frontiers in Education

ORIGINAL RESEARCH
published: 28 February 2020
doi: 10.3389/feduc.2020.00017

Differences Between Students' and Teachers' Fairness Perceptions: Exploring the Potential of a Self-Administered Questionnaire to Improve Teachers' Assessment Practices

Philipp Sonnleitner[1] and Carrie Kovacs[2]**

[1] Luxembourg Centre for Educational Testing, University of Luxembourg, Esch-sur-Alzette, Luxembourg, [2] School of Informatics, Communications and Media, University of Applied Sciences Upper Austria, Hagenberg, Austria

OPEN ACCESS

Edited by:
Gavin T. L. Brown,
The University of Auckland,
New Zealand

Reviewed by:
Robin Dee Tierney,
Independent Researcher, San Jose,
United States
Roseanna Bourke,
Massey University, New Zealand

****Correspondence:***
Philipp Sonnleitner
philipp.sonnleitner@uni.lu
Carrie Kovacs
carrie.kovacs@fh-hagenberg.at

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *31 August 2019*
Accepted: *04 February 2020*
Published: *28 February 2020*

Citation:
Sonnleitner P and Kovacs C (2020) Differences Between Students' and Teachers' Fairness Perceptions: Exploring the Potential of a Self-Administered Questionnaire to Improve Teachers' Assessment Practices. Front. Educ. 5:17. doi: 10.3389/feduc.2020.00017

The ability to assess learning outcomes is vital to effective teaching. Without understanding what students have learned, it is impossible to tailor information, tasks or feedback adequately to their individual needs. Thus, assessment literacy has been increasingly recognized as a core teacher competency in educational research, with many empirical studies investigating teachers' abilities, knowledge and subjective views in relation to classroom assessment. In contrast, relatively few studies have focused on students' perspectives of assessment. This is surprising, since gathering students' feedback on their teachers' assessment practices seems a logical step toward improving those practices. To help fill this gap, we present an explorative study using the recently developed Fairness Barometer as a tool to help identify specific strengths and weaknesses in individual teachers' assessment methods. Viewing assessment through the lens of classroom justice theory, the Fairness Barometer asks students and teachers to rate aspects of procedural and informational justice in their own (teachers') assessment practices. We examined the resulting fairness discrepancy profiles for 10 Austrian secondary school classes (177 students). Results showed wide variation in profile pattern, evidence that both students and teachers can differentiate between different aspects of assessment fairness. Further exploration of the resulting discrepancy-profiles revealed certain problem types, with some teachers differing from their students' perception in almost every rated aspect, some showing specific assessment-related behaviors that require improvement (e.g., explaining grading criteria of oral exams), and others demonstrating almost identical responses as their students to the addressed fairness aspects. Results clearly indicate the potential of the Fairness Barometer to be used for teacher training and teacher self-development within the domain of teacher assessment literacy.

Keywords: teacher assessment literacy, classroom justice, perceived fairness, profile interpretations, Fairness Barometer

INTRODUCTION

One of the core tasks of teachers is assessing their students' competencies (Ainscow et al., 2013). Although assessment is not a favorite activity among teachers (Blount, 2016), pinpointing where students stand on specific competencies provides a basis for formative and summative student evaluation, reveals individual learning trajectories, and indicates whether content has been mastered or needs to be repeated.

Despite the importance of assessment in teaching, teachers themselves report being insufficiently prepared for this task through teacher education programs and mostly learning "on the job" how to best assess their students (Volante and Fazio, 2007; Battistone et al., 2019). This uncertainty seems warranted, given numerous studies that experimentally show flaws in teachers' diagnostic competencies (e.g., Kaiser et al., 2017; Tobisch and Dresel, 2017) or that identify a substantial lack of teacher assessment literacy (for an excellent review of the last three decades, see Xu and Brown, 2016).

Most such studies focus squarely on teachers' abilities to assess accurately and on their subjective views of assessment. These issues are important and hold a deserved place in educational research. One might also, however, ask how assessment unfolds on the other side of the exam sheet – in other words, how students perceive the assessment situation. Can such a shift in perspective deliver new insights into teachers' assessment practices or provide novel tools for professional development? The present study examines the potential of the Fairness Barometer (Sonnleitner and Kovacs, 2018), a self-administered questionnaire grounded in classroom justice theory (Duplaga and Astani, 2010), to reveal discrepancies between students' and their teachers' perception of assessment practices. Beyond the relevance of this information for large-scale research, uncovering such differences might also provoke "cognitive conflict" (Cobb et al., 1990) that requires teachers to reflect upon their actions and thus improve aspects of their assessment literacy.

Teachers' Problems With Assessment Literacy

The last decades have seen an intense discussion of what teachers should know about assessment. Their knowledge of assessment practices, their ability to conduct assessments, to validly interpret assessment outcomes, and to communicate the resulting inferences to various educational stakeholders (students, parents, school principals, educational ministries, etc.) have all played a part in driving diverse conceptions of teacher assessment literacy (AL; e.g., Shin, 2015; Xu and Brown, 2016; Pastore and Andrade, 2019). With the rise of a broad palette of assessment methods (e.g., reflection logs, case studies, classroom experimental activities) and new educational concepts, such as assessment for learning or standards-based grading, models of AL have become increasingly diverse and complex. In the last few years, the focus has shifted toward situational aspects of AL, including the various socio-cultural and institutional contexts in which assessment is carried out, as well as cross-cultural differences in teachers' ideas of the purpose(s) and function(s) of assessment (Xu and Brown, 2016; Brown et al., 2019).

What remains at the core of all these models since the beginning of AL research, however, are knowledge and skills related to measurement theory and practice. AL includes, for example, the ability to design an accurate assessment process that allows a teacher to draw valid inferences about student learning; it also encompasses the ability to explain the purpose and the content of this process to the assessed students (Stiggins, 1991). This focus on educational measurement practices was heavily inspired by the Standards for Teacher Competence in Educational Assessment of Students published in 1990 by the (United States-) American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association (American Federation of Teachers [AFT] et al., 1990). A 2019 Delphi study by Pastore and Andrade, in which experts were asked to define AL, shows that this focus continues to be relevant: participants showed highest agreement on "praxeological" aspects of assessment, such as conducting assessments, interpreting results, and giving feedback to students. Especially in European countries (e.g., Austria, France, Germany, and Switzerland), teachers deal with these aspects almost on a daily basis, since many national grading regulations stipulate several written or oral examinations as the basis for grading one term (e.g., Klieme et al., 2007). Contrary to countries such as the United States or Canada, and excluding school transfer decisions, these assessments are not standardized and provided by external testing agencies but designed, administered, and scored by the teachers themselves.

When teachers were asked whether they fulfill these praxeological criteria of AL, results showed that the majority felt unprepared for assessment and grading upon leaving college, and that they only became comfortable with this task through learning on the job (Battistone et al., 2019). This self-doubt is echoed in empirical studies trying to operationalize and measure teachers' AL-levels (e.g., Mertler and Campbell, 2005; Alkharusi, 2011; Gotch and French, 2014). Although most of the instruments used in this research lack psychometric evaluations (Gotch and French, 2014), results on teachers' AL provide a consistent picture showing that teachers struggle or feel uncomfortable with certain activities that are central to the assessment process. This uncertainty is only partly warranted, as meta-analyses and literature reviews researching teacher judgment accuracy have shown fairly robust relationships between teacher judgments and independently assessed student performance (e.g., mean $r = 0.63$ based on 75 studies; Südkamp et al., 2012).

Students and Their Perceptions of Assessment

The second crucial stakeholder in assessment are the students themselves: They either benefit from accurate judgments of their learning progress or are substantially disadvantaged by flawed evaluations of their competencies. For example, student satisfaction, motivation, and affective learning have been shown to relate to perceived fairness of teachers' grading procedures;

these, in turn, strongly relate to professional satisfaction of the teachers themselves (Chory-Assad, 2002; Wendorf and Alexander, 2005). Apparently, accurate and transparent assessment can benefit both students and teachers.

Despite such findings, the question of how students perceive the quality of the whole assessment process they are involved in or, in other words, how they rate its fairness, is a relatively novel field of research. Despite being among the highest ideals in educational assessment, the concept of fairness was mostly treated from a mere measurement point of view (i.e., differential item functioning (DIF) or test bias; see for example the standards for educational and psychological testing, American Educational Research Association, [AERA] et al., 2014) but hardly defined or described in a comprehensive theory. This is especially true for the highly dynamic (and socio-cultural) context of classroom assessment, which would require a multifaceted understanding of fairness (e.g., Tierney, 2013, 2014; Rasooli et al., 2018). Since the traditional approach to (test and assessment) fairness through measurement theory falls short in accomodating the reality of classroom assessment, several authors have suggested enhancing the concept of fairness by building upon Organizational Justice (Lizzio and Wilson, 2008; Rasooli et al., 2019) or Classroom Justice Theory (Chory-Assad and Paulsel, 2004; Chory, 2007; Duplaga and Astani, 2010). Fairness, as a general student perception, would then follow from a subjective and evaluative judgment of applied and enacted justice rules (Rasooli et al., 2019). Note that in the following, we will focus on this procedural, hence pragmatic understanding of Fairness as a working definition. For broader and multifaceted conceptualizations of Fairness that question the utility of a single umbrella term and that explicitly include socio-cultural aspects, we recommend comprehensive treatments of this topic by Gipps and Stobart (2009), Camilli (2013), Tierney (2013, 2014), and Rasooli et al. (2018).

More specifically, Classroom Justice (CJ) differentiates between four types of justice that concern different aspects or phases of the assessment process (Tata, 1999; Colquitt, 2001; Chory-Assad, 2002; Chory-Assad and Paulsel, 2004). *Distributive Justice* occurs when students feel there is a balance between what they invest into a course or an exam and the grade they receive for it (e.g., Tata, 1999; Chory-Assad, 2002). *Procedural Justice* focuses on the process of assessment and grading and emphasizes the definition of clear assessment or grading standards that are effectively communicated and followed. Additionally, it encompasses students' ability to understand the rationale behind their grades and the freedom to discuss them with their teacher. This assumes that the assessment process is applied equally to every student and that student feedback on it is welcomed (Colquitt, 2001; Chory-Assad and Paulsel, 2004). *Interpersonal Justice* mainly concerns teacher–student interactions and the extent to which they are guided by mutual respect and integrity. A teacher's propensity to repeatedly interrupt students or make derogatory personal remarks about them would clearly violate this type of CJ. Finally, *Informational Justice* indicates the extent to which assessment and grading criteria are transparent and communicated in a timely manner to students. This also encompasses clear and comprehensive explanations of exam results when students ask for them. Evidently, *Informational* and *Procedural Justice* are closely related: Whereas *Informational Justice* concerns the transparency and clarity of assessment and grading rules, *Procedural Justice* relates to whether the teacher plays by these rules in everyday classroom interactions (see **Table 1** for questions capturing both domains).

These four different facets of CJ were found to be highly correlated but empirically distinct (Colquitt, 2001), with differential relations to various student or instructor characteristics. Empirical results identify *Procedural Justice* as the most important facet, being substantially linked to students' evaluation of their teachers, their aggression (r = –0.63) and hostility toward their teachers (r = –0.48), their motivation (r = 0.35), their satisfaction with their grades (r = 0.35), and their performance ($0.30 \leq r \leq 62$; Tata, 1999; Colquitt, 2001; Chory-Assad, 2002; Chory-Assad and Paulsel, 2004; Nesbit and Burton, 2006; Vallade et al., 2014). Although *Distributive Justice* has shown similar but smaller correlations with these variables, these relationships have generally vanished when controlling for *Procedural Justice* in multiple regressions. It seems that setting and clearly communicating the conditions and rules for assessment is not enough; students also need to feel that teachers are reliable and stick to these rules. If they do, *Procedural Justice* seems to be a powerful lever on several crucial student variables ranging from students' motivation to students' performance. Most importantly, teachers can intentionally influence this dimension by applying consistent and clear assessment procedures – core aspects of teacher assessment literacy.

Despite such suggestive results, however, it is important to note that the vast majority of studies on this topic have been carried out either in a university or an organizational/vocational context, examining participants of 20 years or older. Whether the factorial structure of CJ and its link to these relevant outcomes can be confirmed in younger elementary or secondary school samples remains to be seen.

Using Perception Discrepancies as Opportunity to Learn: The Fairness Barometer

Evidently, today's teachers face a huge challenge. Methods for assessing student learning have become increasingly diverse and new concepts of learning have been introduced, a trend reflected in ever more diversified models of AL. At the same time, teachers feel insufficiently prepared after their university education to deal with these concepts while being criticized by scientific studies showing lack of assessment knowledge or inaccurate diagnostic competencies. Moreover, research on CJ underscores the high impact of perceived fairness of assessment and grading procedures on key student variables, rendering AL even more important. But can students' sensitivity to suboptimal assessment procedures be leveraged as situated feedback to help teachers improve their AL? Allowing students to report how they perceive concrete assessment-related behaviors of their teachers is not only a way to recognize them as valuable stakeholders of this process; it might also provide valuable insights to be used within reflective practice (Schön, 1983). Especially if student and teacher

TABLE 1 | Item content and descriptives within the student sample (n = 168).

	Item content	Mean	Median	*SD*	Min	Max
Informational fairness						
fb1	The content of the exam is announced on time.	8.92	10	1.63	2	10
fb2	I know what criteria are used to assess oral exams.	7.27	8.00	2.69	1	10
fb3	I understand my own grades on oral exams.	7.53	8.00	2.49	3	10
fb4	I know what criteria are used to assess written exams.	8.74	9.00	1.59	1	10
fb5	I understand my own grades on written exams.	8.92	10	1.50	3	10
fb6	If I ask, my teacher will explain my grade to me.	8.79	10	1.87	1	10
Procedural fairness						
fb7	My teacher is open to comments about his/her grading system.	8.49	9.00	1.74	3	10
fb8	Grading criteria are applied equally to everyone in my class (unless there is a justified exception).	8.34	9.00	2.30	1	10
fb9	My current achievements will be graded independently of the grades I have had in the past.	8.51	10	2.06	2	10
fb10	The oral exams in class include enough questions for me to show what I know and what I can do.	7.23	8.00	2.70	1	10
fb11	The written exams in class include enough questions for me to show what I know and what I can do.	8.61	9.00	1.84	1	10
fb12	During written exams I have enough time to complete the given questions/tasks.	8.65	9.00	1.79	1	10
fb13	The questions/tasks included in exams are an accurate reflection of the material that has been taught in class.	8.83	9.00	1.51	2	10
fb14	The difficulty of exam questions/tasks is appropriate.	8.38	9.00	1.83	1	10
fb15	The exams only test material that has been taught in class.	9.05	10	1.54	3	10
General ratings						
fb16	How strong is your interest in this school subject?	6.68	7.00	2.35	1	10
fb17	How fairly do you think performance is graded in this subject?	8.65	9.00	1.70	3	10

perceptions differ substantially on specific aspects of fairness, this is likely to provoke some cognitive conflict within the teacher and thus provide a unique learning opportunity (Cobb et al., 1990).

The present study explores the potential of the Fairness Barometer (**Table 1**, Sonnleitner and Kovacs, 2018), a self-administered questionnaire, to capture students' as well as teachers' perceived fairness of the assessment process in order to reveal discrepancies that stimulate critical reflection and learning in teachers. Development of the Fairness Barometer (FB) aimed at providing a short and easy to administer tool that is psychometrically reliable and didactically actionable. The self-directed nature of the FB is intended to empower teachers in the debate on their assessment competencies and introduces a completely new take on teacher AL by considering students' perspectives. Fairness, as a construct, is indirectly measured by the students' evaluation and judgment of how their teachers implement principles of classroom justice (see above; Colquitt and Shaw, 2005).

Importantly, the FB only addresses aspects of assessment that students are able to observe and judge themselves. Questions were designed to be as concrete and realistic as possible: their intent was neither to raise unrealistic expectations among students nor to set impossibly high standards for teachers or to act as a judgment of their personal integrity. By focusing on changeable assessment behavior, the FB was developed to provide results that suggest concrete courses of action for teachers to improve their classroom assessment practices. Asking students about their satisfaction with the assessment process without being able or willing to change that process could only be expected to worsen classroom climate. Thus, the Fairness Barometer only contains statements related to Informational Justice (6 items, see **Table 1**) and Procedural Justice (9 items, see **Table 1**) that are rated in terms of agreement on a 10-point Likert scale by the students as well as their teachers. Empirical studies have shown the central importance of introducing transparent assessment and grading procedures and reliably following them. Thus, behavior related to Informational (e.g., announcing the content of an exam on time) and Procedural Justice (e.g., only testing content that was taught in class) were seen as particularly relevant. These aspects seem reasonably easy for students to rate and for teachers to change. Note that laws or school (district) guidelines and conventions also often regulate such behaviors.

In contrast, aspects related to Distributive Justice might be (more) prone to misjudgments due to individual student differences (e.g., in ability or effort) and to strategic student responding (e.g., trying to lower the amount of course work). Students may also be hesitant to report problematic aspects of Interpersonal Justice due to politeness or fear of reprisals. From the teacher's perspective, this last dimension is quite likely to include behaviors that are difficult to change merely based on feedback coming from a questionnaire; improving interpersonal interaction styles or changing attitudes that drive interpersonal discrimination would require further (more intense) training. Thus, although Distributive and Interpersonal Justice are theoretically and empirically (as several studies have shown; Colquitt, 2001; Chory-Assad and Paulsel, 2004) relevant, the inclusion of these topics in a questionnaire intended for self-directed teacher development seemed problematic.

First empirical results in two student samples ($n > 800$) were promising. The Fairness Barometer displayed high reliability with Cronbach's α and McDonald's ω ranging from.90 to.93 for the subjects of mathematics, German, and English (Sonnleitner and Kovacs, 2018). Surprisingly, confirmatory factor analysis favored a one-dimensional model of a general Perceived Fairness

factor over a two-dimensional model including Informational and Procedural Justice. This general factor was substantially correlated with the students' interest in the subject ($r = 0.31$) and liking for the teacher ($r = 0.60$) but not to students' age ($r < 0.10$), demonstrating a certain stability.

The Present Study

In order to use differing perceptions of assessment procedures between students and their teachers to improve AL, those differences need to be reliably identified. The present study therefore centers on the following explorative research questions:

1. Does the Fairness Barometer, administered to teachers and their students within a class, produce meaningful, interpretable discrepancy profiles? In the ideal case, both profiles would overlap at the highest possible rating (10). Other possibilities include (a) identical teacher and student ratings on lower levels, (b) general or specific discrepancies with the teacher overestimating the fairness of his or her actions, and (c) general or specific discrepancies with the teacher underestimating the fairness of his or her actions (in case of a low self-concept). We set out to produce and interpret a small sample of real discrepancy profiles in order to better understand and illustrate the potential as well as the practical limits of the Fairness Barometer as a practical (self-)assessment tool.
2. Can profile discrepancies be meaningfully quantified? In other words, is it possible to derive a single measure of profile discrepancy to allow for further classification and research on relevant student and teacher characteristics?

MATERIALS AND METHODS

Participants and Procedure

Participants were recruited by graduate students of Business and Economics Education as part of their methodological training. Students and their teachers responded anonymously via an online questionnaire at school during regular class time. From the resulting *ad hoc* sample of 177 Austrian students placed in 10 different classes, 9 were excluded due to a high number of missing values (>33% of the questions not answered). The final sample consisted of 168 students (91 girls; $M = 16.3$ years, $SD = 1.03$) within the instructional context of commercial schools (64%), or technical colleges (21%), and one class of a polytechnic school (25 students). All students were enrolled in upper secondary grades in Grade 9 (15.1%), Grade 10 (19.9%), Grade 11 (55.4%), and Grade 12 (9.6%).

These 10 different classes were taught by 9 teachers (8 female), with one of them teaching two classes (see headers of **Annex 3**). Class subjects ranged from languages (German, English) to business related (business administration, accounting, text processing) and profession-oriented topics (crop production, home management skills). Mean age of the teachers was 51.3 years ($SD = 9.68$) with a mean teaching experience of 22 years ($SD = 11.3$).

Students and their teachers were matched by a single (randomly created) code per class guaranteeing full anonymity for the students. Prior to the study, all participants were informed about its purpose and background and given the option to withdraw their participation at any point.

Variables

Perceived Fairness

To measure students' perceived fairness of their teachers' assessment practices, we administered a student and a teacher online version of the Fairness Barometer (Sonnleitner and Kovacs, 2018; see **Table 1** for the student version, **Annex 1** for the teacher version, and **Annex 2** for the original German student and teacher version). In total, it consists of 6 items assessing *Informational Fairness* (IF), and 9 items asking aspects of *Procedural Fairness* (PrF). These 15 items were answered for the respective subject that the students' teachers were teaching them. All items of the Fairness Barometer were positively phrased statements for which participants indicated their agreement on a 10-point Likert scale (1 – I don't agree to 10 – I fully agree). Student and teacher versions differed in the specific phrasing of the questions but addressed the same aspects of fairness. In addition, students were asked to make an overall judgment of their teacher's assessment practices' general *Fairness* on a 10-point Likert scale ranging from unfair (1) to fair (10), whereas teachers had to rate themselves concerning their overall Fairness. To arrive at the discrepancy profiles, students' ratings per class were averaged and plotted against their teacher's responses (see **Figure 1**). **Table 1** presents items as well as their descriptive statistics within the student sample. For the student sample, internal consistency of the scale was found to be high, with Cronbach's $\alpha = 0.90$ and McDonald's $\omega = 0.91$. Note that sample size of the teachers was too small ($n = 9$) for reliability analysis or meaningful descriptive statistics.

Distributional Justice and Interest

To get a better understanding of additional factors influencing Perceived Fairness, we asked students to report their grades for their last exam, as well as their last school report card. For both occasions, we asked them to report the grade they felt they deserved in order to enable the computation of a proxy for *Distributive Justice* (DJ = obtained grade – self-assigned grade). Comparing this difference for the last exam (DJe) to that for the last term (DJt) might allow us to screen for short-term or long-term effects on perceived (un)fairness. Note that grades in the Austrian school system range from 1 (highest) to 5 (lowest). The vast majority felt adequately rated on their last exam, and showed no difference on DJe (67.8%). However, a substantial proportion of the students (14.1%) reported overpayment, getting better grades than expected, and 18% felt they had received worse grades than they should have. Satisfaction with the last term grade was higher (77.7%) but nevertheless, 5.7% felt over- and 16.5% underpaid for their efforts in the given subject.

Students' *Interest* in a given subject has been shown to be closely related to their academic performance in that subject (e.g., Köller et al., 2001). Thus, students were also asked to rate

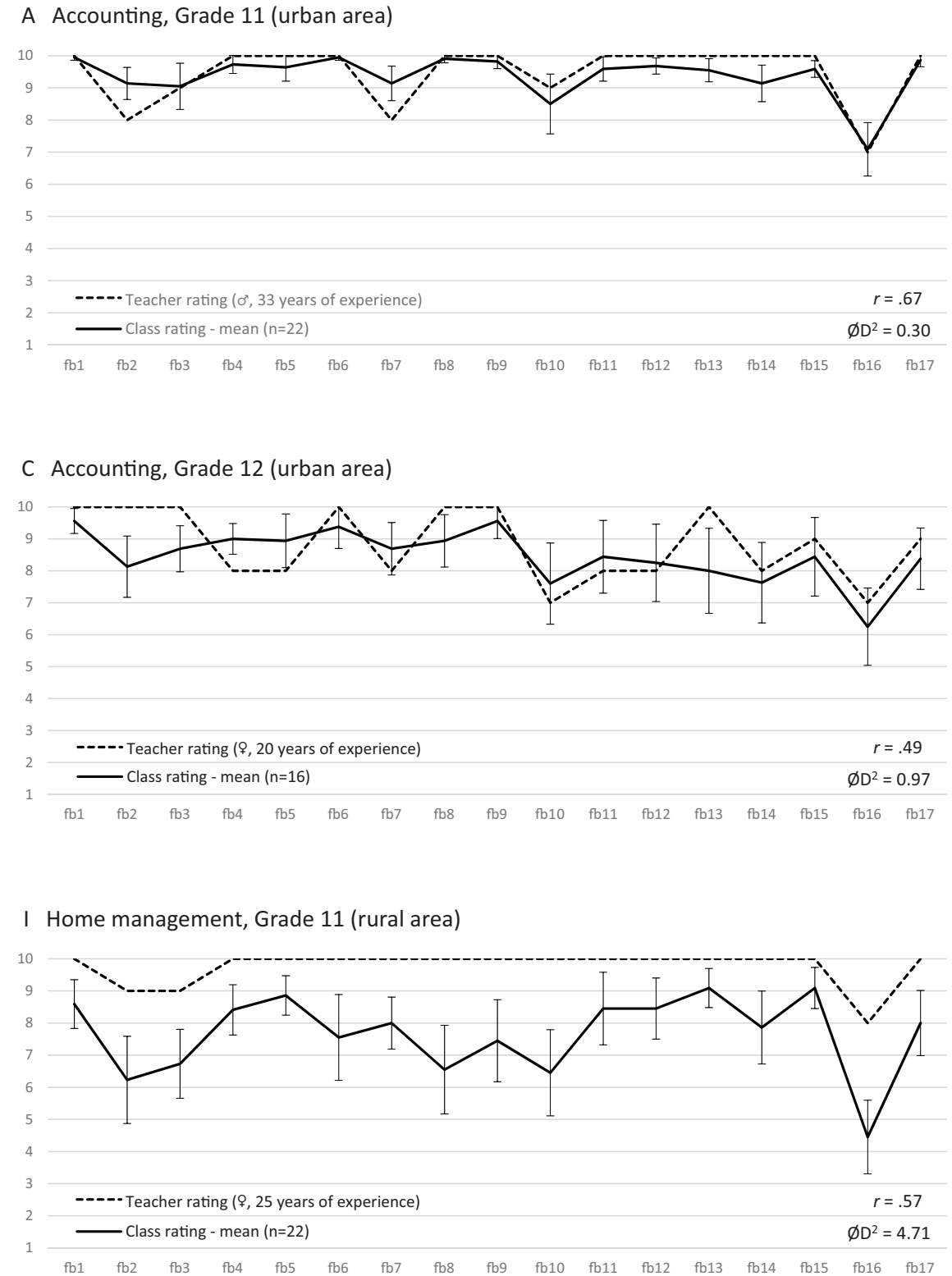


FIGURE 1 | Profile discrepancies ($ØD^2$) and correlations (*r*) between teacher and student ratings on all 15 items of the Fairness Barometer. Illustrated are prototypic profiles showing almost no differences (Class **A**, top image), specific discrepancies (Class **C**, middle image), and general perception differences (Class **I**, bottom image). Error bars indicate 95% CI. Interest in the subject (fb16) and perceived general fairness (fb17) were excluded from computation of $ØD^2$ and *r*.

their Interest (Int) for the given subject on a 10-point Likert scale ranging from not interested (1) to very interested (10). The teacher version of the questionnaire requested the participating educators to rate their students' overall interest as well as their competency level.

Methodological Approach

Qualitative Profile Discrepancy Interpretation

Resulting discrepancy profiles for the 10 classes were analyzed by the authors in terms of (a) overall congruency between students and their teachers, (b) similarities between profiles, (c) groups of items behaving similarly and thus allowing for clustering and diagnosing specific problems, and (d) representativeness of the class mean indicated by the 95% confidence interval. In a first step, these aspects were analyzed by the authors independently from each other. Following the individual interpretations, a joint discussion resulted in agreement on the main conclusions drawn from each profile.

Profile Similarity Scores

To complement the qualitative profile discrepancy interpretation, we additionally quantified the observed differences between teacher and student ratings within each class. To this end, we drew on profile similarity scores (PSIs) that are widely used for profile congruency identification in person-organization fit research. For every question of the Fairness Barometer (fb1-fb15), we computed the difference D between the individual teacher rating and the mean student rating in the class and squared this value to neutralize negative values and emphasize large deviations. We then averaged D^2 across all 15 items to arrive at a global PSI per class. While D^2 is assumed to give an indication of the overall congruency of the teacher-student ratings, we additionally computed Pearson's r as a consensus estimate (Kozlowski and Hattrup, 1992). Although the calculation of D^2 can be considered standard (e.g., Kristof, 1996), it is insensitive to the source of profile differences, and per design ignores whether a teacher is under- or overestimating the fairness of his or her assessment practices.

Construct Evidence

The relationships between perceived fairness and other central student variables could provide further insights into the validity of inferences drawn from the Fairness Barometer and its discrepancy profiles. We therefore investigated perceived fairness' construct evidence by looking at correlations between the sum score of Informational Fairness items, Procedural Fairness items, the total Fairness Barometer sum score, and students' interest in the subject, and perceived Distributive Justice of the last exam and last term. Moreover, we controlled for the impact of students' ability, using the last obtained grades as proxy, and investigated whether older students would be more critical concerning their teachers' fairness. All statistical analyses were conducted with jamovi version 1.0.7.0. (The Jamovi Project, 2019) and level of significance was set to alpha < 0.05.

RESULTS

Students' Perceived Fairness of Their Teachers' Assessment Procedures

Descriptive statistics of students' responses to the Fairness Barometer, presented in **Table 1**, clearly indicate a ceiling effect for almost every item, with item means between 7.23 (fb10; "The oral exams in class include enough questions for me to show what I know and what I can do") and 9.05 ("Exams only test material that has been taught in class") and medians ranging between 8 and 10. Thus, the vast majority of students perceived most of their teachers' behaviors related to informational and procedural aspects of assessment as fair. This was reflected in the consistently high general rating of their teachers' grading behavior (fb17; M = 8.65, SD = 1.70). Nevertheless, some students used the whole range of the scale, clearly indicating their dissatisfaction with their teachers' actions (scale minima ranging from 1 to 3). Differing opinions, reflected in the standard deviations, indicate the most homogeneous opinion on fb5 "I understand my own grades on written exams" (SD = 1.50), and the most disagreement concerning item fb10 asking whether oral exams contain enough questions to sufficiently show students' competencies (SD = 2.70). This variance between students' ratings may be interpreted not simply as measurement error but as an individually suboptimal fit between a student and his or her teacher's assessment related behaviors.

In general, students indicated the most problems with oral exams (fb2, fb3, and fb10): criteria used to rate them are not exactly transparent (M = 7.27), obtained grades are unclear (M = 7.53), and, as stated above, oral exams are perceived as being too short to demonstrate abilities (M = 7.23). These issues were not identified as problems in written exams. However, exam content overall seems to be announced on time (M = 8.92), and exams only tackle content that has been previously taught in class (M = 9.05). Students' interest in the respective subjects was rather moderate (M = 6.68, SD = 2.35), suggesting honest response behavior on the questionnaire.

Profile Discrepancies

Inspection of the resulting discrepancy profiles showing the mean rating of each item per class and teacher, resulted in a ranking of congruency for the 10 investigated classes. The profiles clearly differed in terms of class-teacher similarity (see **Figure 1** for three prototypic discrepancy profiles; **Annex 3** shows profiles for all 10 participating classes). The two classes taught by the same teacher (A and B) showed almost no differences (**Figure 1**, top) while Profile I showed substantial teacher-student disagreement (**Figure 1**, bottom). While four profiles (C to F) seemed to have only specific deviations (**Figure 1**, middle image), another four profiles (G to J) revealed a more general difference, with teachers rating their behaviors much higher than the corresponding students. With few exceptions, teachers judged their actions more positively than their students, even when the 95% confidence interval of the students' rating was considered. These cases clearly indicate substantial differences in perception of assessment-related behavior. Interestingly, even when teachers

showed an obvious "positive bias" toward their own assessment practices, the relative judgment of single aspects were often parallel to their students' ratings. Remarkably, 8 out of 10 teachers gave a differentiated judgment of the assessed aspects that more or less resembled their students'. Only Profiles G and I (but also with exceptions on oral exams) failed to show a differentiated judgment of the assessed aspects. This might be due to Teacher G's brief teaching experience of only 2 years. Thus, in most cases, teachers exhibit strong awareness of behaviors that could be improved (also from a students' perspective).

As already seen in the items' descriptive statistics, questions related to oral exams (fb2, fb3, and fb10) caused considerable disagreement, with substantial differences between student and teacher ratings in 6 out of 10 profiles. Especially for teachers who showed the highest perception differences (e.g., Class I, **Figure 1** bottom image), oral exams seemed to be a main source of disagreement. Note, however, that three teachers (A, D, F) rated themselves more critically on oral exam practices than their students did, pointing to the imprecise nature of oral exams in general. Profile C, however, showed the same phenomenon for written exams, with students feeling better informed than their teacher expected (**Figure 1**, middle image).

The question of whether teachers were open to comments about their grading system (fb7) caused substantial differences in all profiles. However, three teachers (four profiles, with one teaching two classes) judged themselves less open for feedback than their students did. Thus, regardless of whether there were almost no (A, B), partial (F), or general (G) deviations, or if teachers' level of experience was high (A, B), medium (F), or low (G), this very personal though central aspect of Procedural Fairness seemed to be difficult to judge.

One interesting result concerned perceived interest of the students in the given subject (fb16). Only two profiles (G and I) showed substantial deviations between students' and teachers' estimates of student interest, with Teacher G vastly underestimating her students' interest in the subject. Thus, the teachers in our study appraised their students fairly accurately. When asked about the overall fairness of the teachers' grading behavior, teachers and students mostly agreed on a very high level, showing self-confidence of the involved teachers and suggesting high satisfaction of the vast majority of their students.

To quantify profile discrepancies, we relied on two proven profile similarity scores, *r* (Kozlowski and Hattrup, 1992) and D^2 (e.g., Kristof, 1996). Note that contrary to the depicted profiles in **Figure 1** and **Annex 3**, those scores only included items of the Fairness Barometer (fb1–fb15). General interest and perceived fairness were omitted since they do not constitute distinct aspects of fairness. The consensus estimate *r* ranged from –0.27 (G; due to two underestimates on side of the teacher) to.84 (J). In most cases, it seemed to appropriately represent the similarity of ratings but appeared to be sensitive when several underestimates of the teacher occurred (C), especially when they were huge (G). Generally, few large synchronous differences inflated *r*: Although Teacher I almost always chose the highest rating, thus clearly deviating from the students' judgment, similar deviations on two items (fb2 and fb3) positively influenced the consensus score to be at $r = 0.57$.

D^2 reflected the qualitative classification of the profiles quite well. Profiles with almost no deviations (A, B) showed mean D^2 ranging from 0.30 to.40. Profiles with several specific differences (C to F) led to D^2 between 0.97 and 2.05, and huge profile discrepancies (G to J) featured D^2 from 1.56 to 5.29. Thus, it seems as if D^2 could be used as representative proxy of overall profile discrepancies. Note, however, that over- as well as underestimates on the side of the teacher equally contribute to this measure and that a look at the specific deviations is required to understand the situation in a specific class.

Crucially, both scores appeared to be independent from each other: Profiles A and B, showing almost identical curves, exhibited only medium sized consensus estimates (*r* between 0.55 and 0.67), whereas Profiles F (specific deviations) and J (largest D^2) produced estimates between $r = 0.79$ and 0.84. This, however, might be caused by restriction of range due to the ceiling effect on all items. As a consequence, *r* might be a valid consensus estimate only for profiles showing medium to large variance.

Relations Between Perceived Fairness and Student Variables

How students' perceived fairness is correlated with other central variables within its nomological net is presented in **Table 2**. Manifest measures of Informational Fairness (IF) and Procedural Fairness (PrF) as captured by the Fairness Barometer were substantially correlated ($r = 0.72$) but clearly distinct, replicating results from Colquitt (2001) and Sonnleitner and Kovacs (2018). Subscores of the Fairness Barometer (IF and PrF), as well as the total sum score (PF) were substantially related to students' one-item general fairness judgment of their teacher (*r* ranging from 0.49 to 0.63), but again, far from identical. This suggests that the nuanced evaluation of specific assessment-related aspects goes far beyond a diffuse student judgment of teacher fairness. This is somewhat supported by students' interest in the subject showing the highest relation to the global fairness judgment ($r = 0.44$) and lowest to IF ($r = 0.24$), rendering at least some aspects of the Fairness Barometer more robust against a positive bias grounded in a higher interest in the subject.

Concerning Distributive Justice, we found a small but significant relationship between short-term over-/underpayment (DJe) and perceived fairness, ranging from $r = -0.21$ (PrF) to $r = -0.24$ (IF), whereas no long-term effects (DJt) proved statistically significant. Thus, there could be a (plausible but not necessarily conscious) tendency of students to "pay back" perceived distributive injustice when rating their teacher's fairness. Note, however, that analysis of short-term effects of Distributive Justice included 14% "overpaid" students with no reason to bear a grudge against the teacher. Yet results replicate previously documented ties between Distributive Justice, Informational and Procedural Justice, as well as perceived fairness (e.g., Tata, 1999; Chory-Assad and Paulsel, 2004; Wendorf and Alexander, 2005).

Using the grades on the last exam and last term as a proxy for students' abilities in the given subjects, results show that higher student ability was related to a more positive rating of all assessment-related behaviors. However, since correlations

TABLE 2 | Correlations between Perceived Fairness, Distributive Justice, Interest in the Subject, and Teacher's experience (n = 168).

		IF	PrF	PF	GF	Int	DJe	DJt	Ge	Gt	age
(IF) Informational Fairness	Pearson's r	–									
	p-value	–									
(PrF) Procedural Fairness	Pearson's r	0.72	–								
	p-value	< 0.01	–								
(PF) Perceived Fairness sum	Pearson's r	0.90	0.95	–							
	p-value	< 0.01	< 0.01	–							
(GF) General Fairness	Pearson's r	0.49	0.63	0.62	–						
	p-value	< 0.01	< 0.01	< 0.01	–						
(Int) Interest in Subject	Pearson's r	0.24	0.41	0.36	0.44	–					
	p-value	< 0.01	< 0.01	< 0.01	< 0.01	–					
(DJe) Distributive Justice exam	Pearson's r	– 0.24	– 0.21	– 0.22	– 0.16	0.04	–				
	p-value	< 0.01	0.01	< 0.01	0.05	0.67	–				
(DJt) Distributive Justice term	Pearson's r	0.01	– 0.12	– 0.07	– 0.02	– 0.03	0.12	–			
	p-value	0.95	0.16	0.38	0.81	0.75	0.14	–			
(Ge) Grade last exam	Pearson's r	– 0.19	– 0.32	– 0.29	– 0.31	– 0.17	0.25	0.18	–		
	p-value	0.02	< 0.01	< 0.01	< 0.01	0.04	< 0.01	0.03	–		
(Gt) Grade last term	Pearson's r	– 0.14	– 0.23	– 0.22	– 0.20	– 0.17	0.22	0.28	0.86	–	
	p-value	0.09	0.01	0.01	0.01	0.05	0.01	< 0.01	< 0.01	–	
(age) Students' age	Pearson's r	0.11	0.05	0.08	0.06	– 0.01	– 0.09	– 0.15	– 0.04	– 0.03	–
	p-value	0.18	0.54	0.31	0.47	0.98	0.27	0.06	0.60	0.77	–

between grades and procedural assessment aspects (PrF) were higher (r = –0.32 and –0.23) compared to informational activities (IF, r = –0.19 and –0.14), this result could also be read as students' being (truly) rated below their ability, and now pinpointing the reasons for this unfair treatment in the questionnaire. Another line of interpretation could be that better students are better able to judge certain aspects of their teachers' efforts to guarantee fair assessment procedures.

Importantly, similar to the consensus index r, these correlations might be a lower bound estimate due to the restriction of range caused by the observed ceiling effects on the items of the Fairness Barometer.

DISCUSSION

The present explorative study set out to investigate the potential of the Fairness Barometer (Sonnleitner and Kovacs, 2018), a self-administered questionnaire for teacher self-improvement. Comparing teachers' perspectives on assessment-related behaviors with the perspectives of their students should deliver useful leads on what could be improved in daily classroom assessment. Such an easy-to-use tool could help improve at least some aspects of teachers' assessment literacy (e.g., Xu and Brown, 2016; Pastore and Andrade, 2019) and would leverage students' sensitivity to unfair evaluation and grading practices (e.g., Chory-Assad and Paulsel, 2004; Wendorf and Alexander, 2005). To this end, we administered the Fairness Barometer in 10 classes of upper secondary schools, thus gathering the perceptions of 9 different teachers and comparing them to their students' estimates (n = 168) in specific discrepancy profiles (**Figure 1**).

Potential of Assessing Differences in Fairness Perceptions

The leading research question addressed whether the resulting profiles would yield meaningful interpretations that could be used by teachers to improve aspects of their assessment practices. Results clearly show the potential of this approach. Although students in this sample overall were satisfied or highly satisfied with most of their teachers' assessment and grading related behavior, 8 out of 10 profiles indicated areas of improvement by showing substantial deviations in perception at least on some aspects. Teachers observing such discrepancies might reflect on these behaviors or even discuss and clarify reasons for the differing perceptions with their students or colleagues. Thus, a directed dialogue or search for strategies to improve the behaviors in question could be triggered by the results. Even teachers receiving optimal results with an overlap of perceptions at a very high rating, like the teacher shown at the top of **Figure 1**, still show areas of discrepancy which could be pedagogically interesting (i.e., this teacher seems to see himself as less open to comments than his students perceive him to be). The high congruency between the two classroom profiles of this specific teacher (see **Annex 3**, Profiles A and B) might also point to a certain stability of his assessment practices, leading to equally positive evaluations in different classes. This result points to a highly interesting question in its own right: whether teachers' assessment behavior is stable across classes and leads to similar perceptions in students.

The case of teachers showing general perception differences (e.g., Teacher J, shown in **Figure 1** bottom image) is more complex, since such negative feedback could be viewed as a threat to their assessment-related self-concept. Thus, they might easily dismiss these differences, externally attributing

them to misperceptions or negative attitudes of their students. Whether teachers view such profiles as constructive feedback remains to be seen. One reason for creating the Fairness Barometer as a self-assessment tool is that teachers who are interested in using the questionnaire at all are also likely to be motivated to improve their own assessment practices by interpreting results constructively, not defensively. It seems likely that more in-depth, guided reflection than can be offered by the Fairness Barometer (e.g., coaching, workshops, or peer-related interventions) would be needed to raise awareness of extremely problematic assessment behaviors.

As seen in the correlations of perceived fairness with Distributional Justice and students' grades, there might be a small effect of students (externally) attributing their weaker performance to flawed teacher behavior. Such attribution strategies are well documented but could also be explained by reciprocal causation (for an example on the relation between boredom and academic achievement, see Pekrun et al., 2014): If students perceive grading procedures to be unfair, they might reduce their efforts in resignation or protest. This in turn, however, would lead to even poorer performance, leading to worse ratings that might be perceived as even more unjust. Another explanation is that stronger students might better be able to accurately judge the quality of their own performance (e.g., Kruger and Dunning, 1999) and thus accurately judge teachers' efforts to guarantee fairness. In any case, teachers should be critically aware of this effect when receiving low fairness ratings. Clearly, further research extending students' self-reported data by considering multiple sources (e.g., grades gathered from school administration) is needed to clarify the Fairness Barometer's sensitivity to these issues.

The second research question explored whether derived profile similarity scores would be meaningful representations of the overall profile. The consensus estimate *r*, as well as D^2 reflected the qualitative interpretation of the profiles quite well. Both indicators reasonably accounted for profile variance in similarity and distance of perceptions, although the ceiling effect on most items might have restricted the range of *r*. Given sufficiently trustworthy benchmarks (i.e., based in a larger, representative sample of classes), such profile similarity indices could help teachers compare their own profiles to typical consensus or distance scores in the respective grade or subject they are teaching. From a research perspective, it might be interesting to discover whether these scores are related to teacher experience or student interest and motivation. The impact of interventions or training programs on teacher assessment literacy might be observed on a larger scale using such scores. Such large-scale use of the Fairness Barometer is quite interesting from an educational research perspective, though it also carries very real practical dangers. On the one hand, operationalizing a few (important but limited) aspects of fair classroom assessment through a 15-item questionnaire runs the risk of reducing the issue of classroom justice to one limited quantitative measure, not to mention the potential negative effects of using such a measure normatively as a lone indicator for assessment fairness in an accountability context (e.g., Nichols and Berliner, 2005). On the other hand, developing a "quick and dirty" tool to assess fairness makes it more likely that this construct will be considered at all in the broad social debates and system reforms resulting from large-scale educational research. A full discussion of these issues is beyond the scope of the current paper, but we do feel that large-scale research applications of the Fairness Barometer should at least be empirically explored.

Limitations and Outlook

As already indicated, one limitation of the current study is the small number of classrooms studied. Due to the convenience sampling strategy, the final results also cannot be seen as representative. However, as an explorative study intended to gain a first impression of the profile patterns to be expected when administering the Fairness Barometer to students and teachers within the same classroom, we felt this to be acceptable. We see two directions in which future studies need to travel in order to test and strengthen the tentative claims we make based on our current results. On the one hand, it is necessary to gather more detailed information from teachers on the practical usefulness of discrepancy-profile feedback. Though the Fairness Barometer was developed with a strong view to practical considerations and in communication with (preservice and practicing) teachers, our assumption that it can stimulate reflection among teachers about their own assessment practices and thus improve assessment is still hypothetical. Future studies need to explore how teachers really respond to such profile feedback and how they are able – or even want – to navigate interpretation of results on their own. Such research might explore strategies that go beyond quantitative feedback in encouraging teachers to reflect on their own assessment practices and helping students and teachers negotiate a fair classroom.

In the course of such research, it is also important to explore the limitations of the Fairness Barometer in capturing the full breadth of fair assessment practices in the classroom. Due to their importance in determining students' overall grades in several European countries, the FB focuses on formal exam contexts (written and oral). The frequency of such exams and the extent to which their results determine final grades, however, is highly dependent on the specific educational system studied. Teachers themselves also have some leeway in determining how strongly different types of assessment are employed, though this freedom also varies by educational system and country. Future revisions or translations of the Fairness Barometer must consider typical assessment practices of the country being studied so that questions about rare assessment practices (e.g., oral exams) might be replaced with more relevant content (e.g., grading of homework or portfolios). Thus, future studies might explore whether the Fairness Barometer and hence the measured concept of fairness can be expanded while still focusing only on aspects that students can reliably evaluate and teachers can change. We also want to stress that our working definition of fairness mostly focused on classroom justice theory and that broader conceptualizations are possible, and possibly–depending on the context–preferable (e.g., Gipps and Stobart, 2009; Camilli, 2013); Tierney (2013, 2014) and Rasooli et al. (2018).

A second direction of possible future research involves the attempt to integrate aspects of fairness perception into large-scale educational studies in order to facilitate more general conclusions about the role of assessment fairness in education. As an indicator on the student level, variations in perceived fairness seem likely to predict various relevant outcomes (as suggested both by the present study and by the classroom justice research cited in Section "Students and Their Perceptions of Assessment"). However, on the classroom level, not only overall fairness perceptions but also the degree of similarity between students' and teachers' fairness perception profiles are likely to have substantial impact on those outcomes. After all, effectively communicating about assessment practices as well as understanding and responding to students' perceptions of those practices are all aspects of assessment literacy and thus can be seen, in a broad sense, as an aspect of teaching competency. Profile similarity indicators such as D^2 or r might be quite useful as new indicators of an important facet of instructional quality. In addition to this simplified, variable-based approach, a person-centered statistical approach might also be quite fruitful. By gathering difference profiles from a large number of classrooms, it may be possible to identify typical profile patterns using cluster analysis or latent mixture modeling. Such a "fairness typology" might capture interactions ignored by single–variable discrepancy indicators such as D^2 in predicting relevant outcomes and might shed further light on the dynamics between teacher maturity, teaching experience, and their impact on students' perceived fairness.

In sum, looking at teachers' assessment and grading behaviors from two perspectives–the teachers' as well as their students'–showed promising potential for improving aspects of assessment literacy. We want to stress, however, that using the Fairness Barometer beyond discrepancy profile interpretations would require further research and evidence on the validity of its scores. We have made the Fairness Barometer freely available online in the hopes that interested teachers and researchers will use the questionnaire in their own classrooms and studies, leading to a better understanding of the instrument's practical as well as scientific uses and limitations.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

PS planned the study, coordinated the data collection, and took the lead in writing the manuscript. PS and CK jointly worked on analyzing and interpreting the results. CK substantially contributed to writing the manuscript.

ACKNOWLEDGMENTS

We want to thank all participating students and teachers, and our colleagues Dr. Tanja Baudson, and Ulrich Keller for their helpful and valuable comments. We also want to explicitly thank RT for offering substantial and constructive feedback on earlier versions of the manuscript.

REFERENCES

Ainscow, M., Beresford, J., Harris, A., Hopkins, D., Southworth, G., and West, M. (2013). *Creating the Conditions for School Improvement: A Handbook of Staff Development Activities*. New York, NY: Routledge.

Alkharusi, H. A. (2011). An analysis of the internal and external structure of the teacher assessment literacy questionnaire. *Int. J. Learn.* 18, 515–528. doi: 10.18848/1447-9494/CGP/v18i01/47461

American Educational Research Association, [AERA], National Council on Measurement in Education [NCME], American Psychological Association [APA], and Joint Committee on Standards for Educational and Psychological Testing (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

American Federation of Teachers [AFT], National Council on Measurement in Education, [NCME], and National Education Association [NEA] (1990). Standards for teacher competence in educational assessment of students. *Educ. Meas.* 9, 30–32.

Battistone, W., Buckmiller, T., and Peters, R. (2019). Assessing assessment literacy: are new teachers prepared to assume jobs in school districts engaging in grading and assessment reform efforts? *Stud. Educ. Eval.* 62, 10–17. doi: 10.1016/j.stueduc.2019.04.009

Blount, H. P. (2016). The keepers of numbers: teachers' perspectives on grades. *Educ. Forum.* 61, 329–334. doi: 10.1080/00131729709335278

Brown, G. T. L., Gebril, A., and Michaelides, M. P. (2019). Teachers' conceptions of assessment: a global phenomenon or a global localism. *Front. Educ.* 4:1–13. doi: 10.3389/feduc.2019.00016

Camilli, G. (2013). Ongoing issues in test fairness. *Educ. Res. Eval.* 19, 104–120. doi: 10.3402/meo.v20.28821

Chory, R. M. (2007). Enhancing student perceptions of fairness: the relationship between instructor credibility and classroom justice. *Commun. Educ.* 56, 89–105. doi: 10.1080/03634520600994300

Chory-Assad, R. M., and Paulsel, M. L. (2004). Classroom justice: student aggression and resistance as reactions to perceived unfairness. *Commun. Educ.* 53, 253–273. doi: 10.1080/0363452042000265189

Chory-Assad, R. M. (2002). Classroom justice: perceptions of fairness as a predictor of student motivation, learning, and aggression. *Commun. Q.* 50, 58–77. doi: 10.1080/01463370209385646

Cobb, P., Wood, T., and Yackel, E. (1990). Chapter 9: classrooms as learning environments for teachers and researchers. *J. Res. Math. Educ. Monogr.* 4, 125–210.

Colquitt, J. A. (2001). On the dimensionality of organizational justice: a construct validation of a measure. *J. Appl. Psychol.* 86, 386–400. doi: 10.1037/0021-9010.86.3.386

Colquitt, J. A., and Shaw, J. (2005). "How should organizational justice be measured," in *Handbook of Organizational Justice*, Vol. 1, eds J. Greenberg, and J. A. Colquitt (Mahwah, NJ: Erlbaum), 113–152.

Duplaga, E. A., and Astani, M. (2010). An exploratory study of student perceptions of which classroom policies are fairest. *Decision Sci. J. Innov. Educ.* 8, 9–33. doi: 10.1111/j.1540-4609.2009.00241.x

Gipps, C., and Stobart, G. (2009). "Fairness in assessment," in *Proceeding of the Educational Assessment in the 21st Century* (Dordrecht: Springer), 105–118. doi: 10.1007/978-1-4020-9964-9_6

Gotch, C. M., and French, B. F. (2014). A systematic review of assessment literacy measures. *Educ. Meas. Issues Pract.* 33, 14–18. doi: 10.1111/emip.12030

Kaiser, J., Südkamp, A., and Möller, J. (2017). The effects of student characteristics on teachers' judgment accuracy: disentangling ethnicity, minority status, and achievement. *J. Educ. Psychol.* 109, 871–888. doi: 10.1037/edu0000156

Klieme, E., Döbert, H., van Ackeren, I., Bos, W., Klemm, K., Lehmann, R., et al. (2007). *Vertiefender Vergleich der Schulsysteme ausgewählter PISA-Teilnehmerstaaten: Kanada, England, Finnland, Frankreich, Niederlande, Schweden. [In-Depth Comparison of School Systems of Selected PISA Participating Countries: Canada, England, Finland, France, The Netherlands, Sweden.]* Berlin, Germany: BMBF.

Köller, O., Baumert, J., and Schnabel, K. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *J. Res. Math. Educ.* 32, 448–470.

Kozlowski, S. W. J., and Hattrup, K. (1992). A disagreement about within-group agreement: disentangling issues of consistency versus consensus. *J. Appl. Psychol.* 77, 161–167. doi: 10.1037/0021-9010.77.2.161

Kristof, A. L. (1996). Person-organization fit: an integrative review of its conceptualizations, measurement, and implications. *Pers. Psychol.* 49, 1–49. doi: 10.1111/j.1744-6570.1996.tb01790.x

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Personal. Soc. Psychol.* 77, 1121–1134. doi: 10.1037/0022-3514.77.6.1121

Lizzio, A., and Wilson, K. (2008). Feedback on assessment: students' perceptions of quality and effectiveness. *Assess. Eval. High. Educ.* 33, 263–275. doi: 10.1080/02602930701292548

Mertler, C. A., and Campbell, C. (2005). Measuring teachers' knowledge and application of classroom assessment concepts: development of the assessment literacy inventory. *Paper Presented at the Annual Meeting of the American Educational Research Association*, Montreal, Canada.

Nesbit, P. L., and Burton, S. (2006). Student justice perceptions following assignment feedback. *Assess. Eval. High. Educ.* 31, 655–670. doi: 10.1080/02602930600760868

Nichols, S. L., and Berliner, D. C. (2005). *The Inevitable Corruption of Indicators and Educators through High-Stakes Testing*. Available at: http://files.eric.ed.gov/fulltext/ED508483.pdf (accessed December 20, 2019).

Pastore, S., and Andrade, H. L. (2019). Teacher assessment literacy: a three-dimensional model. *Teach. Teach. Educ.* 84, 128–138. doi: 10.1016/j.tate.2019.05.003

Pekrun, R., Hall, N. C., Goetz, T., and Perry, R. P. (2014). Boredom and academic achievement: testing a model of reciprocal causation. *J. Educ. Psychol.* 106, 696–710. doi: 10.1037/a0036006

Rasooli, A., Zandi, H., and DeLuca, C. (2018). Re-conceptualizing classroom assessment fairness: a systematic meta-ethnography of assessment literature and beyond. *Stud. Educ. Eval.* 56, 164–181. doi: 10.1016/j.stueduc.2017.12.008

Rasooli, A., Zandi, H., and DeLuca, C. (2019). Conceptualising fairness in classroom assessment: exploring the value of organisational justice theory. *Assess. Educ.* 26, 584–611. doi: 10.1080/0969594x.2019.1593105

Schön, D. (1983). *The Reflective Practitioner: How Professional Think in Action*. New York, NY: Basic Books.

Shin, L. W. (2015). Teachers ' assessment literacies and practices: developing a professional competency and learning framework. *Adv. Scholarsh. Teach. Learn.* 2, 1–20.

Sonnleitner, P., and Kovacs, C. (2018). "How fairly am I assessing my students? Equipping teachers with a tool to learn about their own assessment practices: Theory and development of the Fairness Barometer," in *Paper Presented at the 11th Conference of the International Test Commission*, Montreal, MTL.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan* 72, 534–539.

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

Tata, J. (1999). Grade distributions, grading procedures, and students' evaluations of instructors: a justice perspective. *J. Psychol.* 133, 263–271. doi: 10.1080/00223989909599739

The Jamovi Project (2019). Jamovi. (Version 1.0) [Computer Software]. doi: 10.1080/00223989909599739

Tierney, R. (2013). "Fairness in classroom assessment," in *Sage Handbook of Research on Classroom Assessment*, ed. J. H. McMillan (Thousand Oaks, CA: SAGE Publications), 125–144.

Tierney, R. (2014). Fairness as a multifaceted quality in classroom assessment. *Stud. Educ. Eval.* 43, 55–69. doi: 10.1016/j.stueduc.2013.12.003

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

Vallade, J. I., Martin, M. M., and Weber, K. (2014). Academic entitlement, grade orientation, and classroom justice as predictors of instructional beliefs and learning outcomes. *Commun. Q.* 62, 497–517. doi: 10.1080/01463373.2014.949386

Volante, L., and Fazio, X. (2007). Exploring teacher candidates' assessment literacy: implications for teacher education reform and professional development. *Can. J. Educ.* 30, 749–770.

Wendorf, C. A., and Alexander, S. (2005). The influence of individual- and class-level fairness-related perceptions on student satisfaction. *Contemp. Educ. Psychol.* 30, 190–206. doi: 10.1016/j.cedpsych.2004.07.003

Xu, Y., and Brown, G. T. L. (2016). Teacher assessment literacy in practice: a reconceptualization. *Teach. Teach. Educ.* 58, 149–162. doi: 10.1016/j.tate.2016.05.010

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.

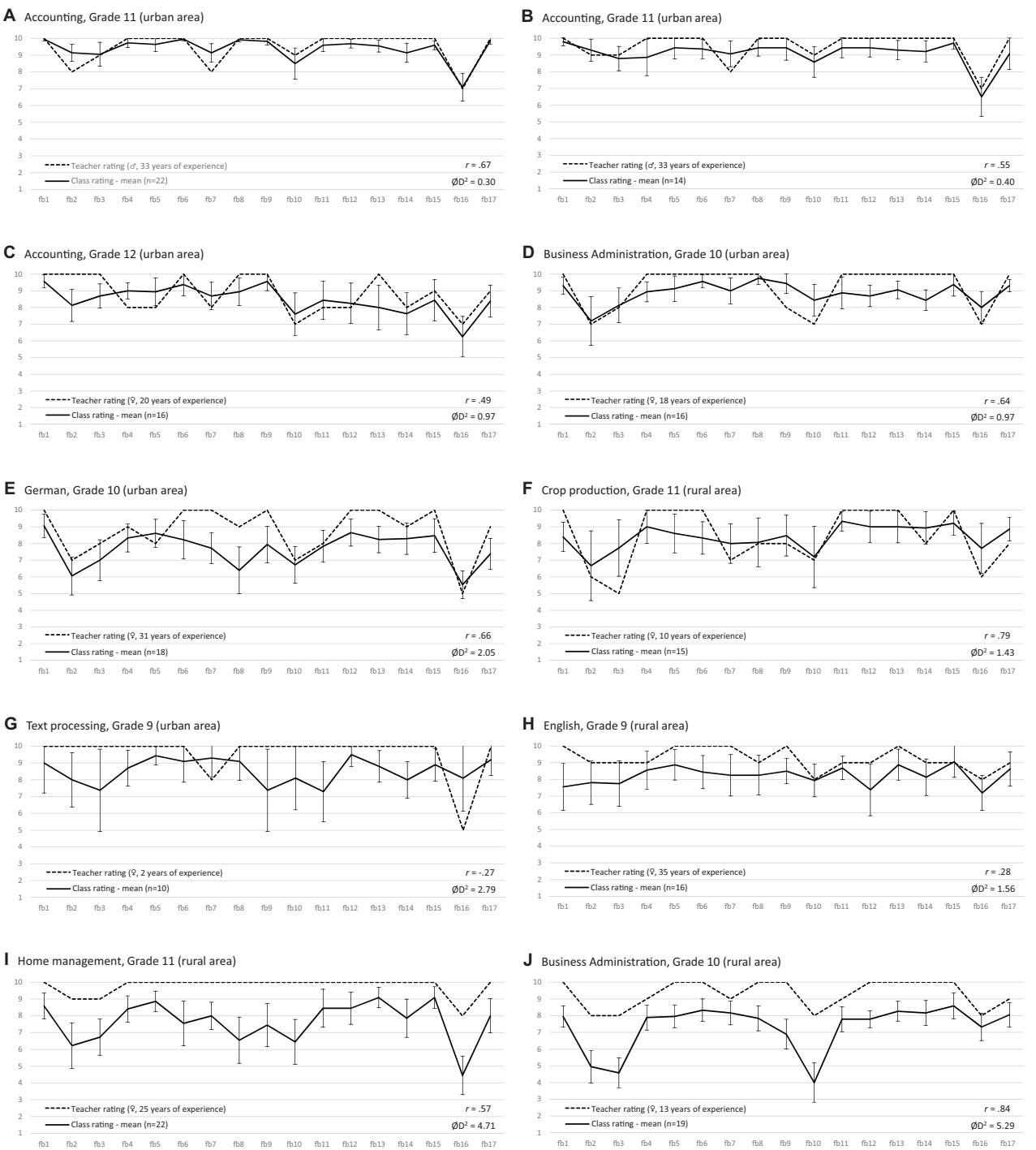
ANNEX

ANNEX 1 | Teacher version of the Fairness Barometer.

	Item content
Informational fairness	
fb1t	The content of the exam is announced on time.
fb2t	My students know what criteria are used to assess oral exams.
fb3t	My students understand their own grades on oral exams.
fb4t	My students know what criteria are used to assess written exams.
fb5t	My students understand their own grades on written exams.
fb6t	If students ask, I will explain their grade to them.
Procedural fairness	
fb7t	I am open to comments about my grading system.
fb8t	Grading criteria are applied equally to everyone in the class (unless there is a justified exception).
fb9t	Students' current achievements are graded independently of the grades they have had in the past.
fb10t	The oral exams in class include enough questions for students to show what they know and what they can do.
fb11t	The written exams in class include enough questions for students to show what they know and what they can do.
fb12t	During written exams I allow enough time to complete the given questions/tasks.
fb13t	The questions/tasks included in exams are an accurate reflection of the material that has been taught in class.
fb14t	The difficulty of exam questions/tasks is appropriate.
fb15t	The exams only test material that has been taught in class.
General ratings	
fb16t	How strong do you think your students' interest in this school subject is?
fb17t	How fairly do you think you grade student performance in this subject?

ANNEX 2 | German version of the Fairness Barometer.

	Item content – student version	Item content – teacher version
Informationale fairness		
fb1	Der Prüfungsstoff wird jeweils rechtzeitig bekannt gegeben.	fb1t Der Prüfungsstoff wird jeweils rechtzeitig bekannt gegeben.
fb2	Die Beurteilungskriterien für mündliche Prüfungen sind mir bekannt.	fb2t Die Beurteilungskriterien für mündliche Prüfungen sind meinen SchülerInnen bekannt.
fb3	Meine Beurteilungen (Noten/erreichte Punkteanzahl) auf mündliche Prüfungen sind für mich nachvollziehbar.	fb3t Die Beurteilungen (Noten/erreichte Punkteanzahl) auf mündliche Prüfungen sind für meine SchülerInnen nachvollziehbar.
fb4	Die Beurteilungskriterien für schriftiche Prüfungen sind mir bekannt.	fb4t Die Beurteilungskriterien für schriftiche Prüfungen sind meinen SchülerInnen bekannt.
fb5	Meine Beurteilungen (Noten/erreichte Punkteanzahl) auf schriftliche Prüfungen sind für mich nachvollziehbar.	fb5t Die Beurteilungen (Noten/erreichte Punkteanzahl) auf schriftliche Prüfungen sind für meine SchülerInnen nachvollziehbar.
fb6	Wenn ich nachfrage, wird mir meine Beurteilung (Note/erreichte Punkteanzahl) erklärt.	fb6t Bei Nachfragen, erkläre ich die jeweilige Beurteilung (Note/erreichte Punkteanzahl).
Prozedurale fairness		
fb7	Meine Lehrerin/mein Lehrer ist offen gegenüber Anmerkungen zur Leistungsbeurteilung.	fb7t Ich bin offen gegenüber Anmerkungen zur Leistungsbeurteilung.
fb8	Die Beurteilungskriterien werden auf alle in meiner Klasse gleich angewandt (ausgenommen begründete Ausnahmen)	fb8t Die Beurteilungskriterien werden auf alle in der Klasse gleich angewandt (ausgenommen begründete Ausnahmen)
fb9	Meine aktuellen Leistungen (Prüfungsnote/erreichte Punkteanzahl) werden unabhängig von früheren Leistungen beurteilt.	fb9t Aktuelle Leistungen (Prüfungsnote/erreichte Punkteanzahl) werden unabhängig von früheren Leistungen beurteilt.
fb10	In mündlichen Prüfungen werden ausreichend Fragen gestellt um zu zeigen, was ich kann und weiß.	fb10t Bei mündlichen Prüfungen stelle ich ausreichend viele Fragen, damit die SchülerInnen zeigen können, was sie können und wissen.
fb11	In schriftlichen Prüfungen werden ausreichend Fragen gestellt um zu zeigen, was ich kann und weiß.	fb11t Bei schriftlichen Prüfungen stelle ich ausreichend viele Fragen, damit die SchülerInnen zeigen können, was sie können und wissen.
fb12	Bei schriftlichen Prüfungen habe ich ausreichend Zeit, die gestellten Aufgaben zu bearbeiten.	fb12t Bei schriftlichen Prüfungen gebe ich ausreichend Zeit, um die gestellten Aufgaben zu bearbeiten.
fb13	Die in Prüfungen gestellten Fragen/Aufgaben spiegeln den unterrichteten Stoff gut wider.	fb13t Die in Prüfungen gestellten Fragen/Aufgaben spiegeln den unterrichteten Stoff gut wider.
fb14	Die Schwierigkeit der in den Prüfungen gestellten Fragen/Aufgaben ist angemessen.	fb14t Die Schwierigkeit der in den Prüfungen gestellten Fragen/Aufgaben ist angemessen.
fb15	Bei den Prüfungen wird nur Unterrichtsstoff abgefragt, den wir bereits durchgenommen haben.	fb15t Bei den Prüfungen wird nur Unterrichtsstoff abgefragt, den wir bereits durchgenommen haben.



ANNEX 3 | Profile discrepancies ($ØD^2$) and correlations (*r*) between teacher and student ratings on all 15 items of the Fairness Barometer for 10 different classes **(a–j)**. Error bars indicate 95% CI. Interest in the subject (fb16) and perceived general fairness (fb17) were excluded from computation of $ØD^2$ and *r*.