Running head: CONSENSUS IN STUDENT'S PERCEPTIONS

Is everyone in class in agreement and why (not)? Using student and teacher reports to predict within-class consensus on goal structures

Lisa Bardach[1], Takuya Yanagida[2], Alexandre J. S. Morin[3], Marko Lüftenegger[2,4 *]

[1] Lisa Bardach, Department of Education, University of York, United Kingdom
[2] Takuya Yanagida, Marko Lüftenegger,
Faculty of Psychology, Department of Developmental and Educational Psychology, University of Vienna, Austria
[3] Alexandre J.S. Morin, Substantive-Methodological Synergy Research Laboratory, Department of Psychology, Concordia University, Canada
[4] Marko Lüftenegger, Centre for Teacher Education, Department for Teacher Education, University of Vienna, Austria

Acknowledgements: The 3rd author was supported by a grant from the Social Science and Humanity Research Council of Canada (435-2018-0368).

*Correspondence concerning this article should be addressed to Marko Lüftenegger, University of Vienna, Faculty of Psychology, Department of Developmental and Educational Psychology, and Centre for Teacher Education, Department for Teacher Education, Porzellangasse 4, Wien, 1090, Phone: +43-1-4277-60030, Email: marko.lueftenegger@univie.ac.at

This document is a pre-publication version of the following manuscript:

Bardach, L., Yanagida, T., Morin, A.J.S., & Lüftnegger, M. (2021; accepted 31 August 2020). Is everyone in class in agreement and why (not)? Using student and teacher reports to predict within-class consensus on goal structures. *Learning & Instruction, 71*, 101400. doi: 10.1016/j.learninstruc.2020.101400

Abstract

Students in the same class can differ considerably in their perceptions of teaching quality, but little is known about the drivers of this lack of agreement. In this study, we relied on within-class consensus as a measure of agreement in students' perceptions and explored both student-reported predictors (Sample 1) and teacher-reported predictors (Sample 2) of consensus in students' ratings of six goal structures dimensions (task, autonomy, recognition, grouping, evaluation, time). Classroom-level results from multi-level models indicated that student-perceived differential teacher treatment negatively predicted consensus regarding evaluation, as well as achievement heterogeneity, whereas instructional clarity positively predicted consensus for recognition, grouping, and evaluation, as well as class-average achievement (Sample 1). Mediating effects of achievement heterogeneity and achievement were not statistically significant. In Sample 2, negative effects of teacher-reported emotional exhaustion and teaching-related anxiety on consensus were observed for several dimensions. Teaching-related enjoyment had no effect.

Keywords: Within-class consensus, goal structure, teaching quality, differential teacher treatment, teacher emotion, emotional exhaustion, achievement

In educational research, student surveys capturing aspects of teaching quality have become widely used, both for research purposes and to guide educational practice and policy (e.g., as part of teacher evaluation systems). Relying on student surveys to gain insights into teachers' instructional practices seems to be a worthwhile strategy, as students can offer unique information that is difficult to glean from other sources (e.g., Schenke, Ruzek, Lam, Karabenick, & Eccles, 2018) and can be considered experts due to their exposure to different teachers in different subjects over the course of their school careers (e.g., Clausen, 2002; De Jong & Westerhof, 2001). However, prior research suggests that students in the same class who rate the same teacher tend to vary greatly in their perceptions (e.g., Bardach, Lüftenegger, Yanagida, Spiel, & Schober, 2019b; Schweig, 2017), so that most of the variance in students' perceptions exists within classes rather than across different classes (e.g., Kunter et al., 2008). This represents a major dilemma, as both research on student-perceived teaching quality and practical applications such as the use of student surveys to determine teacher effectiveness typically use class-level aggregates supposed to map the shared perceptions of students from a given class (e.g., Lüdtke, Robitzsch, Trautwein, & Kunter, 2009). Hence, if (aggregated) student perceptions of teaching quality continue to be used in research and as policy-relevant markers, it is imperative to learn more about within-class variability in students' perceptions, and to identify factors that drive this variability (see also e.g., Schenke et al., 2018; Schweig, 2016).

One way to approach within-class variability in students' perceptions is to focus on the level of agreement among students from the same class (within-class consensus, e.g., Lüdtke, Trautwein, Kunter, & Baumert, 2006; Schweig, 2016). While some empirical evidence has been obtained on factors related to within-class consensus, such as educational track (Lüdtke et al., 2006) or age (Gärtner, 2010), our knowledge of factors that influence consensus remains fairly limited, particularly in regard to (more) malleable factors such as teachers' instructional approaches, as well as in relation to teacher characteristics. The present study aims to bridge this gap. Relying on goal structures as a theoretical framework to study teaching quality, we investigated a range of factors potentially related to within-class consensus on the six goal structure dimensions of task, autonomy, recognition, grouping, evaluation, and time (e.g., Ames, 1992; Patrick, & Ryan, 2008) in mathematics classes. First, drawing on a large sample of secondary school students (Sample 1), we examined student-perceived differential teacher treatment of higher vs. lower achieving students (e.g., Rubie-Davies, 2007) and the clarity of teachers' instruction (e.g., Hattie, 2012) as two potential predictors of within-class consensus on goal structures ratings (e.g., Bardach, Yanagida, Schober, & Lüftenegger, 2018; Lüdtke et al., 2006; Schweig, 2016). In addition to the direct effects of differential teacher treatment and instructional clarity, we further explored whether (a) the relation between clarity and consensus was mediated by class average achievement (e.g., Bardach et al., 2019b) and whether (b) variability in achievement among students from the same class serves as a mediator of the relation between differential teacher treatment and consensus. Second, we also explored the role of teacher characteristics (Sample 2), such as teacher-reported emotional exhaustion (Maslach, Schaufeli, & Leiter, 2001; Seiz, Voss, & Kunter, 2015) and teaching-related emotions of enjoyment and anxiety (e.g., Frenzel et al., 2016), in the prediction of within–class consensus. To complement the focus on student-reported differential teacher treatment, we also considered teachers' perceptions of student heterogeneity, especially with regard to achievement (e.g., Skaalvik & Skaalvik, 2016).

Theoretical Framework: Goal Structures

Following previous research on within-class consensus (e.g., Bardach et al., 2018, 2019b), this study assesses consensus in relation to goal structures. The concept of goal structures has evolved to represent a comprehensive and application-oriented aspect of teaching quality (e.g., Bergsmann, Lüftenegger, Jöstl, Schober, & Spiel, 2013). According to a recent meta-analysis (Bardach, Oczlon, Pietschnig, & Lüftenegger, 2019), goal structures can encompass teachers' policies and practices within a learning environment that helps to make different achievement goals more or less salient, the goal-related messages teachers explicitly communicate to their students, and the more general goal-related climate that pervades a learning environment (see also Ames, 1992; Urdan, 2010). In the present study, we rely on a

conceptualization of goal structures aligned with the first component of the above definition, referring to specific sets of teachers' practices and policies (Bardach, Oczlon et al., 2019; see also Ames, 1992; Benning et al., 2019; Lüftenegger, Tran, Bardach, Schober, & Spiel, 2017).

Nowadays, most researchers distinguish between three types of goal structures. In a *mastery goal structure*, teachers emphasize the importance of learning, understanding, and trying hard, and emphasize competence development as the main objective. In contrast, when a *performance avoidance goal structure* prevails, teachers communicate to students that it is essential not to demonstrate lower performance than their peers. Finally, a *performance approach goal structure* describes a classroom setting in which students are valued for outperforming peers or showcasing their abilities (e.g., Midgley et al., 2000; Murayama & Elliot, 2011; Miller & Murdock, 2007; Patrick, Kaplan, & Ryan, 2011; Schwinger & Stiensmeier-Pelster, 2011). Over the past 30 years, substantial information on these different types of goal structures and their correlates has been accumulated. For instance, research has shown mastery goal structures to be most beneficial for students, as indicated by associations with adaptive motivation patterns, positive emotions, and positive student-teacher relationships (e.g., Baudoin & Galand, 2017; Lüftenegger, van de Schoot, Schober, Finsterwald, & Spiel, 2014; Polychroni, Hatzichristou, & Sideridis, 2012). In contrast, performance avoidance goal structures have been identified as being least adaptive for students, displaying associations with a wide array of detrimental outcomes, such as maladaptive motivation patterns (e.g., Schwinger & Stiensmeier-Pelster, 2011). Conclusions are not as clear for performance approach goal structures. Thus, whereas performance approach goal structures have been shown to mainly produce maladaptive effects (e.g., Murayama & Elliot, 2009), benefits have also been reported (e.g., autonomous motivation: Peng, Cherng, Chen, & Ling, 2013). Nonetheless, the present study focuses on mastery goal structures, given the fact that such structures are unambiguously the most advantageous for students (for reviews, see e.g., Bardach, Oczlon, et al., 2019; Meece et al., 2006; Urdan & Schönfelder, 2006).

In an attempt to identify core dimensions of instructional practices involved in the establishment of a mastery goal structure, researchers have proposed the multi-dimensional TARGET framework, which encompasses six dimensions (task, autonomy, recognition, grouping, evaluation, time). *Task* refers to the design of tasks, which should be personally meaningful, challenging, and stimulate students' curiosity to learn more. *Autonomy* describes the involvement of students in decision-making regarding learning and interpersonal issues, such as the establishment of classroom rules. *Recognition* practices include, for instance, feedback and praise. The *grouping* dimension represents procedures used to group students, which should be flexible and designed in a way that values and fosters cooperation. *Evaluation* refers to the practices employed to assess and monitor students' individual progress, rather than to compare students with one another. Finally, *time* captures the use of time, such as devoting time to student questions or to in-depth explanations (Ames, 1992; Epstein, 1988; Meece et al., 2006; Lüftenegger et al., 2014; Lüftenegger et al., 2017).

Within-Class Consensus on Teaching Quality

Methodologically, a focus on students' evaluations of the quality of a teacher's instruction necessitates a multilevel representation to allow teaching quality to be modelled as a feature that varies between classes (e.g., Lüdtke et al., 2009). The units of interest are the ratings of individual students within each class, aggregated to the classroom level to represent students' shared perceptions of teaching quality. However, if we treat the class as the unit of analysis, then there should be overlap among the perceptions of those comprising this unit, i.e., the students within a class. As they all rate the same classroom teacher (or another classroom feature and thus the same target), their mental representation of this target should share at least some similarities (Marsh et al., 2009; Marsh et al., 2012; Morin, Marsh, Nagengast, & Scalas, 2014). Specifically, sufficient overlap in perceptions of teaching quality, which can be quantified by consensus measures, has been discussed as a precondition to the aggregation of individual student ratings at the classroom level. As such, several researchers highlight the need to consider consensus when investigating teaching quality and other group-level phenomena (e.g., Lüdtke et al., 2006; Nelson & Christ,

2016). In prior studies of within-class consensus focusing on goal structures and using the same measure of consensus as those used in the present investigation, within-class consensus was generally found to vary from weak ($r^*_{wg(J)} \geq .30$) to moderate ($r^*_{wg(J)} \geq .50$) (see Bardach et al., 2018, 2019b; Bardach, Lüftenegger, Yanagida, Schober, & Spiel, 2019a). This begs the important question of what contributes to higher or lower levels of consensus.

Nevertheless, our knowledge about factors that potentially shape the presence of consensus in student ratings is rather limited[1]. There is thus a need for studies exploring such factors in order to expand existing knowledge and provide evidence that could benefit both research and educational practice/policy. For example, in educational practice, it is typically advantageous to identify teachers who receive high or low teaching quality ratings for feedback and development purposes. In such settings, it might also be relevant to understand whether the students assigning these ratings agree in their perceptions, as differing levels of agreement could arise for substantial reasons. For instance, a low level of consensus could indicate that a teacher is not very successful at meeting the individual needs of the various students forming the class (e.g., Lüdtke et al., 2006).

Factors Related to Within-Class Consensus

Factors that potentially influence consensus can be grouped into at least three categories, namely (aggregated) student characteristics, classroom/school characteristics, and teacher characteristics. To date, several studies have investigated how consensus regarding the assessment of different classroom features is related to (aggregated) student characteristics or classroom/school characteristics. For example, Gärtner (2010) showed that class size and subject had no substantial impact on students' consensus on instructional quality. Findings regarding differences between academic tracks are more inconsistent. Thus, Lüdtke and colleagues (2006) reported higher consensus regarding instructional quality for Gymnasium schools, i.e., the highest track school in Germany. In contrast, a later study conducted in Germany found no such differences (Gärtner, 2010). With regard to student characteristics, Gärtner's (2010) results indicated that consensus might increase with students' age. Several studies have also looked at relations between aggregated levels of achievement and consensus on instructional quality (e.g., Schweig, 2016; Wittwer, 2009) and goal structures (Bardach et al., 2018; Bardach et al., 2019b), and have generally reported positive associations between achievement and consensus.

In sum, some evidence exists regarding factors that might influence consensus. However, so far, research conducted in this area has mainly focused on student characteristics (e.g., achievement, age) or classroom/school characteristics (e.g., class size, academic track). Potential effects of teacher-related factors, assessed both from the student and teacher perspective, have remained largely unexplored. This is surprising given that teachers' characteristics and behaviors are likely to play a major role in driving consensus on students' perception of their own practices. In the present study, we turn our attention to two types of teacher-related factors: (a) students' perception of their teachers' instructional practices and differential treatment of students, and (b) teachers-rated characteristics and perceptions of their students.

Teachers' Approaches to Instruction and Treatment of Students

Teachers are key in establishing the goal structures within a classroom; hence, variability in students' perceptions may also be affected by what teachers do in class and how they interact with students and different groups of students. First, it has been proposed that *clarity of instruction* (e.g., Hattie, 2012) might be associated with greater within-class consensus (e.g., Bardach et al., 2018). More precisely, in classes where students receive relatively unambiguous messages, it is more likely that these messages will be perceived and interpreted in a more homogeneous manner, leading to more similar or overlapping mental images of what happens in class among students. Furthermore, instructional clarity itself has been shown

[1] Although methodological advancements, such as doubly latent multi-level structural equation models, correct for disagreement among students regarding teaching quality at the classroom level as part of the latent aggregation process (e.g., Morin et al., 2014), they do not address potential sources of (dis-)agreement.

to be often accompanied by higher levels of teacher immediacy and perceived relevance of information among students (e.g., Mottet et al., 2008), two elements that could themselves increase consensus among students. Thus, if the teacher communicates his or her positive affect to students (teacher immediacy), and if the content of the lessons matters for students (perceived relevance of information), students are more likely to pay attention to the teacher, which is a precondition for instructional clarity to unfold its effects on consensus. Conversely, low clarity of instruction might confuse and demotivate some students, whereas others might still be able or willing to follow teachers' instruction for a variety of other reasons (e.g., importance or interest in the course content), leading to differences in students' perceptions of what happens in the classroom. In addition, even students who try to follow teachers' explanations might find it difficult to stay focused if the clarity of instruction is low, which could add "noise" to students' perceptions and further hamper consensus.

Moreover, teachers can also differ in the instructional approaches they employ for higher vs. lower achieving students (*differential teacher treatment*, Zhu, Urhahne, & Rubie-Davies, 2018), which might lead to heterogenous perceptions of instruction between students depending on their achievement level, and thus to lower within-class consensus (e.g., Bardach et al., 2018; Schenke et al., 2017; Schweig, 2016). It is important to note that for our work we follow a conceptualization of differential teacher treatment as entailing maladaptive teacher behavior (see also e.g., Rubie-Davies, 2007). Therefore, in this study, our representation of differential teacher treatment differs from other adaptive approaches that sometimes also fall under the same label (e.g., tailoring instruction to the differing needs of individual students, providing personalized learning support, see e.g., Nurmi & Kiaru, 2015). Our conceptualization of differential teacher treatment is rather embedded in a theoretical tradition of research focusing on the effects of teacher expectancies, which stems from Rosenthal and Jacobson's (1968) seminal Pygmalion effect. From a theoretical perspective, teachers are expected to form different expectations in relation to different students. These expectations can be based on students' prior achievement levels or individual characteristics, as well as on any other impressions a teacher can form through interacting with these students. These differential expectations then translate into a differential treatment of students expected to function or perform better or more poorly, leading teachers to offer more attention and support to students expected to perform more favorably. These differential expectations thus lead to unequal learning opportunities via which students expected to perform better come to be exposed to more favorable learning environments than students expected to perform more poorly, which confirms teachers' initial expectations (e.g., Rubie-Davies, 2007; Jussim, 2009; see also e.g., Gentrup, Lorenz, Kristen, & Kogan, 2020). In our study, we restrict our focus to differential teacher treatment, as this specific teacher behavior could be useful in explaining consensus. If teachers treat some students in a class, i.e., the higher achiever, better than other students (i.e., lower achievers), then consensus should be lower: Those students receiving favorable treatment should provide more favorable ratings of the class environment than their peers who were treated more poorly.

Potential Mediators of the Effects of Instructional Clarity and Differential Teacher Treatment

To achieve a more complete understanding of the mechanisms involved in the effects of instructional clarity and differential teacher treatment on consensus, we consider two moderators. First, the role of instructional clarity as a driver of consensus might be partially mediated by achievement. On the one hand, instructional clarity has been shown to facilitate student learning and achievement (Hattie, 2012; Titsworth, Mazer, Goodboy, Bolkan, & Myers, 2015). In their meta-analysis, Titsworth and colleagues (2015) summarize prominent theoretical perspectives explaining the effect of clarity on learning and achievement. Early research on teacher clarity was dominated by information-processing theory (e.g., Mayer, 1996), viewing teachers as distributors of information and learners as information processors. Higher clarity should contribute to student learning by enabling them to better manage, store, process, and retrieve information. Other researchers have complemented this view by suggesting that teachers adapt their clarity to students through communication. Hence, clarity becomes a process of communication where meanings are reciprocally negotiated (e.g., Civikly, 1992; see Titsworth et al., 2015 for more details),

creating a space for sense making and high-quality learning that promotes achievement. On the other hand, higher class-average achievement has itself been consistently shown to predict higher consensus (e.g., Bardach et al., 2019b; Schweig, 2016), because higher achieving students (i.e., those who possess more subject-related knowledge) might be able to more accurately describe their classroom functioning, and thus to provide more homogenous ratings (e.g., Wittwer, 2008). It therefore seems plausible to positions achievement as a possible mediator of the relation between instructional clarity and consensus.

Second, the role of differential teacher treatment as another, negative, predictor of consensus could itself be partially mediated by the level of heterogeneity in the achievement of students from the same classroom. More precisely, better treatment generally translates into higher achievement for students able to benefit from this more positive treatment, whereas the opposite is true for students exposed to less optimal forms of treatments (e.g., Wang, Rubie-Davies, & Meissel, 2018). For this reason, differential treatment should contribute to increase the variability of student achievement in the class, leading to an increasing gap in students' levels of achievement. Just like we were able to assume that higher levels of achievement should help increase consensus in classroom perceptions, greater achievement variability should similarly serve to decrease consensus.

Teacher Features and Teachers' Perceptions of Students

Whereas investigating students' perceptions of teacher-related factors would represent a crucial extension of existing work on potential predictors of consensus, it is arguably equally critical to consider teachers' own perspectives. First, if the constructs assessed from the perspective of the teacher share similarities with the constructs assessed from the student perspective, it becomes possible to develop a more comprehensive understanding of the phenomena under study in a way that encapsulates the perspective of these two important types of informants. In the present study, we focus on teachers' *perceptions of student heterogeneity*, particularly in terms of achievement. These perceptions would serve as an important complement to students' perceptions of differential teacher treatment based on achievement in the prediction of within-class consensus. Ideally, teachers should equally support all students, irrespective of their background and achievement levels. However, this might be difficult to achieve in a class where students differ a lot, particularly if we account for the fact that teachers are typically under substantial pressure to cover all components of the class curriculum within a limited time frame. Hence, asking teachers about their perceptions of the student heterogeneity could provide a more nuanced view on why some teachers might fail to equally support all students, with potential consequences for consensus in students' perceptions.

Second, following the rich line of research concerned with teachers' characteristics and their effects on students (e.g., Barr, 1957; Keller, Hoy, Goetz, & Frenzel, 2016; Zee & Koomen, 2016), a further critical question becomes whether specific teacher characteristics influence consensus. Relative to students' characteristics or students' perceptions of teachers' practices and behaviors, teacher characteristics are distal factors in relations to students' perceptions, particularly if the teachers themselves provide the ratings. However, these factors also have the advantage of being arguably more directly under the control of the teachers from an intervention perspective. Finding that these more distal factors are able to influence variability in students' perception of instruction would add a new layer to current knowledge regarding the key drivers of consensus, as well as to our knowledge regarding the impact of teacher' characteristics on students. Current conceptions of teaching quality understand classroom processes as a complex interplay between teachers' "offers" in terms of specific instructional strategies and students' uptake on, and reactions to, these strategies, resulting in a dynamic interaction process whereby each actor influences the other. Both teacher' "offers" and the uptake of these "offers" by students are furthermore affected by the characteristics teachers and students bring to the situation and develop (e.g., Creemers, Kyriakides, & Antoniou, 2013; Göllner et al., 2018; Seidel, 2006). Specific teacher features have been found to affect students' perception of the quality of instruction—probably because they enable teachers to provide high-quality "offers" to students or because they enhance teachers' ability to create conditions that help students to fully benefit

from quality instruction. We expand on this by suggesting that the variability of perceptions of instruction should also be prone to influences of teacher features, potentially operating on different levels: As such, teacher features might affect the "use"-side, e.g., in that they hamper teachers' provision of consistent instructional support. They might also indirectly impact on the willingness of some groups of students to make use of learning opportunities and teachers' "offers", e.g., because these teacher features contribute to creating a distance between teachers and some (more sensitive) students. In the present study, we focus on the role played by teachers' levels of emotional exhaustion, enjoyment and anxiety.

Teachers' *emotional exhaustion* has previously been linked to students' perception of instructional quality (for empirical studies see e.g., Klusmann, Kunter, Trautwein, Lüdtke, & Baumert, 2008; Kunter et al., 2008; Shen et al., 2015; Wong, Ruble, Yu, & McGrew, 2017; see also e.g., Chang, 2009; Maslach & Leiter, 1999). However, insights on whether emotional exhaustion might also predict consensus on instruction is still lacking. Nonetheless, the instructional practices, such as goal structures, of teachers who feel emotionally drained might be less consistent over time and across students, which could in turn negatively affect student's shared image of these practices.

Likewise, ample research has been conducted on the relevance of specific teaching-related emotions, such as enjoyment and anxiety (e.g., Frenzel et al., 2016; Seiz et al., 2015) for the prediction of a variety of student and teacher outcomes. Importantly, teachers' *enjoyment* (e.g., Frenzel, 2014), has been found to affect students' ratings of teachers' instruction (e.g., Frenzel, Goetz, Stephens, & Jacob, 2009), but has never been studied in relation to students' consensus regarding instructional quality. However, teachers' enjoyment typically goes along with enthusiasm (e.g., Frenzel, Goetz, Lüdtke, Pekrun, & Sutton, 2009; Keller et al., 2019; Kunter et al., 2008), and enthusiastic teachers might be better able, or willing, to reach out to a majority of students within their class, thus creating a more similar mental image of teaching quality among students. In contrast, teaching-related *anxiety* could interfere with the establishment of consensus, as heightened self-consciousness stemming from anxiety might divert a teacher's attention from instruction, leading to the communication of more ambiguous messages to students (e.g., Coates & Thoresen, 1976; see also e.g., Chang, 2009; Frenzel et al., 2016).

Research Goals and Hypotheses

The present research seeks to improve our understanding of what drives within-class consensus in students' perceptions of teaching quality. Relying on a large sample of secondary school students (Sample 1[2]), we assumed that higher levels of student-perceived instructional clarity will positively predict consensus (*Hypothesis 1a-f*), and that higher levels of student-perceived differential teacher treatment will negatively predict consensus (*Hypothesis 2a-f*) regarding the six goal structures dimensions task, autonomy, recognition, grouping, evaluation, and time. We further hypothesized that higher class-average levels of achievement will mediate the association between instructional clarity and consensus (*Hypothesis 3a-f)*, and that higher levels of within-class variability in achievement levels will mediate the relation between differential teacher treatment and consensus (*Hypothesis 4a-f*). In addition, class-average levels of achievement will also be directly, and positively, related to consensus (*Hypothesis 5a-f*), whereas achievement heterogeneity will also be directly, and negatively related to consensus (*Hypothesis 6a-f*). Finally, we expected instructional clarity to positively predict class-average levels of achievement (*Hypothesis 7*), and differential teacher treatment to positively predict achievement heterogeneity (*Hypothesis 8*).

In Sample 2, we focused on the role of teacher characteristics, and assumed that teachers' perceptions of student heterogeneity, as well as their self-reported feelings of emotional exhaustion and anxiety, will negatively predict consensus (*Hypotheses 9a-f, 10a-f,* and *11a-f*, respectively). Likewise, we

[2] As explained in the method section, Sample 1 and Sample 2 partially overlap, because Sample 2 included teacher data for some but not all of the classes used in Sample 1. We use the terms 'Sample 1' and 'Sample 2' to avoid unnecessarily lengthy descriptions, but emphasize that these two samples are not independent.

also hypothesized their feelings of enjoyment to positively predict consensus (*Hypothesis 12a-f*). As age has been found to be related to consensus (Gärtner, 2010), we included school grade as a control variable in all analyses. Figure 1 gives an overview of the hypotheses tested in this study (hypotheses for sample 1 in the upper part, hypotheses for sample 2 in the lower part).

Method

Samples and Procedures

The current studies were carried out as part of a larger research project focused on (social) motivation among Austrian secondary school students' (Authors, 2018, 2019, 2020 [blinded for peer review]). Students were recruited within Austrian secondary schools (*Gymnasium*) to participate in this study. Gymnasium schools are the highest track of secondary schools in Austria. Sample 1 comprised 1,743 students (mean age of = 14.38 years, *SD* = 1.41 years, 53.8% females) from 89 mathematics classes[3]. Of those students, 28.9% were in Grade 7, 27.9% were in Grade 8, 20.1% were in Grade 9, 11.4% were in Grade 10, and 11.8% were in Grade 11. The number of students per teacher ranged from 10 to 29.

For a sub-sample of 37 classes, we also gathered teacher data from their respective mathematics teachers (Sample 2). These 37 teachers (48.6% female) were 49.33 years old on average (*SD* = 12.31) and had, on average, 24.23 years (*SD* = 12.34) of teaching experience. The corresponding classes (37 classes with a total of 726 students, 11 to 27 students per teacher) also included students from all grade levels (Grade 7: 35.0%; Grade 8 = 27.3%; Grade 9: 18.9%; Grade 10: 10.6%; Grade 11: 8.3%). These students were 14.19 years old on average (*SD* = 1.38), and 51.4% were females.

Participating students and teachers responded to paper-and-pencil questionnaires during regular classroom hours, supervised by trained research assistants. All students and teachers participated voluntarily in this study. For students, active parental consent was required. The consent rate was above 99%, and, as such, less than 1% of students were not allowed to participate by their parents or directly refused to participate. No compensation for participation was provided.

Measures: Students' Reports

All student variables referred to mathematics and to the class's mathematics teacher. Whereas the items assessing students' perceptions of the six dimensions of mastery goal structures were used to calculate consensus levels regarding the six dimensions (see 'Analysis') for Sample 1 and Sample 2, the other student scales were only employed in the analyses relying on Sample 1, which focused on student-reported predictors of consensus. A 6-point Likert scale ranging from 1 (strongly disagree) to 6 (strongly agree) was used as the response format for all student scales.

Six dimensions of mastery goal structures. Mastery goal structures were measured with items from the Goal Structure Questionnaire (GSQ, Lüftenegger et al., 2017). Students were reminded to think about mathematics and about their current mathematics teachers when answering the items, which began with the phrase 'In mathematics class, …'. Five items measured task, (e.g., '…we should set learning goals for ourselves'; $\alpha_{Sample1}$ = .66; $\alpha_{Sample2}$ = .63), six items measured autonomy (e.g., '…we make important decisions about the learning process together with the teacher'; $\alpha_{Sample1}$ = .84; $\alpha_{Sample2}$ = .84), six items measured recognition (e.g., '… we get feedback concerning our learning progress'; $\alpha_{Sample1}$ = .81; $\alpha_{Sample2}$ = .81), four items measured grouping (e.g., '…we can work on tasks together with classmates if we want'; $\alpha_{Sample1}$ = .73; $\alpha_{Sample2}$ = .71), six items measured evaluation (e.g., '... the teacher points out when someone has improved'; $\alpha_{Sample1}$ = .85; $\alpha_{Sample2}$ = .85), and five items measured time (e.g., '…the teacher makes enough time for explanations'; $\alpha_{Sample1}$ = .82; $\alpha_{Sample2}$ = .81).

Instructional clarity. To gauge instructional clarity, we relied on items based on Chesebro and McCroskey (1998). Five items were used (e.g., "In mathematics class, the teacher gives us clear,

[3] As our work focus on within-class consensus, and as clusters of at least 10 raters (here: students) have been recommended for the consensus measure used ($r^*_{wg(J)}$; see the Analyses section; Lindell, Brandt, & Whitney, 1999), we only included classes including at least 10 participants for analyses.

unambiguous instructions'; α = .85).

Differential teacher treatment. We measured differential teacher treatment with five items developed for purposes of this study based on a consultation of the research literature on differential treatment (e.g., Rubie-Davies, 2007; Zhu et al., 2018). The items focused specifically on students' perceptions of their teachers' behavior in mathematics class (e.g., 'In mathematics class, the teacher treats higher and lower achieving students differently'; α = .83).

Achievement. Teacher assigned mathematics grades taken from the students' most recent report card were used as indicators of achievement. In the Austrian school system, '1' represents the best grade and '5' the lowest grade. Thus, analyses were carried out after recoding grades so that higher values reflect higher achievement.

Achievement heterogeneity. To reflect within-class achievement heterogeneity, we used the standard deviation of the class average achievement, yielding one 'achievement heterogeneity index' for each class.

Measures: Teachers' Reports

The measure of emotional exhaustion referred to teachers' job more generally, whereas the other measures focused on teaching mathematics in the specific class for which we also gathered student data. All teacher-reported measures were rated using a 6-point Likert scale, ranging from 1 (strongly disagree) to 6 (strongly agree).

Emotional exhaustion. We assessed emotional exhaustion using items based on the German version (Enzmann & Kleiber, 1989) of the Maslach Burnout Inventory (Maslach, Jackson, & Leiter, 1996). Four items were employed (e.g., 'I often feel exhausted due to my job as a teacher', α = .91).

Teachers' enjoyment and anxiety. Teachers emotions in terms of enjoyment and anxiety when teaching mathematics in this specific class were measured using scales developed by Frenzel and colleagues (2016). Four items referred to enjoyment (e.g., 'In general, I enjoy teaching this class in mathematics', α = .86). Anxiety was measured with four items (e.g., 'I generally feel tense and nervous when teaching this class in mathematics'; α = .88).

Teacher perceptions of student achievement heterogeneity. To assess teachers' perceptions of the achievement-related heterogeneity of students in their class, we employed items adapted from Skaalvik and Skaalvik (2016). Four items were used (e.g., 'In this class, there is a huge difference between the best and the poorest students'; α = .86).

Analyses

The analyses were performed with Mplus Version 8.2 (Muthén & Muthén, 2017) using the robust maximum likelihood estimator (MLR). To deal with missing data present at the item level (between 0% and 5.9% for student scales, and between 0% and 5.4% for teacher scales), full information maximum likelihood estimation (FIML) was employed (Enders, 2010).

As a measure of within-class consensus regarding the six mastery goal structures dimensions, we used the inter-rater agreement index for multiple items, $r^*_{wg(J)}$. The continuous measure of agreement $r^*_{wg(J)}$ was developed by Lindell and colleagues (Lindell & Brandt, 1997; Lindell et al., 1999) and represents an inverse linear function of the ratio of the average obtained variance to the variance of uniformly distributed random error (Lindell et al., 1999). In the case of six response categories, $r^*_{wg(J)}$ can range between -1.14 and 1, with higher values representing greater within-classroom consensus. Moreover, a positive $r^*_{wg(J)}$ value indicates that consensus within a group is stronger than would be expected by chance, whereas a negative $r^*_{wg(J)}$ value indicates that consensus is less than would be expected by chance (Lindell et al. 1999). $r^*_{wg(J)}$ is a correlation coefficient. Hence, Fisher's z-transformations were applied to all $r^*_{wg(J)}$ values prior to using them in the analyses described below (although no value lower than -1 was observed in the present study, such values would have to be recoded to -1 prior to this transformation).

Preliminary Measurement Models

A multilevel confirmatory factor analytic model (ML-CFA) including all constructs (instructional

clarity, differential teacher treatment, the two mediators of achievement and achievement heterogeneity, consensus regarding the six goal structures dimensions, and the control variable grade) was first estimated to verify the adequacy of the measurement model underlying all constructs and to assess cross-level metric invariance in measurement. Cross-level tests of metric invariance are necessary to ensure comparability of measurement across levels and, when supported, contribute to the efficiency and accuracy of multilevel analyses by introducing greater parsimony in the estimated models (e.g., Lüdtke et al., 2011; Marsh et al., 2009; Morin et al., 2014). Prior to estimation, all variables were standardized to enhance interpretability and limit non-essential multi-collinearity (Arens, Morin, & Waterman, 2015; Marsh et al. 2012; Morin et al. 2014). The ML-CFA was first estimated freely at both levels and then re-estimated with invariant factor loadings across levels (i.e., metric invariance).

We assessed the goodness of fit for all models using the comparative fit index (CFI), the Tucker-Lewis Index (TLI) and the root mean square error of approximation (RMSEA). Typical cut-off scores taken to reflect excellent and adequate fit to the data, respectively, were considered, namely (a) CFI and TLI > .95 and .90; (b) RMSEA < .06 and .08 (e.g., Hu & Bentler, 1999). For tests of metric invariance, models were contrasted considering that drops in CFI or TLI > .01 and increases in RMSEA > .015 can be used to reject the invariance hypothesis (e.g., Chen, 2007; Cheung & Rensvold, 2002).

Analyses of Student-Reported Predictors (Sample 1)

To investigate the effects of differential teacher treatment and instructional clarity and the mediating effects of achievement and heterogeneity in achievement, we estimated a multilevel structural equation model (ML-SEM) building upon the aforementioned ML-CFA model of metric invariance (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Marsh et al., 2009; Marsh et al., 2012; Morin et al., 2014). In this model, the multiple items assessing differential teacher treatment and instructional clarity completed by the students were used to define latent factors at the student level (L1; reflecting inter-individual differences in student perceptions) and the class level (L2; reflecting aggregated class perceptions). We assessed students' mathematics achievement with a single indicator at L1 and aggregated it to L2 using a latent aggregation process in order to obtain an indicator of class average achievement. As a measure of the class-level achievement heterogeneity, we used the standard deviation of mathematics achievement within each class, yielding one value for each class that differed between but not within classes. This manifest heterogeneity measure was included at L2 in the ML-SEM. Consensus is also a naturally L2 construct, so consensus for the six mastery goal structures dimensions was entered at L2 in the ML-SEM. In these models, differential teacher treatment and instructional clarity were automatically group-mean centered as part of the ML-SEM estimation procedure, whereas the L2 variables (consensus on the six dimensions, achievement heterogeneity) were grand-mean centered (Marsh et al., 2012; Morin et al., 2014).

As age has been found to be related to consensus (Gärtner, 2010), we included grade-level as a control variable at L2. We then regressed consensus on the six mastery goal structures dimensions on differential teacher treatment and instructional clarity and additionally tested for partial mediation of the former (with achievement heterogeneity as mediator) and latter (with achievement as mediator) relations. It should be noted that bootstrapping (the recommended approach for tests of mediation) are currently not implemented for ML-SEM in Mplus 8.2. Therefore, in order to test indirect effects, the distribution of product method (MacKinnon, Fritz, Williams, & Lockwood, 2007) was used to compute 95% asymmetric confidence intervals in R version 3.6.2 (R Core Team, 2019) using the package RMediation version 1.1.4 (Tofighi & MacKinnon, 2011). This method considers the correct distribution of the product term instead of assuming a normal distribution or applying an approximation method. Simulation studies have shown that the distribution of product method performs well as compared to bootstrapping (e.g. Preacher & Selig, 2012). Figure 2 displays the ML-SEM investigated in this study.

Analyses of Teacher-Reported Predictors (Sample 2)

To examine the effects of teacher-reported factors (emotional exhaustion, enjoyment, anxiety, achievement heterogeneity) on consensus, we relied on a manifest single level (i.e., estimated at the

classroom level) path analytic model, due to the considerably smaller sample size (37 classes with teacher data). Hence, manifest mean scores capturing teachers' levels of emotional exhaustion, enjoyment, anxiety, and perceptions of achievement heterogeneity were used as predictors. All of these predictors were grand-mean centered. As in Sample 1, consensus scores regarding all six mastery goal structures dimensions were included as outcomes. Once again, we controlled for the effects of school grade on consensus regarding all mastery goal structures dimensions.

Results

Preliminary Analyses

Multi-level modeling with variables that are entered at both levels requires the presence of variability at L2. This can be determined by calculating the intraclass correlation coefficient, ICC(1). Ideally, ICC(1) are close to or higher than .10 (Lüdtke et al. 2008, 2011; see also Arens et al. 2015). In Sample 1, ICC(1) values were high for instructional clarity (.45) and differential teacher treatment (.37), but lower for achievement (.04). An additional consideration is related to the reliability of the L2 aggregates (i.e., inter-rater agreement in ratings of the L2 construct). Reliability of the L2 aggregates is measured with the ICC(2), which can be interpreted just like other reliability measures (Lüdtke et al. 2009; Marsh et al. 2012). ICC(2) values in Sample 1 were high for instructional clarity (.94) and differential teacher treatment (.92), and, again, lower for achievement (.45). Importantly, doubly latent ML-SEMs, such as those used in Sample 1, are naturally controlled for this source of measurement error as part of the latent aggregation process (Arens et al. 2015; Marsh et al. 2012; Morin et al. 2014). Table 1 reports ICC(1) and ICC(2) values for Sample 1. Note that all variables for the analyses for Sample 2 were only located at the classroom level, meaning that ICC values could not be calculated.

In this study, mean within-class consensus values (i.e., $r^*_{wg(J)}$ values) for the six dimensions ranged between .45 and .54 (Sample 1) and between .44 and .52 (Sample 2). Common interpretation guidelines suggest that values above .30 indicate weak consensus, values above .50 indicate moderate consensus, values above .70 indicate strong consensus, and values above .90 indicate very strong consensus (LeBreton & Senter, 2008). Thus, consensus values observed in this study reflect a generally moderate amount of consensus. The values in our study were on average higher than the consensus values found in previous studies using the same consensus measure and focusing on goal structures, which typically reflected weak to moderate levels of consensus (e.g., Bardach et al., 2018, 2019b). Table 1 provides descriptive information (means and standard deviations for all variables used in the analyses), and correlations among all variables for Sample 1. Table 2 provides the same information for Sample 2. Tables S1 (Sample 1) and S2 (Sample 2) of the online supplements shows the proportion of $r^*_{wg(j)}$ values corresponding to each of the proposed cut-offs in the literature.

Preliminary Measurement Models (Sample 1)

We estimated a ML-CFA for Sample 1 to verify the adequacy of the measurement model and to test cross-level metric invariance (e.g., Morin et al., 2014). The results from these analyses are reported in Table 3 and support the adequacy of the measurement model, as well as its cross-level metric invariance. This model was thus used to set-up the ML-SEM predictive model, which was also able to achieve a adequate level of model fit (see Table 3). Table S3 in the online supplements reports the standardized factor loadings at both levels for the multiple-item measures differential teacher treatment and instructional clarity.

Student-Reported Predictors (Sample 1)

Effects of differential teacher treatment and instructional clarity on consensus. Instructional clarity did not significantly predict consensus regarding task (*Hypothesis 1a*, $\beta = 0.16$, $p = .177$) and autonomy (*Hypothesis 1b*, $\beta = 0.22$, $p = .105$). However, the results revealed significant and positive effects of instructional clarity on consensus regarding recognition (*Hypothesis 1c*, $\beta = 0.26$, $p = .049$), grouping (*Hypothesis 1d*, $\beta = 0.35$, $p = .033$) and evaluation (*Hypothesis 1d*, $\beta = 0.39$, $p = .002$). Instructional clarity was not significantly related to consensus regarding time (*Hypothesis 1f*, $\beta = 0.14$, $p = .236$). No statistically

significant effects emerged for differential teacher treatment predicting consensus regarding task (*Hypothesis 2a*, β = -0.19, *p* = .143), autonomy (*Hypothesis 2b*, β = -0.16, *p* = .217), recognition *(Hypothesis 2c*, β = -0.23, *p* = .088), grouping *(Hypothesis 2d*, β = 0.24, *p* = .851*), and time (Hypothesis 2f*, β = 0.22, *p* = .872*).* Differential teacher treatment significantly and negatively predicted consensus regarding evaluation (*Hypothesis 2e*, β = -0.33, *p* = .014). All effects are shown in Table 4. In addition, Figure 3 provides a graphical representation of the effects found for instructional clarity and differential teacher treatment.

Effects of achievement and achievement heterogeneity on consensus. As shown in Table 4, achievement did not statistically significantly predict consensus regarding the six mastery goal structures dimensions (*Hypotheses 5a-f*, β's ranging between -0.29 for grouping and 0.18 for task, *p*'s ranging between .139 and 913). Likewise, achievement heterogeneity did not statistically significantly predict consensus regarding the six mastery goal structures dimensions (*Hypotheses 6a-f*, β's ranging between -0.15 for grouping and 0.07 for evaluation, *p*'s ranging between .088 and .870).

Effects of differential teacher treatment and instructional clarity on achievement and achievement heterogeneity. Instructional clarity significantly and positively predicted achievement (*Hypothesis* 7, β = 0.07, *p* = .020), and differential teacher treatment significantly and positively predicted achievement heterogeneity (*Hypothesis 8*, β = 0.24, *p* = .005).

Mediating effects of achievement and achievement heterogeneity. Achievement did not significantly mediate the effects of instructional clarity on consensus regarding task (*Hypothesis 3a*, 0.12, *p* = .158), autonomy (*Hypothesis 3b*, 0.02, *p* = .408), recognition (*Hypothesis 3c*, -0.03, *p* = .603), grouping (*Hypothesis 3d*, -0.19, *p* = .857), evaluation (*Hypothesis 3e*, -0.01, *p* = .551), and time (*Hypothesis 3f*, -0.11, *p* = .845). Similarly, achievement heterogeneity did not significantly mediate any of the effects of differential teacher treatment predicting consensus (*Hypothesis 4a* for task: 0.04, *p* = .363; *Hypothesis 4b* for autonomy: -0.05, *p* = .653; *Hypothesis 4c* for recognition: 0.12, *p* = .142; *Hypothesis 4d* for grouping: -0.15, *p* = .834, *Hypothesis 4e* for evaluation: 0.07, *p* = .258, *Hypothesis 4f* for time: -0.13, *p* = .863). All mediating effects including confidence intervals can be consulted in Table 4.

Effects of Grade Level (control variable). Grade did not significantly predict consensus regarding any of the mastery goal structures dimensions (β's ranging between -0.07 for recognition and 0.19 for time, *p*'s ranging between .056 and .959, see Table 4 for all coefficients).

Teacher-Reported Predictors (Sample 2)

In the MLM, teachers' perceptions of student heterogeneity were not significantly related to consensus regarding task (*Hypothesis 9a*, β = 0.02, *p* = .574), autonomy (*Hypothesis 9b*, β = -0.20, *p* = .076), recognition (*Hypothesis 9c*, β = -0.17, *p* = .117), grouping (*Hypothesis 9d*, β = -0.05, *p* = .368), and time (*Hypothesis 9f*, β = -0.02, *p* = .404). A significant effect emerged for the evaluation dimension, with teacher perceived student heterogeneity being negatively related to consensus regarding evaluation *(Hypothesis 9e*, β = -0.41, *p* < .001). Emotional exhaustion negatively predicted consensus regarding task (*Hypothesis 10a*, β = -0.46, *p* = .020) and autonomy (*Hypothesis 10b*, β = -0.34, *p* = .006), while none of the other effects for emotional exhaustion attained statistical significance (*Hypothesis 10c* for recognition: β = 0.20, *p* = .868; *Hypothesis 10d* for grouping: β = -0.16, *p* = .119, *Hypothesis 10e* for evaluation: β = -0.10, *p* = .288, *Hypothesis 10f* for time: β = 0.00, *p* = .502). Teachers' reported anxiety did not significantly predict consensus regarding task (*Hypothesis 11a*, β = -0.32, *p* = .054), autonomy *Hypothesis 11b*, β = -0.11, *p* = .315), grouping *Hypothesis 11d*, β = -0.03, *p* = .444), evaluation *Hypothesis 11e*, β = -0.07, *p* = .366), and time *Hypothesis 11f*, β = -0.02, *p* = .467), However, for recognition, anxiety negatively predicted consensus (*Hypothesis 11c*, β = -0.50, *p* = .013). None of the effects for enjoyment were statistically significant (*Hypothesis 12a* for task: β = -0.22, *p* = .848; *Hypothesis 12b* for autonomy: β = -0.22, *p* = .833; *Hypothesis 12c* for recognition: β = -0.16, *p* = .799; *Hypothesis 12d* for grouping: β = -0.07, *p* = .643, *Hypothesis 12e* for evaluation: β = -0.06, *p* = .631, *Hypothesis 12f* for time: -0.24, *p* = .893). With the sole exception of recognition (β = -0.42, *p* = .019), the control variable grade did not statistically significantly

predict consensus regarding mastery goal structures dimensions (β's ranging between -0.02 for evaluation and 0.14 for grouping and time, *p*'s ranging between .392 and 866). Results from the predictive MLM model are reported in Table 5. Figure 4 shows a graphical representation of the effects of the teacher-reported predictors.

Discussion

The perceptions of teaching quality among students from the same class who are exposed to the same teacher can diverge substantially. However, our understanding of the factors involved in driving this variability in perceptions is very limited. In this study, we addressed this question by focusing on the extent of agreement, or consensus, among students from a given class. Relying on two samples, we explored both student-reported predictors and teacher-reported predictors of within-class consensus on teaching quality in terms of goal structures. First, with regard to student-reported predictors (Sample 1), our study suggests that, as discussed extensively in prior research on heterogeneity in students' ratings of instruction (e.g., Bardach et al., 2018, 2019b; Lüdtke et al., 2006; Schweig, 2016; Schenke et al., 2017), maladaptive differential teacher treatment negatively predicts consensus, at least for the evaluation dimension (*Hypothesis 2e*). In other words, the more students perceived that their teacher treated students differently based on their achievement levels (i.e., with a better treatment for higher achievers), the less they shared perceptions of their teachers' evaluation practices. However, and unexpectedly, no such effect was found for the remaining facets of classroom mastery goal structure (i.e., task, autonomy, recognition, grouping [*Hypothesis 2a-d*], or time [*Hypothesis f*]). A potential explanation for why the effects of differential teacher treatment were restricted to the evaluation dimension could be related to the fact that this dimension is the one most intimately connected to achievement (i.e., to the assessment of achievement). In contrast, by being more disconnected from achievement, the remaining facets thus seem to be perceived in a way that is not impacted by perceptions of differential treatment. In addition to replicating the current findings to verify the extent to which they would generalize across samples and contexts, we encourage future researchers to directly test this potential explanation for why differential teacher treatment effects solely occurred for evaluation.

Moreover, maladaptive differential teacher treatment did, as assumed, predict achievement heterogeneity (*Hypothesis 8*). This indicates that when teachers' support is unequally distributed in a class in a maladaptive manner (i.e., favoring higher achievers), achievement gaps tends to enlarge. However, no statistically significant effect for achievement heterogeneity was documented in the prediction of consensus regarding mastery goal structures dimensions (*Hypotheses 6a-f*). The hypothesized mediating mechanism, involving achievement heterogeneity mediating the relation between differential teacher treatment and consensus, was not empirically supported for any of the mastery goal structures dimensions (*Hypotheses 4a-f*). In this study, we used the standard deviation of achievement levels within classes as a very global measure of variability in achievement. Hence, it might have been desirable to directly ask students about their perceptions of achievement-related heterogeneity in class in addition to the focus on actual heterogeneity, which our measure intended to assess. This more subjective measure might have had more explanatory power as a mediator of the relation between subjectively perceived differential teacher treatment and consensus on goal structures.

Furthermore, instructional clarity, the second student-rated predictor, was found to positively predict consensus regarding recognition (*Hypothesis 1c*), grouping (*Hypothesis 1d*), and evaluation (*Hypothesis 1f*), with higher levels of clarity positively related to consensus on these dimensions. With regard to recognition, for example, which comprises instructional strategies such as the provision of feedback (Ames, 1992; Lüftenegger et al., 2017), we propose that clarity might be essential for students to extract the meaning of teacher feedback. If recognition practices are loosely constructed and lack clarity, perceptions of these practices necessarily become blurred, decreasing consensus. Clarity also positively predicted consensus regarding evaluation. The vital importance of explicit, transparent, and clear assessment and evaluation strategies has long been emphasized with regard to promoting student learning

(e.g., Nicol & Macfarlane-Dick, 2006; see also e.g., Evans & Waring, 2011; Balloo, Evans, Hughes, Zhu, & Winstone, 2018). Our findings extend these considerations and indicate that clarity of instruction might also be relevant in shaping students' perceptions of teachers' evaluation strategies, and could play a role in influencing whether the students within a class share overlapping perceptions of their teacher's approaches to evaluation. As the results indicated an effect of differential teacher treatment on evaluation, evaluation seems to be the dimension most strongly affected by both of the student-rated teacher factors investigated here. Further research is now necessary to verify the generalizability of our results, and to investigate the effects of additional student-perceived teacher characteristics on various forms of consensus.

Contrary to our hypotheses and prior research, which has consistently demonstrated that achievement does play a role for consensus (e.g., Wittwer, 2008), no significant effects of achievement on consensus were obtained (*Hypotheses 5a-f)*. Class average achievement also did not mediate any of the associations between clarity and consensus (*Hypotheses 4a-f*). Notwithstanding, as expected based on prior research and meta-analytical findings (e.g., Hattie, 2012; Titsworth et al., 2015), the results revealed a positive effect of instructional clarity on achievement (*Hypothesis 7*). The reasons for the absence of effects of achievement on consensus as well as mediating effects of achievement are hard to establish. However, it could be that characteristics of our sample attending the highest academic track at the secondary level in the Austrian school system may have influenced the findings. In our sample of generally well-performing students, features other than achievement (e.g., motivation, personality, emotions, quality of the student-teacher relationship, e.g., Göllner, Wagner, Eccles, & Trautwein, 2018; Prewett, Bergin, & Huang, 2019) might have played a greater role in influencing variability in students' perceptions of teachers' instruction. For example, classes in which students perceived that the teacher's grading practices were unfair or that the teacher did not sufficiently support students—independent of the actual class-average achievement level—might have developed collective feelings of resentment towards the teacher. These negative emotions could then cause low—but more consistent—ratings in these classes. In addition, it should be kept in mind that this is one isolated finding, and, without doubt, more research is needed. For example, longitudinal studies might yield essential insights on the interplay among consensus, clarity, and achievement. Existing work in the goal structure domain has already shown that the relation between consensus on goal structures and achievement unfolds over time (Bardach et al., 2019b), and the same might apply to mediating effects.

Another important feature of the present research was our examination of teacher-reported predictors (Sample 2). In this regard, our findings are noteworthy because they present, to the best of our knowledge, the first attempt to link teachers' perceptions to within-class consensus. All in all, teacher reports turned out to be less predictive of consensus than student reports. While this might be due to the teacher characteristics selected, it could also simply be a matter of perspective. Arguably, if we are striving to explain students' shared perceptions of instruction, student ratings per se might be a more important source of data than ratings by teachers, as 'outsiders'. Nevertheless, in our study, several effects of teacher-reported factors were still aligned with our hypotheses, which highlights the utility of teacher ratings in research on within-class consensus. First, the results indicated negative relations between emotional exhaustion and consensus regarding task and autonomy (*Hypothesis 10a* and *10b*, respectively). We propose that the practices of emotionally drained teachers might be characterized by less stability and could thus fluctuate to a greater extent between lessons and across students. In particular, it requires effort to create the autonomy-supportive opportunities (e.g., Reeve, 2009) captured in our study by the autonomy dimension as well as the task dimension to some extent due to the inclusion of autonomous and self-regulated task-related work (see e.g., Lüftenegger et al., 2017). Teachers who feel emotionally exhausted might not systematically manage to actively support student autonomy, potentially giving rise to heterogenous perceptions. Second, we found that teachers' anxiety when teaching negatively predicted consensus regarding recognition (*Hypothesis 11c*). Recognition practices require the teacher to tailor his or her feedback to each student and provide individual suggestions for improvement. Teachers who experience higher levels of anxiety when teaching a class might be less able or have less cognitive capacity to offer

such personalized feedback in a way that shows equity and uniformity across all students in a class. These inconsistencies might then explain the finding that anxiety lowers consensus regarding recognition. Combined with the lack of effects found for teaching-specific enjoyment (see *Hypotheses 12a-f* stating positive effects), the findings for emotional exhaustion and anxiety emphasize that negative emotional factors on the teacher side seem to carry more weight in explaining consensus than positive ones (i.e., enjoyment), at least if we use teacher self-ratings. Finally, teacher-reported achievement-related heterogeneity among the students in a class negatively predicted within-class consensus regarding evaluation (*Hypothesis 9d*), which squares well with the effect of student-perceived differential treatment on consensus regarding evaluation. As our study is the first to explore these teacher-reported predictors and their impact on consensus, the findings rather point towards promising vs. potentially less promising paths for future research. At this first stage of research on teacher features and their relations to consensus, a single study cannot provide definite answers.

Limitations and directions for further research

The present study adds to the existing body of literature on heterogeneity in students' perceptions; nonetheless, it is also limited in several aspects. First, we note that the reliability of the scales assessing task and grouping study were rather low. Importantly, our study relied on a cross-sectional research design, meaning that we are not able to verify the directionality of the associations or test reciprocal relations (e.g., Bardach et al., 2019b). Further research should overcome this limitation by conducting longitudinal studies. Studies with multiple measurement points (e.g., diary studies, studies applying experience sampling methods, e.g., Goetz, Sticca, Pekrun, Murayama, & Elliot, 2016) would be particularly well-suited to this topic and could address further relevant questions, e.g., regarding the stability vs. variability of consensus and its relations with student- and teacher-reported factors. In addition, it must be acknowledged that other features and teacher characteristics not considered here might also play a critical role in consensus. For instance, it must be kept in mind that teaching is highly complex and interactive (e.g., Seidel, 2006). Certain students' attributes and teachers' perceptions of these attributes might hamper teaching and create 'noise' in students' perceptions of instruction, leading to lower consensus regarding instructional practices. One such attribute could be teachers' perception of low student motivation (e.g., Skaalvik & Skaalvik, 2016), as low motivation might act as an interpretative filter interfering with students' perception of teachers' actual practices. Put differently, in classes in which teachers perceive that students are rather uninterested, consensus regarding instruction might be low as well, as students do not pay sufficient attention to the teacher's teaching and thus can only provide 'noisy' ratings. Therefore, the inclusion of teachers' perceptions of low student motivation, but also further characteristics such as teachers' and students' own motivation (e.g., Janke, Bardach, Oczlon, & Lüftenegger, 2019), represents a fascinating avenue for future studies.

References

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261-271.

Arens, A.K., Morin, A.J.S., & Watermann, R. (2015). Relations between classroom disciplinary problems and student motivation: Achievement as a potential mediator? *Learning and Instruction, 39*, 184-193.

Authors (2018)

Authors (2019)

Authors (2020)

Balloo, K., Evans, C., Hughes, A., Zhu, X., & Winstone, N. (2018). Transparency isn't spoon-feeding: how a transformative approach to the use of explicit assessment criteria can support student self-regulation. *Frontiers*, 3:69.

Bardach, L., Oczlon, S., Pietschnig, J., & Lüftenegger, M. (2019). Has achievement goal theory been right? A meta-analysis of the relation between goal structures and personal achievement goals. *Journal*

of Educational Psychology. Advanced Online Publication. https://doi.org/10.1037/edu0000419

Bardach, L., Lüftenegger, M., Yanagida, T., Schober, B., & Spiel, C. (2019a). The role of within-class consensus on mastery goal structures in predicting socio-emotional outcomes. *British Journal of Educational Psychology, 89*, 239-258.

Bardach, L., Lüftenegger, M., Yanagida, T., Spiel, C., & Schober, B. (2019b). Achievement or agreement - Which comes first? Clarifying the temporal ordering of achievement and within-class consensus on classroom goal structures. *Learning and Instruction, 61*, 72-83.

Bardach, L., Yanagida, T., Schober, B., & Lüftenegger, M. (2018). Within-class consensus on classroom goal structures - Relations to achievement and achievement goals in mathematics and language classes. *Learning and Individual Differences, 67*, 68-90.

Baudoin, N., & Galand, B. (2017). Effects of classroom goal structures on student emotions at school. *International Journal of Educational Research, 86*, 13-22.

Benning, K., Praetorius, A. K., Janke, S., Dickhäuser, O., & Dresel, M. (2019). Das Lernen als Ziel: Zur unterrichtlichen Umsetzung einer Lernzielstruktur [Learning as an objective: instructional implementation of a mastery goal structure in classroom]. *Unterrichtswissenschaft, 47*(4), 523-545.

Chang, M. L. (2009). An appraisal perspective of teacher burnout: Examining the emotional work of teachers. *Educational Psychology Review, 21*(3), 193-218.

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive?* [Instructional quality: A question of perspectives?] Münster, Germany: Waxmann.

Coates, T., & Thoresen, C. (1976). Teacher Anxiety: A Review with Recommendations. *Review of Educational Research, 46*(2), 159–184

Creemers, B.P.M., Kyriakides, L., & Antoniou, P. (2013). Establishing theoretical frameworks to describe teacher effectiveness. In B.P.M. Creemers, L. Kyriakides, & P. Antoniou (Eds.), *Teacher professional development for improving quality of teaching* (pp. 101–135). Dordrecht, NL: Springer.

De Jong, R., & Westerhof, K.J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*(1), 51–85.

Enders, C.K. (2010). *Applied missing data analysis*. New York: Guilford.

Enzmann, D. & Kleiber, D. (1989). *Helfer-Leiden: Stress und Burnout in psychosozialen Berufen*. Heidelberg: Roland Asanger Verlag.

Epstein, J.L. (1988). Effective schools or effective students: Dealing with diversity. In R. Haskins & D. MacRae (Eds.), *Policies for America's Public Schools: Teacher, Equity and Indicators* (pp. 89–126). Norwood, NJ: Ablex.

Evans, C., & Waring, M. (2011). Student teacher assessment feedback preferences: the influence of cognitive styles and gender. *Learning & Individual Differences*, 21, 271–280.

Frenzel, A.C., Goetz, T., Stephens, E.J., & Jacob, B. (2009). Antecedents and effects of teachers' emotional experiences: An integrated perspective and empirical test. In P.A. Schutz & M. Zembylas (Eds.), *Advances in teacher emotion research: The impact on teachers' lives* (pp. 129– 151). New York, NY: Springer.

Frenzel, A.C., Goetz, T., Lüdtke, O., Pekrun, R., & Sutton, R.E. (2009). Emotional transmission in the classroom: exploring the relationship between teacher and student enjoyment. *Journal of Educational Psychology, 101*, 705-716.

Frenzel, A.C. (2014). *Teacher emotions*. In E.A. Linnenbrink-Garcia & R. Pekrun (Eds.), International Handbook of Emotions in Education (pp. 494-519). New York, NY: Routledge.

Frenzel, A.C., Pekrun, R., Goetz, T., Daniels, L.M., Durksen, T.L., Becker-Kurz, B., & Klassen, R.M. (2016). Measuring teachers' enjoyment, anger, and anxiety: The Teacher Emotions Scales (TES). *Contemporary Educational Psychology, 46*, 148-163.

Gentrup, S., Lorenz, G., Kristen, C., & Kogan, I. (2020). Self-fulfilling prophecies in the classroom:

Teacher expectations, teacher feedback and student achievement. *Learning and Instruction*, *66*, 101-296.

Gärtner, H. (2010). Wie Schülerinnen und Schüler ihre Lernumwelt wahrnehmen [How Students Perceive Their Learning Environment: A Comparison of Four Indices of Interrater Agreement]. *Zeitschrift für Pädagogische Psychologie*, *24*, 111-122.

Göllner, R., Wagner, W., Eccles, J. S., & Trautwein, U. (2018). Students' idiosyncratic perceptions of teaching quality in mathematics: A result of rater tendency alone or an expression of dyadic effects between students and teachers? *Journal of Educational Psychology*, *110*(5), 709-725.

Goetz, T., Sticca, F., Pekrun, R., Murayama, K., & Elliot, A. J. (2016). Intraindividual relations between achievement goals and discrete achievement emotions: an experience sampling approach. *Learning and Instruction*, *41*, 115-125.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. London, UK: Routledge.

Hu, L.-T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Janke, S., Bardach, L., Oczlon, S., & Lüftenegger, M. (2019). Enhancing feasibility when measuring teachers' motivation: A brief scale for teachers' achievement goal orientations. *Teaching and Teacher Education,* 1–11.

Keller, M. M., Hoy, A. W., Goetz, T., & Frenzel, A. C. (2016). Teacher enthusiasm: Reviewing and redefining a complex construct. *Educational Psychology Review*, *28*(4), 743-769.

Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Teachers' occupational well-being and quality of instruction: The important role of self-regulatory patterns. *Journal of Educational Psychology*, *100*(3), 702-715.

Kunter M., Tsai Y.-M., Klusmann U., Brunner M., Krauss S., Baumert J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction, 18*, 468-482.

LeBreton, J.M., & Senter, J.L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815-852.

Lindell, M.K., & Brandt, C.J. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement, 21*, 271-278.

Lindell, M.K., Brandt, C.J., & Whitney, D.J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, *23*, 127-135.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research*, *9*, 215-230.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modelling. *Contemporary Educational Psychology*, *34*, 120-131.

Lüftenegger, M., van de Schoot, R., Schober, B., Finsterwald, M., & Spiel, C. (2014). Promotion of students' mastery goal orientations: does TARGET work? *Educational Psychology*, *34*, 451-469.

Lüftenegger, M., Tran, U., Bardach, L., Schober, B., & Spiel, C. (2017). Measuring a classroom mastery goal structure using the TARGET dimensions: Development and validation of a classroom goal structure scale. *Zeitschrift für Psychologie, 225*, 64-75.

MacKinnon, D.P., Fritz, M.S., Williams, J., & Lockwood, C.M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, *39*, 384-389.

Marsh, H.W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*, 764-802.

Marsh, H.W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A.J.S., Abduljabbar, A., & Köller, O. (2012). Classroom climate effects: Methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47,* 106-124.

Maslach, C., & Leiter, M. P. (1999). Teacher burnout: A research agenda. In R. Vandenberghe & A.M. Huberman (Eds.), *Understanding and preventing teacher burnout: A sourcebook of international research and practice* (pp. 295–303). Cambridge, UK: Cambridge University Press.

Maslach, C., Jackson, S.E. & Leiter, M.P. (1996). *The Maslach Burnout Inventory* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Maslach, C., Schaufeli, W.B., & Leiter, M.P. (2001). Job burnout. *Annual Review of Psychology, 52,* 397-422.

Meece, J.L., Anderman, E.M., & Anderman, L.H. (2006). Classroom goal structures, student motivation, and academic achievement. *Annual Review of Psychology, 57,* 487-503.

Midgley, C., Maehr, M.L., Hruda, L.Z., Anderman, E., Anderman, L., Freeman, K. E., Gheen, M., Kaplan, A., Kumar, R., Middelton, M.J., Nelson, J., Roeser, R., & Urdan, T. (2000). *Manual for the patterns of adaptive learning scales (PALS)*. Ann Arbor, MI: University of Michigan.

Miller, A.D., & Murdock, T.B. (2007). Modeling latent true scores to determine the utility of aggregate student perceptions as classroom indicators in HLM: The case of classroom goal structures. *Contemporary Educational Psychology, 32,* 83-104.

Morin, A.J.S., Marsh, H.W., Nagengast, B., & Scalas, L.F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *Journal of Experimental Education, 82,* 143-167.

Mottet, T. P., Garza, R., Beebe, S. A., Houser, M. L., Jurrells, S., & Furler, L. (2008). Instructional communication predictors of ninth-grade students' affective learning in math and science. *Communication Education, 57,* 333-355.

Murayama, K., & Elliot, A.J. (2009). The joint influence of personal achievement goals and classroom goal structures on achievement-relevant outcomes. *Journal of Educational Psychology, 101,* 432-447.

Muthén, L.K., & Muthén, B.O. (1998-2017*). Mplus User's Guide: Statistical Analysis with Latent Variables* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nelson, P.M., & Christ, T.J. (2016). Reliability and agreement in student ratings of the class environment. *School Psychology Quarterly, 31,* 419-430.

Nicol, D.J., & Macfarlane-Dick, D. (2006) Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education, 31,* 199-218.

Nurmi, J. E., & Kiuru, N. (2015). Students' evocative impact on teacher instruction and teacher–child relationships: Theoretical background and an overview of previous research. *International Journal of Behavioral Development, 39*(5), 445-457.

Patrick, H., & Ryan, A.M. (2008). What do students think about when evaluating their classroom's mastery goal structure? An examination of young adolescents' explanations. *Journal of Experimental Education, 77,* 99-124.

Patrick, H., Kaplan, A., & Ryan, A. (2011). Positive classroom motivational environments: Convergence between mastery goal structure and classroom social climate. *Journal of Educational Psychology, 103,* 367-382.

Peng, S.L., Cherng, B.L., & Chen, H.C. (2013). The effects of classroom goal structures on the creativity of junior high school students. *Educational Psychology, 33,* 540-560.

Polychroni, F., Hatzichristou, C., & Sideridis, G. (2012). The role of goal orientations and goal structures in explaining classroom social and affective characteristics. *Learning and Individual Differences, 22,* 207-217.

Preacher, K.J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6,* 77–98.

Prewett, S. L., Bergin, D. A., & Huang, F. L. (2019). Student and teacher perceptions on student-teacher relationship quality: A middle school perspective. *School Psychology International, 40*(1), 66-87.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reeve, J. (2009). Why teachers adopt a controlling motivating style toward students and how they can become more autonomy supportive. *Educational Psychologist, 44*, 159-175.

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review*, 16–20.

Rubie-Davies, C.M. (2007). Classroom interactions: Exploring the practices of high- and low-expectation teachers. *British Journal of Educational Psychology, 77*, 289-306.

Schenke, K., Ruzek, E., Lam, A.C., Karabenick, S.A., & Eccles, J.S. (2018). To the means and beyond: Understanding variation in students' perceptions of teacher emotional support. *Learning and Instruction, 55*, 13-21.

Schenke, K.C., Ruzek, E.A., Lam, A.S., Karabenick, S., & Eccles, J. (2017). Heterogeneity of student perceptions of the classroom climate: A latent profile approach. *Learning Environments Research, 20*, 289-306.

Schweig, J. (2016). Moving beyond means: Revealing features of the learning environment by investigating the consensus among student ratings. *Learning Environments Research, 19*, 441-462.

Schwinger, M., & Stiensmeier-Pelster, J. (2011). Performance-approach and performance-avoidance classroom goals and the adoption of personal achievement goals. *British Journal of Educational Psychology, 81*, 680-699.

Seidel, T. (2006). The role of student characteristics in studying micro teaching-learning environments. *Learning Environments Research, 9*, 253-271.

Seiz, J., Voss, T., & Kunter, M. (2015). When knowing is not enough: The relevance of teachers' cognitive and emotional resources for classroom management. *Frontline Learning Research, 3*, 54-75.

Shen, B., McCaughtry, N., Martin, J., Garn, A., Kulik, N., & Fahlman, M. (2015). The relationship between teacher burnout and student motivation. *British Journal of Educational Psychology, 85*, 519–532.

Skaalvik, E.M., & Skaalvik, S. (2016). Teacher stress and teacher self-efficacy as predictors of engagement, emotional exhaustion, and motivation to leave the teaching profession. *Creative Education, 7*, 1785-1799.

Titsworth, S., Mazer, J. P., Goodboy, A. K., Bolkan, S., & Myers, S. A. (2015). Two meta-analyses exploring the relationship between teacher clarity and student learning. *Communication Education*, *64*(4), 385-418.

Tofighi, D. and MacKinnon, D.P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods, 43*, 692–700.

Urdan, T., & Schönfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology, 44*, 331-349.

Urdan, T. (2010). The challenges and promise of research on classroom goal structures. In J. Meece & J. Eccles (Eds.), *Handbook of research on classroom motivation* (pp. 92–108). Mahwah, NJ: Erlbaum.

Wang, S., Rubie-Davies C.M., & Meissel K. (2018). A systematic review of the teacher expectation literature over the past 30 years. *Educational Research and Evaluation, 24*, 124-179.

Wong, V. W., Ruble, L. A., Yu, Y., & McGrew, J. H. (2017). Too stressed to teach? Teaching quality, student engagement, and IEP outcomes. *Exceptional Children*, *83*(4), 412-427.

Wittwer, J. (2008). What influences the agreement among student ratings of science instruction? *Zeitschrift für Erziehungswissenschaft, Sonderheft 10*, 205-220.

Zee, M., & Koomen, H. M. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: A synthesis of 40 years of research. *Review of Educational Research*, *86*(4), 981-1015.

Zhu, M., Urhahne, D., & Rubie-Davies, C.M. (2018). The longitudinal effects of teacher judgement and different teacher treatment on students' academic outcomes. *Educational Psychology, 38*, 648-668.

Table 1

Class-Level Descriptive Statistics, Intraclass Correlations, and Bivariate Correlations (Sample 1)

Variable	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. Instructional clarity											
2. Differential teacher treatment	**-.69**										
3. Consensus task (z-transformed)	**.39**	**-.35**									
4. Consensus autonomy (z-transformed)	**.36**	**-.33**	**.53**								
5. Consensus recognition (z-transformed)	**.40**	**-.35**	**.43**	**.52**							
6. Consensus grouping (z-transformed)	.08	.02	.14	.21	.17						
7. Consensus evaluation (z-transformed)	**.63**	**-.58**	**.47**	**.57**	**.67**	.06					
8. Consensus time (z-transformed)	-.02	.11	.21	**.49**	**.40**	**.31**	**.25**				
9. Achievement	**.32**	-.20	.25	.15	.04	-.19	.14	-.14			
10. Achievement heterogeneity	-.11	**.27**	-.07	-.12	.03	-.01	-.04	.01	-.31		
11. Grade	**.19**	-.08	.10	.03	.02	.12	.13	.15	-.03	**.29**	
M	4.18	3.57	.50	.50	.55	.64	.61	.57	3.46	1.04	4.61
SD	.64	.67	.16	.16	.19	.20	.24	.17	.23	.14	1.37
ICC(1)	.45	.37							.04		
ICC(2)	.94	.92							.45		

Note. N = 1,743 students from 89 classrooms; ICC(1) = intraclass correlation coefficient 1 (proportion of between-classroom variance in total variance); ICC(2) = intraclass correlation coefficient 2 (reliability of aggregated variable). In the analyses, the multiple-item measures instructional clarity and differential teacher treatment were modeled as latent variables. All other measures (consensus regarding the six dimensions of goal structures, achievement, achievement heterogeneity, the control variable school grade) were based on single indicators. Statistically significant correlation coefficients at $p < .05$ are in boldface.

Table 2

Class-Level descriptive Statistics and Bivariate Correlations (Sample 2)

Variable	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. Emotional exhaustion											
2. Enjoyment	-.09										
3. Anxiety	.08	**-.69**									
4. Teachers' perceptions of student heterogeneity	.09	-.32	.21								
5. Consensus task (z-transformed)	**-.43**	.03	-.23	-.01							
6. Consensus autonomy (z-transformed)	-.33	-.05	-.05	-.17	**.53**						
7. Consensus recognition (z-transformed)	.05	.22	**-.30**	-.25	**.41**	**.47**					
8. Consensus grouping (z-transformed)	-.13	-.02	-.04	-.04	.21	.20	**.26**				
9. Consensus evaluation (z-transformed)	-.14	.13	-.12	**-.42**	**.55**	**.61**	**.53**	.10			
10. Consensus time (z-transformed)	.06	-.22	.10	.07	.26	**.51**	**.35**	.16	.19		
11. Grade	.27	.01	**-.26**	.10	.07	.01	-.25	.10	-.07	.14	
M	3.10	4.69	1.78	4.66	.50	.49	.52	.61	.60	.56	4.43
SD	1.13	.90	.77	.79	.18	.16	.15	.20	.21	.16	1.33

Note. $N = 37$ teachers and their 726 students from 37 classrooms; please note that we do not report ICC(1) and ICC(2) values for Sample 2 as all variables were only measured at the classroom level. Statistically significant correlation coefficients at $p < .05$ are in boldface.

Table 3

Fit Indices For the Multi-Level Confirmatory Factor Analyses and Predictive Model (Sample 1)

	$\chi 2$	*df*	CFI	TLI	RMSEA	Model description
1.Model	527.78	169	.95	.93	.035	ML-CFA, free factor loadings (configural invariance)
2.Model	552.91	177	.95	.93	.035	ML-CFA, factor loadings invariant across levels (metric invariance)
3.Model	476.28	158	.96	.94	.034	ML-SEM (from the metric invariance ML-CFA)

Note. $\chi 2$ = chi square test of model fit; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; ML-CFA = multi-level confirmatory factor analysis; ML-SEM = multi-level structural equation model.

Table 4

Direct (Top) and Indirect (Bottom) Effects the Multi-Level Mediation Analyses (Sample 1)

Independent Variable	Mediating Variable	Dependent Variable	Est. (*SE*)	Std. Est.
Direct Effects				
Instructional clarity	Achievement		**0.13** (0.07)	**0.07**
Instructional clarity		Consensus task	0.33 (0.36)	0.16
Instructional clarity		Consensus autonomy	0.45 (0.36)	0.22
Instructional clarity		Consensus recognition	**0.54** (0.33)	**0.26**
Instructional clarity		Consensus grouping	**0.74** (0.40)	**0.35**
Instructional clarity		Consensus evaluation	**0.79** (0.27)	**0.39**
Instructional clarity		Consensus time	0.29 (0.40)	0.14
Differential teacher treatment	Heterogeneity achievement		**0.96** (0.37)	**0.24**
Differential teacher treatment		Consensus task	-0.79 (0.74)	-0.19
Differential teacher treatment		Consensus autonomy	-0.66 (0.85)	-0.16
Differential teacher treatment		Consensus recognition	-0.92 (0.68)	-0.23
Differential teacher treatment		Consensus grouping	1.03 (0.98)	0.24
Differential teacher treatment		Consensus evaluation	**-1.29** (0.58)	**-0.33**
Differential teacher treatment		Consensus time	0.88 (0.76)	0.22
	Achievement	Consensus task	0.88 (0.81)	0.18
	Achievement	Consensus autonomy	0.16 (0.68)	0.03
	Achievement	Consensus recognition	-0.22 (0.81)	-0.05
	Achievement	Consensus grouping	-1.42 (1.05)	-0.29
	Achievement	Consensus evaluation	-0.09 (0.67)	-0.02
	Achievement	Consensus time	-0.85 (0.67)	-0.18
	Heterogeneity achievement	Consensus task	0.04 (0.11)	0.04
	Heterogeneity achievement	Consensus autonomy	-0.05 (0.12)	-0.05
	Heterogeneity achievement	Consensus recognition	0.12 (0.11)	0.12
	Heterogeneity achievement	Consensus grouping	-0.15 (0.14)	-0.15
	Heterogeneity achievement	Consensus evaluation	0.07 (0.11)	0.07
	Heterogeneity achievement	Consensus time	-0.14 (0.10)	-0.14
Grade		Consensus task	0.05 (0.06)	0.01
Grade		Consensus autonomy	0.00 (0.08)	0.01
Grade		Consensus recognition	-0.05 (0.08)	-0.07
Grade		Consensus grouping	0.10 (0.08)	0.13
Grade		Consensus evaluation	0.03 (0.06)	0.04
Grade		Consensus time	0.14 (0.07)	0.19
Independent Variable	Mediating Variable	Dependent Variable	Est. (*SE*)	95% CI
Indirect effect				
Instructional clarity	Achievement	Consensus task	0.12 (0.12)	[-0.10, 0.43]
Instructional clarity	Achievement	Consensus autonomy	0.02 (0.09)	[-0.18, 0.24]
Instructional clarity	Achievement	Consensus recognition	-0.03 (0.11)	[-0.30, 0.21]
Instructional clarity	Achievement	Consensus grouping	-0.19 (0.18)	[-0.62, 0.09]
Instructional clarity	Achievement	Consensus evaluation	-0.01 (0.09)	[-0.22, 0.19]
Instructional clarity	Achievement	Consensus time	-0.11 (0.11)	[-0.38, 0.06]
Differential teacher treatment	Heterogeneity achievement	Consensus task	0.04 (0.11)	[-0.18, 0.28]
Differential teacher treatment	Heterogeneity achievement	Consensus autonomy	-0.05 (0.11)	[-0.30, 0.19]
Differential teacher treatment	Heterogeneity achievement	Consensus recognition	0.12 (0.11)	[-0.09, 0.39]
Differential teacher treatment	Heterogeneity achievement	Consensus grouping	-0.15 (0.15)	[-0.49, 0.11]
Differential teacher treatment	Heterogeneity achievement	Consensus evaluation	0.07 (0.11)	[-0.14, 0.32]
Differential teacher treatment	Heterogeneity achievement	Consensus time	-0.13 (0.12)	[-0.40, 0.06]

Note. N = 1,743 students from 89 classrooms; Est. = unstandardized parameter estimate; Std. Est. = standardized estimate; 95% CI = 95% asymmetric confidence intervals based on the distribution of the product method (MacKinnon et al., 2007). Statistically significant results at $p < .05$ are in boldface.

Table 5

Results of the Predictive Model Relying on Sample 2

Predictor	Consensus											
	Task		Autonomy		Recognition		Grouping		Evaluation		Time	
	Est. (*SE*)	Std.	Est. (*SE*)	Std.	Est. (*SE*)	Std.	Est. (*SE*)	Std.	Est. (*SE*)	Std.	Est. (*SE*)	Std.
Emotional exhaustion	**-.07 (.04)**	**-.46**	**-.05 (.02)**	**-.34**	.03 (.02)	.20	-.03 (.03)	-.16	-.02 (.03)	-.10	.00 (.02)	.00
Enjoyment	-.04 (.04)	-.22	-.04 (.04)	-.22	-.03 (.03)	-.16	-.02 (.04)	-.07	-.02 (.05)	-.06	-.04 (.03)	-.24
Anxiety	-.08 (.05)	-.32	-.02 (.05)	-.11	**-.10 (.05)**	**-.50**	-.01 (.05)	-.03	-.02 (.06)	-.07	-.00 (.05)	-.02
Teachers' perceptions of student heterogeneity	.01 (.03)	.02	-.04 (.03)	-.20	-.03 (.03)	-.17	-.01 (.04)	-.05	**-.11 (.04)**	**-.41**	-.00 (.04)	-.02
Grade (control)	.02 (.02)	.11	.01 (.02)	.09	**-.05 (.02)**	**-.42**	.02 (.03)	.14	-.00 (.02)	-.02	.02 (.02)	.14

Note. N = 37 teachers and their 726 students from 37 classes; Est. = unstandardized parameter estimate; Std. = standardized estimate. Statistically significant results at $p < .05$ are in boldface.

.

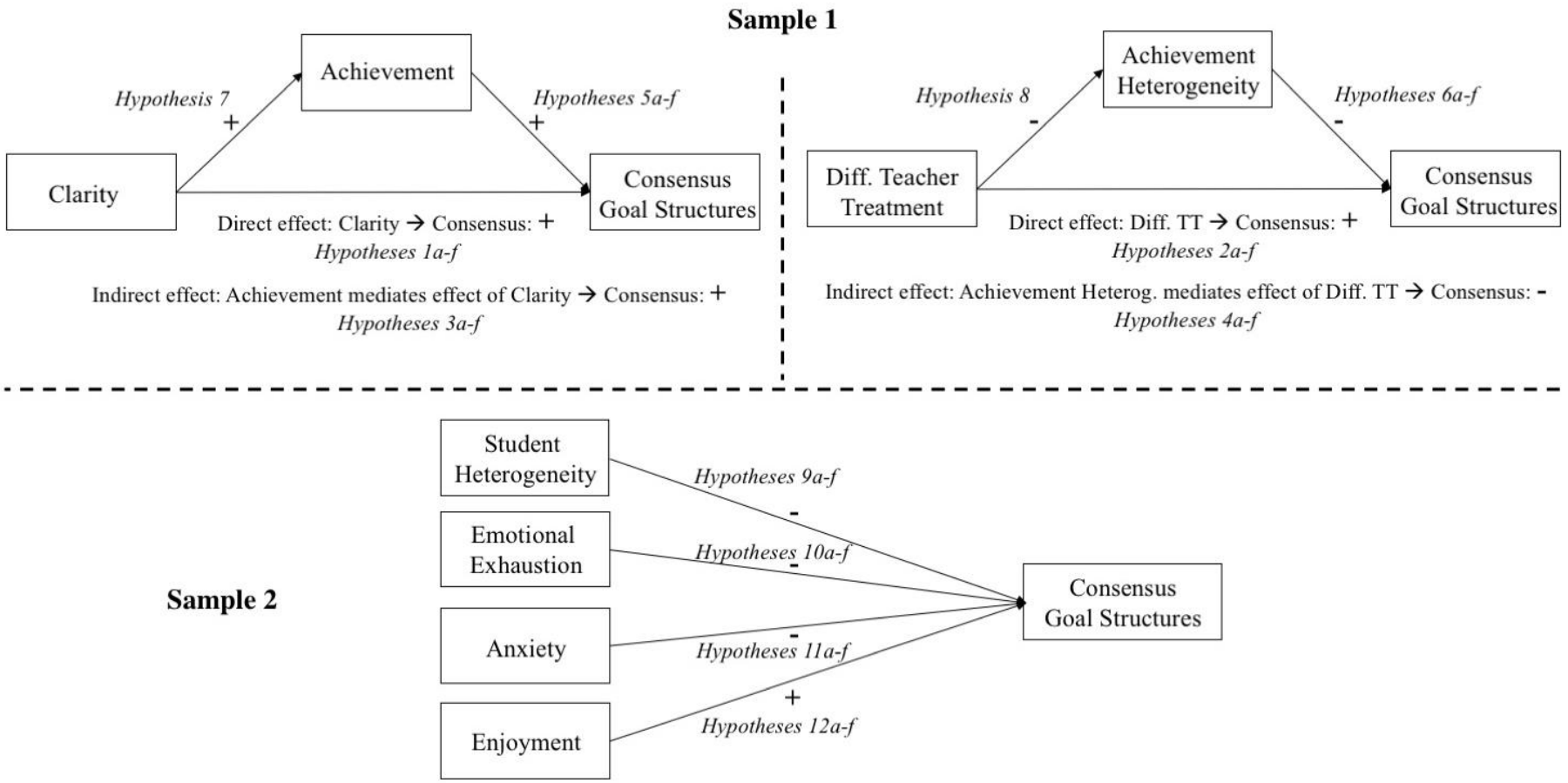

Figure 1. Overview of the hypotheses and assumed directions (positive effects, negative effects) in this study. The hypotheses for the analyses relying on sample 1 are displayed in the upper part, the hypotheses for the analyses relying on sample 2 are displayed in the lower part. As the hypotheses were the same for consensus regarding all six dimensions of goal structures, "consensus goal structures" refers to all six dimensions.

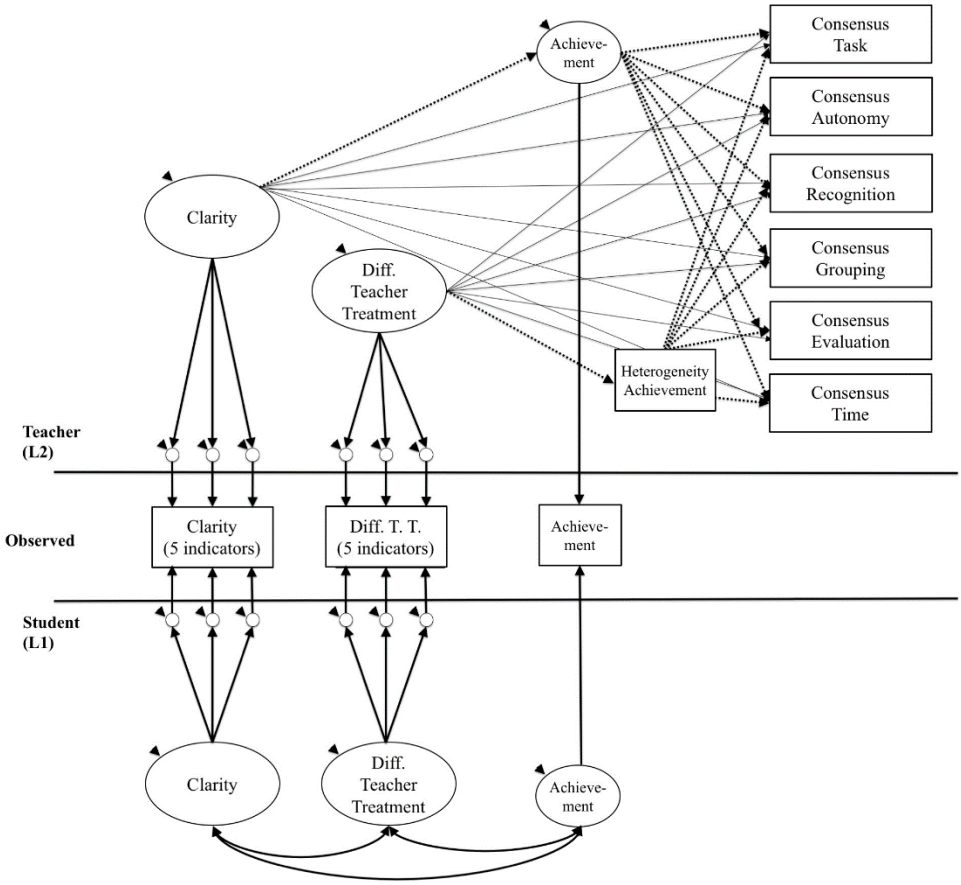

Figure 2. ML-SEM tested for Sample 1. Dotted lines represent the tested partial mediations, with achievement partially mediating the relation between instructional clarity and within-class consensus and achievement heterogeneity partially mediating the relation between differential teacher treatment and within-class consensus. For parsimony, correlations between mediators and the estimated paths of the control variable school grade on within-class consensus regarding the six mastery goal structures dimensions are not displayed.

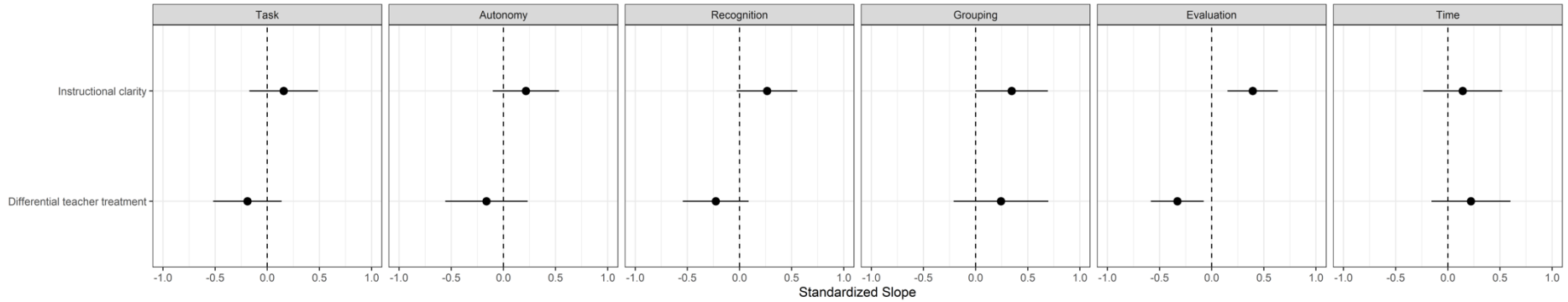

Figure 3. Graphical representation of the effects in terms of 95% confidence intervals of the regression slopes of the two main predictors, instructional clarity and differential teacher treatment, predicting consensus regarding six goal structure dimensions in Sample 1. It should be noted that the confidence intervals are based on two-tailed tests, whereas we conducted one-tailed tests. Hence, the effects of instructional clarity predicting recognition and grouping, which include zero in the confidence intervals, were still statistically significant in our study.

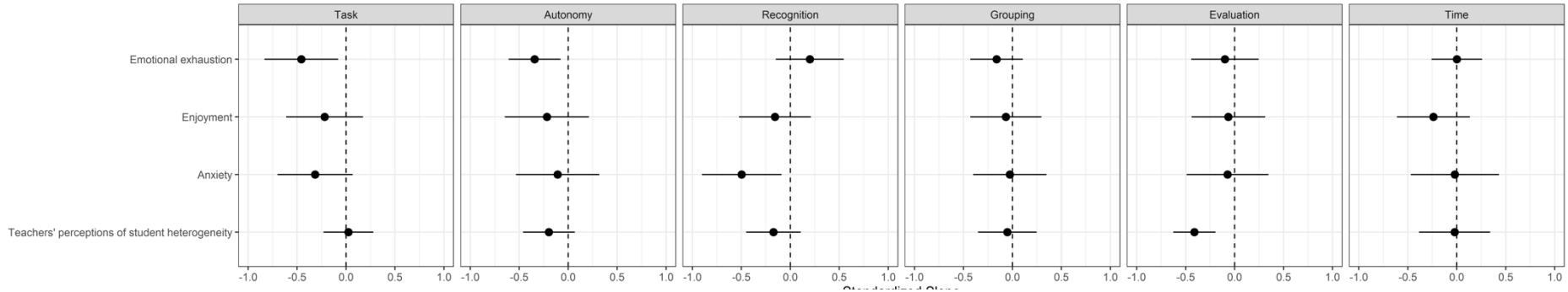

Figure 4. Graphical representation of the effects in terms of 95% confidence intervals of the regression slopes of the teacher-reported predictors, emotional exhaustion, enjoyment, anxiety, and teachers' perceptions of student heterogeneity, predicting consensus regarding six goal structure dimensions in Sample 2.