the participants who received the error table actually used it. This was the case for the majority, 89.01% of these participants. We therefore excluded 20 participants, who did not use the error table.

Some participants raised the suspicion of not having worked carefully because they counted very few errors. We decided to exclude participants who counted less than 10 errors (3 participants). In addition, we excluded two participants for whom no *D*-score measuring implicit associations could be calculated due to too many processing errors. This resulted in a total of 31 excluded participants.

With the final sample consisting of 157 participants, we calculated a hierarchical regression analysis using IBM SPSS Statistics 29.0 (IBM Corp, 2022) to analyze the effects of the error table, migration background, and implicit associations on the dependent variables grade and counted errors. We specified three models for each dependent variable and tested whether each model revealed significantly more variance. In Model 1, we included the main effects error table, migration background, and implicit associations. In Model 2 we added the two-way interactions between migration background × error table, between migration background × implicit associations, and between error table × implicit associations. Model 3 also included the three-way interaction between migration background × error table × implicit associations. Given our directional hypotheses, we applied a 95% one-tailed confidence interval level. For a better interpretation of the regression results we mean centered the continuous variable implicit associations.

Our main research interest was the three-way- interaction effect of migration background x error table x implicit associations (Model 3), as we expected that only negative implicit attitudes should affect the grading and only if no error table was present (Hypothesis 1–2). We calculated a sensitivity analysis with G*Power 3.1 (Faul et al., 2007) to determine the size of the effect, which we were able to determine with a power of 0.80 with our sample size of 157. With a one-sided alpha level of 0.05 and seven predictors, the analysis showed that we were able to detect an effect equal to or greater than $f^2 = 0.040$. This corresponds to a small effect according to Cohen (1988) classification.

To determine the direction of the three-way-interaction effect, we also calculated simple slope (Cohen et al., 2013) and slope differences tests (Dawson and Richter, 2006).

3 Results

3.1 Descriptive

See Table 1 for the descriptive.

3.2 Implicit associations

The average *D*-score in this sample, which measured implicit associations, indicated that the participants associated people with a migration background with lower performance levels than

TABLE 1 Descriptive for the dependent variables grade and counted errors.

		Grade	Errors
Condition	*n*	*M* (*SD*)	*M* (*SD*)
Error table Migration background	37	5.83 (0.53)	28.81 (4.86)
Error table No migration background	32	5.90 (0.41)	29.62 (4.52)
No error table Migration background	44	4.16 (1.04)	29.89 (5.34)
No error table No migration background	44	3.88 (1.08)	32.02 (4.03)

people without a migration background ($D = 0.32$, $SD = 0.37$), $t(156) = 10.81$, $p < 001$, $d = 0.86$. The order of the blocks showed no differences concerning the D-score. Therefore, both conditions were evaluated jointly.

3.3 Grade

To analyze the main effects on the dependent variable grade, we looked at Model 1 more closely. Model 1, which included error table, migration background, and implicit associations predicted 52.91% (corrected R^2) of the variance of the grade, $F(3, 153) = 59.43$, $p < 0.001$. The error table had a significant main effect on the grade, $\beta = 1.84$, $p < 0.001$. With addition of the error table, the dictations ($M = 5.86$, $SD = 0.47$) were rated worse than dictations without addition of the error table ($M = 4.02$, $SD = 1.07$).

Model 2 did not significantly explain more of the variance, $F(3, 150) = 1.89$, $p = 0.067$, $R^2 = 0.54$ $\Delta R^2 = 0.02$. Also, no significant two-way interaction was found between error table × migration background, $\beta = -0.14$, $p = 0.078$.

Model 3 explained a significantly higher proportion of variance of the grade, $F(1, 149) = 3.64$, $p = 0.029$, $R^2 = 0.55$, $\Delta R^2 = 0.01$. A significant three-way interaction effect was found between migration background, error table, and implicit associations, $\beta = -0.22$, $p = 0.029$.

A detailed analysis of the significant three-way interaction effect can be found in Figure 3. In addition, we calculated the predicted values of the grade under certain conditions (error table: yes vs. no, migration background vs. no migration background). The implicit associations were plotted for a standard deviation above the mean (indicating positive associations) and for one standard deviation below the mean (indicating negative associations). Subsequently, we performed a simple slope analysis to determine whether the slopes differed significantly from zero (Cohen et al., 2013), and we used a slope differences test to analyze whether two slopes differed significantly from each other (Dawson and Richter, 2006). If no error table was provided, the relationship between migration background and the grade differed depending on the valence of implicit associations (positive vs. negative associations), $t(157) = 2.75$, $p = 0.007$. Negative associations toward people with a migration background were associated with worse grades for the student with a migration background., $t(157) = 3.14$, $p = 0.002$. No such relationship was found for positive implicit associations, $t(157) = -0.74$, $p = 0.460$.

If an error table was available, the relationship between migration background and the grade did not differ depending on the valence of implicit associations (positive vs. negative associations), $t(157) = -0.03$, $p = 0.974$. The migration background did not affect grading for negative associations, $t(157) = -0.26$, $p = 0.799$, and positive associations, $t(157) = -0.2$, $p = 0.842$.

3.4 Counted errors

Model 1 predicted 4.43% (corrected R^2) of the variance of the counted errors, F (3, 153) = 3.41, $p = 0.019$. The error table had a significant main effect, $\beta = -0.18$, $p = 0.024$. If an error table was provided, the participants identified fewer errors ($M = 29.19$, $SD = 4.69$) than those who did not have an error table ($M = 30.95$, $SD = 4.82$).

Model 2 did not reveal a statistically significant increase in the predicted variance of the counted errors, $F(3, 150) = 0.26$, $p = 0.856$, $R^2 = 0.03$ $\Delta R^2 = 0.01$. None of the two-way interaction effects were statistically significant (see Table 2).

Also Model 3 did not increase the predicted variance significantly, $F(1, 168) = 0.19$, $p = 0.890$, $R^2 = 0.02$, $\Delta R^2 = 0.00$. There was no significant three-way interaction (see Table 2).

4 Discussion

We tested the effectiveness of using objective assessment criteria during grading to reduce the influence of student characteristics on grading (e.g., bias against students with migration background).

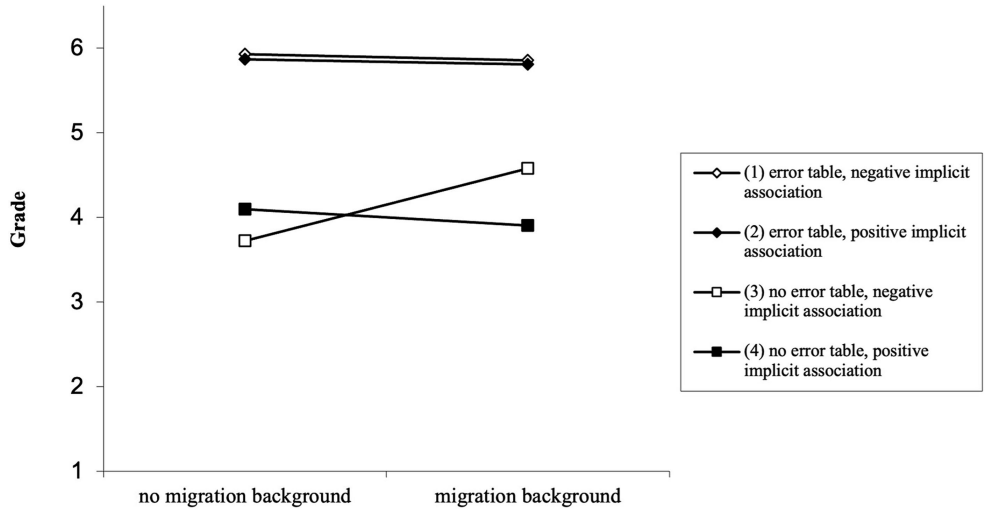


FIGURE 3
Grade as a function of migration background, error table, and implicit associations. Coding: positive implicit associations = Mean − 1 SD, negative implicit associations = Mean + 1 SD.

10.3389/feduc.2024.1386016

TABLE 2 Regression analysis for dependent variables grade and errors.

Model	Variable	Grade			Errors		
		b	*SE*	β	*b*	*SE*	β
Model 1	MB	0.139	0.139	0.056	−1.485	0.760	−0.154
	ET	1.836***	0.139	0.729***	−1.730*	0.760	−0.178*
	IA	0.062	0.188	0.018	0.770	1.031	0.059
	R^2	0.538			0.063		
Model 2	MB	0.319	0.186	0.128	−1.990	1.037	−0.207
	ET	2.018***	0.199	0.801***	−2.360*	1.109	−0.243*
	IA	−0.224	0.302	−0.066	0.635	1.685	0.049
	MB × ET	−0.397	0.278	−0.135	1.177	1.549	0.104
	MB × IA	0.747	0.379	0.155	0.619	2.114	0.033
	ET × IA	−0.140	0.380	−0.028	−0.522	2.118	−0.027
	R^2	0.555			0.067		
Model 3	MB	0.334	0.185	0.134	−1.983	1.041	−0.206
	ET	1.989***	0.198	0.790***	−2.374*	1.117	−0.245*
	IA	−0.508	0.335	−0.150	0.503	1.887	0.039
	MB × ET	−0.399	0.275	−0.136	1.176	1.554	0.104
	MB × IA	1.419**	0.515	0.295**	0.931	2.906	0.050
	ET × IA	0.592	0.538	0.118	−0.187	3.033	−0.009
	MB × ET × IA	−1.437*	0.753	−0.220	−0.668	4.250	−0.027
	R^2	0.566			0.068		

MB, migration background (0 = no migration background, 1 = migration background); ET, error table (0 = no error table, 1 = error table); IA, implicit association (mean centered). *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

We assumed that students with a migration background would only be disadvantaged if no objective assessment criteria were provided, as the judge would then draw on invalid information (migration background and implicit associations) to make his or her assessment.

Contrary to our expectations and previous research on this topic (e.g., Bonefeld et al., 2017; Bonefeld and Dickhäuser, 2018), we found no disadvantage for students with a migration background across conditions when no error table was provided. Nevertheless, the significant three-way-interaction indicated a moderation effect of implicit association. As expected, negative associations toward people with a migration background were associated with worse grades for the student with a migration background. This association was not evident if an error table was provided. This finding shows that implicit associations play an important role in decision making when there are no given assessment criteria. This is in line with other studies which have found an impact of implicit associations on judgments (e.g., Peterson et al., 2016; Bonefeld and Dickhäuser, 2018). The implementation of evaluation criteria has also been found to be a successful approach to making judgments more objective in other studies (Quinn, 2020).

Pre-service teachers have little teaching experience and may therefore be insecure about grading. This might be the reason why their implicit associations explain judgment differences in the process of grading. The error table provided clear rules for grading; therefore, they did not need to resort to implicit associations. Furthermore, positive implicit associations did not seem to influence the relationship between migration background and the grade when no error table was provided. This is in line with the expectation that stereotypes and associations are applied especially if they correspond with the performance (Glock et al., 2013; Glock and Krolak-Schwerdt, 2013; Kleen and Glock, 2018). Interestingly, the error table not only reduced differences in judgment related to migration background but also influenced the grade across conditions. When working with the error table, the participants gave a worse grade overall. The dictation used in the study had 30 errors. In the error table used in this study, the worst-possible grade of 6 (insufficient) was assigned for 30 errors (see Figure 1). This could be attributed to the increased accuracy of the grading judgments. Thus, an error table provides information on the appropriate grade for a certain number of errors and therefore increases the accuracy of grading judgments. The ambiguity regarding the criteria to be applied is reduced. Accuracy was not, however, the focus of this study, especially because the criterion (a correct grade) is difficult to determine, considering that another error table could assign a different grade to the same number of errors.

Furthermore, as expected, counting errors as a rule-based procedure was not affected by migration background and implicit associations. Surprisingly, we found a main effect of the error table. The participants counted fewer errors if the error table was provided. We assume that the specific characteristics of the error table used might explain this finding. It assigned the worst grade for more than 16 errors (see Figure 1). As the dictation used in the study had considerably more than 16 errors, it could be that some participants stopped counting the errors early on, as the grade was already fixed. It would be important to find out whether this effect was caused by the study or whether teachers would behave in the same way in everyday

school life. The latter would be unfortunate, because they should not only consider the students' errors in terms of grades but also in terms of their development.

In view of the very high number of errors, it is questionable why not all assessors assign the worst possible grade. However, this pattern of mild assessment has already been shown in past studies (Bonefeld and Dickhäuser, 2018). This shows that even with such an unambiguous performance, an error table can be helpful since rule-based judging without it does not seem to be obvious or easy for all judges.

4.1 Limitations

Some limitations of our research should be kept in mind. First, the participants had to judge students based on limited information. They had no frame of reference in the sense of other students' work because they only judged one dictation. Further, they judged the student without knowing about the students' prior development at school. As they were not provided with any additional information about the student or the performance of other students in the class, the external validity of the results might be limited (Christensen et al., 2014). At the same time, this is not a completely artificial situation, especially when thinking about the transition to a secondary school or teachers taking over a new class. However, an experimental approach, such as in the present study, has a clear benefit. Different grades can, of course, be rooted in actual differences in performance. The experimental approach allowed us to control for such differences and helped us to clearly identify the sole effect of migration background and of the error table on the different types of performance rating. If participants had judged two dictations from a student with and without a migration background (within), there would have been more of a direct comparison, but we would not have been able to present the identical dictation.

Second, we operationalized migration background in this study by using student names. The risk of doing this is that names might not only transfer information about migration background, but also give rise to several other associations or assumptions (e.g., socioeconomic status: Tobisch and Dresel, 2017). Therefore, our results could stem from associations with further student characteristics besides their migration background. For this reason, we intentionally selected two names that are strongly connected with the respective backgrounds and are comparable in terms of other variables (intelligence, attractiveness, gender) as a previous study (Bonefeld and Dickhäuser, 2018) suggested.

Third, migration background was only manipulated with a male name to control for gender stereotypes. Bonefeld et al. (2021) showed that characteristics of migration background and gender can interact. It could be possible that the male name activated also other stereotypes regarding spelling skills. However, these stereotypes should have been activated for both a student with and without a migration background. Further, Kleen and Glock (2018) demonstrated assessment disparities between underperforming female students with and without a migration background. Therefore, future research should replicate the effect for female students in this context.

Fourth, the dictation at hand had many spelling mistakes, but since the dictation was a real student performance, we decided not to change it to keep the authenticity. It is possible, however, that the high number of errors made it more difficult to prove the teacher bias, since with such a high number of errors a bad grade must be given so clearly (also to the German student) that differences cannot be proven. In these cases, teachers have little to no scope for discretion. This might lead to a very conservative measurement in our study. If differences are already apparent with such clear student results, as is the case with pre-service teachers, then it is to be expected that differences will become even more prevalent with better student results. However, past studies have shown differences according to migration background, especially in the poor performance level (Bonefeld and Dickhäuser, 2018). A conclusive answer can only be provided by further research that takes different performance levels into account.

Fifth, the participants completed the implicit association test right after assessing the dictation. Therefore, the assessment might have influenced (preservice) teachers' implicit associations. However, we deliberately chose this order, taking into account that the previous completion of an implicit association test in which performance stimuli were assigned to Turkish and German individuals might influence the dictation assessment. Since implicit association tests based on reaction times are, in our view, more difficult to influence than the dictation assessment, we decided on this order after careful consideration.

Furthermore, participants saw images of Turkish-looking adults, not students in the implicit association test. We were interested to see how general associations toward the performance of Turkish compared to German people can influence teachers in their assessment. Future research should investigate whether teachers respond differently to images of students and whether an adjusted IAT can explain assessment differences even better. Moreover, the pre-service teachers in our sample had little experience in grading dictations (only about 17 percent reported that they had already graded dictations). This is not surprising, since most of the pre-service teachers do not yet teach and are only being trained to do so. However, this must be considered when interpreting the results and at the same time suggests that it is important to train future teachers well and to prepare them adequately for the challenges they will face, since they need to be able to assess students at an early stage in their careers and to do this as well as possible. Therefore, it is important to provide adequate practice opportunities already during their studies.

Another aspect to consider is the use of one-sided hypothesis testing, which may affect the interpretation of p-values. While we adjusted the p-values accordingly, it's worth noting that multiple testing could potentially increase these values and impact the significance of results, particularly concerning the three-way interaction effect. This limitation underscores the need for caution when interpreting the findings. Additionally, given the potential influence of multiple factors on the outcomes, conducting acceptably powered replication studies is essential to validate and extend the results obtained in this study.

4.2 Implications and future research

Our results have implications for teacher education. Our study highlights the importance of focusing on strategies for avoiding biases associated with student characteristics in teacher education. The present findings demonstrate the importance of working with objective assessment criteria such as error tables and to conduct further research on this topic. Based on theoretical assumptions and these initial findings one can conclude that pre-service teachers

should be encouraged to prepare grading schemes to be used for making rule-based judgments or, more precisely, judgments based on deduction and thus assign grades based on these fixed standards. It is essential to reinforce our findings with further research. For assessing performance in other tasks, it is clearly more difficult to provide an objective table. An example of such tasks is essays or other more creative tasks (Banta, 2008). Here tables based on objective criteria must leave more scope for interpretation, which leads to higher uncertainty. However, they are still an approximate means for achieving higher consistency, e.g., within a class. Examples of how to objectify essays in English (Jansen et al., 2019) or oral examinations (Westhoff et al., 2002) already exist in the literature and therefore can provide initial orientation for the establishment of own objective assessment criteria in class.

Another important point is the issue of whether teachers actually use the objective assessment criteria in practice. Use is related to motivation: the motivation to create objective assessment criteria, use them, stick to them, and not deviate from them. Therefore, future research should take motivational factors of the judge into account.

5 Conclusion

The assessment of students is a central component of the teaching profession. It usually involves assigning grades, which should be an indicator of school performance and learning success and are important instruments in school. Among other things, they guide decisions on access to further education and professions (Magno, 2010). For this reason, objective, reliable, and valid performance assessments are essential in the school context (Fiscal, 2019). It is important to keep in mind the fact that an error table even if it can reduce the influence of student characteristics in grading, does not, of course, prevent teachers' negative associations. And precisely these associations can have a variety of (negative) effects (e.g., in the sense of a self-fulfilling prophecy) for students with a migration background. Nevertheless, objective assessment criteria prevent effects of student characteristics on evaluation and, therefore, are of great importance from a practical perspective. Judgment errors play a crucial role when one considers that school grades in elementary school influence future school careers and also greatly impact subsequent educational access, such as admission to different courses or different careers (Glock and Krolak-Schwerdt, 2013). Thus, it is necessary that teacher judgments are as accurate as possible to ensure the best possibilities for all students. Our research provides important insights in relation to the usefulness of objective assessment criteria like error tables. The present research work follows the assumption that objective assessment criteria, such as an error table, reduce the influence of implicit associations. Linked to this assumption, it could be shown that an error table reduces the influence of student characteristics. This finding has important implications for the education and training of pre-service teachers.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the studies involving humans because the studies were conducted in full accordance with the Ethical Guidelines of the German Association of Psychologists (DGPs) and the American Psychological Association (APA). At the time the data were acquired, it was not customary at most German universities to seek ethics approval for survey studies on such a subject. The study exclusively makes use of anonymous questionnaires. No identifying information was obtained from participants. We had no reason to assume that our survey would induce persisting negative states (e.g., clinical depression) in the participants. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

SP: Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Methodology. KK: Writing – original draft, Writing – review & editing. MB: Formal analysis, Methodology, Writing – original draft, Writing – review & editing, Conceptualization, Investigation, Project administration, Supervision, Visualization.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Banta, T. W. (2008). Sadly, rubrics are not for everyone. *Assess* 20:3. doi: 10.1002/au.204

Birkel, P., and Tarnai, C. (2018). "Zensuren und verbale Schulleistungsbeurteilung [Grades and verbal school performance assessment]" in *Handwörterbuch Pädagogische Psychologie*. eds. D. H. Rost, J. R. Sparfeldt and S. R. Buch (Beltz: Weinheim), 904–917.

Bonefeld, M. (2019). *Messung leistungsbezogener und allgemeiner Einstellungen gegenüber türkischen und deutschen Personen. Ein impliziter Assoziationstest mit Wort und Bildstimuli [Measuring achievement-related and general attitudes toward Turkish and German individuals. An implicit association test with word and picture stimuli] [Unpublished manuscript]. Department of Psychology, University of Mannheim*. Available

at: https://www.sowi.unimannheim.de/media/Lehrstuehle/sowi/Karst/Dateien/IAT_MeikeBonefeld2019.pdf

Bonefeld, M., and Dickhäuser, O. (2018). (Biased) Grading of students' performance: Students' names, performance level, and implicit attitudes. *Front Psychol.* 9:481. doi: 10.3389/fpsyg.2018.00481

Bonefeld, M., Dickhäuser, O., Janke, S., Praetorius, A.-K., and Dresel, M. (2017). Migrationsbedingte Disparitäten in der Notenvergabe nach dem Übergang auf das Gymnasium [Migration-related disparities in grade assignment after the transition to high school]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 49, 11–23. doi: 10.1026/0049-8637/a000163

Bonefeld, M., Dickhäuser, O., and Karst, K. (2020). Do preservice teachers' judgments and judgment accuracy depend on students' characteristics? The effect of gender and immigration background. *Soc. Psychol. Educ.* 23, 189–216. doi: 10.1007/s11218-019-09533-2

Bonefeld, M., Kleen, H., and Glock, S. (2021). The Effect of the Interplay of Gender and Ethnicity on Teachers Judgements: Does the School Subject Matter?. *The Journal of Experimental Education* 23, 1–21. doi: 10.1080/00220973.2021.1878991

Bundesregierung. (2023). Lagebericht: Rassismus in Deutschland [Report: Racism in Germany]. Available at: https://www.integrationsbeauftragte.de/resource/blob/1864320/2157012/77c8d1dddeea760bc13dbd87ee9a415f/lagebericht-rassismus-komplett-data.pdf?download=1

Chaiken, S., Liberman, A., and Eagly, A. H. (1989). "Heuristic and systematic information processing within and beyond the persuasion context" in *Unintended thought*. eds. J. S. Uleman and J. A. Bargh (New York: Guilford Press), 212–252.

Christensen, L. B., Johnson, R. B., and Turner, L. A. (2014). *Research methods, design, and analysis. 12th* Edn. Boston: Pearson.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences *(2nd ed.). Routledge.* doi: 10.4324/9780203771587

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences* (New York: Routledge).

Dawson, J. F., and Richter, A. W. (2006). Probing three-way interactions in moderated multiple regression: development and application of a slope difference test. *J. Appl. Psychol.* 91, 917–926. doi: 10.1037/0021-9010.91.4.917

Dee, T. S. (2005). A teacher like me: does race, ethnicity, or gender matter? *Am. Econ. Rev.* 95, 158–165. doi: 10.1257/000282805774670446

DESTATIS (2019). *Bevölkerung und Erwerbstätigkeit: Ausländische Bevölkerung Ergebnisse des Ausländerzentralregisters [Population and employment. Foreign population results of the central register of foreigners]* Statistisches Bundesamt (Destatis).

Devine, P. G. (1989). Stereotypes and prejudice: their automatic and controlled components. *J. Pers. Soc. Psychol.* 56, 5–18. doi: 10.1037/0022-3514.56.1.5

Drüke-Noe, C. (2014). *Aufgabenkultur in Klassenarbeiten im Fach Mathematik: Empirische Untersuchungen in neunten und zehnten Klassen [Task culture in mathematics assignments: Empirical studies in ninth and tenth grades]*. Berlin: Springer Spektrum. doi: 10.1007/978-3-658-05351-2

Faul, F., Erdfelder, E., and Lang, A. G. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods.* 39, 175–191. doi: 10.3758/BF03193146

Fazio, R. H. (2001). On the automatic activation of associated evaluations: an overview. *Cognit. Emot.* 15, 115–141. doi: 10.1080/0269993004200024

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Soc. Cogn.* 25, 603–637. doi: 10.1521/soco.2007.25.5.603

Fazio, R. H. (1986). "How do attitudes guide behavior?" in *The handbook of motivation and cognition: Foundations of social behavior*. ed. R. M. Sorrentino, and E. T. Higgins, 204–243.

Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., and Sherman, J. W. (2006). Automatic and controlled components of judgment and decision making. *J. Pers. Soc. Psychol.* 91, 797–813. doi: 10.1037/0022-3514.91.5.797

Fiscal, Z. J. M. (2019). *Educational Psychology* Ashland, Bern: Society Publishing and EBSCO Industries Inc.

Fiske, S. T., Lin, M., and Neuberg, S. L. (1999). "The continuum model" in *Dual-process theories in social psychology*. eds. S. Chaiken and Y. Trope (London: Guilford Press), 231–254.

Fiske, S. T., and Neuberg, S. L. (1990). "A continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation" in *Advances in experimental social psychology*. ed. M. P. Zanna, vol. *23* (New York, London: Academic Press), 1–74.

Froehlich, L., Martiny, S. E., Deaux, K., and Mok, S. Y. (2016). It's their responsibility, not ours. *Soc. Psychol.* 47, 74–86. doi: 10.1027/1864-9335/a000260

Gawronski, B., and Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychol. Bull.* 132, 692–731. doi: 10.1037/0033-2909.132.5.692

Gebhardt, M., Rauch, D., Mang, J., Sälzer, C., Klieme, E., and Köller, O. (2013). "Mathematische Kompetenz von Schülerinnen und Schülern mit Zuwanderungshintergrund [mathematical competence of students with immigrant backgrounds]" in *PISA 2012*. eds. M. Prenzel, C. Sälzer, E. Klieme and O. Köller (Münster: Waxmann), 275–308.

Glock, S. (2016). Does ethnicity matter? The impact of stereotypical expectations on inservice teachers' judgments of students. *Soc. Psychol. Educ.* 19, 493– 509. doi: 10.1007/s11218-016-9349-7

Glock, S., and Krolak-Schwerdt, S. (2013). Does nationality matter? The impact of stereotypical expectations on student teachers' judgments. *Soc. Psychol. Educ.* 16, 111–127. doi: 10.1007/s11218-012-9197-z

Glock, S., Krolak-Schwerdt, S., Klapproth, F., and Böhmer, M. (2013). Beyond judgment bias: how students' ethnicity and academic profile consistency influence teachers' tracking judgments. *Soc. Psychol. Educ.* 16, 555–573. doi: 10.1007/s11218-013-9227-5

Glock, S., Krolak-Schwerdt, S., and Pit-ten Cate, I. M. (2015). Are school placement recommendations accurate? The effect of students' ethnicity on teachers' judgments and recognition memory. *Eur. J. Psychol. Educ.* 30, 169–188. doi: 10.1007/s10212-014-0237-2

González-Rodríguez, D., Vieira, M.-J., and Vidal, J. (2019). Factors that influence early school leaving: a comprehensive model. *Educ. Res.* 61, 214–230. doi: 10.1080/00131881.2019.1596034

Greenwald, A. G., and Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychol. Rev.* 102, 4–27. doi: 10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* 74, 1464–1480. doi: 10.1037//0022-3514.74.6.1464

Henschel, S., Heppt, B., Rjosk, C., and Weirich, S. (2022). "Zuwanderungsbezogene Disparitäten [Immigration-related disparities]" in IQB-Bildungstrend 2021. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im dritten Ländervergleich. Eds. P. Stanat, S. Schipolowski, R. Schneider, K. A. Sachse, S. Weirich, and S. Henschel (Münster, New York: Waxmann Verlag GmbH), 181–219.

Hewstone, M., Hantzi, A., and Johnston, L. (1991). Social categorization and person memory: the pervasiveness of race as an organizing principle. *Eur. J. Soc. Psychol.* 21, 517–528. doi: 10.1002/ejsp.2420210606

Hoffman, C., and Hurst, N. (1990). Gender stereotypes: perception or rationalization? *J. Pers. Soc. Psychol.* 58, 197–208. doi: 10.1037/0022-3514.58.2.197

Holder, K., and Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: a new look from a shifting standards perspective. *Soc. Psychol. Educ*, 20, 471–490. doi: 10.1007/s11218-017-9384-z

Hübner, N., Jansen, M., Stanat, P., Bohl, T., and Wagner, W. (2024). Alles eine Frage des Bundeslandes? Eine mehrebenenanalytische Betrachtung der eingeschränkten Vergleichbarkeit von Schulnoten [Is it all a question of the federal state? A multi-level analysis of the limited comparability of school grades]. *Zeitschrift für Erziehungswissenschaft*. doi: 10.1007/s11618-024-01216-9

IBM Corp (2022). *IBM SPSS statistics for windows [computer program] (version 29.0)* Armonk, NY: IBM Corp.

Jansen, T., Vögelin, C., Machts, N., Keller, S., and Möller, J. (2019). Empirische Arbeit: Das Schülerinventar ASSET zur Beurteilung von Schülerarbeiten im Fach Englisch. Drei experimentelle Studien zu Effekten der Textqualität und der Schülernamen [Empirical work: The ASSET student inventory for assessing student work in English. Three experimental studies on effects of text quality and student names]. *Psychol. Erzieh. Unterr.* 66, 303–315. doi: 10.2378/peu2019.art21d

Jussim, L., and Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. *Personal. Soc. Psychol. Rev.* 9, 131–155. doi: 10.1207/s15327957pspr0902_3

Kahraman, B., and Knoblich, G. (2000). Stechen statt Sprechen: Valenz und Aktivierbarkeit von Stereotypen über Türken [stabbing instead of talking: valence and activability of stereotypes about Turks]. *Z. Sozialpsychol.* 31, 31–43. doi: 10.1024//0044-3514.31.1.31

Kleen, H., and Glock, S. (2018). A further look into ethnicity: the impact of stereotypical expectations on teachers' judgments of female ethnic minority students. *Soc. Psychol. Educ.* 21, 759–773. doi: 10.1007/s11218-018-9451-0

Kristen, C., Reimer, D., and Kogan, I. (2008). Higher education entry of Turkish immigrant youth in Germany. *Int. J. Comp. Sociol.* 49, 127–151. doi: 10.1177/0020715208088909

Lorenz, G. (2018). *Selbsterfüllende Prophezeiungen in der Schule [Self-fulfilling prophecies in school]*. Wiesbaden: Springer Fachmedien Wiesbaden.

Macrae, C. N., and Bodenhausen, G. V. (2000). Social cognition: thinking categorically about others. *Annu. Rev. Psychol.* 51, 93–120. doi: 10.1146/annurev.psych.51.1.93

Macrae, C. N., Milne, A. B., and Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: a peek inside the cognitive toolbox. *J. Pers. Soc. Psychol.* 66, 37–47. doi: 10.1037/0022-3514.66.1.37

Magno, C. (2010). "The functions of grading students" in *The Assessment Handbook*, *vol.* 3, 50–58.

McKown, C., and Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *J. Sch. Psychol.* 46, 235–261. doi: 10.1016/j.jsp.2007.05.001

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dyn. Theory Res. Pract.* 6, 101–115. doi: 10.1037/1089-2699.6.1.101

Nosek, B. A., Greenwald, A. G., and Banaji, M. R. (2005). Understanding and using the implicit association test: II. Method variables and construct validity. *Pers. Soc. Psychol. Bull.* 31, 166–180. doi: 10.1177/0146167204271418

OECD (2019). *PISA 2018 Results (Volume II)*. Paris: OECD Publishing.

Olson, M. A., and Fazio, R. H. (2009). "Implicit and explicit measures of attitudes: the perspective of the MODE model" in *Attitudes*. eds. R. E. Petty, R. H. Fazio and P. Briñol (New York: Psychology Press), 19–63.

Parks, F. R., and Kennedy, J. H. (2007). The impact of race, physical attractiveness, and gender on education majors' and teachers' perceptions of student competence. *J. Black Stud.* 37, 936–943. doi: 10.1177/0021934705285955

Peterson, E. R., Rubie-Davies, C., Osborne, D., and Sibley, C. (2016). Teachers' explicit expectations and implicit prejudiced attitudes to educational achievement: relations with student achievement and the ethnic achievement gap. *Learn. Instr.* 42, 123–140. doi: 10.1016/j.learninstruc.2016.01.010

Petty, R. E., and Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Adv. Exp. Social Psychol.* 19, 123–205. doi: 10.1016/S0065-2601(08)60214-2

Petty, R. E., and Cacioppo, J. T. (1983). "Central and peripheral routes to persuasion: application to advertising" in *Advertising and consumer psychology*. eds. L. Percy and A. G. Woodside (Lexington, MS: Lexington Books), 3–23.

Pit-ten Cate, I. M., and Glock, S. (2018). Teacher expectations concerning students with immigrant backgrounds or special educational needs. *Educ. Res. Eval.* 24, 277–294. doi: 10.1080/13803611.2018.1550839

Quinn, D. M. (2020). Experimental evidence on teachers' racial Bias in student evaluation: the role of grading scales. *Educ. Eval. Policy Anal.* 42, 375–392. doi: 10.3102/0162373720932188

Roskos-Ewoldsen, D. R., and Fazio, R. H. (1992). On the orienting value of attitudes: attitude accessibility as a determinant of an object's attraction of visual attention. *J. Pers. Soc. Psychol.* 63, 198–211. doi: 10.1037/0022-3514.63.2.198

Smith, E. R., and DeCoster, J. (1999). "Associative and rule based processing" in *Dual-process theories in social psychology*. Eds. S. Chaiken and Y. Trope (New York, London: Guilford Press), 323–336.

Spencer, K. B., Charbonneau, A. K., and Glaser, J. (2016). Implicit Bias and policing. *Soc. Personal. Psychol. Compass* 10, 50–63. doi: 10.1111/spc3.12210

Südkamp, A., Kaiser, J., and Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *J Educ Psychol.*104, 743–762. doi: 10.1037/a0027627

Swim, J., Borgida, E., Maruyama, G., and Myers, D. G. (1989). Joan McKay versus John McKay: do gender stereotypes bias evaluations? *Psychol. Bull.* 105, 409–429. doi: 10.1037/0033-2909.105.3.409

Taylor, S. E. (1981). "A categorization approach to stereotyping" in *Cognitive processes in stereotyping and intergroup behaviour*. ed. D. L. E. Hamilton (Hillsdale: Lawrence Erlbaum Associates), 83–211.

Tobisch, A., and Dresel, M. (2017). Negatively or positively biased? Dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds. *Soc. Psychol. Educ.* 20, 731–752. doi: 10.1007/s11218-017-9392-z

Trapmann, S., Hell, B., Weigand, S., and Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs-eine Metaanalyse [The validity of school grades for predicting academic success-a meta-analysis.]. *Zeitschrift für pädagogische Psychologie*, 21, 11–27. doi: 10.1024/1010-0652.21.1.11

Triventi, M. (2020). Are children of immigrants graded less generously by their teachers than natives, and why? Evidence from student population data in Italy. *Int. Migr. Rev.* 54, 765–795. doi: 10.1177/0197918319878104

Uhlmann, E., and Cohen, G. L. (2005). Constructed criteria: redefining merit to justify discrimination. *Psychol. Sci.* 16, 474–480. doi: 10.1111/j.0956-7976.2005.01559.x

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., and Holland, R. W. (2010). The implicit prejudiced attitudes of teachers. *Am. Educ. Res. J.* 47, 497–527. doi: 10.3102/0002831209353594

van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Econ. Educ. Rev.* 30, 1045–1058. doi: 10.1016/j.econedurev.2011.05.008

Walton, G. M., and Spencer, S. J. (2009). Latent ability: grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychol. Sci.* 20, 1132–1139. doi: 10.1111/j.1467-9280.2009.02417.x

Westhoff, K., Hagemeister, C., and Eckert, H. (2002). On the objectivity of oral examinations in psychology. *Zeitschrift Differentielle Diagnostische Psychol.* 23, 149–157. doi: 10.1024//0170-1789.23.2.149

Wiggan, G. (2007). Race, school achievement, and educational inequality: toward a student-based inquiry perspective. *Rev. Educ. Res.* 77, 310–333. doi: 10.3102/003465430303947