Uncovering Dynamics in Students' Academic Experiences in Everyday School Life

Irma Talić

Vollständiger Abdruck der von der Fakultät für Humanwissenschaften der Universität der Bundeswehr München zur Erlangung des akademischen Grades eines

Doktors der Philosophie

genehmigten Dissertation.

Gutachter:

1. Dr. Christoph Niepel

2. Prof. Dr. Karl-Heinz Renner

Die Dissertation wurde am 08.12.2022 bei der Universität der Bundeswehr München eingereicht und durch die Fakultät für Humanwissenschaften am 15.06.2023 angenommen. Die mündliche Prüfung fand am 29.06.2023 statt.

Irma Talić

Uncovering Dynamics in Students' Academic Experiences in Everyday School Life

Dissertation an der Fakultät für Humanwissenschaften der Universität der Bundeswehr München

Erstgutachter: Dr. Christoph Niepel, Universität Luxemburg

Zweitgutachter: Prof. Dr. Karl-Heinz Renner, Universität der Bundeswehr München

“Isn’t it funny how day by day nothing changes
but when you look back, everything is different?”

C. S. Lewis

Acknowledgments

I would like to thank my supervisors Dr. Christoph Niepel and Prof. Karl-Heinz Renner for their extensive and continuous support during my Ph.D. journey. They have helped me grow both as a researcher and as a person. I want to thank Prof. Wolfgang Mack for being the chair of my examination committee. I would further like to thank my co-authors Prof. Samuel Greiff, Prof. Ronny Scherer, Prof. Jens Möller, Prof. Herb Marsh, Prof. Jörn Sparfeldt, and Prof. John Rauthmann for their support in building and shaping our joint articles into their present form, as well as many colleagues from the University of Luxembourg and Universität der Bundeswehr München who have accompanied me and offered manifold support and guidance. Importantly, I also want to thank all students who duly participated in our study and with this became a fundamental piece of this work.

I want to dedicate this thesis to my parents Snežana and Sedad. They, Damir, Elif, and Devran Talić have always given their unconditional support. I want to thank Christina Kostić, who is inexpressibly appreciated, for everything she has done. I also want to thank Tami and Dennis Nguyen, Miryam Yousef, Melanie Haschberger, Dr. Patrick Schwaferts, my mentor Sigrid Bidell, as well as my loyal, four-legged companions, who all have constantly provided innovative perspectives on life and change. Without you, this work would not have been possible.

Zusammenfassung

Die Erfassung psychologischer Konstrukte geht oft mit mehreren, konfundierten Varianzquellen einher. Diese Konfundierung kann die strukturelle Konstruktrepräsentation sowie Zusammenhänge zu anderen Konstrukten verzerren. In drei Forschungsartikeln verfolgte die vorliegende kumulative Dissertation das Ziel, spezifische von generellen Varianzkomponenten in Erfahrungen von Schülerinnen und Schülern im Klassenzimmer auf drei unterschiedliche Arten aufzuteilen. Alle Forschungsarbeiten verwendeten Daten des intensiven longitudinalen Projektes DynASCEL, das Daten zu Schülerinnen und Schüler der Sekundarstufe (neunte und 10. Klasse) im deutschen Gymnasium erhob. Die spezifischen (versus generellen) Komponenten, die aufgeteilt wurden, variierten über die Artikel hinweg, d.h., Domänenspezifität (versus Domänengeneralität) in Artikel 1, Situationsspezifität (versus Habitualität) in Artikel 2, und Personenspezifität (versus Konsens) in Artikel 3. Genauer gesagt wird in Artikel 1 ein latentes Modell mit genesteten Faktoren verwendet, um domänenspezifische von domänen-übergreifenden Komponenten in zwei Dimensionen selbstberichteter *Trait* Prüfungsängstlichkeit (d.h., Besorgnis und Aufgeregtheit) in den zwei Domänen Mathe und Deutsch sowie über mehrere Domänen hinweg zu trennen ($N = 348$ Schülerinnen und Schüler). In Artikel 2 wurden situative (*State*) Wahrnehmungen dreier Basisdimensionen von Lehrqualität (Unterstützung durch die Lehrkraft, kognitive Aktivierung, Klassenführung) über drei Wochen in vier Fächern erfasst, und in Zwei-Ebenen konfirmatorischen Faktoranalysen situationsspezifische von habituellen Komponenten unterschieden ($N = 372$ Schülerinnen und Schüler, and $n_{\text{Mathematik}} = 2{,}681$, $n_{\text{Physik}} = 1{,}555$, $n_{\text{Deutsch}} = 2{,}026$, $n_{\text{Englisch}} = 1{,}835$ Beobachtungen). Schließlich wurden in Artikel 3 situative Wahrnehmungen von Lehrqualität von Schülerinnen und Schülern ($N = 372$ Schülerinnen und Schüler, and $n_{\text{Mathematik}} = 2{,}681$ Beobachtungen) in personenspezifische, idiosynkratische Varianzkomponenten und konsensuelle Klassenwahrnehmungen in gemischten Modellen unterteilt. Zusammenhänge zu wichtigen Konstrukten (z.B. Schulnoten) wurden in jedem Artikel zweifach dargestellt, d.h., (a) mit und (b) ohne die respektive Varianzzerlegung, sodass der Effekt dieser unmittelbar sichtbar wurde. Die vorliegende Dissertation bereicherte also das Verständnis struktureller Repräsentation zentraler Konstrukte der Bildungsforschung sowie derer Implikationen.

Abstract

Ratings obtained in psychological assessment often confound multiple sources of variances. This confounding can distort constructs' structural representation as well as their relations to other constructs. In three research articles, the present cumulative dissertation thus aimed at disentangling specific from general components in students' academic experiences in the classroom in three different ways. All articles drew on parts of the larger intensive longitudinal DynASCEL project data on German secondary school students attending the ninth and 10th grades of the highest ability track. The specific (versus general) components that were disentangled varied across articles, that is, domain-specificity (versus domain-generality) in Article 1, situation-specificity (versus habituality) in Article 2, and person-specificity (versus consensus) in Article 3. Specifically, in Article 1, a latent nested factor modeling approach was used to differentiate domain-specific from domain-general components in two dimensions of self-reported trait test anxiety (i.e., worry and emotionality) in the two domains of math and German as well as across domains ($N = 348$ students). In Article 2, state perceptions of three basic dimensions of instructional quality (teacher support, cognitive activation, classroom management) were assessed across a three-week period in four subjects, where situation-specific variance components were disentangled from habitual, trait-like components ($N = 372$ students, and $n_{\text{mathematics}} = 2{,}681$, $n_{\text{physics}} = 1{,}555$, $n_{\text{German}} = 2{,}026$, $n_{\text{English}} = 1{,}835$ observations) in two-level confirmatory factor analyses. Finally, Article 3 disentangled person-specific, idiosyncratic from classroom, consensual variance components in students' state perceptions of instructional quality in math ($N = 372$ students, and $n_{\text{mathematics}} = 2{,}681$ observations) in linear mixed effects models. Relations to crucial constructs (e.g., school grades) were displayed in each article in two ways, that is (a) with and (b) without disentangling the different variance components such that the effect of differentiating the different components becomes apparent immediately. In doing so, the present dissertation enhanced the understanding of key educational constructs' representation and their implications.

Table of Contents

Chapter 1

Introduction

1. Introduction

Psychological assessment is targeted at latent constructs that are not directly visible but need to be inferred from manifest, observable indicators (e.g., responses to a questionnaire; Ziegler & Bühner, 2012). If we were, for instance, interested in an individual's (let's call him Joe) self-esteem, we might assess his self-reported self-esteem on a validated and reliable scale, and subsequently use this rating to test relations to other psychological constructs (e.g., stress), to be able to draw conclusions on the association of these two constructs (e.g., higher self-esteem is related to lower stress) in individuals or across many individuals. Yet, such ratings do not exclusively reflect true score variance (i.e., the true manifestation of that person's self-esteem), but rather entail variance from multiple other variance sources whose consideration or non-consideration, respectively, might distort relations to outcome criteria (Brunner et al., 2009).

First, ratings could confound multiple domain areas. Joe might be a great athlete and musician, but not a great math student. Thus, his self-esteem ratings will likely differ from domain to domain to an overall self-esteem rating, where he might average his self-esteem across many domains. Yet, if we only assess Joe's self-esteem in math, we cannot tell how much of that rating is affected by his general self-esteem. We also would not know whether his lower math self-esteem or the part in his math self-esteem that is higher because of his general self-esteem, is the decisive one for the relation to Joe's stress ratings. Vice versa, if we only assess Joe's general self-esteem, we do not know which domain areas he used as basis of information and how the different domains affect his general self-esteem rating and the relation to his stress rating.

Second, ratings might entail a large situation-specific component. We might assume that Joe's self-esteem is relatively stable, such that it might seem to suffice to assess it once with regard to his *usual, habitual* self-esteem. On that trait level, self-esteem might be negatively related to stress. Yet, even though his self-esteem might be relatively stable across time, there will likely be some moments when he might feel better or worse than usual. The relation between such momentary self-esteem and stress might be positive on the state level (e.g., in situations where Joe's self-esteem is higher, his stress is also higher), for instance, due to performing exciting athletic challenges successfully. If we only assessed Joe's self-esteem once, we would miss out on these daily fluctuations and relations to outcome criteria at the state level. We would not be able to tell how much of his self-esteem is habitual and how much is situation-specific, although the latter might reveal new insights into the dynamic of his experiences across situations in daily life.

Third, situation ratings could be heavily influenced by subjective perceptions rather than the actual situation. For instance, Joe might be interpreting a situation differently than other people who were in that same situation. Being a mediocre math student, Joe might interpret receiving his recent math test score from his teacher as potentially self-esteem threatening because he perceives his teacher acting funny while handing out the graded tests, whereas his classmates do not share this impression of their teacher and might see this situation as an opportunity to improve their self-esteem by receiving good performance feedback (or at least an opportunity to identify mistakes to avoid in the future). Thus, if we only assessed Joe's perspective on that situation, we would not be able to tell the extent to which his rating is subjective or objectively true. Was the teacher truly acting funny (that is, did multiple students agree on this perception?) or was this merely Joe's perception who is generally anxious in the math classroom? Analogously, we would not be able to tell if Joe's subjective perception or the objective situation is more crucial with regard to changes in his momentary self-esteem.

Thus, the present dissertation disentangles different variance components in three different ways in an educational context with the goal of improving the preciseness of psychological constructs' structural representations and associations to crucial, related constructs. We disentangle (a) domain-specific from general, (b) situation-specific from habitual, and (c) person-specific from consensual construct manifestations. The examination of (b) and (c) was enabled by employing the experience sampling method (ESM; Bolger & Laurenceau, 2013), where momentary perceptions were assessed repeatedly. Self-esteem and stress were only used in the introductory example and are not part of the dissertation. This dissertation consists of three different empirical articles on a sample of German secondary school students of the ninth and tenth grades of the highest ability track, thus comparable to many thousands of youth that are included in international large-scale assessments such as the Programme for International Student Assessment (PISA; OECD, 2014). Thus, this dissertation draws on data from the larger intensive longitudinal project DynASCEL ("Dynamics of Academic Self-Concept in Everyday Life"), where students participated in a three-week experience sampling study that assessed their momentary academic experiences in everyday school life via e-diaries, as well as an exhaustive trait pre- and post-assessment in paper-and-pencil format. We focus on the two key educational constructs of (trait) test anxiety (Article 1) and (state) students' perceptions of instructional quality (Articles 2 and 3). We tested associations to the crucial, related constructs of school grades as achievement indicators (all Articles), academic self-concept (Article 1), trait students' perceptions of instructional quality (Article 2), subject-specific interest (Article 2), perceived lesson-specific learning achievement (Article 3), personality traits (Article 3), and reasoning ability (Article 3). In doing this, the present dissertation illustrates the disentanglement of different variance components by three examples to refine the current understanding of students' test anxiety and perceptions of instructional quality and their relations to key student outcome criteria.

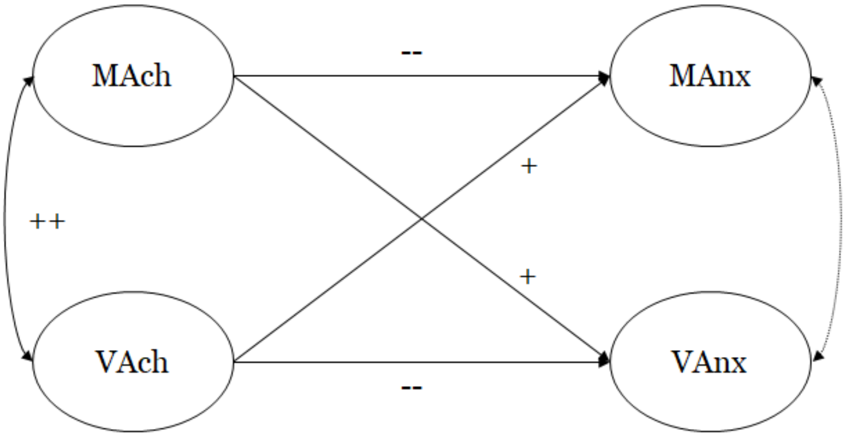
Test Anxiety: Domain-Specificity versus Generality

Test anxiety (TA) is a key educational construct that is negatively associated with students' well-being (Steinmayr et al., 2016) and achievement (Chapell et al., 2005; von der Embse et al., 2018) and can thus threaten long-term educational opportunities (Zeidner, 2020). TA can be subsumed within a set of detrimental reactions to possible failure in test or evaluative situations (Zeidner, 1998), and is thus also conceptualized as situation-specific or contextualized personality trait (Zeidner, 2020). The two TA dimensions of worry (e.g., failure-related ruminations) and emotionality (e.g., rapid heartbeat) can be distinguished (Liebert & Morris, 1967). To predict the formation of TA, the generalized internal/external frame of reference (GI/E) model can be used (Arens, Becker, & Möller, 2017; Möller et al., 2016) that stems from research on the academic self-concept (Marsh, 1986). Specifically, the GI/E model assumes domain-specific achievement-based comparison processes to play a crucial role in the formation of TA—both within and across different domains (e.g., the school subjects math and German). First, students engage in social comparisons, externally comparing their own achievement with their peers' achievements. Social comparison processes in the GI/E model are visible in within-domain relations. For instance, if students reach the conclusion that their own achievement is better than their peers' achievements, this will likely have a beneficial effect on their test anxiety in the same domain (e.g., higher math achievement is related to lower math test anxiety). The GI/E model also assumes that students engage in dimensional comparisons, internally comparing their own achievements in one domain with their own achievement in another domain. Dimensional comparison processes in the GI/E model are reflected in cross-domain relations. These cross-domain relations can show to be of the same algebraic sign as within-domain relations (reflecting assimilation effects, e.g., higher math achievement is related to lower German test anxiety) or the opposite algebraic sign (reflecting contrast effects; e.g., higher math achievement is related to *higher* German test anxiety; Möller et al., 2016). Whether dimensional comparisons show as assimilation of contrast effect is still a matter of debate, yet domain similarity is one possible moderator (Möller et al., 2020). In the case of math and German, domain dissimilarity is assumed to be the strongest, thus leading to dimensional contrast effects (Möller et al., 2020). See a simple illustration of the GI/E model on a unidimensional anxiety measure in Figure 1.[1] The two processes of social and dimensional comparisons interact in such a way that anxiety across domains is hardly correlated even though achievement across domains is highly correlated.

[1] In the article, we implemented a two-dimensional TA measure, that assessed the dimensions of worry and emotionality. For the sake of simplicity, here we only display relations on a unidimensional measure.

Figure 1

Hypothesized GI/E relations between achievement and test anxiety in two domains

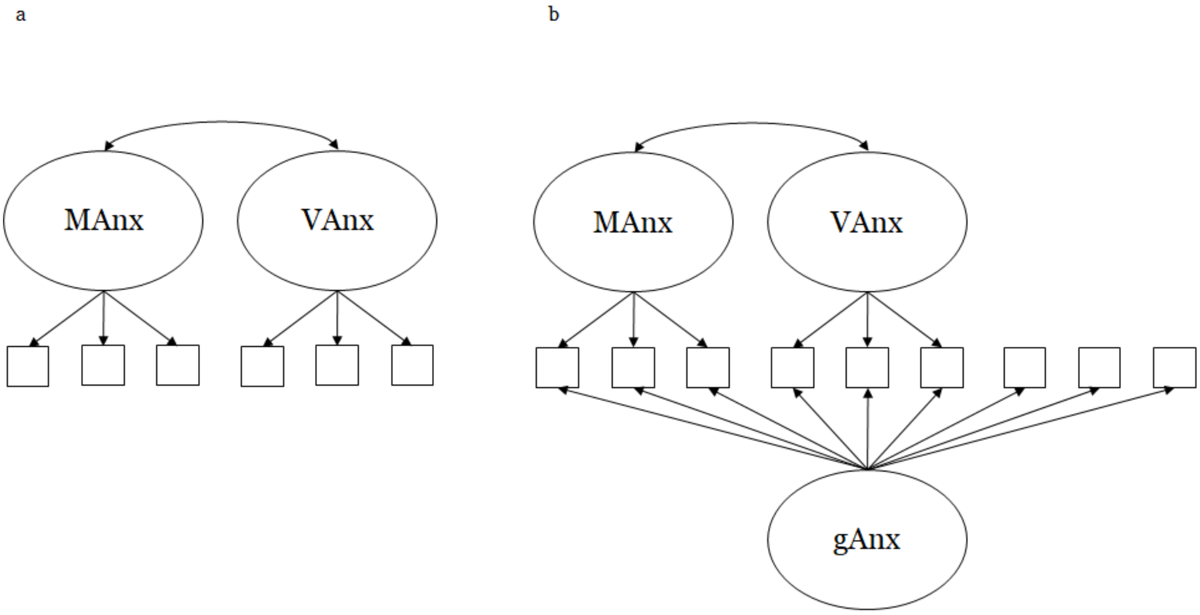


Note. MAch = Math achievement; VAch = Verbal achievement; MAnx = Math anxiety; VAnx = Verbal anxiety. ++ = strong positive relation, + = positive relation, -- = strong negative relation.

Clearly, the GI/E model draws on domain-specific processes. At the same time, TA has been identified as a hierarchical construct that entails both domain-specific and domain-general components (Gogol et al., 2016; Gogol et al., 2017). In other words, students differ with regard to how anxious they generally are across domains in comparison to other students, and students differ with regard to how anxious they are in one domain in comparison to another domain. Yet, traditionally, studies on the GI/E model use first-order factor (FOF) models for representing TA and testing its domain-specific relations to achievement (Arens, Becker, & Möller, 2017; Marsh, 1988; Schilling et al., 2005) although the FOF model cannot consider general construct components at the apex of the hierarchy (Brunner et al., 2010). Thus, domain-specific TA entails not only domain-specific but domain-general variance components in the FOF model that might distort relations to other variables (Brunner et al., 2009; Devine et al., 2012). See Figure 2a for an illustration of the FOF model, where domain-specific anxiety items are loading on their domain-specific factor only. To avoid confounding domain-specific and general construct components, a nested factor (NF) modeling strategy is suggested. The NF model represents the construct's structure adequately, with a general component at the apex of the hierarchy, influencing all domain-specific components (Gogol et al., 2016; Gogol et al., 2017). See Figure 2b for an illustration of the NF model, where domain-specific items load on their domain-specific factors and the general factor. An additional set of general items load on the general factor exclusively, defining its meaning as the reference domain (Eid et al., 2017). Importantly, this modeling strategy disentangles domain-specific from general construct components, and with this, purifies the domain-specific components by variance from the general anxiety factor.

Figure 2

Contrasting First-Order-Factor and Nested-Factor Modeling Approaches



Note. MAnx = Math anxiety; VAnx = Verbal anxiety; gAnx = general anxiety.
Item residual variances are not displayed for enhanced clarity.

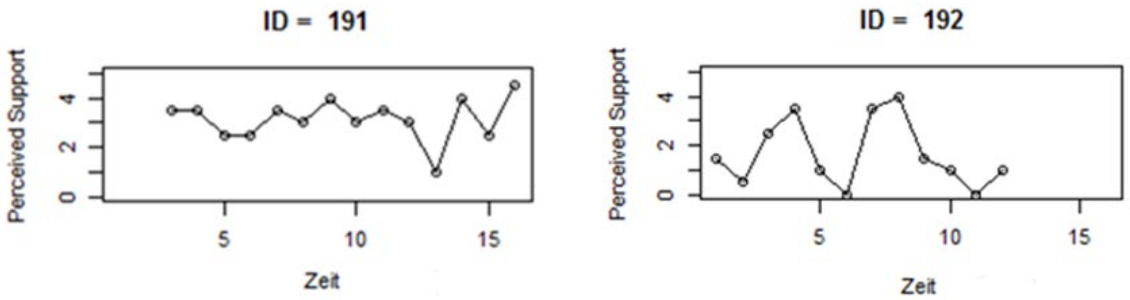
The first contribution (see Chapter 2) illustrates how employing the NF modeling strategy changes result patterns within the GI/E model (that is based on domain-specific relations) in contrast to the FOF modeling strategy. Thus, the first contribution takes on the first introductory example of handling imprecise ratings with regard to domain-specificity versus generality and its effect on predicting two dimensions of test anxiety within the GI/E model.

SPIQ: Situation-Specificity versus Habituality

Students' perceptions of instructional quality (SPIQ) are crucial determinants of students' achievement and motivation (Scherer & Nilsen, 2016). The framework of Three Basic Dimensions (TBDs, Klieme et al., 2001) describes SPIQ in a parsimonious and robust way by the dimensions of teacher support (e.g., avoiding achievement pressure), cognitive activation (e.g., posing challenging tasks), and classroom management (e.g., handling classroom disruptions effectively). The TBDs are empirically distinguishable, yet interrelated dimensions that show relations to student achievement and motivation (Praetorius et al., 2018). The majority of studies on SPIQ are between-person research designs, assessing SPIQ at one point in time and aggregating them to higher levels (e.g., class, school, or even country levels in large-scale assessments; Praetorius et al., 2018), thus drawing on interindividual variation when examining relations to other variables such as student achievement. Individual SPIQ are hereby considered noise (Lüdtke et al., 2009). Often, however, the agreement and the reliabilities of SPIQ are low within classrooms and substantial amounts of variance are attributable to student characteristics (Feistauer & Richter, 2017; Wagner et al., 2016). Additionally, assessing SPIQ at one point in time also implicitly assumes that SPIQ are highly stable across time, although the classroom is a dynamic context (Curby et al., 2011; Praetorius et al., 2014). A longitudinal study has identified a substantial time-specific component within SPIQ of the same students rating the same teacher (Wagner et al., 2016) which was corroborated by two experience sampling studies (Goetz et al., 2013; Goetz et al., 2020). This intraindividual variation (i.e., differences in SPIQ within students across points in time) stands in contrast to the interindividual variation (i.e., differences in SPIQ between different students) described above (Molenaar, 2004; Murayama et al., 2017). See Figure 3 for an illustration of interindividual and intraindividual variance. Stemming from real data assessed in the project, the Figure displays students' perceptions of teacher support across a maximum of 16 measurement points in mathematics lessons. Both students (IDs 191 and 192) were rating the same teachers' behavior in the same lessons on a scale from 0 to 5. Interindividual variance is visible in differences between the students, for instance, regarding measurement point 11, where the student with the ID 191 rates their teachers' support with 3.5 rather high whereas their classmate with the ID 192 perceives zero support in that same lesson. Intraindividual variance pertains to within-student processes from lesson to lesson. Here we see a rather stable rating pattern in student 191 until lesson 13, where they start fluctuating in their SPIQ. In contrast, student 192 is fluctuating in their perceptions from the beginning and shows missing data in the last few measurement points.

Figure 3

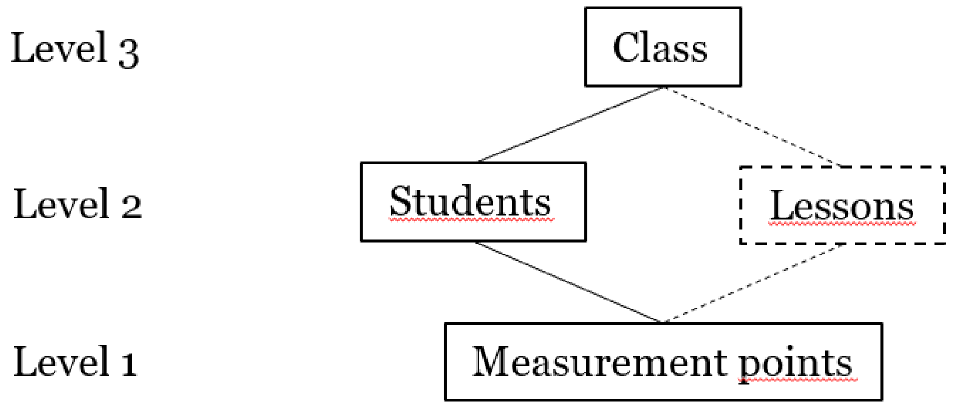
Two Classmates' Perceived Teacher Support Across 16 Measurement Points



To uncover such within-student dynamics across time, the experience sampling method (ESM; Bolger & Laurenceau, 2013) was used that assessed lesson-specific, *state* SPIQ in the classroom (in contrast to habitual, *trait* SPIQ). ESM entails the repeated measurement of momentary individuals' daily life experiences in their natural environments, thus eliminating retrospective biases and enhancing ecological validity, and, importantly, enabling the investigation of psychological processes (Trull & Ebner-Priemer, 2014). Due to repeated measurements within individuals, this method creates dependencies in the data that call for methods to adequately handle such hierarchical data at different levels when it comes to analysis (i.e., multilevel models; Hox et al., 2018). Assessing state SPIQ in shared situations (lessons) for all study participants created a cross-classified data structure, where measurement points (Level 1) are nested both within students (Level 2a) and lessons (Level 2b), that are, in turn, nested within classrooms (Level 3). See Figure 4 for an illustration of this hierarchical, cross-classified data structure. Adequate methods disentangle the variance components attributable to each level.

Figure 4

Hierarchical, Cross-Classified Data Structure of State SPIQ



Note. State SPIQ assessed in repeated measurement points (Level 1) are nested within students (Level 2a; i.e., measurement points within students are more similar than measurement points across different students) and lessons (Level 2b; i.e., measurement points within lessons (across students) are more similar than measurement points across lessons), which are both nested within classes (Level 3). Nesting within the Between-Lesson Level (Level 2b) are represented in dashed lines because they were controlled for, but not explicitly examined.

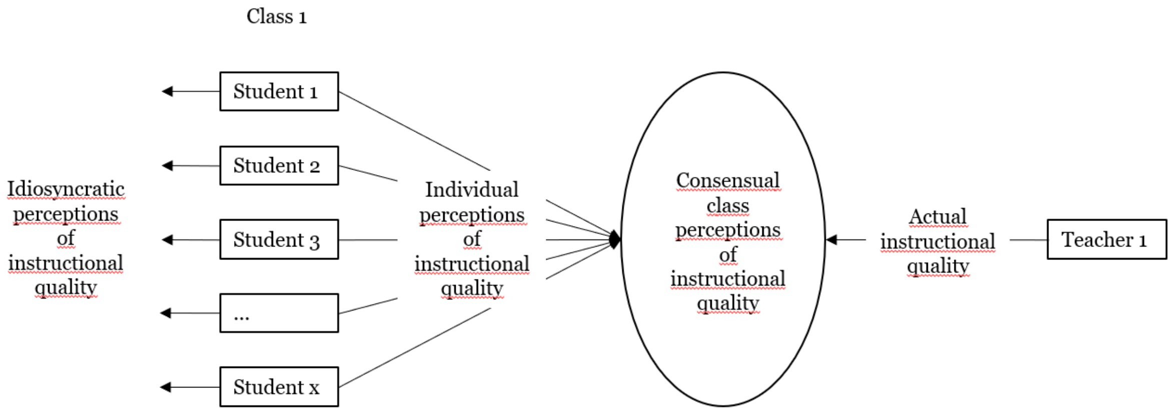
Thus, the traditional way of assessing SPIQ cross-sectionally and relating aggregated SPIQ means to other variables of interest neglects the student and their perceptions' dynamics. In more general terms, assessing ratings at one point in time does not allow for the differentiation of situation-specific and habitual components within that rating. Further, the construct structure and associations to other variables can differ across the types of variation (Molenaar, 2004). The second contribution (see Chapter 3) illustrates how ESM data provides the opportunity of assessing situation-specific, state ratings and disentangling them from habitual, trait ratings to gain insight into the extent of construct stability and variability across time. With this, the second contribution takes on the second introductory example of handling imprecise ratings with regard to situation-specificity versus habituality in SPIQ and relations to student achievement and motivation.

SPIQ: Person-Specificity versus Consensus

The finding that SPIQ, that are supposed to reflect teachers' instructional behavior, are influenced by the student rater (Feistauer & Richter, 2017; Wagner et al., 2016) casts some doubts on SPIQ's validity, especially when keeping in mind that SPIQ are one of the most important sources of information on teaching effectiveness at the country level (OECD, 2014). For instance, Lazarides and Ittel (2012) identified four distinct perception patterns in mathematics instruction (e.g., students who perceived overall low instructional quality across all assessed dimensions) that were differentially related to gender, subject-specific interest and self-concept (e.g., students who belonged to the overall low quality perception cluster reported lower interest, self-concept and were more likely to be female). Thus, SPIQ cannot be implicitly assumed to reflect reality. To estimate the degree of subjectivity versus objectivity within SPIQ, and with this, to corroborate the validity of SPIQ, usually, other information sources are needed, for instance, video recordings that are subsequently rated by external raters (e.g., Praetorius et al., 2014) or teachers' self-perceptions (e.g., Wisniewski et al., 2022). Due to economic reasons, such designs are not always feasible. Gathering state perceptions of all students in shared lessons in the classroom using ESM, however, provides the unique opportunity to estimate the degree of subjectivity (or person-specificity) within SPIQ in relation to the class mean (or consensus in the classroom). Specifically, one can draw on works from situation research, where relations between people and the situations, they find themselves in, are of special interest (Rauthmann, 2021). By assessing shared perceptions of the situation (in our case, shared perceptions of instructional quality) by multiple agents that are present in that situation (in our case, lesson), shared, overlapping perceptions can be disentangled from idiosyncratic perceptions that are usually confounded within the raw rating (Rauthmann & Sherman, 2019). Figure 5 illustrates how individual SPIQ entail components that overlap with their classmates' (i.e., consensual perceptions) and ones, that differ from their classmates' (i.e., person-specific, idiosyncratic perceptions) in a class with x students.

Figure 5

Illustration of Idiosyncratic and Consensual Components within SPIQ


Class 1
Student 1
Student 2
Student 3
...
Student x
Idiosyncratic perceptions of instructional quality
Individual perceptions of instructional quality
Consensual class perceptions of instructional quality
Actual instructional quality
Teacher 1

The differentiation between idiosyncratic and consensual SPIQ is possible by employing ESM, as we do not deal with data that is assessed with a vague time reference (e.g., "usually") but targeted at lesson-specific instructional behavior. The overlapping, consensual class perceptions best approximate actual instructional quality (i.e., if all students in the classroom agree on something, it is intersubjective and therefore approximates true instructional quality). The parts of individual SPIQ that do not overlap with class SPIQ are purely idiosyncratic SPIQ (i.e., person-specific). Instead of using the raw state SPIQ ratings that confound consensual and idiosyncratic components, these different components are disentangled from each other. Relations to a subjective lesson-specific student achievement indicator are tested separately for each component to examine possible differential relations. For instance, if only the idiosyncratic perception of students is related to achievement, but not the consensual perception, this suggests that the individual interpretation of instructional behavior in a certain way is more crucial for one's achievement than what all students agree on. To find out if certain groups of students differ in their perceptions (e.g., more agreeable students might tend to agree more with their classmates) and if certain groups of students differ in their relations between SPIQ and achievement (e.g., the relation between SPIQ and achievement might be stronger for students higher in negative emotionality), student personality traits are investigated as predictors and moderators. This approach allows for an estimation of the relative role of the personal reality versus the social reality (Rauthmann & Sherman, 2019) on student's perceived short-term learning achievement.

While the former contribution (Chapter 3) acknowledged the important role of the student rater in SPIQ, it did so by focusing on within-student, temporal fluctuations. The present, third contribution (see Chapter 4) does so by zooming into specific situations (i.e., lessons) where lesson-specific dynamics are examined with the classroom and the classmates as frame of reference. By doing this, the third contribution takes on the third introductory example of handling imprecise ratings with regard to person-specificity versus consensus and its relations to perceived lesson-specific achievement and personality traits.

The Goal of the Present Dissertation

The present dissertation thus aims at demonstrating how raw ratings entail many different variance components that traditional means of modeling or data assessment cannot disentangle. This is done by three examples in three different research articles (out of which one is published, one is accepted for publication and one is to be submitted) that are organized within chapters in logical order. The first example pertained to the assessment of only domain-specific construct manifestations. In the case of hierarchical constructs, these domain-specific manifestations include domain-general components whose non-consideration can distort relations to outcome criteria. This is demonstrated by the use of nested-factor modeling (in contrast to first-order factor modeling) on a two-dimensional (worry and emotionality; Liebert & Morris, 1967) measure of test anxiety in two core subjects (mathematics and German). Relations to academic self-concept and student achievement are illustrated in dependence on the modeling strategy (Chapter 2). The second example pertained to the assessment of construct manifestations at one point in time that are aggregated to higher levels of analyses. This procedure lacks temporal and individual information. Traditional longitudinal studies show some advantages over cross-sectional studies, yet, they are also unable to differentiate situation-specific from habitual components. This problem is resolved by an intensive longitudinal study via experience sampling that enables the differentiation of state versus trait-like components in three basic dimensions (teacher support, cognitive activation, classroom management; Klieme et al., 2001) of students' perceptions of instructional quality in four subjects (math, physics, German, English). Aggregated state and trait relations to student achievement and motivation are illustrated (Chapter 3). The third example pertained to the assessment of only one situation perception that does not allow for an estimation of person-specificity versus consensus, where usually, both components are confounded in one raw rating. To resolve this, perceptions of all students within a class regarding the same instructional quality in the same lessons are assessed and consensual perceptions were disentangled from idiosyncratic perceptions in mathematics instruction. Relations to perceived lesson-specific achievement and personality traits for each SPIQ component are illustrated (Chapter 4).

The data was assessed within the intensive longitudinal DynASCEL project ("Dynamics of Academic Self-Concept in Everyday Life") that investigated academic experiences of German secondary school students (ninth and tenth grades) of the highest-ability track in their daily school life. This data is particularly rich in the way that the student sample is comparable to students assessed in large-scale assessments worldwide (OECD, 2014). Further, not only exhaustive trait variables were assessed at two measurement points, but an experience sampling phase was incorporated that produced many thousands of measurement points in four core school subjects that provide valuable insight into lesson-to-lesson variation of key educational constructs such as perceived instructional quality. 18 entire classrooms from six different schools from four German states were assessed that allowed for the inclusion of class-specific frames of reference in a multicenter design. Taken together, this project considerably advances insight into students' stable and dynamic experiences in the classroom.

Chapter 2

Social and Dimensional Comparison Effects in General and Domain-Specific Test Anxiety: A Nested Factor Modeling Approach

This contribution is accepted for publication as:

Talić, I., Sparfeldt, J. R., Möller, J., Renner, K.-H., Greiff, S., & Niepel, C. (accepted). Social and dimensional comparison effects in general and domain-specific test anxiety: A nested factor modeling approach. *Current Psychology*.

2. Social and Dimensional Comparison Effects in General and Domain-specific Test Anxiety: A Nested Factor Modeling Approach

Abstract

The generalized internal/external frame of reference (GI/E) model assumes social and dimensional achievement comparisons to form self-perceptions. These domain-specific comparisons have been shown to shape two facets of test anxiety (i.e., worry and emotionality) both directly and indirectly through academic self-concepts. However, examinations of such domain-specific relations have rarely integrated general components, although the hierarchical nature of both test anxiety and academic self-concept is well-known. Thus, the present study implemented a nested factor modeling approach. We examined social and dimensional comparison effects on worry and emotionality as well as mediation effects of academic self-concepts in the math and verbal domains while controlling for general components. We contrasted this approach with the conventionally used first-order factor model where general components were not considered. Data from N = 348 German secondary school students (M_{age} = 15.3 years, Grades 9-10) were analyzed using structural equation models. Direct negative within-domain and positive cross-domain achievement-anxiety relations emerged, yet, the pattern of cross-domain relations changed across modeling approaches. Only the nested factor model showed indirect cross-domain mediation relations. Our findings suggest the importance of structural representations of hierarchical constructs. The nested factor model approach enhanced predictions within the GI/E model, particularly those related to dimensional comparisons.

Keywords: Test anxiety; worry; emotionality; structural equation modeling; nested factor model; generalized internal/external frame of reference model

Introduction

Test anxiety (TA) comprises a set of detrimental reactions to potential failure in evaluative situations (Zeidner, 1998). Such reactions can ultimately lead to severe educational disadvantages (Zeidner, 2020). Major research efforts have been devoted to predicting TA with the goal of preventing this experience and creating effective interventions (von der Embse et al., 2013). To predict TA, domain-specific achievement-based comparison processes can be drawn on (Arens, Becker, & Möller, 2017; Marsh, 1988; Schilling et al., 2005; Streblow, 2004). Using the internal/external frame of reference (I/E) model, social (e.g., "*How good am I in math compared with my classmates?*") and dimensional (e.g., "*How good am I in math compared with German?*") comparison processes were postulated to shape the formation of the domain-specific academic self-concept (ASC; Marsh, 1986). ASC is typically defined as students' self-perceptions of their own competence (Marsh & Craven, 2006). In extending the I/E model to the generalized I/E (GI/E) model (Möller et al., 2016), TA was included as an outcome variable. Generalizing results from ASC research to TA research is facilitated on the basis of a link between the two constructs that comprises causality (i.e., ASC as determinant of TA; Marsh, 1988; see also Schilling et al., 2005) and structural similarities (i.e., hierarchical structure with general component at the apex; Gogol et al., 2016).

Domain-specificity plays a pivotal role in these social and dimensional comparison processes within the GI/E model. Regarding TA, however, general (e.g., Cassady & Johnson, 2002) and domain-specific (e.g., Sparfeldt et al., 2005) approaches seem to coexist in parallel. These are rarely combined, even though the consideration of general TA can alter domain-specific achievement-anxiety relations (Devine et al., 2012). Methodologically, the consideration of a hierarchical general factor can be achieved with a nested factor (NF) model (Gogol et al., 2016; Gogol et al., 2017, for ASC see also Arens et al., 2021; Brunner et al., 2009; Brunner et al., 2010). In this model, general variance is distinguished from domain-specific variance, which seems suitable with regard to the domain-specific processes that the GI/E model examines.

The overarching objective of the present study is therefore the application of the NF model within the GI/E framework. Specifically, domain-specific social and dimensional comparison processes regarding the TA facets (i.e., worry and emotionality) in the math and verbal domains are investigated, while controlling for general TA. In doing so, we examined both direct achievement-anxiety paths and indirect mediation paths through the ASC while controlling for general ASC. To illustrate differences between this model and the conventional choice of modeling strategy, we contrasted our NF model against a first-order factor (FOF) model in which general components were not considered and general variance was subsumed in domain-specific components. To achieve this, we first give an overview on the structure of TA and reiterate theoretical and empirical considerations for the inclusion of TA (facets) as outcome in the GI/E model, before introducing the NF model and proposing its application in investigating social and dimensional comparison effects in the GI/E model.

The Structure of Test Anxiety (TA)

Exam- or test-related concerns about failure or negative consequences can be seen in a set of detrimental phenomenological, physiological, and behavioral responses denoted as TA (Zeidner, 1998). TA is of high interest to researchers and practitioners due to its detrimental associations with academic achievement (Barroso et al., 2021; Hembree, 1988; von der Embse et al., 2018), and subjective well-being (Steinmayr et al., 2016). In the light of these relations, efforts have been directed towards investigating TA to enhance the understanding of its structure and antecedents.

Multidimensionality

In general, multidimensionality can be conceived with regard to different aspects, for instance, multidimensionality in terms of domain-specificity and multidimensionality in terms of different construct components or facets (Arens et al., 2011). Multidimensionality in terms of domain-specificity is discussed in detail in the paragraph on TA's hierarchy (see Section Hierarchy: Domain-Specificity and Generality), contrasting domain-specific and general approaches to TA. Here, we discuss multidimensionality in terms of different construct facets. Liebert and Morris (1967) identified two fundamental facets of TA, worry and emotionality. To date, there is a wide agreement on these two facets (e.g., Zeidner, 2020). Worry as the cognitive facet encompasses negative self-talk and failure-focused expectations or cognitions and generally shows stronger negative relations with academic achievement (e.g., Hembree, 1988; Steinmayr et al., 2016; von der Embse et al., 2018). Emotionality refers to perceived autonomic hyperarousal (e.g., rapid heartbeat or sweating) or feelings of nervousness (Morris et al., 1981). Worry and emotionality tend to be moderately correlated with each other but are conceptualized as two separate TA dimensions that respond to different stimuli in evaluative situations (Morris et al., 1981). This two-dimensional facet distinction has received consistent empirical support (e.g., Gogol et al., 2017; Hembree, 1988; Sparfeldt et al., 2013).

Hierarchy: Domain-Specificity and Generality

Multidimensionality in terms of domain-specificity has been investigated in TA, where both general approaches, assessing TA with regard to test situations *in general* (e.g., Cassady & Johnson, 2002) and domain-specific approaches, assessing TA with regard to specific domains or subjects (e.g., Sparfeldt et al., 2005) are present in the literature. Without discounting the general nature of TA, the importance of considering different anxiety contexts has been emphasized throughout, visible in the conceptualization of TA as a *situation-specific* or *contextualized* personality trait (see Zeidner, 2020). Accordingly, distinct domain (i.e., school subject) factors need to be considered in both TA facets, worry and emotionality (Sparfeldt et al., 2005). Yet, employing distinct domain-specific factors only (i.e., not considering general manifestations) cannot adequately represent the hierarchical structure of TA. Pointing out the importance of considering general *and* domain-specific manifestations simultaneously, Devine et al. (2012) reported changes in the pattern of results in achievement-anxiety relations when controlling for general anxiety levels.

The Link Between TA and ASC

Some of these construct characteristics (i.e., multidimensionality, domain-specificity, generality, hierarchy) apply not only to TA (i.e., phenomenological, physiological, and behavioral reactions associated with possible failure in tests or other evaluative situations; Zeidner, 1998), but also ASC (i.e., students' self-perceptions of their own competence; Marsh & Craven, 2006)—a key determinant of student achievement (Möller et al., 2020). The link between TA and ASC is discussed with regard to (a) structural similarities between TA and ASC, (b) causal relations between TA and ASC, and (c) the transfer of ASC to TA research. Similar to TA, (a) the structure of ASC has been subject to investigation (e.g., Brunner et al., 2010; for an overview of different structural models and their implications see Arens et al., 2021). Some recent works have placed emphasis on structural similarities of TA and ASC: Specifically, both constructs were paralleled with regard to their domain-specific structure with a general component at the apex of the hierarchy (Gogol et al., 2016; Gogol et al., 2017)[1]. Further, (b) a causal link between TA and ASC has been discussed with ASC as predictor of TA (i.e., low [high] ASC leading to higher [lower] TA in the same domain; Marsh, 1988). Empirical evidence for negative relations between ASC and TA has repeatedly been reported (Ahmed et al., 2012; Gogol et al., 2017; von der Embse et al., 2018). Despite presumed reciprocal relations between ASC and TA, the effect of ASC *on* TA seems to be more crucial (Schilling et al., 2005). Correspondingly, achievement-TA relations have been shown to be mediated through ASC (Arens, Becker, & Möller, 2017; Schilling et al., 2005). Finally, (c) on the basis of a causal relationship between TA and ASC, Marsh (1988) first suggested a possible application of ASC research-derived results to TA: "If self-concept is a causal determinant of anxiety, then processes affecting self-concept should also affect anxiety" (p. 139). In particular, this has been done with regard to the GI/E model.

The Generalized Internal/External Frame of Reference (GI/E) Model

The observation that math and verbal ASCs were nearly uncorrelated despite the substantial correlations of math and verbal achievement indicators led to the development of the I/E model that aimed to explain the formation of ASCs through achievement-based social and dimensional comparison processes (Marsh, 1986). The I/E model was later extended to the so-called the GI/E model (Möller et al., 2016) such that constructs other than ASC (i.e., in our case TA) could be considered as outcome variables

[1] In addition and analogously to TA, one could also distinguish an affective and a cognitive facet in ASC as another similarity between the two constructs (see Arens et al., 2011). In the present study, however, we defined ASC as self-perceptions of competence, emphasizing its cognitive nature.

of these comparisons. Within the GI/E model, where multiple domains are considered, the social and dimensional comparisons show in the direct within and cross-domain paths.

Direct Paths: Social and Dimensional Comparisons

When engaging in social comparisons, students draw on an external frame of reference and compare their achievement in one domain with relevant others' achievements in the same domain (Möller et al., 2016). Social comparison effects thereby appear as negative within-domain achievement-anxiety relations (e.g., higher math achievement is related to lower math TA; Arens, Becker, & Möller, 2017). Dimensional comparisons require an internal frame of reference as students compare their own achievements across domains (Möller et al., 2016; Möller & Marsh, 2013). Dimensional comparison effects appear as positive cross-domain achievement-anxiety relations (e.g., higher math achievement is related to higher German TA; Arens, Becker, & Möller, 2017).

With regard to the ASC, empirical support for GI/E-hypothesized comparison processes is ample (see meta-analysis by Möller et al., 2020). With regard to TA, evidence for GI/E-hypothesized social comparison effects on worry and emotionality has been consistent, whereas evidence for dimensional comparison effects has been tied to math but not verbal TA (Arens, Becker, & Möller, 2017; Schilling et al., 2005). In another study, dimensional comparison effects were observed in the verbal domain , yet applied to a measure of TA that did not differentiate between the worry and emotionality components (Marsh, 1988). Another study identified social comparison effects in math and two verbal domains, as well as dimensional comparison effects between one verbal domain (i.e., French) but not another verbal domain (i.e., German) and math anxiety, and between the two verbal domains in a multilingual context (van der Westhuizen et al., 2022). These findings were derived using a unidimensional measure of TA, that is not differentiating different facets. Thus, there is some evidence for the relevance of social and dimensional comparison processes in the formation of domain-specific TA. To gain further insight into the mechanism of these processes, mediation analyses have been conducted using the ASC as mediator.

Indirect Paths: Mediation via Academic Self-Concept (ASC)

Based on theoretical assumptions (i.e., school grades as source for self-perceptions of academic competence that in turn influence further socio-affective variables such as TA) and empirical relations between achievement and ASC on the one hand (Möller et al., 2020), and between ASC and TA on the other hand (see Section The Link Between TA and ASC), the possible mediation of achievement-anxiety relations through ASC seems straightforward. Accordingly, some empirical support has been reported (Arens, Becker, & Möller, 2017; Schilling et al., 2005). In the GI/E model, the mediation of social comparisons is indicated by substantial indirect within-domain paths (e.g., between achievement in math and TA in math through the math ASC) and the algebraic sign of the indirect path would be negative (i.e., multiplying a positive within-domain achievement-ASC relation with a negative within-domain ASC-TA relation). Inversely, the mediation of dimensional comparisons is indicated by substantial indirect cross-domain paths (e.g., between achievement in math and TA in German through the German ASC) and the sign of this indirect path would be positive (i.e., multiplying a negative cross-domain achievement-ASC relation with a negative within-domain ASC-TA relation).

Arens, Becker, and Möller (2017)reported significant indirect paths for both social and dimensional comparisons in both the math and verbal domains. (Schilling et al., 2005) reported significantly lower direct achievement-anxiety relations after including ASCs, that did not reach statistical significance in the verbal domain, suggesting differential mediation effects in both domains. Hence, evidence for a mediation of GI/E-based achievement-anxiety relations through ASCs is provided, yet worth replicating. In addition, even though all hypothesized relations within the GI/E model relate to domain-specific construct manifestations, the structural representation of TA and ASC implemented in these studies (Arens, Becker, & Möller, 2017; Marsh, 1988; Schilling et al., 2005; van der Westhuizen et al., 2022) did not allow for a precise disentanglement of different domain-specific construct components (e.g., a clear differentiation of math anxiety from verbal anxiety or general anxiety). To prevent a conglomeration of different domain-specific and general construct components that blur the examination of strict within- and cross-domain relations within the GI/E model, we argue for adapting the constructs' structural representations.

Structural Representation Within Nested Factors

To best represent the multidimensional and hierarchical construct structure of both TA and ASC, we chose a nested factor (NF) model (Gustafsson & Balke, 1993). This model can be used to decompose a given manifestation (e.g., math worry) into its general component, its domain-specific component, and measurement error (Eid et al., 2017). A general factor is specified to influence all (general and domain-specific) items, whereas domain-specific factors are specified to additionally influence their respective domain-specific items. The general factor serves as the reference domain, that is, the general items do not form their own domain-specific factor but load directly on the general factor along with all the other domain-specific items. Domain-specific factors are interpreted as residual factors (i.e., the part of the domain-specific manifestation that is not accounted for by the general component) and are thus uncorrelated with the general factor. Different domain-specific factors can correlate. This model is also referred to as the bifactor (S-1)-model because it has one domain-specific factor less than the number of domains that are included (Eid et al., 2017).

The NF model has been validated with regard to the ASC (i.e., Nested Marsh/Shavelson Model; Arens et al., 2021; Brunner et al., 2009; Brunner et al., 2010). Given the structural similarities of ASC and TA, it has been applied to TA as well (Gogol et al., 2016; Gogol et al., 2017). The NF model takes account of general manifestations operating at the apex of domain-specific manifestations (Brunner et al., 2010). In contrast to a higher-order factor model, where all domain-specific latent factors load on a higher-order latent general factor, the NF model shows superior model fit when correlations between the domain-specific factors are low (Arens et al., 2021). The NF model thus offers flexibility in representing relations between domain-specific factors as positive, negative, or zero, while retaining the meaning of the general factor due to its defined reference domain irrespective of the number and scope of the domains that are included (Eid et al., 2017).

Yet, within the GI/E model, the predominant modeling strategy is the first-order factor (FOF) model in which domain-specific items load on their respective domain-specific factors only. Hierarchical structures cannot be represented in this model (Arens et al., 2021). The domain-specific factors in the FOF model represent a mixture of general and domain-specific variances, which can distort relations with correlates (Brunner et al., 2009). In contrast to this mixture, in the NF model, domain-specific factors have a clear meaning and operate independently from general levels. This separation seems fruitful particularly in the GI/E framework in which the domain-specific relations are of upmost interest.

The Present Study

In the present study, we therefore demonstrated the application of the NF modeling approach in investigating relations in the GI/E model in contrast to the widely used FOF models. Specifically, we examined the role of social and dimensional comparisons in the formation of the TA facets worry and emotionality while disentangling general and domain-specific components to purify domain-specific relations—the core of the GI/E model. Contrasting the FOF to the NF models within the GI/E framework allowed us to examine the difference in result patterns concerning (mediated) social and dimensional comparison effects on the two TA facets in dependence on the modeling strategy. In other words, we controlled for the influence of general TA and ASC levels on domain-specific TA and ASC manifestations and with this, potentially draw a more complete picture of social and dimensional comparisons—comparisons that students naturally engage in and thus hold important implications both on theoretical (e.g., investigating the strength of dimensional comparisons when general levels are controlled for) and practical grounds (e.g., adjusting psychoeducation in TA interventions).

A careful synthesis of the current literature indicated that our study is the first one to examine direct and ASC-mediated social and dimensional comparison effects on the TA facets worry and emotionality in the math and verbal domains in the GI/E model (Arens, Becker, & Möller, 2017; Schilling et al., 2005), that controlled for general TA and ASC using an NF modeling strategy (Gogol et al., 2016; Gogol et al., 2017).

In our first research question (RQ), we aimed to replicate findings reported in previous studies (Arens, Becker, & Möller, 2017; Schilling et al., 2005) that used FOF models to observe social comparison effects of grades as achievement indicators on facets of TA in both math and German and dimensional comparison effects on facets of TA in German secondary school students. Further, in Arens, Becker, and Möller (2017), all direct achievement-TA paths were fully mediated by ASC in both domains, whereas Schilling et al. (2005) reported full mediations in German and partial mediations in math.

RQ1: Replication. Can prior work (Arens, Becker, & Möller, 2017; Schilling et al., 2005) be replicated with regard to (a) FOF GI/E model relations and (b) FOF mediated GI/E model relations?

Second, we addressed our focal RQ, which aimed at investigating social and dimensional comparison effects on facets of TA when controlling for general manifestations in the NF model. In doing so, we examined direct paths between grades and facets of TA as well as indirect, ASC-mediated paths. Hereby, general TA and ASC were controlled for within nested factors. Hence, this RQ addressed the question of whether domain-specific social and dimensional comparisons influence facets of TA irrespective of general TA and whether these relations are mediated by domain-specific ASCs when controlling for general ASC.

RQ2: Extension. Can (a) GI/E model relations and (b) mediated GI/E model relations be detected when employing the NF modeling approach? How will (c) statistical predictions differ across the NF versus FOF models?

Third, the NF models include additional paths that are not formalized in the original GI/E model. Transferring domain-specific processes to general processes, we examined relations between achievement and general facets of TA, as well as their mediation by general ASC.

RQ3: Ancillary. How do additional paths between grades and general worry and emotionality show in the NF GI/E model, and are these paths mediated by general ASC in the NF mediated GI/E model?

Method

Procedure and Participants

The present work is part of the larger “Dynamics of Academic Self-Concept in Everyday Life” (DynASCEL) project (Niepel et al., 2022) on students’ perceptions of academic competence and learning environments, where an intensive longitudinal experience sampling design was embedded in a paper-and-pen pre- and post-assessment. In the present study, only selected data from the pre-assessment were relevant for our research questions.[2] We recruited a convenience sample of N = 348 German secondary school students (43.1% of whom were male students based on n = 340 students with available gender information) attending the ninth (n = 288) and 10th (n = 60) grades of the highest ability track (i.e., the German *Gymnasium*). Students were nested within 18 classrooms from six schools located in four different German federal states (i.e., Baden-Württemberg, Mecklenburg-Vorpommern, Nordrhein-Westfalen, Rheinland-Pfalz). Participants reported a mean age of 15.3 years (SD = 0.66, Range = 13.3 to 17.4 years; based on n = 335). Student clusters within classrooms were stable across school subjects and across school grades, such that students were asked to refer to the same math and German test situations. The APA Ethics Code (American Psychological Association, 2020) was considered in all stages of the research process to ensure scientific accuracy whilst protecting the rights and welfare of the minor participants. Specifically, student participation was voluntary, participants could withdraw from the study at any time without stating any reasons and without facing any negative consequences, and written parental consent was obtained for all participating students. Students, parents, and schools were exhaustively informed on the study's purposes and subsequent data processing. All measures and procedures were approved by the local ethics review panel of the University of Luxembourg and by all involved education authorities in the respective four German federal states.

Measures

Test Anxiety (TA)

TA was assessed with general (i.e., school in general) and domain-specific (i.e., math and verbal domains) adaptations of worry and emotionality items based on the German Test Anxiety Inventory (TAI-G; Hodapp, 1991; Hodapp et al., 2011). Following the introduction "*In evaluative situations (e.g., tests, written or oral examinations),*"

[2] Data from the larger research project, including data used in the present study, have been and will be used in other manuscripts, yet addressing different research questions (e.g., see Dörendahl et al., 2021; Franzen et al.2022; Hausen et al., 2022).

students responded to the five parallel-worded item stems for worry (e.g., "*I worry about my results*") and emotionality, each (e.g., "*I feel anxious*"). The items were presented in a grid format as first introduced by Rost and Sparfeldt (2002), where the item stems were presented in rows with a placeholder "…" (for the target domain), and the target domains (i.e., school, math, German) were presented in columns. The students related the items stems from the rows to the target domain in the column and responded using a 6-point Likert scale in the cells of the grid, ranging from 1 (*almost never*) to 6 (*almost always*) such that higher scores represented higher TA (see also Sparfeldt et al., 2005; Sparfeldt et al., 2013). Domain-specific worry and emotionality ratings presented in this format have been shown to be reliable with ω coefficients ≥ .91 and measurement invariant across school subjects (Schneider et al., 2022).

Report Card Grades

Students reported their math and German grades from their last report card, which we used as academic achievement indicators. Self-reported and actual grades tend to be highly correlated in German school student samples, indicating the reliability and validity of self-reported grades ($r \geq .91$, Sparfeldt et al., 2008; see $r \geq .76$ for grades 9 and 10 in a German-speaking Swiss sample reported by Sticca et al., 2017 and $r = .88$ for grades 7 and 8 across three German school tracks reported by Dickhäuser & Plenter, 2005). School grades in Germany are measured on a 6-point Likert scale, which we recoded so that higher values corresponded with higher achievements, ranging from 1 (*insufficient*) to 6 (*excellent*).

Academic Self-Concept (ASC)

General (i.e., school in general) and domain-specific (i.e., math and verbal domains) ASCs were assessed using six parallel-worded items each, which were based on the well-validated and reliable Self-Description Questionnaire (SDQ; Marsh et al., 1983) and the short scale by Gogol et al. (2014). An example item is "*I am good at [most school subjects] / [math] / [German].*" Gogol et al. (2014) reported reliability coefficients of ω ≥ .75 for their three-item short scales. Items were rated on a 6-point Likert scale ranging from 0 (*does not apply at all*) to 5 (*completely applies*) such that higher values indicated higher ASCs.

Statistical Analyses

Statistical analyses were performed within the structural equation modeling (SEM) framework using the software package Mplus 8.3 (L. K. Muthén & Muthén, 1998-2017). To adjust standard errors for the nonindependence of observations because students were clustered in classrooms, we used the "TYPE = COMPLEX" option. Correlated uniqueness was considered by allowing for correlated residual variances between parallel-worded items. We used the MLR estimator to obtain robust standard errors and deal with missing data (Kaplan, 2009; L. K. Muthén & Muthén, 1998-2017). The percentages of missing values ranged from 1.7% to 3.2% for worry, 2.6% to 4.9% for emotionality, and 2.3% to 3.7% for ASC across domains. 4.3% and 5.2% were missing in grades in math and German, respectively.

To reduce the complexity and support the power of the model, we (a) reduced the number of indicators per factor (Rick H. Hoyle & Gottfredson, 2015), selecting three items for each TA facet and ASC out of the larger item sets based on the size of the factor loadings (see also Marsh et al., 2006).[3] Such short scales have been shown to measure TA and ASC reliably (Gogol et al., 2014). Further, we (b) adopted a two-step approach (Anderson & Gerbing, 1988), in which we, first, conducted confirmatory factor analyses based on which we extracted values for factor loadings, item residual variances, and exogenous factor variances. Second, we fixed these parameters to these values when specifying the structural models. To enter school grades as latent single-item factors, we followed the procedure illustrated by Kline (2016), fixing factor loadings to 1 and fixing residual variances to constant values that were derived from the indicators' empirical sample variance and the reliability estimate reported in previous work (Sparfeldt et al., 2008).

[3] To test the robustness of our results, we additionally ran all models with the respective full item sets. Descriptively, the model fits were lower compared with the models using three-item scales (i.e., CFI ≥ .939, RMSEA ≤ .061, SRMR ≤ .062). The pattern of significant within- and cross-domain paths was identical across models using three items versus models using the full item sets.

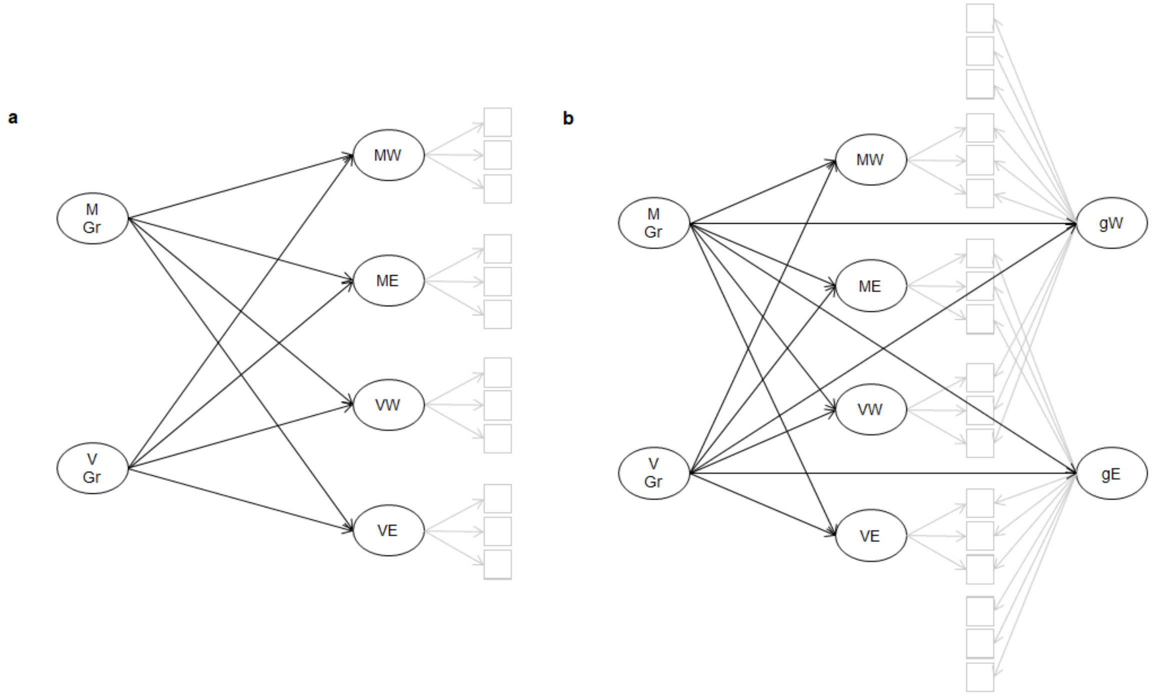
To estimate the replicability of result patterns in the FOF model (RQ1), we specified the FOF GI/E model with domain-specific grades as predictors, domain-specific worry and emotionality as criteria, and within- or cross-domain regression paths between each predictor and each criterion. Importantly, domain-specific factors influenced their corresponding domain-specific indicators only (see Figure 1a). In the FOF mediated GI/E model, domain-specific ASCs were included as mediator variables. Indirect relations were requested using the "MODEL INDIRECT" option in Mplus. Analogous to Arens, Becker, and Möller (2017), we focused on the paths for which the ASC and TA facets belonged to the same domain. To estimate the sizes of the indirect effects, we calculated squared standardized indirect path coefficients as measures of explained variance in accordance with Lachowicz et al.'s (2018) recommendations. Thus, cut-off criteria for proportions of explained variance were applied (i.e., small = 2%, medium = 15%, large = 25%; Cohen, 1988).

To investigate GI/E-hypothesized relations while controlling for general components (RQ2), we next implemented the NF modeling approach by adding general factors (the S-1 specification according to Eid et al., 2017). To this end, domain-specific (i.e., math and German) worry and emotionality items were specified to load on their respective domain-specific latent factors. In addition, we specified general factors, which influenced all worry or emotionality domain-specific and general items. Correlations between the general and its domain-specific factors (e.g., general worry and math worry) were fixed to zero. Relations among domain-specific factors were allowed. In the NF GI/E model, domain-specific grades were entered as predictors, and domain-specific and general worry and emotionality were entered as criteria (see Figure 1b). In the NF mediated GI/E model, we added domain-specific and general ASCs as mediator variables. Here, ASC was represented analogously within nested factors.

Clearly, the NF modeling approach yielded relations that had not been formalized in the original GI/E model (i.e., paths between grades and general TA and their mediation via general ASC), which we additionally addressed in RQ3. For model evaluation, we followed the recommended cut-off criteria in the absolute goodness-of-fit indices CFI, RMSEA, and SRMR, where values of CFI ≥ .95, RMSEA ≤ .06, and SRMR ≤ .08 are considered to indicate a good fit to the data (Hu & Bentler, 1999).

Figure 1

Different Modeling Approaches in Testing GI/E Relations



Note. Testing GI/E relations in different modeling approaches: (a) the first-order factor (FOF) GI/E model, where facets of test anxiety (TA; i.e., worry and emotionality) are predicted by grades via within- and cross-domain paths in two domains (Model 1b), and (b) the nested factor (NF) GI/E model, where domain-specific and general facets of TA are predicted by grades via within- and cross-domain paths in two domains (Model 3b). M Gr = Math grade; V Gr = German grade; MW = Math worry; ME = Math emotionality; VW = German worry; VE = German emotionality; gW = general worry; gE = general emotionality. For better clarity of presentation, measurement models are displayed in grey, and item residual variances, factor variances, and correlational paths have been omitted.

Results

Preliminary Analyses

All preliminary confirmatory factor analyses showed a good fit to the data for both the FOF and NF modeling approaches (see Models 1a, 2a, 3a, and 4a in Table 1).[4] Table 2 presents the standardized factor loadings as well as the corresponding McDonald's ω reliability coefficients.[5] All factor loadings differed significantly from zero and were moderate to large in both modeling approaches. The reliability coefficients of the three-item scales were ω ≥ .88 in the FOF model and ω ≥ .70 in the NF model across factors and domains (see Table 2).

[4] To ensure the applicability of the NF models compared with higher-order factor models in which general factors load on domain-specific latent factors, we also computed higher-order factor models. These models showed inadequate model fits when we considered the combination of the aforementioned fit indices.

[5] Note that in Table 2, only Models 2a and 4a are reported because these include all examined constructs. Standardized factor loadings differed between Model 1a versus 2a and between Model 3a versus 4a only negligibly with Δλ ≤ .006 in all cases.

Table 1

Goodness-of-Fit Indices for Tested Models

Model		MLR χ2 (*df*)	CFI	RMSEA [90 % CI]	SRMR
	Preliminary confirmatory factor analyses				
1a	First-order factor measurement model with six factors (i.e., worry, emotionality, and grades in math and German)	72.915 (58)	0.995	0.027 [0.000; 0.045]	0.024
2a	First-order factor measurement model with eight factors (i.e., worry, emotionality, academic self-concepts, and grades in math and German)	294.289 (135)	0.970	0.058 [0.049; 0.067]	0.032
3a	Nested factor measurement model with eight factors (i.e., worry and emotionality in math, German and general, and grades in math and German)	151.966 (118)	0.993	0.029 [0.012; 0.041]	0.025
4a	Nested factor measurement model with 11 factors (i.e., worry, emotionality, and academic self-concept in math, German and general, and grades in math and German)	490.707 (285)	0.974	0.046 [0.039; 0.052]	0.033
	First-order factor models				
1b	First-order factor GI/E model with six factors (i.e., worry, emotionality, and grades in math and German)	66.227 (80)	1.000	0.000 [0.000; 0.017]	0.024
2b	First-order mediated GI/E model with eight factors (i.e., worry, emotionality, academic self-concept, and grades in math and German)	278.925 (168)	0.979	0.044 [0.034; 0.052]	0.034
	Nested factor models				
3b	Nested factor GI/E model with eight factors (i.e., worry and emotionality in math, German and general, and grades in math and German)	140.231 (162)	1.000	0.000 [0.000; 0.012]	0.028
4b	Nested factor mediated GI/E model with 11 factors (i.e., worry, emotionality, and academic self-concept in math, German and general, and grades in math and German)	493.847 (359)	0.983	0.033 [0.025; 0.040]	0.041

Note. MLR = Maximum likelihood estimation with robust standard errors; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; SRMR = Standardized Root Mean Square Residual.

Table 2

Standardized Factor Loadings and McDonald's ω Reliability Coefficient

Factor	First-Order Factor measurement model (Model 2a)				Nested Factor measurement model (Model 4a)			
	Item 1	Item 2	Item 3	ω	Item 1	Item 2	Item 3	ω
Worry$_{\text{math}}$	.807	.892	.880	.897	.401	.498	.458	.702
Worry$_{\text{German}}$	.798	.867	.867	.883	.560	.638	.597	.792
Emotionality$_{\text{math}}$	.858	.931	.892	.923	.440	.472	.485	.763
Emotionality$_{\text{German}}$	.860	.888	.853	.900	.600	.592	.577	.807
ASC$_{\text{math}}$	.909	.952	.916	.947	.688	.815	.741	.924
ASC$_{\text{German}}$	.889	.936	.922	.940	.733	.850	.791	.924
	Ranges Item 1 through Item 9			ω	Ranges Item 1 through Item 9			ω
Worry$_{\text{general}}$	NA			NA	$.570 \leq \lambda \leq .875$			.939
Emotionality$_{\text{general}}$	NA			NA	$.616 \leq \lambda \leq .888$			.954
ASC$_{\text{general}}$	NA			NA	$.407 \leq \lambda \leq .856$			.944

Note. ASC = Academic self-concept; NA = Not applicable in first-order factor modeling.
All reported standardized factor loadings were significant at $p < .001$

The latent factor correlations across the modeling approaches can be found in Table 3. Significantly negative within-domain relations between TA and grades and ASCs were observed in all domain-specific TA facets except for German worry, where only one relation to German ASC differed significantly from zero in the NF model. Relations within a facet changed considerably across modeling approaches (i.e., $\rho = .72$, $p < .001$ [$\rho = .33$, $p = .016$] for math and German worry in the FOF [NF] model, and $\rho = .59$, $p < .001$ [$\rho = -.11$, $p = .408$] for math and German emotionality in the FOF [NF] model). Across domains and modeling approaches, domain-specific worry and emotionality were moderately correlated with each other (i.e., $\rho = .45$ to $\rho = .58$). In the NF model, general worry and emotionality were positively correlated at $\rho = .55$ ($p < .001$), and each was negatively related to the math grade (i.e., general worry: $\rho = -.15$, $p = .044$; general emotionality: $\rho = -.21$, $p = .001$).

Table 3

Latent Factor Correlations in the First-Order Factor and Nested Factor Modeling Approach

	Factor Correlations										
	1	2	3	4	5	6	7	8	9	10	11
1. Worry$_{\text{math}}$	-	.723***	.575***	.303***	-.293***	.007	-.269***	.158***	NA	NA	NA
2. Worry$_{\text{German}}$	.334*	-	.349***	.471***	-.111	.034	-.087	.063	NA	NA	NA
3. Emotionality$_{\text{math}}$	.454***	-.055	-	.586***	-.423***	-.138	-.515***	.086	NA	NA	NA
4. Emotionality$_{\text{German}}$	-.023	.448***	-.107	-	-.085	-.196*	-.116*	-.273***	NA	NA	NA
5. Grade$_{\text{math}}$	-.306***	-.004	-.449***	.115*	-	.483***	.736***	.028	NA	NA	NA
6. Grade$_{\text{German}}$	-.058	.006	-.064	-.156*	.482***	-	.207***	.634***	NA	NA	NA
7. ASC$_{\text{math}}$	-.242*	-.041	-.427***	.212***	.453***	-.202***	-	-.028	NA	NA	NA
8. ASC$_{\text{German}}$	.123	-.140*	.342***	-.430***	-.327***	.352***	-.436***	-	NA	NA	NA
9. Worry$_{\text{general}}$	.000[a]	.000[a]	.098	-.057	-.149*	.045	-.161*	.200***	-	NA	NA
10. Emotionality$_{\text{general}}$	.047	.029	.000[a]	.000[a]	-.212*	-.105	-.253***	.096	.545**	-	NA
11. ASC$_{\text{general}}$	.087	.126*	-.199***	-.009	.628***	.651***	.000[a]	.000[a]	-.061	-.199*	-

Note. Correlations below the diagonal represent correlations within the nested factor model containing all examined factors (i.e., Model 4a), whereas correlations above the diagonal represent correlations within the first-order factor model containing all examined factors (i.e., Model 2a). ASC = Academic self-concept; NA = Not applicable in first-order factor modeling. [a] Fixed to zero due to nested factor modeling.

* $p < .05$. ***$p < .001$.

The First-Order Factor (FOF) GI/E Model

First, we addressed RQ1 to replicate prior findings with (a) the FOF GI/E and (b) the FOF mediated GI/E model. The (a) FOF GI/E model showed an excellent fit to the data (Model 1b in Table 1). Table 4 presents the standardized path coefficients and standard errors. Negative within-domain paths between grades and facets of TA, indicating social comparison effects, differed significantly from zero for worry and emotionality in math (math grade → math worry, $\beta = -.38$ and math grade → math emotionality, $\beta = -.47$, *p*s < .001) and emotionality in German (German grade → German emotionality, $\beta = -.21$, $p = .001$). Statistically significant positive cross-domain paths between math [German] grades and facets of TA in German [math], indicating dimensional contrast effects, were only observed between the German grade and worry in math (German grade → math worry, $\beta = .19$, $p = .005$). Hence, prior work could be replicated regarding social comparison effects on all domain-specific TA facets except for worry in German. Dimensional comparison effects on both TA facets in math (Arens, Becker, & Möller, 2017; Schilling et al., 2005) were only replicated with regard to worry in math.

Table 4

Standardized Path Coefficients for the First-Order Factor and Nested Factor GI/E Model

	First-Order Factor GI/E Model (Model 1b)		Nested Factor GI/E Model (Model 3b)	
	β	*SE*	β	*SE*
Within-domain paths from grades to test anxiety				
$\text{Grade}_{\text{math}}$ → $\text{Worry}_{\text{math}}$	-0.381***	0.071	-0.372**	0.108
$\text{Grade}_{\text{math}}$ → $\text{Emotionality}_{\text{math}}$	-0.466***	0.049	-0.560***	0.082
$\text{Grade}_{\text{German}}$ → $\text{Worry}_{\text{German}}$	0.105	0.086	0.001	0.111
$\text{Grade}_{\text{German}}$ → $\text{Emotionality}_{\text{German}}$	-0.209**	0.065	-0.286***	0.051
Cross-domain paths from grades to test anxiety				
$\text{Grade}_{\text{math}}$ → $\text{Worry}_{\text{German}}$	-0.164	0.095	-0.015	0.120
$\text{Grade}_{\text{math}}$ → $\text{Emotionality}_{\text{German}}$	0.014	0.056	0.246***	0.058
$\text{Grade}_{\text{German}}$ → $\text{Worry}_{\text{math}}$	0.186**	0.067	0.116	0.101
$\text{Grade}_{\text{German}}$ → $\text{Emotionality}_{\text{math}}$	0.082	0.062	0.201*	0.081
Additional paths				
$\text{Grade}_{\text{math}}$ → $\text{Worry}_{\text{general}}$	NA	NA	-0.204**	0.077
$\text{Grade}_{\text{math}}$ → $\text{Emotionality}_{\text{general}}$	NA	NA	-0.205**	0.074
$\text{Grade}_{\text{German}}$ → $\text{Worry}_{\text{general}}$	NA	NA	0.140*	0.067
$\text{Grade}_{\text{German}}$ → $\text{Emotionality}_{\text{general}}$	NA	NA	-0.021	0.073

Note. *SE* = Standard error; NA = Not applicable in first-order factor modeling.

* *p* < .05. ** *p* < .01. *** *p* < .001.

To examine (b) the FOF mediated GI/E model, we added math and German ASCs as mediator variables. The model showed a good fit to the data (Model 2b in Table 1). Table 5 presents the standardized direct path coefficients along with standard errors. Only one direct path between grades and facets of TA, both within and across domains, was statistically significantly different from zero (i.e., math grade → math worry, β = -.28, *p* = .035). The direct relations between grades and ASC were all significantly positive within matching domains (math grade → math ASC, β = .82 and German grade → German ASC, β = .80, *p*s < .001) and negative across nonmatching domains (i.e., math grade → German ASC, β = -.34, *p* < .001, and German grade → math ASC, β = -.18, *p* = .008), replicating the original I/E pattern. Direct paths between ASC and facets of TA reached statistical significance in a few cases, and if so, they were negative within domains (i.e., math ASC → math emotionality, β = -.46, *p* < .001, and German ASC → German emotionality, β = -.27, *p* = .006) and positive across domains (i.e., German ASC → math worry, β = .11, *p* = .042, and German ASC → math emotionality, β = .15, *p* = .029). Finally, indirect paths (see Table 6) were significantly negative in two out of four cases within matching domains (i.e., math grade → math ASC → math emotionality, β = -.38, *p* = .001 and German grade → German ASC → German emotionality, β = -.22, *p* = .006). None of the four indirect paths across nonmatching domains were significantly different from zero. Thus, in contrast to prior work, we found within-domain mediations that were related to emotionality only.

Table 5

Standardized Direct Path Coefficients and Standard Errors for the First-Order Factor and the Nested Factor GI/E Mediation Model

	First-Order Factor GI/E Mediation Model (Model 2b)		Nested Factor GI/E Mediation Model (Model 4b)	
	β	*SE*	β	*SE*
Direct within-domain paths from grades to test anxiety				
Grade$_{\text{math}}$ → Worry$_{\text{math}}$	-0.280*	0.133	-0.349**	0.112
Grade$_{\text{math}}$ → Emotionality$_{\text{math}}$	-0.027	0.136	-0.208	0.111
Grade$_{\text{German}}$ → Worry$_{\text{German}}$	0.124	0.126	0.170	0.168
Grade$_{\text{German}}$ → Emotionality$_{\text{German}}$	-0.012	0.117	0.056	0.107
Direct cross-domain paths from grades to test anxiety				
Grade$_{\text{math}}$ → Worry$_{\text{German}}$	-0.190	0.138	-0.150	0.182
Grade$_{\text{math}}$ → Emotionality$_{\text{German}}$	0.041	0.128	-0.083	0.100
Grade$_{\text{German}}$ → Worry$_{\text{math}}$	0.086	0.092	0.117	0.110
Grade$_{\text{German}}$ → Emotionality$_{\text{math}}$	-0.123	0.089	-0.088	0.101
Additional paths				
Grade$_{\text{math}}$ → Worry$_{\text{general}}$	NA	NA	-0.224**	0.078
Grade$_{\text{math}}$ → Emotionality$_{\text{general}}$	NA	NA	-0.162	0.088
Grade$_{\text{German}}$ → Worry$_{\text{general}}$	NA	NA	0.186*	0.076
Grade$_{\text{German}}$ → Emotionality$_{\text{general}}$	NA	NA	0.138	0.090
Direct paths from grades to academic self-concept (mediator)				
Grade$_{\text{math}}$ → ASC$_{\text{math}}$	0.817***	0.050	0.738***	0.050
Grade$_{\text{math}}$ → ASC$_{\text{German}}$	-0.342***	0.080	-0.655***	0.077
Grade$_{\text{math}}$ → ASC$_{\text{general}}$	NA	NA	0.396***	0.066
Grade$_{\text{German}}$ → ASC$_{\text{math}}$	-0.175**	0.067	-0.579***	0.064
Grade$_{\text{German}}$ → ASC$_{\text{German}}$	0.798***	0.039	0.665***	0.062
Grade$_{\text{German}}$ → ASC$_{\text{general}}$	NA	NA	0.477***	0.052
Direct paths from academic self-concept (mediator) to test anxiety				
ASC$_{\text{math}}$ → Worry$_{\text{math}}$	-0.078	0.111	-0.053	0.078
ASC$_{\text{math}}$ → Emotionality$_{\text{math}}$	-0.461***	0.119	-0.325***	0.088
ASC$_{\text{math}}$ → Worry$_{\text{German}}$	0.026	0.099	-0.014	0.118
ASC$_{\text{math}}$ → Emotionality$_{\text{German}}$	-0.151	0.108	0.044	0.083
ASC$_{\text{German}}$ → Worry$_{\text{math}}$	0.112*	0.055	-0.036	0.066
ASC$_{\text{German}}$ → Emotionality$_{\text{math}}$	0.149*	0.068	0.149	0.090
ASC$_{\text{German}}$ → Worry$_{\text{German}}$	-0.006	0.065	-0.234**	0.071
ASC$_{\text{German}}$ → Emotionality$_{\text{German}}$	-0.273**	0.100	-0.458***	0.089
ASC$_{\text{general}}$ → Worry$_{\text{general}}$	NA	NA	-0.034	0.104
ASC$_{\text{general}}$ → Emotionality$_{\text{general}}$	NA	NA	-0.215	0.126

Note. ASC = Academic self-concept; *SE* = Standard error; NA = Not applicable in first-order factor modeling.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 6

Standardized Indirect Path Coefficients, Standard Errors and Effect Sizes for the First-Order Factor and the Nested Factor GI/E Mediation Model

	First-Order Factor GI/E Mediation Model (Model 2b)			Nested Factor GI/E Mediation Model (Model 4b)		
	β_{ind}	*SE*	β_{ind}^2	β_{ind}	*SE*	β_{ind}^2
Indirect within-domain paths						
Grade$_{\text{math}}$ → ASC$_{\text{math}}$ → Worry$_{\text{math}}$	-0.064	0.093	0.004	-0.039	0.057	0.001
Grade$_{\text{math}}$ → ASC$_{\text{math}}$ → Emotionality$_{\text{math}}$	-0.377**	0.112	0.142	-0.240***	0.068	0.058
Grade$_{\text{German}}$ → ASC$_{\text{German}}$ → Worry$_{\text{German}}$	-0.005	0.052	0.000	-0.156**	0.046	0.024
Grade$_{\text{German}}$ → ASC$_{\text{German}}$ → Emotionality$_{\text{German}}$	-0.217**	0.080	0.047	-0.304***	0.068	0.092
Indirect cross-domain paths						
Grade$_{\text{math}}$ → ASC$_{\text{German}}$ → Worry$_{\text{German}}$	0.002	0.023	0.000	0.153**	0.050	0.023
Grade$_{\text{math}}$ → ASC$_{\text{German}}$ → Emotionality$_{\text{German}}$	0.093	0.048	0.009	0.300***	0.070	0.090
Grade$_{\text{German}}$ → ASC$_{\text{math}}$ → Worry$_{\text{math}}$	0.014	0.022	0.000	0.031	0.046	0.001
Grade$_{\text{German}}$ → ASC$_{\text{math}}$ → Emotionality$_{\text{math}}$	0.081	0.042	0.007	0.188***	0.049	0.035
Additional indirect paths						
Grade$_{\text{math}}$ → ASC$_{\text{general}}$ → Worry$_{\text{general}}$	NA	NA	NA	-0.014	0.040	0.000
Grade$_{\text{math}}$ → ASC$_{\text{general}}$ → Emotionality$_{\text{general}}$	NA	NA	NA	-0.085	0.049	0.007
Grade$_{\text{German}}$ → ASC$_{\text{general}}$ → Worry$_{\text{general}}$	NA	NA	NA	-0.016	0.050	0.000
Grade$_{\text{German}}$ → ASC$_{\text{general}}$ → Emotionality$_{\text{general}}$	NA	NA	NA	-0.103	0.059	0.010

Note. ASC = Academic self-concept; *SE* = Standard error; NA = Not applicable in first-order factor modeling.

* *p* < .05. ** *p* < .01. *** *p* < .001.

The Nested Factor (NF) GI/E Model

Applying the Nested Factor (NF) Approach to the GI/E Model

Subsequently, we addressed RQ2, aimed at applying the NF modeling approach to (a) the GI/E model and (b) the mediated GI/E model and (c) contrasting statistical predictions in the NF versus FOF models. The (a) NF GI/E model showed an excellent fit to the data (Model 3b in Table 1). Statistically significant negative within-domain paths were found in three out of four cases (i.e., math grade → math worry, β = -.37, *p* = .001, math grade → math emotionality, β = -.56, *p* < .001 and German grade →

German emotionality, β = -.29, *p* < .001), indicating social comparison effects. Statistically significant positive cross-domain paths were found only for emotionality (i.e., math grade → German emotionality, β = .25, *p* < .001, and German grade → math emotionality, β = .20, *p* = .013), indicating dimensional comparison effects (see Table 4).

The (b) NF mediated GI/E model, including general and domain-specific ASCs as mediator variables, showed a good fit to the data (see Model 4b in Table 1). Standardized direct path coefficients (Table 5) indicated one significantly negative within-domain path (i.e., math grade → math worry, β = -.35, *p* = .002) and no significant direct cross-domain paths between grades and facets of TA. Concerning indirect (Table 6) within-domain paths, all except for the path between math grade and math worry via math ASC were significantly negative (i.e., math grade → math ASC → math emotionality, β = -.24, German grade → German ASC → German worry, β = -.16, and German grade → German ASC → German emotionality, β = -.30, all *p*s < .01). Indirect cross-domain paths were significantly positive for all except for the path between German grade and math worry via math ASC (i.e., math grade → German ASC → German worry, β = .15, math grade → German ASC → German emotionality, β = .30, and German grade → math ASC → math emotionality, β = .19, all *p*s < .01). Thus, a mediation of social and dimensional comparison effects on facets of TA via ASCs was supported for math emotionality and German worry and emotionality.

Finally, we (c) contrasted the statistical predictions made by the FOF versus NF models. In the nonmediated set of models (Model 1b vs. Model 3b; Table 4), the pattern of significant within-domain relations, indicative of social comparison effects, did not change, but the strength of grade-emotionality associations was descriptively higher in the NF model. However, the pattern of results regarding positive cross-domain paths, indicative of dimensional comparison effects, changed in the FOF model (i.e., one path to math worry) versus the NF model (i.e., two paths to emotionality in both domains). In the mediated set of models (Model 2b vs. 4b; Tables 5 and 6), the resulting pattern of significant and nonsignificant direct paths was virtually the same. Concerning the indirect paths, a pronounced change was visible, with one additional significant indirect within-domain and three additional indirect cross-domain paths in the NF mediated GI/E model as opposed to the FOF mediated GI/E model. Thus, particularly cross-domain paths (both direct and indirect) changed with respect to the modeling strategy.

General Paths

The NF models yielded additional general paths, which we addressed in our ancillary RQ3. In the NF GI/E model (i.e., Model 3b, see Table 4), the math grade was significantly negatively related to general worry and emotionality (math grade → general worry, β = -.20, *p* = .008 and math grade → general emotionality, β = -.21, *p* = .006), whereas the German grade was positively related to general worry (German grade → general worry, β = .14, *p* = .037). In the NF mediated GI/E model (i.e., Model 4b, see Table 5), two of these relations were significantly different from zero (i.e., math grade → general worry, β = -.22, *p* = .004, and German grade → general worry, β = .19, *p* = .015; see Table 5). With regard to the general ASC, significant positive relations with both grades were found (math grade → general ASC, β = .40, and German grade → general ASC, β = .48, *p*s < .001). However, there were neither significant direct paths between general ASC and general worry or emotionality nor significant indirect paths between grades and general facets of TA via general ASC (see Table 6). In conclusion, direct paths to general TA were found with regard to worry, but they were not mediated by the general ASC.

Discussion

The present study combined a general with a domain-specific approach using the NF model to examine the role of social and dimensional comparison effects on the formation of two facets of TA (i.e., worry and emotionality) in two different domains (i.e., math and verbal) within the GI/E model. The overarching aim was to apply an NF modeling strategy to the GI/E model to control for general proportions of TA within domain-specific TA, ultimately purifying domain-specific relations to academic achievement and self-concept. We investigated these relations in NF models and also contrasted them against relations identified with conventional modeling strategies that do not consider hierarchical construct structures (i.e., FOF models). In doing so, we examined domain-specific achievement-based relations—the core of the GI/E model—while controlling for general components. To the best of our knowledge, this is the first study to do so with regard to social and dimensional comparison effects on facets of TA.

First-Order Factor (FOF) Models: A Replication

First, when employing the FOF modeling strategy, we were able to replicate prior findings to a large extent (RQ1). Specifically, previous studies had reported social comparison effects on both facets in the math and verbal domains and dimensional comparison effects on both facets in math only (Arens, Becker, & Möller, 2017; Schilling et al., 2005; see Marsh, 1988, who also found dimensional comparison effects on an English TA measure that did not differentiate worry and emotionality, and van der Westhuizen et al., 2022, who found dimensional comparison effects of French achievement on math TA, and of French [German] achievement on German [French] TA, again using TA measures that did not differentiate worry and emotionality). In the present study, we found social comparison effects on both facets in math and on emotionality in German. In addition, we replicated the dimensional comparison effects on worry in math. The effect on math emotionality did not reach statistical significance in our case but was almost identical in its effect size (Arens, Becker, & Möller, 2017) such that the lack of significance was likely a matter of statistical power. TA in German (which was the native language for most students) did not seem to be subject to dimensional comparisons in either study, possibly due to the fact that self-perceptions in the verbal domain are not as restricted to the school context as in math domains. Accordingly, students have more sources of self-evaluation (other than academic math achievement) that impact their verbal-domain TA (Arens, Becker, & Möller, 2017).

Concerning the role of domain-specific math and German ASCs as mediators of the achievement-anxiety relations, in the FOF model, we were only able to replicate a considerably smaller proportion of relations as compared with Arens, Becker, and Möller (2017), who reported full mediations on both facets in both domains, and Schilling et al. (2005), who reported full mediations on both facets in German and partial mediations on both facets in math. In the present study, we observed only within-domain mediations on emotionality in both domains. With regard to worry, we did not find significant indirect paths at all, even though the large overlap between the cognitive facet of worry and ASC has been discussed (Arens, Becker, & Möller, 2017). However, our observed effect sizes were in the expected direction.

Nested Factor (NF) Models: An Extension

Second, we successfully applied the NF framework to the GI/E model, clearly showing GI/E-hypothesized relations in both direct and mediated NF models (RQ2). We demonstrated social comparison effects on all facets except for worry in German (analogous to social comparison effects found in the FOF model). However, in contrast to the FOF model, we found dimensional comparison effects on emotionality in both domains, suggesting that the emotionality component (as opposed to worry) is more susceptible to dimensional comparisons when general TA is controlled for. In all comparisons between the FOF and NF model results, it is crucial to keep in mind that the interpretation of the domain-specific factors varies according to the modeling strategy, with domain-specific factors in the NF model representing residual variance that is not explained by the general factor (Arens et al., 2021). When we examined the factor correlations, it became clear that the domain-specific worry factors were more strongly related to each other in the FOF model than the domain-specific emotionality factors were. Indeed, when we employed the NF model, the correlation between the domain-specific worry factors remained significantly different from zero (after the general worry factor was included), whereas the correlation between the domain-specific emotionality factors did not differ from zero (after the general emotionality factor was included). General factors in the NF model are defined by their unique items, that is, their reference domain (Eid et al., 2017). The significantly positive relation between the domain-specific worry factors thus indicated substantial common variance even after general school-specific worry was controlled for. The remaining relation between the math and German worry factors that is not tied to the school context, suggests lower school-specificity for general worry as opposed to general emotionality. On the one hand, this still significant correlation between the math and German worry factors might describe worry cognitions that are tied specifically to math and German (and that are independent to school in general) as two core school subjects (e.g., high emphasis placed on these subjects or higher amount of weekly lessons). On the other hand, it might also be possible that worry entails more general components not restricted to evaluative situations within the school context but rather school-unrelated other life domains (e.g., generalized worry). In line with this reasoning, Hock et al., submitted (submitted) showed that the stable, trait manifestation of worry was prevalent in situations that are generally perceived as threatening and aversive even when a state scale is administered that assesses worry with the instruction "*right now, at this very moment*" using latent state-trait analyses. Emotionality could be more tied to evaluative situations in the school context also due to its temporal proximity to the evaluative situation, whereas worry might also occur days or weeks prior to the situation during exam or test preparations (e.g., Sparfeldt et al., 2005), thus possibly mixing with other worries unrelated to the school context. Furthermore, worry that is associated with future-directed "What if"-type of questions (e.g., "What if I fail in this exam?") may collapse with "Why"-type of questions typical for rumination and directed to the past (e.g., "Why did I fail in past exams?", "Why am I such a failure?") in the course of processing an upcoming exam, highlighting the time-overarching character of worry compared to emotionality (Renner et al., 2018). To conclude, further research is needed to clarify the psychological meaning of this significant correlation.

With regard to the NF mediated model, considerable changes were evident compared with the FOF mediated model—particularly concerning indirect cross-domain paths (i.e., dimensional comparison effects). Three out of four indirect within-domain and three out of four indirect cross-domain paths reached statistical significance (in contrast to a total of two out of eight paths in the FOF model). Specifically, both TA facets in German were mediated by German ASC, and emotionality in math was mediated by math ASC, both within and across domains. To understand why dimensional comparisons in particular are affected by the modeling strategy, one has to keep in mind that dimensional comparisons are internal comparisons across domains (i.e., students comparing their own abilities across domains). By applying the NF model, domain-specific manifestations are purified, ultimately meaning that they truly reflect domain-specific manifestations (i.e., students perceiving different levels of TA across domains) and not a mixture of domain-specific and general components (i.e., students perceiving themselves to be generally more or less anxious than others across school domains). In other words, the removal of confounded (i.e., domain-specific and general) variance enabled the detection of intraindividual (dimensional) relations.

Third and finally, we found significant relations in the NF models with regard to general TA (RQ3). Both facets were negatively predicted by the math grade, whereas general worry was positively predicted by the German grade. One reason for this finding might be that the portion of worry previously attributed to math worry is more dominant in the general worry factor than the portion previously attributed to German worry. In other words, the salience of worry attributed to math for general worry might be higher than that of worry attributed to German. This difference would explain negative [positive] relations with math [German] grades. Indeed, factor loadings of items loading on their domain-specific worry factor were descriptively lower for math than for German after the general factor was included.

No indirect path was visible when general ASC was considered as a mediator (i.e., no paths between math and German grades and general worry and emotionality mediated by general ASC). The concept of domain-specific mediations of achievement-anxiety relations does not seem to translate to general relations. It might be the case that the (respectively negative and positive) effects of math and German grades on general worry through general ASC (which captures the variance shared between math and German ASCs, which are, in turn, positively related to math and German grades, respectively) was canceled out due to relations that went in opposite directions. Such an occurrence emphasizes the caution researchers need to exercise when interpreting the general factor and its relations. Yet, one advantage of the NF model (e.g., in contrast to the higher-order factor model) includes the invariance of the meaning of the general factor due to its ties to the reference domain (Eid et al., 2017). Also, neither the FOF nor the NF model considers item cross-loadings. Given these rather strict requirements, the good fit to the data is all the more convincing. The NF modeling strategy and its implications have been examined in contrast to other modeling strategies with regard to a number of psychological constructs outside of TA and ASC, for instance in the individual clinical assessment of depressive symptoms (Heinrich et al., 2020) and attention deficit hyperactivity disorder (Eid, 2020), highlighting its importance in more applied settings.

Limitations and Future Research

Our study has some limitations. First, we employed cross-sectional data and thereby cannot draw conclusions about causality. Yet, we chose to refer to social and dimensional comparison *effects* to remain in line with prior research on the GI/E model. Longitudinal and experimental studies that were designed to infer causality have supported GI/E-based assumptions for the ASC (see Niepel et al., 2014; Wolff et al., 2020). Further, our sample was limited to ninth- and 10th-grade students from the highest ability track in Germany. In order to improve the generalizability of our results, further research is needed across different age groups, school tracks, and countries. School-track specific differences in achievement-anxiety relations have previously been identified when comparing the highest ability track to other school tracks (Penk et al., 2014). Another study found that achievement-anxiety relations differed across grade levels with the strongest negative relations in the middle (sixth through eighth) grades and the lowest negative relations in the higher (ninth through 12th) grades (von der Embse et al., 2018), where our sample was located. A recent meta-analysis identified small to moderate, statistically significant negative achievement-anxiety relations in math across 747 effect sizes, also identifying grade level as one moderator of the strength of these relations (Barroso et al., 2021).

Finally, the present findings are restricted to the math and one verbal (i.e., German as language of instruction) domain, such that further research incorporating multiple other domains is warranted. If researchers are particularly interested in (cross-domain) dimensional comparison effects, the inclusion of other domains is recommendable. When assuming a continuum with math and verbal subjects as contrary endpoints, the perceived subject similarity is thought to moderate dimensional comparisons, such that they can even be found to work in the opposite direction (i.e., so-called assimilation effects as opposed to contrast effects; Möller & Marsh, 2013). Yet, a recent meta-analysis on the GI/E model with ASC as the outcome variable did not find such assimilation effects across 505 data sets (Möller et al., 2020). Employing an NF model to purify domain-specific relations considerably advances insights into dimensional comparisons. In addition, if multiple domains from the math-verbal continuum are included, the NF model enables closer examinations of contrast and assimilation effects and their occurrences when extracting the variance that is shared across domain-specific manifestations. For instance, nonsignificant relations between math grades and German TA might be revealed in the NF model as a result of positive relations between math grades and pure German TA in combination with negative relations between math grades and the portion of German TA that is positively correlated with math TA (i.e., general TA).

Implications and Conclusion

In their everyday school lives, students encounter various peers whose abilities in domain-specific domains serve as references (i.e., social comparisons) as well as various different domains in which a student's own abilities serve as a reference (i.e., dimensional comparisons) that shape students' socio-affective experiences and perceptions (e.g., TA, ASC). In the present study, we used NF models to consider (a) general worry and emotionality levels to identify social and dimensional comparison effects on purely domain-specific worry and emotionality and (b) general ASC levels to examine mediation effects of social and dimensional comparison effects through purely domain-specific ASCs.

Our approach facilitated the detection of dimensional comparison effects and differential worry and emotionality characteristics when controlling for general components in NF models—both directly and indirectly via domain-specific ASCs. Our findings thus have several implications. This study combines conceptual considerations (i.e., TA and ASC as domain-specific and hierarchical constructs on the one hand and the interest in domain-specific relations in the GI/E model on the other hand) with methodological considerations (i.e., the NF model as adequate representation of hierarchical constructs). In this study, we therefore argue for matching methodological approaches to conceptual ideas, and thus provide new directions for future research within the GI/E model. One example for the potential incremental value of using NF models in testing GI/E relations is the controversy on assimilation effects in dimensional comparisons (Möller et al., 2020) by suggesting modeling strategy as potential moderator of dimensional comparisons. In this article, we present both the result patterns yielded by the FOF and the NF models such that the effect of modeling strategy on content-related relations is highlighted. Further, the implementation of the NF model offers new insights on the proportions of general and domain-specific components within constructs (e.g., by comparing correlations among domain-specific components before and after the inclusion of a general, overarching factor).

Practically, examining the interplay of social and dimensional comparisons on the formation of the two TA facets worry and emotionality is of interest given TA's undesirability. Achievement feedback to students (e.g., in the form of school grades) is a common occurrence in daily school life. Thus, achievement-based comparison processes within- and across domains could be of interest in applied educational contexts where raising student, teacher or parent awareness for dimensional comparisons might buffer their detrimental impact (i.e., students performing lower in subject A than in another subject B might develop higher TA in subject A if they make dimensional comparisons). Wolff and Möller (2021) demonstrated that minimal interventions may lower such negative influences of dimensional comparison effects. One advantage of combining the NF modeling approach with the GI/E framework is the opportunity to evaluate achievement-TA associations more precisely with regard to general or domain-specific manifestations. Similarly, in an applied setting, TA interventions could be evaluated with regard to their effectiveness concerning general or domain-specific TA.

To conclude, the application of NF models to GI/E models matches conceptual and methodological considerations and offers novel insights on the impact of general TA and ASC levels on domain-specific social and dimensional comparisons.

Chapter 3

Uncovering Everyday Dynamics in Students' Perceptions of Instructional Quality with Experience Sampling

This contribution is published as:

Talić, I., Scherer, R., Marsh, H. W., Greiff, S., Möller, J., & Niepel, C. (2022). Uncovering everyday dynamics in students‘ perceptions of instructional quality with experience sampling. *Learning and Instruction, 81,* 101594. https://doi.org/10.1016/j.learninstruc.2022.101594

3. Uncovering Everyday Dynamics in Students' Perceptions of Instructional Quality with Experience Sampling

Abstract

Within-student dynamics in perceptions of instructional quality have been neglected, although student states constitute a major share of these perceptions. The present study examined the structure and correlates of student state perceptions of the three basic dimensions, teacher support, cognitive activation, and classroom management. We conducted a three-week experience sampling study using state measures in four subjects (observations: $n_{\text{mathematics}} = 2{,}681$, $n_{\text{physics}} = 1{,}555$, $n_{\text{German}} = 2{,}026$, $n_{\text{English}} = 1{,}835$) and analyzed data from 372 German secondary school students ($M_{\text{age}} = 15.3$ years), conducting two-level confirmatory factor analyses. Against more parsimonious solutions, the postulated three-factor structure was confirmed within- and between-students across subjects, entailing 51 % within-student variance on average. Similar to trait-like perceptions, state perceptions were positively related to grades and academic interest. Our results support the factorial and convergent validity of state student perceptions of instructional quality, expanding upon between-person-based literature and uncovering opportunities to enhance teaching effectiveness.

Keywords: Instructional quality; factorial validity; multilevel confirmatory factor analysis; intensive longitudinal data; experience sampling

Introduction

Instructional quality is a key determinant of student learning and motivation (Scherer & Nilsen, 2016). Student ratings are one of the main sources of information on instructional quality, and are incorporated in large-scale assessments and educational effectiveness research accordingly (e.g., OECD, 2014). To inform on teaching effectiveness, individual student perceptions of instructional quality (SPIQ) are typically aggregated to the class or school level where they reflect differences between classes or schools, respectively (Lüdtke et al., 2009). However, these *perceptions* may differ within the same classroom (Wagner et al., 2016) and may affect student learning and motivation above and beyond the teacher's *actual* instructional quality (Lazarides & Ittel, 2012). Therefore, examining these idiosyncratic perceptions by not only considering class or school differences (at the respective class or school level), but also interindividual differences between students (at the between-student level) and intraindividual differences within students between different lessons (at the within-student level) offers fruitful insights into ways to enhance teaching effectiveness which is one of the main assessment goals of SPIQ.

Concurrently, empirical support for the structure and validity of SPIQ is mainly limited to between-person analyses (Bellens et al., 2019; Praetorius et al., 2018; Scherer et al., 2016; Wisniewski et al., 2020). This also holds for one of its most popular frameworks—the three basic dimensions (TBDs) of instructional quality—that describes SPIQ in a parsimonious model of the three essential dimensions of teacher support, cognitive activation, and classroom management (Klieme et al., 2001). However, it cannot be assumed that differences in SPIQ between different students derived from between-person analyses correspond to differences in SPIQ *within* students across points in time (Murayama et al., 2017). For instance, perceived teacher support might be related to higher academic interest on average across students, but might be negatively related across points in time due to situational specifics such as tiredness or annoyance that vary within individual students. Such varying situational *states* are unsurprising given multiple time-varying elements of the everyday dynamic classroom life (Praetorius et al., 2014) and the conception of instruction as a *process* that is inherently evolving (Schmitz, 2006). Complementing existing between-person studies (i.e., examining variance between classes or students) with within-person studies (i.e., examining variance within persons across points in time) accounts for the fact that both student traits (e.g., general rating tendencies) and student states (e.g., situational enthusiasm) heavily influence SPIQ (Wagner et al., 2016). Ultimately, this promotes flexible and situation-oriented teaching, and allows students' and teachers' needs to be addressed in a tailored way. Therefore, we used intensive longitudinal methods to overcome the current lack of within-person studies on SPIQ within the TBDs framework and capture lesson-to-lesson variation in SPIQ in classrooms without a time delay (Bolger & Laurenceau, 2013). In the present study, German secondary school students reported their SPIQ in four core subjects (mathematics, physics, German, and English) over three consecutive weeks of everyday school life. Thus, we applied the TBDs framework in an experience sampling study for the first time, extending current knowledge to the level of specific lessons—the level at which instruction actually occurs.

Three Basic Dimensions of Instructional Quality

The three basic dimensions (TBDs) are the key components of a comprehensive and parsimonious framework of instructional quality. Klieme et al. (2001) extracted the dimensions of teacher support, cognitive activation, and classroom management using a factor-analytical approach based on data from the German 1995 TIMSS video study (Stigler et al., 1999) on 8th-grade mathematics instruction. Although based on math lessons, the dimensions are conceived to be generic and, therefore, applicable across school subjects (Praetorius et al., 2018). Furthermore, the dimensions are conceptually and empirically separable. Hence they represent distinct, yet related factors, where typically, teacher support and cognitive activation show higher correlations with each other than classroom management with either two (e.g., Bellens et al., 2019; Fauth et al., 2014; Kunter & Voss, 2013; Scherer et al., 2016). In addition to its empirical roots, the TBDs have repeatedly been shown to be related to the crucial educational outcomes of student achievement and motivation (where the latter is often approximated by academic interest; e.g., Baumert et al., 1997; Fauth et al., 2014; Scherer et al., 2016; for an overview see Praetorius et al., 2018). These relations are based on well-established psychological factors (i.e., self-determination, cognitive-constructive learning, and learning time), underpinning the framework's conceptual relevance and predictive validity. To date, the TBDs framework is one of the most popular frameworks on instructional quality, and is implemented in educational large-scale assessments, such as the Programme for International Student Assessment (PISA; OECD, 2014).

The first dimension, *teacher support,* comprises various types of support during the learning process, such as adopting a constructive approach to errors, adapting the pace of instruction, or avoiding performance pressure. According to self-determination theory (Ryan & Deci, 2000), fulfilling the basic needs for competence, autonomy, and social relatedness can enhance students' intrinsic learning motivation. This learning motivation is often operationalized by academic interest, which is, in turn, closely related to self-determination theory (Krapp, 2002). Accordingly, positive relations between teacher support and academic interest have been assumed and shown repeatedly (e.g., Fauth et al., 2014). The second dimension, *cognitive activation,* encompasses teaching behavior characterized by posing challenging tasks, provoking students' thinking, or supporting metacognition (Praetorius et al., 2018). Cognitively engaging students results in a constructive learning process related to achievement gains (Hardy et al., 2006). Thus, positive relations between cognitive activation and student achievement can be expected (e.g., Klieme et al., 2001). Yet, positive relations between cognitive activation and academic interest have also been reported (e.g., Fauth et al., 2014). The cognitive activation dimension has been highlighted as the least stable among the three, indicating its high content dependency (Praetorius et al., 2014). The third dimension, *classroom management*, comprises strategies to efficiently transform classroom time into learning time by maintaining clear rules, monitoring students, or effectively dealing with interruptions in class. Classroom management can enhance student achievement by, for instance, increasing the time spent on task as well as academic interest by strengthening student autonomy and competence experiences. Accordingly, positive relations between classroom management and achievement (e.g., Bellens et al., 2019; Scherer et al., 2016; Seidel & Shavelson, 2007; Wang et al., 1993), and to academic interest (e.g., Kunter et al., 2007) have been reported. Thus, the TBDs constitute a parsimonious model of instructional quality with three distinct dimensions that show theoretical and empirical relations to educational outcomes, demonstrating the TBDs' relevance.

Validity of Student Perceptions of Instructional Quality

Student ratings are widely used to assess instructional quality (Marsh, 2007; Praetorius et al., 2018), although validity concerns have been discussed (Gentry et al., 2002; Greenwald, 1997). Recently, Bellens et al. (2019) challenged the notion of instructional quality as merely a set of teacher characteristics (see also Kunter & Baumert, 2007). Instead, they considered instructional quality a student characteristic as well, and SPIQ as an individual's reality—separate from 'true' instructional quality. Accordingly, person-centered approaches have found meaningful interindividual differences in SPIQ—distinct SPIQ student clusters—that were related to between-student differences in relevant educational outcomes (Lazarides & Ittel, 2012). For educational effectiveness research (Scherer et al., 2016), this highlights the importance of a thorough construct validation of *student* perceptions (Kunter & Baumert, 2007), including the underlying factor structure and psychometric properties in addition to convergent validity evidence. Studies of SPIQ's factor structure consistently show the three-factor solution to fit the data best across different operationalizations, rendering substantial support for the TBDs framework's multidimensional conceptualization (e.g., Bellens et al., 2019; Fauth et al., 2014; Kunter & Voss, 2013; Scherer et al., 2016). Additionally, the dimensions are positively related to crucial educational outcomes (e.g., student achievement and academic interest), providing evidence for the framework's convergent validity. However, Praetorius et al. (2018) pointed out that these relations are not consistent across studies. They considered multiple potential explanations for these inconsistencies (e.g., examining different school subjects across studies). In fact, most studies of SPIQ within the TBDs framework are limited to mathematics and mathematics-related subjects. In contrast, SPIQ in verbal or other domains have hardly been examined (Praetorius et al., 2018). Despite these shortcomings, empirical support for TBDs' relation to student achievement and interest is found repeatedly, including large and international validation studies (Baumert et al., 1997). However, these findings are based on between-person designs, and thus tied to the between-person level of analysis.

Level of Analysis and Lesson-to-Lesson Variation in SPIQ

SPIQ inherently encompass variation from distinct, hierarchical levels. Interindividual differences in SPIQ between students (e.g., students perceiving more or less learning support in general) are at the between-student level. We refer to interindividual differences in SPIQ between students in individual situations (e.g., students perceiving more or less learning support in the same lesson) that entail situation-specific shared variance (i.e., the same lesson and its instructional quality) as the between-lesson level. Interindividual differences in instructional quality and other teacher characteristics between teachers (e.g., teachers providing more or less learning support in general) are located at the class level (Marsh et al., 2012). The appropriate level of analysis of SPIQ is tied to the specific research question (Lüdtke et al., 2009). For instance, if students are interchangeable informants about a higher-level construct (i.e., instructional quality of their teacher), between-student variations in SPIQ reflect deviations from shared classroom perceptions and are not of central interest, but ultimately considered a source of unreliability. The construct is considered to be meaningful only at the class or school level (i.e., a shared construct according to Stapleton et al., 2016). Despite the finding that large proportions of SPIQ are idiosyncratic, there is a lack of studies arguing for analyzing intraindividual differences between different points in time within students (e.g., one student perceiving more or less learning support across different lessons) at the within-student level. In everyday school life, the classroom is a highly dynamic interactional system undergoing considerable changes from lesson to lesson (Curby et al., 2011; Praetorius et al., 2014). Changes in objective teaching—due to varying external factors (e.g., time of day) and teacher states (e.g., situational anger) beside teacher traits (e.g., enthusiasm)—are likely mirrored by changes in students' subjective perceptions of teaching, which are in turn affected by both student traits (e.g., openness) and student states (e.g., situational anxiety). Wagner et al. (2016) found that an average of 53.4 % of the total variance in SPIQ within classes (assessed at three measurement points over a period of three months) were *time-specific* (as opposed to time-consistent) ratings. Two experience sampling studies (Goetz et al., 2013; Goetz et al., 2020) highlighted the theoretical advancement immanent in such state SPIQ. While the amount of within-student variance in two dimensions of SPIQ over ten school days was substantial (up to 92.6 %), the authors reported significant within-student effects of SPIQ on situational academic emotions (e.g., pride or boredom). Such results demonstrate the advancements in construct knowledge that can be achieved using experience sampling. Conceptually, such findings are crucial as they expand knowledge to the within-person level. More specifically, the issue with distinct levels of analysis lies in the fact that findings based on analyses at a certain level do not necessarily hold for other levels as well (Molenaar, 2004; Murayama et al., 2017). Implicit generalizations across levels are thus not valid. As Molenaar (2004) pointed out, between-person analyses use differences between persons to describe relations between variables (e.g., “Students who perceive more teacher support tend to exhibit higher interest than students who perceive less teacher support”). In contrast, within-person analyses focus on processes within persons using points in time as a basis (e.g., “Students show higher interest at times when they perceive more teacher support”). These two types of analyses are based on two distinct types of variation (i.e., inter- and intraindividual variation). In other words, there is no strong basis for assuming that properties that describe differences between students also describe differences within students across numerous time points. Thus, assuming the invariance of within- and between-person variation implicitly assumes invariance over time—which seems particularly doubtful in the context of learning and instruction, which are considered *processes* that may change substantially over time (Schmitz, 2006).

The Present Study

To uncover everyday school life dynamics of SPIQ within the TBDs framework and to foreground the *student* in SPIQ, we conducted an intensive longitudinal study in German secondary schools based on experience sampling via e-diaries. We employed state measures to assess students’ situational perceptions of the TBDS at the end of every classroom lesson in four core subjects over a period of three weeks (i.e., 15 school days). Considering the four subjects mathematics, physics, German, and English expands the literature on the TBDs by widening the range of simultaneously examined domains, taking into account the possible domain-specificity of the three dimensions.

Students’ real-time perceptions of teaching have been examined in two experience sampling studies we know of (Goetz et al., 2013; Goetz et al., 2020). In contrast to these prior studies, we aim to systematically and holistically validate an established framework of SPIQ for idiosyncratic variation within- and between students. To examine idiosyncratic student-specific SPIQ, we controlled for teacher effects at the class level by (a) including dummy-coded classroom variables at the between-student level as well as (b) shared situation-specific SPIQ by including a between-lesson level in a second set of models (see Section Statistical Analyses below).

In doing so, we substantially expand the existing (between-person-based) literature on the TBDs to the within-person level. Employing the experience sampling method enabled us to disentangle the situational (i.e., state) and stable (i.e., trait) components of SPIQ and model its within- and between-student factor structures simultaneously. Assessing habitual (i.e., trait) SPIQ enabled us to contrast the within- and between-person approaches and evaluate the distinction between stable components of state SPIQ (i.e., aggregated state SPIQ) and pure trait SPIQ. Furthermore, examining the relations between aggregated state SPIQ and educational outcomes (i.e., school grades and academic interest) allowed us to extend the hypothesized relations based on the TBDs framework (i.e., positive relations between teacher support and interest, between cognitive activation and grades, and between classroom management and both grades and interest) to the within-person level. We applied multilevel modeling, as recommended in studies on SPIQ (Lüdtke et al., 2009; Marsh et al., 2012; Scherer & Gustafsson, 2015) and SPIQ's construct validity (Bellens et al., 2019; Fauth et al., 2014; Wagner et al., 2013; Wisniewski et al., 2020) to address the following research questions:

RQ1. To what extent does the three-factor structure (as hypothesized by the TBDS framework) fit the data both within and between students compared to more parsimonious structures?

RQ2. Do the implemented two-item-based state measures reliably assess the three dimensions of teacher support, cognitive activation, and classroom management in an experience sampling design?

RQ3. How is aggregated state SPIQ within the TBDs related to relevant correlates (i.e., trait SPIQ, reported school grades, and academic interest)?

Method

Procedure and Participants

We conducted a three-week experience sampling study using an e-diary application on smartphones (in study weeks 2 to 4), which was embedded in a pre- and post-assessment (in study weeks 1 and 5, respectively) that obtained student background and trait perceptions in traditional paper-and-pencil format. We collected data within the scope of the larger intensive longitudinal "Dynamics of Academic Self-Concept in Everyday Life" (DynASCEL) project (Niepel et al., 2022) focusing on the dynamics of students' academic experiences in everyday school life.[1] We drew on a sample of N = 372 secondary school students participating in the e-diary (34.1 % boys out of the n = 301 students with available gender information) with a mean self-reported age of 15.3 years (SD = 0.68; range = 13.3-17.4 years; based on n = 298 students with available age information). Students were nested in 18 classrooms with an average of 20.6 (SD = 4.65; range = 13 – 27) students per classroom. They attended the 9th (n = 308) and 10th (n = 64) grades in six highest ability track schools (i.e., German *Gymnasium*) in four German states (Rhineland-Palatinate, North Rhine-Westphalia, Baden-Wuerttemberg, and Mecklenburg-Western Pomerania). Students remained in their class constellations for several years, where generally, different teachers held instruction in the four core subjects for the 18 classes. There were only few combinations of teachers teaching one subject to more than one class (i.e., 1 in math and English, respectively, 3 in physics, and none in German). In total, we examined students' perceptions of 60 teachers' instructional quality (48.3 % male teachers) across the four subjects.

The experience sampling design was event-contingent, depending on the occurrence of the four subjects in the class-specific timetables. Every student was given a smartphone that was pre-programmed to prompt the e-diary assessment three minutes before the scheduled end of each lesson in the four subjects via the application movisensXS (versions 1.3.0-1.3.4; movisens GmbH, Karlsruhe, Germany).

Missing values are a common phenomenon in intensive longitudinal designs. In our case, students were explicitly instructed not to answer the e-diary prompts if they did not attend the lesson (e.g., due to illness) or the lesson did not take place (e.g., due to class trips or teacher illness). Other reasons for missing values include exams and technical issues (e.g., empty batteries). While the absolute number of measurement points

[1] Data from the larger research project have been and will be used in other manuscripts addressing different research questions (Niepel et al., 2022; Dörendahl et al., 2021). The intensive longitudinal data examined in this study have not previously been reported in other manuscripts.

varied between subjects due to differences in the classes' timetables, the relative proportion of missing values did not vary substantially between subjects. Specifically, for the three-week experience sampling period, average numbers of 10.11 (*SD* = 3.39) mathematics, 6.00 (*SD* = 3.01) physics, 7.94 (*SD* = 2.39) German, and 6.44 (*SD* = 2.55) English prompts were pre-programmed per student. Out of these pre-programmed prompts, an average of 7.21 (*SD* = 3.15) mathematics, 4.39 (*SD* = 2.46) physics, 5.49 (*SD* = 2.55) German, and 5.00 (*SD* = 2.15) English prompts were answered per student (i.e., at least one item completed; see Measures Section below). Across all students and lessons, we found that 70.81 % (mathematics), 69.11 % (physics), 69.29 % (German), and 74.38 % (English) prompts of the pre-programmed prompts were answered, equal to 2,681 mathematics, 1,555 physics, 2,026 German, and 1,835 English e-diary prompts in total.[2] Participation was voluntary, and written parental consent was obtained for all participating students. All procedures were approved by the local ethics review panel of the University of Luxembourg and all involved education authorities.

Measures

State measure: State SPIQ

We assessed state SPIQ within the TBDs, teacher support, cognitive activation, and classroom management using state measures only referring to the specific lesson. We adapted the existing German-language TBDs scales used in PISA 2012 (Mang et al., 2018) to meet the specific requirements of intensive longitudinal designs as follows: First, we reduced the response burden on students to enhance compliance (Stone & Shiffman, 2002). For each dimension, we selected two items out of the existing pools of five (for teacher support and classroom management each) and nine items (for cognitive activation) based on the highest item-total correlations in combination with the broadest content validity and practical applicability in an e-diary. Second, we added

[2] Missing values were calculated separately for each of the three dimensions in addition to rule out the possibility of meaningful differences in missing values depending on the dimension. No such differences were detected across all subjects.

"During this lesson" to each item stem and removed trait-like wordings such as "usually" to obtain lesson-specific (as opposed to general) perceptions. Example items include "*During this lesson, the teacher helped students with their learning*" (for teacher support), "*During this lesson, the teacher gave problems that required us to think for an extended time*" (for cognitive activation), and "*During this lesson, there was noise and disorder*" (for classroom management; negative indicator). Third, students responded on a six-point Likert scale ranging from 0 (*false*) to 5 (*true*) with higher values representing higher perceived instructional quality. As opposed to the original scale, where ratings of frequencies of the observed teaching behavior were requested (i.e., ranging from *never [in no lesson]* to *always [in every lesson]*), we changed the target of assessment (i.e., specific situations to be agreed or not agreed with versus an aggregated frequency estimation across multiple situations) as a necessary step to achieve the desired state-trait distinction. All items can be found in German and their English translations in Table S1 in the Online Supplementary Material.

Trait measures

Trait SPIQ. In the pre- and post-assessments, we assessed trait SPIQ for the TBDs, teacher support, cognitive activation, and classroom management, for the four subjects. We used the full five- (for teacher support and classroom management each) and nine-item scales (for cognitive activation) as implemented in PISA 2012 (Mang et al., 2018), where items were introduced with the question "How often do these things happen in your [subject] lessons?" Example items include "*The teacher helps students with their learning*" (for teacher support), "*The teacher gives problems that require us to think for an extended time*" (for cognitive activation), and "*There is noise and disorder*" (for classroom management; negative indicator). Students responded on a six-point Likert scale ranging from 0 [*never (in no lesson)*] to 5 [*always (in every lesson)*]. All items were scored such that higher values represented higher perceived instructional quality. Mang et al. (2018) reported internal consistencies of α = .84 (teacher support), α = .79 (cognitive activation), and α = .89 (classroom management) for the PISA German student sample.

Report Card Grades. In the pre-assessment, students reported their mathematics, physics, German, and English grades from their last report card. Self-reported grades have been shown to be reliable achievement indicators in German student samples as indicated by high relations between self-reported and actual grades ($r \geq .90$; Sparfeldt et al., 2008). Additionally, overestimations were not systematically correlated with achievement or other educational variables (Dickhäuser & Plenter, 2005). School grades in Germany are based on a six-point scale, which was recoded such that higher values represented better grades, ranging from 1 (*insufficient*) to 6 (*very good*).

Academic Interest. In the pre- and post-assessment, we assessed academic interest in the four subjects using six parallel-worded items adapted from Pohlmann et al. (2005). The conceptualization of interest was based on the person-object approach to interest (Krapp, 2002), which is closely linked to self-determination theory (Ryan & Deci, 2000). An example item is "*I look forward to [subject] lessons*". Students responded on a six-point Likert scale ranging from 0 (*false*) to 5 (*true*), with higher values representing higher interest. Pohlmann et al. (2005) reported an internal consistency of $\alpha = .94$ for their interest scale in a German secondary school student sample.

Statistical Analyses

To address our research questions, we conducted a series of multilevel confirmatory factor analyses (MCFA) within the multilevel structural equation modeling (MSEM) framework using the software package Mplus 8.3 (Muthén & Muthén, 1998-2017). Using MCFA enabled us to consider different sources of variance at the same time. Measurement points were nested within students. Thus, to take account of this hierarchical data structure, we specified subject-specific two-level models examining variation between measurement points within students (i.e., Level 1) and variation between students (i.e., Level 2). In MCFA, the total covariance matrix is decomposed into within- and between-level matrices, which form the basis for the within- and between-level structural equation models (Muthén, 1994). Within-student variance represented deviations of scores at specific time points from the student mean, while between-student variance represented differences between student means across students. This approach allowed us to estimate potentially different factor structures across levels (Marsh et al., 2012; Stapleton et al., 2016) to ensure that the hypothesized three-factor structure fit best within and between students against competing, more parsimonious factor solutions (RQ1). We conducted separate analyses for each of the four school subjects to align with prior domain-specific research (see Praetorius et al., 2018). To control for Level 3 (i.e., class level) effects, we included a set of 17 dummy variables (based on 18 classrooms) at Level 2 (Hox et al., 2018). Although three-level modeling would clearly be appropriate in evaluating SPIQ (Marsh et al., 2012), in our case, only the option implementing dummy-coded classroom variables at Level 2 was possible given the small number of classrooms in the sample (Huang, 2016).[3] Cases of teachers teaching the same subject to more than one class could not be considered in the analyses as there were only few such combinations (i.e., a maximum of three; see Section Procedure and Participants). We estimated the two-level models with the MLR estimator to obtain robust standard errors of model parameters and an adjusted chi-square statistic. To estimate the level-inherent variance and extent of data dependency due to the hierarchical structure, we calculated item-based intraclass correlation coefficients (ICC[1]; Kim et al., 2016). After detecting substantial Level 1 variation, we identified the best-fitting level-specific factor structure. Since Mplus mostly provides indices of *global* model fit, which are mainly determined by the fit of the Level 1 model due to its larger sample size, we used a partially saturated model approach in evaluating the factor structures at each level (e.g., Hox et al., 2018; Janis et al., 2016; Ryu & West, 2009). Specifically, we tested both three-factor (as hypothesized by the TBDs framework) and more parsimonious factor solutions at Level 1 and Level 2. We specified a saturated model estimating only item (co)variances at one level (i.e., a model with zero degrees of freedom and perfect model fit), while modeling one-factor and three-factor solutions at the respective other level. The two-level models were evaluated according to recommended cut-off criteria in the absolute model fit indices CFI, RMSEA, and SRMR (i.e., CFI ≥ .95, RMSEA ≤ .06, and SRMR ≤ .08; Hu & Bentler, 1999) as well as the change in the relative model fit indices AIC and BIC, where lower values are preferred (e.g., Kline, 2016). However, we note that these criteria may not fully apply to MSEM and must therefore not be considered “golden rules” (Greiff &

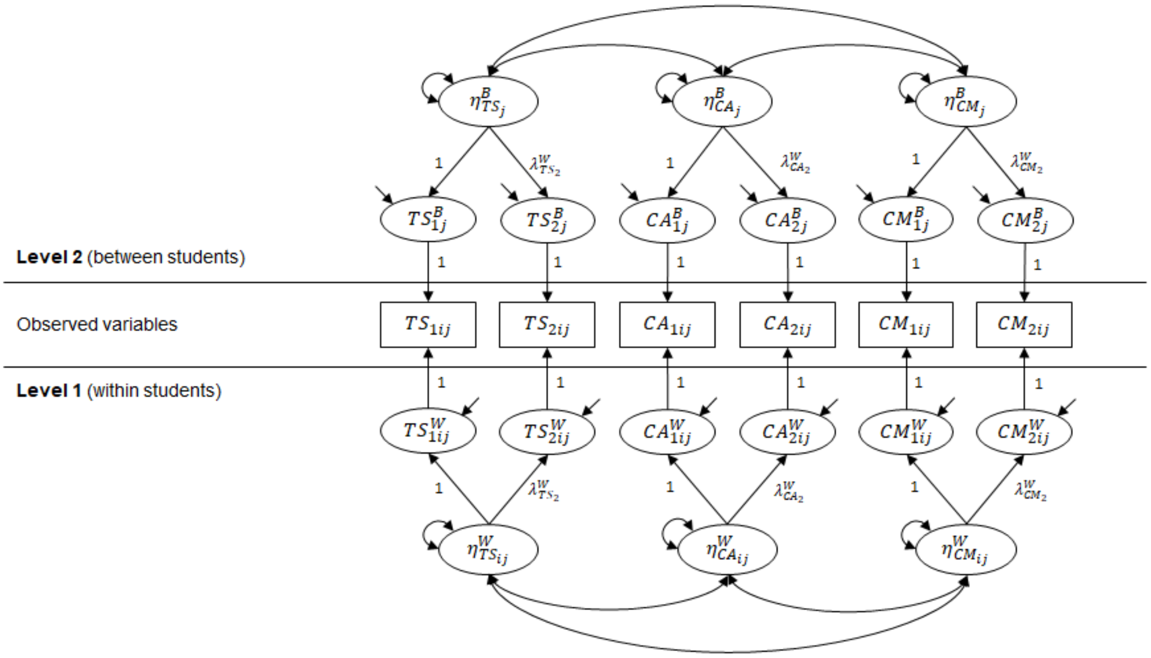
[3] To check the robustness of the results we obtained using the dummy-coding approach, we additionally ran all analyses using the TYPE = COMPLEX option in Mplus, which adjusts standard errors for the nested data structure. The results were virtually the same.

Heene, 2017). As expected with structural equation modeling using only two indicators per factor (in the three-factor solutions), occasional negative error variances occurred, which we fixed to small non-negative values (Kline, 2016).

Having identified the best-fitting level-specific factor structures, we noted that these comprised an identical number of factors across levels, and thus required cross-level measurement invariance for meaningful construct interpretation at both levels (Stapleton et al., 2016). We specified comprehensive two-level models with a structure of three correlated factors represented by six manifest indicators at Level 1, and the item intercepts at Level 2. Indicators were restricted to load on their proposed factor only, and correlations between factors across levels were not allowed (Dyer et al., 2005). See Figure 1 for an illustration of the two-level models. To test for cross-level measurement invariance, we first specified unconstrained two-level models in which factor loadings were estimated freely, and factor variances were fixed to 1. Second, we added cross-level invariance constraints by setting the indicators' factor loadings to equality across both levels while freely estimating factor variances at Level 2 (Jak & Jorgensen, 2017). We compared the unconstrained and constrained models using Satorra-Bentler scaled χ^2-difference tests in addition to common criteria of model fit deterioration (Hox et al., 2018).

Figure 1

Hypothesized Two-level Factor Model for State SPIQ within the TBDs



Note. TS1 and TS2 = Items measuring perceived teacher support; CA1 and CA2 = Items measuring perceived cognitive activation; CM1 and CM2 = Items measuring perceived classroom management (see Table S1 for item wordings). Subscripted indices *i* and *j* denote the measurement points and students, respectively. Superscripted letters W and B indicate the within- and between-student level, respectively. Dummy-coded classroom variables served as predictors at the between-students level but are not explicitly represented for better clarity of presentation. The mean structure is not shown in the model. This model assumes cross-level metric invariance between the levels 1 and 2.

Having established cross-level measurement invariance, we estimated level-specific reliability indices with Level 1 and Level 2 omega coefficients (Geldhof et al., 2014) to determine whether the two-item based state measure can reliably assess situational perceptions of the TBDs (RQ2). As McDonald's ω is calculated based on factor loadings, we removed the cross-level invariance constraints and estimated factor loadings freely, while fixing factor variances to 1 at both levels. Additionally, we examined the reliability of the student means—that is, the extent of agreement between perceptions at different measurement points within students—by estimating ICC[2] indices (Bliese, 2000; see also Lüdtke et al., 2009). The index k_s indicating the number of measurement points within students varied across students and subjects due to our sampling design. Therefore, we used the average number of measurement points within students, which was $k_s = 7.21$ (mathematics), $k_s = 4.39$ (physics), $k_s = 5.49$ (German), and $k_s = 5.00$ (English).

To provide evidence for convergent validity, we examined the relations between aggregated state SPIQ within the TBDs and (a) trait SPIQ within the TBDs at both the pre- and post-assessment, (b) students' reported grades at the pre-assessment, and (c) academic interest at both the pre- and post-assessment in the four subjects (RQ3). To this end, we estimated correlations within the MSEM framework. School grades entered the model as latent single-item factors by fixing the factor loading to 1 and fixing the residual variance to a value based on sample variance and a reliability estimate found in previous studies (Kline, 2016).

Complementary to our MCFA analyses, we conducted additional, cross-classified analyses in a second step. Arguably, SPIQ gathered within shared situations (i.e., all students within classes rating the instructional quality in a given lesson) comprise shared situation-specific variance. Thus, to control for this shared variance (i.e., consensual student perceptions within lessons), we specified two-level cross-classified models examining variation between measurement points (i.e., Level 1) and variation between students (i.e., Level 2a) while additionally estimating variation between shared lessons (i.e., Level 2b). By extracting the overlapping shared lesson variance in these cross-classified models, we obtained estimates for idiosyncratic SPIQ that were controlled for shared lesson variance. We used these to test the robustness of our results obtained by the two-level models. In doing so, our results were confirmed. Due to the widespread and well-established use of MCFA in multilevel factorial validation in general (e.g. Kim et al., 2016) and with regard to the TBDs in particular (e.g., Fauth et al., 2014), we focally report the MCFA results and refer to the Online Supplementary Material for a detailed illustration of the cross-classified model results.

In the cross-classified analyses, we addressed the same three RQs as we did with the MCFA. Specifically, we repeated the entire procedure described above with the following adjustments. We estimated the within-student level (i.e., Level 1) and between-student level (i.e., Level 2a) explicitly, thus testing different factor structures as outlined above. In addition, to consider shared lesson variance, we specified the between-lesson level (i.e., Level 2b) where item (co)variances were estimated. For the cross-classified models, we used the option TYPE = CROSSCLASSIFIED with the Bayes estimator in Mplus. We evaluated models according to the posterior predictive *p* value ($PPP_{\chi 2}$), where values above .05 indicate a good fittig model. We considered the deviance information criterion (DIC) for model comparison, where lower values are preferred (e.g., Kaplan & Depaoli, 2012). To estimate the reliabilities of SPIQ within lessons, we calculated ICC[2] values based on the respective ICC[1] values and the index k_l indicating the average number of student perceptions in a given lesson. These were $k_l = 15.23$ (mathematics), $k_l = 14.81$ (physics), $k_l = 15.46$ (German), and $k_l = 16.68$ (English). All other actions described in the first set of models (i.e., two-level models) were performed analogously in the second set of models (i.e., cross-classified models). See an illustration of the cross-classified models in Figure S10 in the Online Supplementary Material.

Results

Preliminary Analyses

Before addressing our research questions, we examined descriptive and psychometric properties of the trait measures, which were assessed at the pre-assessment only (grades) or the pre- and post-assessment (trait SPIQ and interest), respectively. ,

Single-level McDonald's ω coefficients for trait SPIQ ranged from $\omega = .85$ to $\omega = .94$ for teacher support, $\omega = .65$ to $\omega = .90$ for cognitive activation, and $\omega = .76$ to $\omega = .92$ for classroom management across subjects and the pre- and post-assessment. To examine the class mean reliability of trait SPIQ, we additionally calculated ICC[2] values based on ICC[1] values of trait SPIQ items and the average number of students within classrooms. ICC[2] values ranged from ICC[2] = .81 to ICC[2] = .91 (for teacher support), from ICC[2] = .51 to ICC[2] = .87 (for cognitive activation), and from ICC[2] = .64 to ICC[2] = .94 (for classroom management) across subjects. Test-retest stabilities of trait SPIQ between the pre- and post-assessment ranged between ρ = .62 and ρ = .70 for teacher support, ρ = .55 and ρ = .74 for cognitive activation, and ρ = .47 and ρ = .67 for classroom management across subjects. Coefficients were similar across subjects. Thus, trait SPIQ were measured reliably at each point in time (i.e., the pre- or post-assessment) with sufficient reliability at the class level, yet the correlations between these two assessments over an interval of four weeks were comparably lower, indicating some variability in trait SPIQ. Single-level McDonald's ω for academic interest ranged from ω = .85 to ω = .92 across subjects and the pre- and post-assessment, whereas test-retest stabilities were somewhat lower, ranging between ρ = .76 to ρ = .84.

Factor Structures of SPIQ Within and Between Students (RQ 1 and RQ 2)

Means, Level 1 and Level 2 variances, and ICC[1] values for all state SPIQ items across the four subjects are displayed in Table 1. Notably, the proportions of Level 2 (between-student) and total variance were $M_{\text{ICC[1]}}$ = .49 on average across items and subjects, leaving about 51 % of Level 1 (+ error) variation unexplained in the two-level models.[4]

[4] To estimate variance between teachers (i.e., Level 3), we additionally calculated ICC[1] values for Level 3. These ranged from ICC[1] = .04 to ICC[1] = .20. Level 3 variance reduced the amount of Level 2 but not Level 1 variance, still leaving an average of 51 % of Level 1 variance (i.e., 1- ($M_{\text{ICC[1]Level 2}}$ + $M_{\text{ICC[1]Level 3}}$)) unexplained.

Table 1

Means, Variance Components, ICC[1], and ICC[2] Within Students (Level 1) and Between Students (Level 2) for State SPIQ Items in Four Subjects

Item	*M*	s^2_{within} *(SE)*	s^2_{between} *(SE)*	ICC[1]	ICC[2]	*M*	s^2_{within} *(SE)*	s^2_{between} *(SE)*	ICC[1]	ICC[2]	*M*	s^2_{within} *(SE)*	s^2_{between} *(SE)*	ICC[1]	ICC[2]	*M*	s^2_{within} *(SE)*	s^2_{between} *(SE)*	ICC[1]	ICC[2]
	Mathematics					**Physics**					**German**					**English**				
TS1	3.27	1.11 (.06)	0.89 (.08)	.45	.86	3.14	0.80 (.06)	1.09 (.10)	.57	.85	3.24	0.96 (.06)	1.03 (.10)	.52	.86	3.14	0.92 (.06)	0.98 (.09)	.51	.84
TS2	3.00	1.10 (.06)	1.00 (.08)	.47	.86	2.85	0.87 (.07)	1.13 (.10)	.57	.85	2.99	0.95 (.06)	1.12 (.10)	.54	.87	2.87	0.91 (.06)	1.06 (.09)	.54	.85
CA1	3.06	1.11 (.05)	0.72 (.07)	.39	.82	2.86	1.07 (.07)	0.85 (.09)	.44	.78	2.80	1.25 (.07)	0.85 (.08)	.39	.79	2.73	1.18 (.07)	0.78 (.08)	.40	.77
CA2	3.03	1.08 (.06)	0.79 (.07)	.42	.84	2.94	0.93 (.07)	0.87 (.09)	.48	.80	2.85	1.15 (.07)	0.91 (.09)	.44	.81	2.75	1.08 (.07)	0.86 (.08)	.44	.80
CM1	3.35	1.01 (.06)	1.18 (.09)	.54	.89	3.38	1.06 (.07)	0.97 (.09)	.47	.79	3.24	1.04 (.06)	1.39 (.10)	.57	.88	3.63	1.04 (.07)	0.83 (.09)	.45	.80
CM2	3.45	0.87 (.05)	1.18 (.09)	.57	.91	3.49	0.95 (.07)	0.96 (.09)	.50	.81	3.38	0.91 (.06)	1.33 (.11)	.60	.89	3.74	0.86 (.06)	0.80 (.08)	.48	.82

Note. TS1 and TS2 = Items measuring perceived teacher support; CA1 and CA2 = Items measuring perceived cognitive activation; CM1 and CM2 = Items measuring perceived classroom management (see Table S1 for item wordings). ICC[1] = The proportion of between-student to total variance. ICC[2] = Reliability of the student means as derived from the ICC[1] and the number of measurement points within students.

Means are estimated at the between-student level.

All partially saturated models (i.e., two-level models specifying a one- or three-factor solution at one level, and a saturated model at the other level) showed unacceptable fits in the one-factor solutions across levels and subjects (see Table 2). On the contrary, the three-factor solutions exhibited very good fit to the data, ranging from CFI = .983 to CFI = 1.000, RMSEA = .000 to RMSEA = .033, $\text{SRMR}_{\text{within}}$ = .001 to $\text{SRMR}_{\text{within}}$ = .020, and $\text{SRMR}_{\text{between}}$ = .001 to $\text{SRMR}_{\text{between}}$ = .021. The relative indices AIC and BIC confirmed this finding, displaying lower values in the three-factor solutions than in the one-factor solutions. Combining a one-factor solution at one level with a three-factor solution at the respective other level (Models e and f in Table 2) further supported the misfit of one-factor conceptualizations at both levels. Therefore we discarded the one-factor solutions in all instances and accepted the MCFA models with three correlated factors at each level for further analyses.

To ensure a meaningful construct interpretation at both levels, we applied cross-level measurement invariance constraints to the two-level, three-factor models. Across all subjects, the constrained models fitted the data equally well compared to the unconstrained models (see Table 3). None of the differences in χ^2 statistics were statistically significant, and changes in model fit indices were only minor, ranging from ΔCFI = -.001 to ΔCFI = .001, ΔRMSEA = -.001 to ΔRMSEA = .000, $\Delta\text{SRMR}_{\text{within}}$ = .000 to $\Delta\text{SRMR}_{\text{within}}$ = .002, and $\Delta\text{SRMR}_{\text{between}}$ = .000 to $\Delta\text{SRMR}_{\text{between}}$ = .003. In other words, cross-level measurement invariance held across subjects, indicating that the three dimensions had the same meaning—as reflected by the item-factor relationship—within and between students.

Table 2

Testing Alternative Factor Structures at the Within- and Between-Students Levels for each Subject via the Partially Saturated Modeling Approach

Model	Within	Between	MLR χ2 (*df*)	CFI	RMSEA	SRMR$_w$	SRMR$_b$	AIC	BIC
Mathematics									
1a	saturated	3 factors	108.564 (59)	.991	.018	.001	.019	43436.695	43973.044
1b	saturated	1 factor	947.590 (94)	.843	.058	.024	.120	44250.528	44580.589
1c	3 factors	saturated	6.286 (7)	1.000	.000	.007	.001	43624.108	43865.760
1d	1 factor	saturated	1156.596 (9)	.621	.218	.155	.023	45528.434	45758.298
1e	1 factor	3 factors	1813.726 (68)	.679	.098	.155	.020	45336.060	45819.363
1f	3 factors	1 factor	924.278 (101)	.849	.055	.024	.120	44246.587	44535.391
Physics									
2a	saturated	3 factors	120.882 (58)	.983	.026	.003	.021	24920.470	25412.600
2b	saturated	1 factor	766.509 (94)	.817	.068	.056	.131	25488.297	25787.854
2c	3 factors	saturated	1.502 (7)	1.000	.000	.008	.002	25062.342	25281.660
2d	1 factor	saturated	1249.244 (9)	.443	.298	.155	.028	26109.460	26318.080
2e	1 factor	3 factors	1204.252 (67)	.691	.104	.155	.024	25956.116	26400.102
2f	3 factors	1 factor	741.879 (101)	.826	.064	.056	.131	25475.251	25737.363
German									
3a	saturated	3 factors	122.545 (57)	.986	.024	.002	.019	32969.370	33491.456
3b	saturated	1 factor	960.655 (94)	.821	.067	.044	.116	33721.553	34035.926
3c	3 factors	saturated	15.106 (6)	.997	.027	.016	.003	33200.104	33435.884
3d	1 factor	saturated	1242.710 (9)	.571	.260	.157	.020	34646.688	34865.627
3e	1 factor	3 factors	1631.718 (68)	.677	.107	.157	.022	34420.690	34881.023
3f	3 factors	1 factor	945.371 (101)	.825	.064	.048	.117	33723.078	33998.155
English									
4a	saturated	3 factors	85.493 (58)	.993	.016	.002	.019	29465.353	29972.714
4b	saturated	1 factor	611.300 (94)	.874	.055	.049	.116	29937.977	30246.806
4c	3 factors	saturated	21.316 (7)	.995	.033	.020	.004	29614.720	29840.827
4d	1 factor	saturated	1209.840 (9)	.546	.270	.152	.029	30798.816	31013.893
4e	1 factor	3 factors	1172.004 (68)	.730	.094	.152	.022	30665.094	31117.307
4f	3 factors	1 factor	630.073 (101)	.871	.053	.052	.116	29959.546	30229.771

Note. MLR = Maximum likelihood estimation with robust standard errors; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR$_w$ = Standardized Root-Mean-Square Residual value for within; SRMR$_b$ = Standardized Root-Mean-Square Residual value for between; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion.

Models specified at the between-student level (Models a, b, e, and f) consider classroom effects by including dummy-coded classroom variables. Cases of negative item error variances were fixed to small non-negative values (i.e., 0.0001).

Table 3

Testing Cross-Level Measurement Invariance in Factor Models with 3 Factors at the Within- and Between-Students Levels

Model	Specification	MLR χ2 (*df*)	CFI	RMSEA	SRMR$_w$	SRMR$_b$	ΔCFI	ΔRMSEA	ΔSRMR$_w$	ΔSRMR$_b$	TRd (Δ*df*)	*p*	Decision
Mathematics													
1g	3 / 3 factor model	111.483 (66)	.992	.016	.007	.019	-	-	-	-	-	-	-
1h	3 / 3 cross-level invariant factor model	112.849 (68)	.992	.016	.009	.019	.000	.000	.002	.000	1.566 (2)	.457	Accept
Physics													
2g	3 / 3 factor model	116.751 (65)	.986	.023	.008	.021	-	-	-	-	-	-	-
2h	3 / 3 cross-level invariant factor model	119.811 (67)	.986	.023	.009	.022	.000	.000	.001	.001	3.087 (2)	.214	Accept
German													
3g	3 / 3 factor model	138.412 (64)	.985	.024	.016	.019	-	-	-	-	-	-	-
3h	3 / 3 cross-level invariant factor model	134.161 (66)	.986	.023	.016	.020	.001	-.001	.000	.001	-7.651 (2)	1.000	Accept
English													
4g	3 / 3 factor model	113.195 (65)	.988	.020	.019	.019	-	-	-	-	-	-	-
4h	3 / 3 cross-level invariant factor model	118.328 (67)	.987	.020	.022	.019	-.001	.000	.000	.003	4.861 (2)	.088	Accept

Note. MLR = Maximum likelihood estimation with robust standard errors; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR$_w$ = Standardized Root-Mean-Square Residual value for within; SRMR$_b$ = Standardized Root-Mean-Square Residual value for between; TRd = Satorra-Bentler-scaled χ^2-difference test.

Cases of negative item error variances were fixed to small non-negative values (i.e., 0.0001)

The level-specific reliability coefficients ω for the two-item scales were estimated based on the models without cross-level invariance constraints (Table 4). They ranged between ω_{within} = .69 and ω_{within} = .86 across dimensions and subjects at the within-student level. In other words, 69 % to 86 % of the total variance at Level 1 could be considered true score variance. Reliability coefficients at the between-student level ranged from ω_{between} = .94 to ω_{between} = .98 across dimensions and subjects, indicating that 94 % to 98 % of total between-student variance represents true score variance. ICC[2] indices, which express the reliability of the student means, ranged between ICC[2] = .77 and ICC[2] = .91 across items and subjects (Table 1), indicating sufficient reliability of the student means and thus agreement within students (Lüdtke et al., 2009).

Table 4

McDonald's ω Within and Between Students as Reliability Indices for State SPIQ Across Subjects

	Mathematics			**Physics**			**German**			**English**		
	TS	CA	CM	TS	CA	CM	TS	CA	CM	TS	CA	CM
ω within	.84	.83	.76	.83	.82	.76	.84	.86	.69	.85	.83	.74
ω between	.97	.97	.94	.96	.98	.95	.96	.97	.94	.95	.97	.94

Note. TS = Teacher support; CA = Cognitive activation, CM = Classroom management.

In conclusion, the hypothesized three-factor structure at both levels fitted the data best in all subjects. Model fit indices ranged from CFI = .986 to CFI = .992, RMSEA = .016 to RMSEA = .023, $\text{SRMR}_{\text{within}}$ = .009 to $\text{SRMR}_{\text{within}}$ = .022, and $\text{SRMR}_{\text{between}}$ = .019 to $\text{SRMR}_{\text{between}}$ = .022 (see Models 1h, 2h, 3h and 4h in Table 3). Furthermore, all items loaded significantly ($p < .05$) and substantially on their proposed factor, with factor loadings ranging between λ = .66 to λ = .91 at the within-student level, and λ = .91 to λ = .99 at the between-student level across subjects (see Table 5). Note that between-level factor loadings were considerably higher than within-level factor loadings due to removing most measurement error at the between-level (Zyphur et al., 2008). Three factor loadings at the between-student level had to be fixed to 1 to avoid negative residual variances. Factor correlations for the TBDs as state SPIQ within and between students are displayed in Table 6. In all subjects, a remarkably similar picture emerged: Teacher support and cognitive activation were moderately to highly correlated (i.e., ρ = .50 to ρ = .53 within and ρ = .73 to ρ = .78 between students), whereas classroom management was correlated with teacher support in mathematics and physics between students only (i.e., ρ = .19 and ρ = .21).

Table 5

Standardized Factor Loadings at the Within- and Between-Students Levels in the Four Subjects

	Factor Loadings (Within / Between)			
Item	**Mathematics**	**Physics**	**German**	**English**
Teacher Support				
TS1	.84 / .99	.83 / .97	.86 / .99	.84 / .97
TS2	.87 / .97	.85 / .98	.85 / .95	.88 / .97
Cognitive Activation				
CA1	.78 / .94	.80 / .98	.83 / .98	.79 / .98
CA2	.90 / 1.00*	.88 / .98	.91 / .98	.89 / .98
Classroom Management				
CM1	.74 / 1.00*	.70 / .94	.70 / .97	.66 / .91
CM2	.74 / .93	.78 / 1.00*	.73 / .96	.78 / 1.00*

Note. TS1 and TS2 = Items measuring perceived teacher support; CA1 and CA2 = Items measuring perceived cognitive activation; CM1 and CM2 = Items measuring perceived classroom management (see Table S1 for item wordings).

All reported standardized factor loadings were statistically significant at $p < .05$.

* Factor loadings were fixed to 1 to avoid negative item error variances.

Table 6

Factor Correlations Between Three Basic Dimensions at the Within- and Between-Students Levels in the Four Subjects

		Factor Correlations	
Dimension	Teacher Support	Cognitive Activation	Classroom Management
		Mathematics	
Teacher Support	-	.76**	.19*
Cognitive Activation	.53**	-	.06
Classroom Management	.01	-.03	-
		Physics	
Teacher Support	-	.78**	.21*
Cognitive Activation	.51**	-	.05
Classroom Management	-.01	-.09	-
		German	
Teacher Support	-	.74**	.11
Cognitive Activation	.50**	-	.05
Classroom Management	.09	-.04	-
		English	
Teacher Support	-	.73**	.03
Cognitive Activation	.51**	-	-.11
Classroom Management	-.02	-.06	-

Note. Correlations below the diagonals represent within-students correlations, and correlations above the diagonals represent between-students correlations.

* *p* < .05; ** *p* < .01

The pattern of results was virtually the same when applying the cross-classified models. Detailed results for the cross-classified models are presented in the Online Supplementary Material. ICC[1] and ICC[2] values for the between-student level virtually did not change as compared to the MCFA models (see Table S2). The ICC[1] values for the between-lesson level ranged between .04 and .17, indicating comparably lower shared lesson variance than variance attributable to the student. ICC[2] values for the between-lesson levels were lower than ICC[2] values for the between-student level, suggesting a higher reliability of measurement points within students than that between different SPIQ within lessons, which mostly showed indices below acceptable reliability. The cross-classified models showed the best data fit when a three-factorial solution was specified within- and between-students when considering shared lesson variance (RQ1; Table S3). Both models with and without cross-level measurement invariance constraints showed a good data fit (Table S4). DIC values show only marginal discrepancies contradicting the assumption of a significant decrease of model fit due to cross-level measurement invariance constraints (see also Asparouhov & Muthén, 2020). The ω coefficients indicated that the TBDs were measured reliably within and between students (RQ2; Table S5). Factor loadings and factor correlations did not change noticeably in the cross-classified as compared to the MCFA models (see Tables S6 and S7). However, in the cross-classified models, no negative residual variances occurred. In conclusion, the results of the cross-classified models clearly strengthen the MCFA results in accordance to our RQs, rendering support for the multilevel factorial validation of SPIQ within the TBDs framework when extracting shared perceptions.

Relations to Trait SPIQ, Grades, and Academic Interest (RQ 3)

To obtain convergent validity evidence, we examined aggregated state SPIQ's correlations with trait SPIQ as well as educational outcomes the TBDS framework aims to predict (i.e., school grades and academic interest) at the between-student level (see Table 7). With regard to the four different subjects, we noted that relations were substantially higher within matching subjects than across different subjects, underscoring the TBDs' domain-specificity and, therefore, providing preliminary support for convergent and discriminant validity. Here, we only report significant relations within the same subject for increased clarity of results. However, we present cross-domain relations in Table S8 in the Online Supplementary Material.

Table 7

Latent Correlations Between Aggregated State SPIQ and Trait SPIQ to Trait SPIQ, Grades and Interest at the Pre- and Post-Assessment in Four Subjects

	Aggregated State SPIQ			Trait SPIQ					
Trait Dimension	TS	CA	CM	TS (Pre)	TS (Post)	CA (Pre)	CA (Post)	CM (Pre)	CM (Post)
				Mathematics					
TS (Pre)	.62**	.36**	.21**	--					
TS (Post)	.69**	.43**	.25**	.68**	--				
CA (Pre)	.66**	.54**	.33**	.78**	.60**	--			
CA (Post)	.53**	.62**	.31**	.43**	.63**	.72**	--		
CM (Pre)	.14	.05	.42**	.20**	.15*	.23**	.08	--	
CM (Post)	.21**	.08	.70**	.26**	.22**	.30**	.16**	.67**	--
Grade (Pre)	.23**	.18*	.08	.13*	.14*	.17*	.16*	-.01	.08
INT (Pre)	.41**	.28**	.19**	.29**	.26**	.31**	.18**	.03	.25**
INT (Post)	.39**	.26**	.14	.19**	.26**	.26**	.23**	-.04	.10
				Physics					
TS (Pre)	.69**	.43**	.10	--					
TS (Post)	.70**	.47**	.30**	.70**	--				
CA (Pre)	.63**	.54**	.24**	.63**	.52**	--			
CA (Post)	.59**	.65**	.20*	.38**	.60**	.60**	--		
CM (Pre)	.26**	.15*	.52**	.14*	.12	.25**	.13	--	
CM (Post)	.26**	.13	.76**	.20**	.26**	.28**	.19**	.47**	--
Grade (Pre)	.69**	.43**	.10	.12	.12	.21*	.27**	.15	.03
INT (Pre)	.70**	.47**	.30**	.37**	.37**	.33*	.41**	.19**	.19**
INT (Post)	.63**	.54**	.24**	.33**	.41**	.24**	.37**	.11	.16**
				German					
TS (Pre)	.59**	.36**	.11	--					
TS (Post)	.56**	.39**	.17*	.70**	--				
CA (Pre)	.40**	.33**	.08	.71**	.41**	--			
CA (Post)	.48**	.52**	.20**	.53**	.62**	.55**	--		
CM (Pre)	.17*	.10	.49**	.24**	.14*	.27**	.17*	--	
CM (Post)	.18*	.12	.54**	.19**	.26**	.15*	.21**	.53**	--
Grade (Pre)	.12	.30**	.10	.20**	.24**	.18**	.24**	.15*	.10
INT (Pre)	.35**	.29**	.16*	.46**	.40**	.39**	.40**	.28**	.20**
INT (Post)	.39**	.26**	.16*	.43**	.38**	.33**	.40**	.21**	.20**
				English					
TS (Pre)	.54**	.37**	.07	--					
TS (Post)	.44**	.32**	.29**	.62**	--				
CA (Pre)	.41**	.40**	.07	.69**	.50**	--			
CA (Post)	.39**	.47**	.20**	.56**	.68**	.74**	--		
CM (Pre)	.14	.08	.27*	.26**	.19**	.23**	.14*	--	
CM (Post)	.05	.06	.55**	.15*	.33**	.13*	.22**	.56**	--
Grade (Pre)	.07	.00	.10	.07	.13	.10	.20**	.01	.03
INT (Pre)	.29**	.19**	.26**	.35**	.36**	.47**	.44**	.06	.12
INT (Post)	.20**	.16*	.33**	.34**	.41**	.42**	.45**	.14*	.14*

Note. TS = Teacher support; CA = Cognitive activation; CM = Classroom management; INT = Academic interest; Pre = Pre-assessment (i.e., week 1); Post = Post-assessment (i.e., week 5). State SPIQ were aggregated via latent aggregation, and all correlations were estimated at the between-students level.

* p < .05; ** p < .01

Relations with trait SPIQ (e.g., aggregated state teacher support in mathematics with trait teacher support in mathematics) were positive and moderate to large, ranging between ρ = .27 to ρ = .76 across dimensions and subjects. It is worth noting that the relations between aggregated state SPIQ of teacher support [cognitive activation] and trait SPIQ of cognitive activation [teacher support] often were of similar strength or only somewhat lower as the relations within these dimensions, reflecting the two dimensions' conceptual overlap. To estimate the effect of shortening the SPIQ scale to two items per dimension, we additionally repeated all analyses of the three trait SPIQ scales with short trait SPIQ scales consisting of only the two items that were adapted and used in the e-diary. The results were virtually the same.

Examining the relations to school grades and academic interest made it possible to test essential assumptions of the TBDs framework using the aggregated state measure. As hypothesized by the framework, grades were positively related to state SPIQ of teacher support in two subjects (mathematics and physics; ρ = .23 and ρ = .29) as well as to state SPIQ of cognitive activation in two subjects (mathematics, and German; ρ = .18 and ρ = .30). State SPIQ of classroom management were not related to grades in any subject. Interest was positively related to state SPIQ of teacher support and cognitive activation in all subjects (ranging from ρ = .16 to ρ = .50), and to state SPIQ of classroom management in mathematics, physics, and English (ranging from ρ = .16 to ρ = .34), generally confirming the predictions derived from the TBDs framework.

Again, the same procedure was applied to the cross-classified models. In doing so, the relations of aggregated state SPIQ that were controlled for shared lesson perceptions to trait SPIQ, interest, and achievement were virtually the same (see Table S9).

Discussion

Frameworks on SPIQ are critical to defining, assessing, and ultimately improving instructional quality, a key educational construct that is directly linked to relevant educational outcomes. The TBDs, as one of the most widely employed frameworks, have received considerable attention in factorial and predictive validation studies. Recently, multiple scholars have stressed the importance of multilevel validation studies of SPIQ as a hierarchical construct (e.g., Wisniewski et al., 2020). However, all available multilevel validation studies exclusively examine the between-student and class or school level, even though instruction itself and thus also SPIQ take place within lessons and are subject to substantial within-person dynamics on the part of both teachers and students. To address this shortcoming, we conducted an experience sampling study and assessed state SPIQ within the TBDs in every mathematics, physics, German, and English lesson over three weeks of everyday school life, capturing within-student variation in SPIQ. In doing so, we confirmed the three-factor structure of state SPIQ within- and between students (RQ1) that were assessed reliably in our experience sampling design (RQ2) and showed significant relations to crucial educational outcomes (RQ3).

Factorial Validity of SPIQ Within and Between Students

The approximately 50 % within-person variation found in our study is in line with recent findings across a variety of psychological constructs (e.g., job performance, self-efficacy, or stressors; Podsakoff et al., 2019. Additionally, the relatively low test-retest stabilities in the trait conceptualization of SPIQ over a rather short period of time were of comparable size to prior research (Wagner et al., 2016), hinting at variability in the construct (but see also Marsh, 2007, who found high stabilities of SPIQ at the class level over a long period of time). In the present study, we shed light on within-student variation in SPIQ over time to provide valuable theoretical insights on the TBDs as well as novel practical implications for teaching effectiveness. Note that we focus on the within- and between-student level in the present study, but want to emphasize the need to analyze SPIQ at the class level (see Section Limitations).

Our results yielded strong support for the factor structure hypothesized by the TBDs framework both within and between students (RQ 1). In other words, the three dimensions of teacher support, cognitive activation, and classroom management are distinguishable both in individual lessons and as aggregated perceptions across points in time. Changes in perceptions of these dimensions are thus expressed as changes in state measurement instruments, suggesting the relevance of situational perceptions in real time in addition to trait-like perceptions. As this structure held across the four investigated subjects, the understanding of the TBDs as a generic construct was supported. Goetz et al. (2013; 2020) showed that real-time perceptions of instructional quality are relevant with respect to situational emotions. We added to this reasoning by demonstrating the situational existence and importance of the TBDs. Without this knowledge, one might have assumed, for example, that classroom management, which is considered to be relatively stable, is not prone to situational changes at all and loses its relevance in individual lessons in everyday school life, but rather operates as a trait-like construct. Another misconception might have concerned differing factor structures across subjects other than mathematics. In contrast, our findings indicate that all three dimensions are relevant in individual lessons and similarly associated in mathematics, physics, German, and English. For instance, we found that teacher support and cognitive activation were moderately to highly related to one another; teachers who are perceived to be more supportive also tend to be perceived as more cognitively engaging, both in individual lessons and in general. This relation is consistent with prior research results and theoretical assumptions regarding the conceptual overlap between these two dimensions (Fauth et al., 2014; Jonassen, 2011; Scherer et al., 2016; Wagner et al., 2013), while substantially extending existing findings to a broader range of subject areas.

To assess the trustworthiness of these findings, we investigated the psychometric properties of the implemented state measure (RQ 2). In experience sampling, one of the most prominent methodological issues lies in developing reduced scales suitable for high-frequent assessment, often leading to the use of single-item measures (e.g., Goetz et al., 2020; Klumb et al., 2009). In contrast, we used two items per factor, which limited us in terms of model estimation compared to traditional longer scales (e.g., resulting in occasional negative error variances in the specified MCFA but not in the cross-classified models), but nevertheless enabled latent modeling and thus the elimination of measurement error and the estimation of scale reliability. Furthermore, good reliability estimates were found within and between students, indicating that students reliably differentiated between the three dimensions in individual lessons, and different students reliably differentiated between the three dimensions in their aggregated, overall perceptions. These psychometric properties were especially satisfactory given the rather strict MCFA model assumptions (i.e., perfect item-factor associations)—particularly when including cross-level invariance constraints (i.e., the same strength of item-factor relations in individual lessons as overall). Our results thus seem to provide strong support for the generalizability of the three-dimensional factor structure across levels. Returning to the taxonomy of multilevel constructs introduced by Stapleton and colleagues (2016), we, therefore, argue for considering SPIQ as a configural multilevel construct, displaying a cross-level invariant structure within- and between-students (i.e., Levels 1 and 2 in the two-level models, and Levels 1 and 2a in the cross-classified models, respectively). We have demonstrated substantial, meaningful within-student variation—also after extracting shared situational perceptions—weakening the notion of students' interchangeability as mere informants for a higher-level construct.

Convergent Validity of State SPIQ

To obtain insights into whether assumptions made by the TBDs framework also apply to state SPIQ, we investigated relations with the key educational outcomes of grades and interest in addition to trait SPIQ (RQ 3). Aggregating state SPIQ emphasized its trait-like component, which was reflected in the relations to 'pure' trait SPIQ, which indicated a substantial overlap between these two constructs. Yet, the correlations were far from perfect, reflecting the hypothesized distinction between state and trait we wanted to achieve. However, it is important to note that the state and trait scales differed due to the adaptation to the experience sampling design (i.e., change of reference from multiple lessons to specific lessons and according change of response format along with shortening and wording adaptation). These changes were undertaken to transform the trait scale into the state scale, yet, have to be kept in mind when interpreting the respective scale scores (i.e., responses of state scale reflect the extent of agreement with a statement for a specific situation whereas responses of the trait scale reflect an aggregated evaluation of instructional quality in an unknown number of lessons by the student). We also showed that the selection of two items for the state scale proved to be straightforward when contrasting state relations to trait relations with the same two items versus to trait relations with all items.

As assumed by the TBDs framework, teacher support should predict interest, whereas cognitive activation should predict grades, and classroom management should predict grades and possibly interest. These theory-driven expectations could largely be confirmed, with some differences across subjects as well as other positive relations that had not been hypothesized. In general, we found positive relations between state SPIQ of teacher support and cognitive activation with grades and interest, and between state SPIQ of classroom management with interest only. This indicates that the TBDs are not only present in individual lessons and undergo situational variation, they are also differentially related to crucial outcomes, further underscoring their relevance.

Limitations

The present study has several limitations: First, SPIQ are hierarchical, referring to the teacher or classroom, which is visible in the item wordings accordingly. Although we highlighted the importance of a multilevel approach in analyzing SPIQ, we were not able to explicitly model the class level, as we reached the statistical limits of our sample at Level 3 (i.e., 18 clusters). Note that the present study does not aim to discount teacher / classroom effects in SPIQ, or even argue against analyzing SPIQ on the class level. For instance, class level aggregated SPIQ are crucial to provide information on teaching effectiveness. Instead, the present study seeks to enhance awareness of lesson-to-lesson variation in SPIQ to validate the framework of TBDs at the within-student level. SPIQ are affected by student traits and states, but also teacher traits and states as well as external factors. We did not assess teacher traits and states, and can therefore only speak about *student perceptions* but not actual instructional quality. It is conceivable that we underestimate teacher effects by design, as all variation at Level 1 is attributed to student states in the two-level models. However, the resulting pattern did not change remarkably when employing the cross-classified models, that is when extracting shared classroom perceptions as an approximation to considering actual teaching behavior. Still, the present study clearly focused on the student and not the teacher to raise awareness of meaningful lesson-to-lesson variation in SPIQ and thus also consider the within-student level in studies on instructional quality. To extract idiosyncratic SPIQ and with this, to consider teacher effects, we (a) added dummy-coded classroom variables as predictors at the between-student level and (b) added a cross-classified between-lesson level to clear idiosyncratic situational perceptions by shared, consensual situational perceptions. ICC[1] values for both the class level and the between-lesson level proved to be lower compared to the between-student level, indicating lower SPIQ variance proportions at these levels compared to the between-student level. Still, the approach we realized is not optimal and only a compromise. We acknowledge that (a) we only have a near-perfect match between classrooms and teachers, leaving some double combinations of teachers teaching in multiple classrooms unconsidered (see Fauth et al., 2020). Further, (b) there are still variance components that we cannot explain (e.g., student*teacher interactions; see Feistauer & Richter, 2017). The distribution of these additional variance components and their impact on the current explained variance components remains unclear. Thus, future research is indicated here. Nevertheless, our results show that a large proportion of variance in SPIQ is attributable to the student—even after controlling for the ‘actual’ instructional quality as operationalized by shared lesson perceptions. This further supports research efforts to focus on the *student* in student perceptions of instructional quality (see also Kunter & Baumert, 2007). As such, the present study contributes to the understanding that substantial systematic variance in SPIQ (i.e., within-student variance) is largely ignored if the within-student level is not taken into account (i.e., when conducting between-person analyses relying on the between-student and class levels only).

As another limitation, it is important to keep in mind that our results are correlational. We did not perform experimental manipulations of instructional quality to assess state SPIQ and its effects on outcomes. Therefore, we cannot draw any conclusions concerning causality. In particular, the relations to trait SPIQ, school grades, and interest do not imply a direction and could also potentially be explained by third variables not considered here. Future research could build upon our findings by manipulating state instructional practices or including other situational variables hypothesized to be directly or indirectly linked to SPIQ according to the TBDs framework, thus transforming between-person-based knowledge to the within-person level.

Further, we note that our results are based on a sample of 9th and 10th grade German secondary school students attending six highest ability track schools in four German federal states. As is common in psychological research, we cannot claim that our sample and, with this, our results are representative, and thus call for further research based on different samples (e.g., different cultures, school systems, ability tracks, and age groups).

At the same time, we investigated four different school subjects to enhance generalizability across subjects. Thus, we included other domains of interest to the mathematics-focused TBDS research and found similar result patterns across different domains. Specifically, we established a three-factor structure across the subjects math, physics, German and English. For comparative research on SPIQ across subjects, however, assumptions of measurement invariance need to be tested in future studies (see Schneider et al., 2022). In the present study, student grades and interest as central achievement and motivation indicators were used as convergent validity criteria. One could examine further crucial outcome variables such as academic self-concept (e.g., Scherer et al., 2016) or emotions (e.g., Goetz et al., 2020; see also Praetorius et al., 2018 for an overview of examined criteria).

Finally, note that we used school grades as achievement indicators rather than standardized test scores. The distinction between these two indicators has been discussed in prior work (Arens, Marsh, et al., 2017). On the one hand, school grades are communicated by the teacher, increasing their salience to the student. On the other hand, grades seem more biased (e.g., the class as the frame of reference). We chose grades as achievement indicators due to our idiosyncratic approach, as well as the fact that information on the teacher and classroom are highly salient in SPIQ. Moreover, it has recently been shown that grades affect SPIQ in the same subject and across subjects after controlling for standardized achievement test scores (Jaekel et al., 2021).

Implications and Conclusion

We could confirm the three-dimensional structure of the TBDs within (i.e., from lesson to lesson) and between students in four core school subjects (mathematics, physics, German, and English). Our findings suggest that the student plays a pivotal role when examining (perceptions of) instructional quality. Knowledge of the relevance of all TBDs in individual lessons and the relevance of individual student characteristics in SPIQ could help teachers not only design their lessons but also enhance their teaching effectiveness by raising their awareness of classroom dynamics. Acknowledging the classroom as a highly interactional and situational system makes it possible to identify the circumstances under which certain instructional practices are linked to corresponding outcomes. In this way, teaching can be adaptive instead of fixed. Bringing daily fluctuations to attention also makes it possible to shed light on the roles of student traits and states as well as teacher traits and states in (perceptions of) instruction. Both teachers and students could benefit from a more dynamic view of classroom interactions in these core subjects that weakens rigid dysfunctional attribution styles (e.g., “I am a bad teacher”). Realizing that daily states operate alongside general traits has the potential to encourage the adoption of a more constructive mindset by both students and teachers in which each lesson is treated as a new opportunity. Hence, future research could focus on the effects of SPIQ not only within but also between subjects (Jaekel et al., 2021), which could hold further potential for teacher trainings and school administration.

Overall, our results significantly contributed to existing literature on SPIQ within the TBDs framework by extending the level of analysis to the within-student level. We argue that this framework can inform researchers and practitioners about the perceived quality of instruction over time and across lessons and thus has the potential to shed light on the dynamics and processes that underlie perceptions of instruction.

Online Supplementary Material

Table S1

The German-language Scale to Assess State SPIQ within the TBD and its English Translation

Item	German version	English version
	Teacher Support	
TS1	In der Stunde eben hat unsere Lehrerin / unser Lehrer uns beim Lernen unterstützt.	During this lesson, the teacher helped students with their learning.
TS2	In der Stunde eben hat unsere Lehrerin / unser Lehrer sich für den Lernfortschritt jeder einzelnen Schülerin / jedes einzelnen Schülers interessiert.	During this lesson, the teacher showed an interest in every student's learning.
	Cognitive Activation	
CA1	In der Stunde eben hat unsere Lehrerin / unser Lehrer uns Aufgaben gegeben, bei denen wir einige Zeit darüber nachdenken mussten.	During this lesson, the teacher gave problems that required us to think for an extended time.
CA2	In der Stunde eben hat unsere Lehrerin / unser Lehrer Fragen gestellt, die uns angeregt haben, über die Aufgabe nachzudenken.	During this lesson, the teacher asked questions that made us reflect on the problem.
	Classroom Management	
CM1	In der Stunde eben war es oft laut, und es ging drunter und drüber.	During this lesson, there was noise and disorder.
CM2	In der Stunde eben musste unsere Lehrerin / unser Lehrer lange warten, bis die Schülerinnen und Schüler ruhig wurden.	During this lesson, the teacher had to wait for a long time for students to quiet down.

Note. TS1 and TS2 = Items measuring perceived teacher support; CA1 and CA2 = Items measuring perceived cognitive activation; CM1 and CM2 = Items measuring perceived classroom management.

Table S2

Cross-Classified Models: Variance Components, ICC[1], and ICC[2] Within Students and Within Lessons (Level 1), Between Students (Level 2a), and Between Lessons (Level 2b) for State SPIQ in Four Subjects

	s^2_{within} *(SE)*	$s^2_{\text{between-students}}$ *(SE)*	$s^2_{\text{between-lessons}}$ *(SE)*	Between-Student Level		Between-Lesson Level	
				ICC[1]	ICC[2]	ICC[1]	ICC[2]
Mathematics							
TS1	0.96 (.03)	0.80 (.08)	0.22 (.04)	.40	.83	.11	.65
TS2	0.97 (.03)	0.92 (.08)	0.20 (.03)	.44	.85	.10	.63
CA1	1.02 (.03)	0.70 (.07)	0.12 (.02)	.38	.82	.07	.53
CA2	1.01 (.03)	0.78 (.07)	0.10 (.02)	.41	.83	.05	.44
CM1	0.92 (.03)	1.11 (.09)	0.13 (.03)	.51	.88	.06	.49
CM2	0.77 (.02)	1.08 (.10)	0.15 (.03)	.54	.89	.07	.53
Physics							
TS1	0.75 (.03)	1.00 (.10)	0.14 (.04)	.52	.83	.08	.56
TS2	0.80 (.03)	1.03 (.10)	0.18 (.05)	.51	.82	.09	.59
CA1	0.93 (.04)	0.83 (.09)	0.21 (.05)	.42	.76	.11	.65
CA2	0.90 (.04)	0.89 (.09)	0.07 (.03)	.48	.80	.04	.38
CM1	0.97 (.04)	0.81 (.09)	0.22 (.06)	.41	.75	.11	.64
CM2	0.88 (.04)	0.81 (.09)	0.19 (.06)	.43	.77	.10	.62
German							
TS1	0.90 (.03)	0.95 (.09)	0.14 (.03)	.48	.84	.07	.54
TS2	0.89 (.03)	1.06 (.10)	0.12 (.03)	.51	.85	.06	.50
CA1	1.12 (.04)	0.73 (.08)	0.25 (.05)	.35	.74	.12	.68
CA2	1.04 (.04)	0.81 (.09)	0.20 (.04)	.40	.78	.10	.63
CM1	0.92 (.04)	1.24 (.12)	0.22 (.05)	.52	.86	.09	.60
CM2	0.80 (.03)	1.23 (.11)	0.18 (.04)	.56	.87	.08	.57
English							
TS1	0.80 (.03)	0.73 (.08)	0.32 (.07)	.39	.76	.17	.77
TS2	0.83 (.03)	0.86 (.08)	0.23 (.06)	.45	.80	.12	.69
CA1	1.08 (.04)	0.70 (.07)	0.19 (.05)	.35	.73	.10	.65
CA2	0.99 (.04)	0.76 (.08)	0.19 (.05)	.39	.76	.10	.65
CM1	0.95 (.04)	0.83 (.07)	0.17 (.04)	.43	.79	.09	.62
CM2	0.80 (.03)	0.76 (.07)	0.14 (.04)	.45	.80	.08	.59

Note. TS1 and TS2 = Items measuring perceived teacher support; CA1 and CA2 = Items measuring perceived cognitive activation; CM1 and CM2 = Items measuring perceived classroom management (see Table S1 for item wordings). ICC[1] = The proportion of between-level to total variance. ICC[2] = Reliability of the level means as derived from the ICC[1] and the number of measurement points within students (for the between-student level) or the number of different student perceptions within lessons (for the between-lesson level).

Table S3

Cross-Classified Models: Testing Alternative Factor Structures at the Within- and Between-Students Levels for each Subject via the Partially Saturated Modeling Approach

Model	Within	Between Student Level	Between Lesson Level	PPP_{χ^2}	*pD*	DIC
			Mathematics			
1i	saturated	3 factors	saturated	0.418	1972.817	41326.810
1j	saturated	1 factor	saturated	0.000	2073.550	41677.753
1k	3 factors	saturated	saturated	0.421	1956.117	41295.843
1l	1 factor	saturated	saturated	0.000	1714.140	42662.242
1m	1 factor	3 factors	saturated	0.000	1699.154	42675.699
1n	3 factors	1 factor	saturated	0.000	2057.618	41657.463
			Physics			
2i	saturated	3 factors	saturated	0.438	1452.183	23643.086
2j	saturated	1 factor	saturated	0.002	1450.767	24085.249
2k	3 factors	saturated	saturated	0.465	1467.462	23634.187
2l	1 factor	saturated	saturated	0.000	1273.673	24298.770
2m	1 factor	3 factors	saturated	0.000	1257.933	24282.238
2n	3 factors	1 factor	saturated	0.001	1451.124	24068.202
			German			
3i	saturated	3 factors	saturated	0.390	1743.600	31267.544
3j	saturated	1 factor	saturated	0.000	1809.555	31623.656
3k	3 factors	saturated	saturated	0.344	1764.343	31260.152
3l	1 factor	saturated	saturated	0.000	1515.344	32330.635
3m	1 factor	3 factors	saturated	0.000	1497.365	32334.972
3n	3 factors	1 factor	saturated	0.000	1781.940	31589.293
			English			
4i	saturated	3 factors	saturated	0.371	1630.529	28040.110
4j	saturated	1 factor	saturated	0.000	1651.425	28396.922
4k	3 factors	saturated	saturated	0.177	1624.347	28035.690
4l	1 factor	saturated	saturated	0.000	1438.296	28873.309
4m	1 factor	3 factors	saturated	0.000	1422.122	28831.078
4n	3 factors	1 factor	saturated	0.000	1644.300	28402.065

Note. PPP_{χ^2} = Posterior predictive *p*-value; *pD* = estimated number of parameters; DIC = Deviance Information Criterion. Models specified at the between-student level (models i, j, m, and n) consider classroom effects by including dummy-coded classroom variables.

Table S4

Cross-Classified Models: Testing Cross-Level Measurement Invariance in Factor Models with 3 Factors at the Within- and Between-Students Levels and Saturated Models at the Between-Lessons Level

Model	Specification	PPP$_{\chi^2}$	*pD*	DIC
	Mathematics			
10	3 / 3 / saturated factor model	0.373	1949.358	41299.117
1p	3 / 3 / saturated cross-level invariant factor model	0.395	1965.332	41296.103
	Physics			
20	3 / 3 / saturated factor model	0.475	1441.497	23649.864
2p	3 / 3 / saturated cross-level invariant factor model	0.450	1456.738	23608.595
	German			
30	3 / 3 / saturated factor model	0.314	1717.908	31237.835
3p	3 / 3 / saturated cross-level invariant factor model	0.297	1744.844	31245.894
	English			
40	3 / 3 / saturated factor model	0.191	1592.707	28008.119
4p	3 / 3 / saturated cross-level invariant factor model	0.188	1629.263	28023.023

Note. PPP$_{\chi^2}$ = Posterior predictive *p*-value; *pD* = estimated number of parameters; DIC = Deviance Information Criterion; BCFI = Bayesian Comparative Fit Index; BRMSEA = Bayesian Root Mean Square Error of Approximation.

3 / 3 / saturated factor models specify 3 factors within, 3 factors between students as well as item (co)variances between lessons. 3 / 3 / saturated cross-level invariant factor models specify 3 factors within and 3 factors between students with equal item factor loadings across those levels, as well as item (co)variances between lessons.

Table S5

Cross-Classified Models: McDonald's ω Within and Between Students as Reliability Indices for State SPIQ Across Subjects

	Mathematics			**Physics**			**German**			**English**		
	TS	CA	CM	TS	CA	CM	TS	CA	CM	TS	CA	CM
ω within	.82	.82	.69	.82	.82	.70	.84	.85	.65	.84	.82	.67
ω between students	.86	.89	.87	.91	.89	.86	.90	.85	.88	.85	.87	.86

Note. TS = Teacher support; CA = Cognitive activation, CM = Classroom management.

Table S6

Cross-Classified Models: Standardized Factor Loadings at the Within- and Between-Students Levels in the Four Subjects

	Factor Loadings (Within / Between)			
Item	**Mathematics**	**Physics**	**German**	**English**
Teacher Support				
TS1	.82 / .99	.82 / .97	.84 / .99	.82 / .97
TS2	.85 / .98	.83 / .98	.85 / .96	.87 / .98
Cognitive Activation				
CA1	.78 / .96	.80 / .98	.83 / .98	.78 / .98
CA2	.87 / .99	.85 / .99	.89 / .98	.87 / .98
Classroom Management				
CM1	.68 / .96	.68 / .96	.64 / .97	.64 / .94
CM2	.74 / .97	.73 / .98	.68 / .97	.73 / .99

Note. TS1 and TS2 = Items measuring perceived teacher support; CA1 and CA2 = Items measuring perceived cognitive activation; CM1 and CM2 = Items measuring perceived classroom management (see Table S1 for item wordings).

All reported standardized factor loadings were statistically significant at $p < .05$.

Table S7

Cross-Classified Models: Factor Correlations Between Three Basic Dimensions at the Within- and Between-Students Levels in the Four Subjects

	Factor Correlations		
Dimension	Teacher Support	Cognitive Activation	Classroom Management
Mathematics			
Teacher Support	-	.75**	.18*
Cognitive Activation	.56**	-	.04
Classroom Management	.04	-.02	-
Physics			
Teacher Support	-	.75**	.20*
Cognitive Activation	.52**	-	.05
Classroom Management	-.04	-.14**	-
German			
Teacher Support	-	.73**	.12
Cognitive Activation	.49**	-	.04
Classroom Management	.04	-.06	-
English			
Teacher Support	-	.71**	.02
Cognitive Activation	.50**	-	-.11
Classroom Management	-.04	-.08	-

Note. Correlations below the diagonals represent within-students correlations, and correlations above the diagonals represent between-students correlations.

* *p* < .05; ** *p* < .01

Table S8

Within- and Cross-Domain Factor Correlations Between Aggregated State SPIQ, Trait SPIQ, Grades and Interest at Pre- and Post-Assessment in Two-Level Models

		Aggregated State SPIQ											
	Trait Dimension	**Mathematics**			**Physics**			**German**			**English**		
		TS	CA	CM	TS	CA	CM	TS	CA	CM	TS	CA	CM
Mathematics	TS (Pre)	.62**	.36**	.21**	.34**	.23**	.14	.24**	.25**	.10	.29**	.15**	.07
	TS (Post)	.69**	.43**	.25**	.26**	.22**	.11	.26**	.26**	.16*	.28**	.23**	.17*
	CA (Pre)	.66**	.54**	.33**	.36**	.29**	.30**	.16	.23**	.21*	.20**	.12*	.17*
	CA (Post)	.53**	.62**	.31**	.24**	.37**	.20*	.20**	.31**	.21**	.24**	.37**	.28**
	CM (Pre)	.14	.05	.42**	.00	-.10	.30**	-.05	-.09	.22*	.01	.02	.11*
	CM (Post)	.21**	.08	.70**	.06	-.07	.43**	.13	.03	.26**	-.01	-.04	.35**
	Grade (Pre)	.23**	.18*	.08	.17	.16	.11	.01	.19*	-.06	.04	.01	-.07
	INT (Pre)	.41**	.28**	.19**	.28**	.21**	.14	.26**	.26**	.07	.16*	.05	-.05
	INT (Post)	.39**	.26**	.14	.31**	.26**	.16*	.19**	.20**	.12	.14*	.03	.02
Physics	TS (Pre)	.35**	.24**	-.03	.69**	.43**	.10	.25**	.29**	-.07	.29**	.21**	-.11
	TS (Post)	.36**	.20**	.13	.70**	.47**	.30**	.16*	.07	.04	.25**	.17*	.09
	CA (Pre)	.28**	.26**	.06	.63**	.54**	.24**	.13	.24**	.01	.26**	.25**	.01
	CA (Post)	.35**	.41**	.18	.59**	.65**	.20*	.13	.36**	-.03	.24**	.39**	.00
	CM (Pre)	.10	.04	.26**	.26**	.15*	.52**	.13	.16*	.36**	.14	.05	.24**
	CM (Post)	.09	.05	.40**	.26**	.13	.76**	.09	.06	.48**	-.05	.00	.35**
	Grade (Pre)	.06	-.02	-.06	.69**	.43**	.10	.00	.19*	-.03	-.04	-.11	-.03
	INT (Pre)	.37**	.37**	.07	.70**	.47**	.30**	.27**	.34**	.08	.23**	.24**	-.05
	INT (Post)	.28**	.28**	-.02	.63**	.54**	.24**	.19**	.24**	.03	.16*	.14*	-.03
German	TS (Pre)	.34**	.26**	.04	.28**	.24**	.02	.59**	.36**	.11	.24**	.22**	.13
	TS (Post)	.31**	.25**	.09	.25**	.29**	.04	.56**	.39**	.17*	.19*	.24**	.14
	CA (Pre)	.17*	.28**	-.06	.30**	.34**	.10	.40**	.33**	.08	.15	.27**	.04
	CA (Post)	.35**	.38**	.10	.35**	.38**	.01	.48**	.52**	.20**	.24**	.35**	.09
	CM (Pre)	.06	-.04	.29**	.06	-.02	.31**	.17*	.10	.49**	.06	.09	.30**
	CM (Post)	.12	.02	.35**	.11	-.02	.40**	.18*	.12	.54**	-.03	.04	.33
	Grade (Pre)	.08	.17*	.16*	.13	.15*	.14	.12	.30**	.10	.04	.13	.15
	INT (Pre)	.11	.17*	.07	.07	.10	.09	.35**	.29**	.16*	.12	.23**	.11
	INT (Post)	.12	.16*	-.03	.02	.10	.05	.39**	.26**	.16*	.03	.13	.12
English	TS (Pre)	.31**	.20**	.07	.28**	.20**	.10	.17*	.18*	-.05	.54**	.37**	.07
	TS (Post)	.34**	.26**	.25**	.24**	.17*	.16*	.27**	.21**	.19**	.44**	.32**	.29**
	CA (Pre)	.32**	.37**	.10	.30**	.27**	.10	.13	.19*	-.03	.41**	.40**	.07
	CA (Post)	.32**	.35**	.21**	.17*	.22**	.11	.11	.19*	.13	.39**	.47**	.20**
	CM (Pre)	.15	.06	.33**	.07	-.08	.36**	.11	.15	.16	.14	.08	.27*
	CM (Post)	.09	-.03	.47**	-.02	-.14	.38**	.08	.07	.47**	.05	.06	.55**
	Grade (Pre)	-.03	-.04	-.03	-.03	-.02	.01	-.08	.06	-.05	.07	.00	.10
	INT (Pre)	.07	.05	.18*	-.07	-.05	.21**	-.01	-.02	.20**	.29**	.19**	.26**
	INT (Post)	-.01	.06	.15*	.02	.02	.24**	-.02	-.02	.20**	.20**	.16*	.33**

Note. TS = Teacher support; CA = Cognitive activation; CM = Classroom management; INT = Academic interest; Pre = Pre-assessment (in week 1); Post = Post-assessment (in week 5). State SPIQ were aggregated via latent aggregation, and all correlations were estimated at the between-students level.

* *p* < .05; ** *p* < .01

Table S9

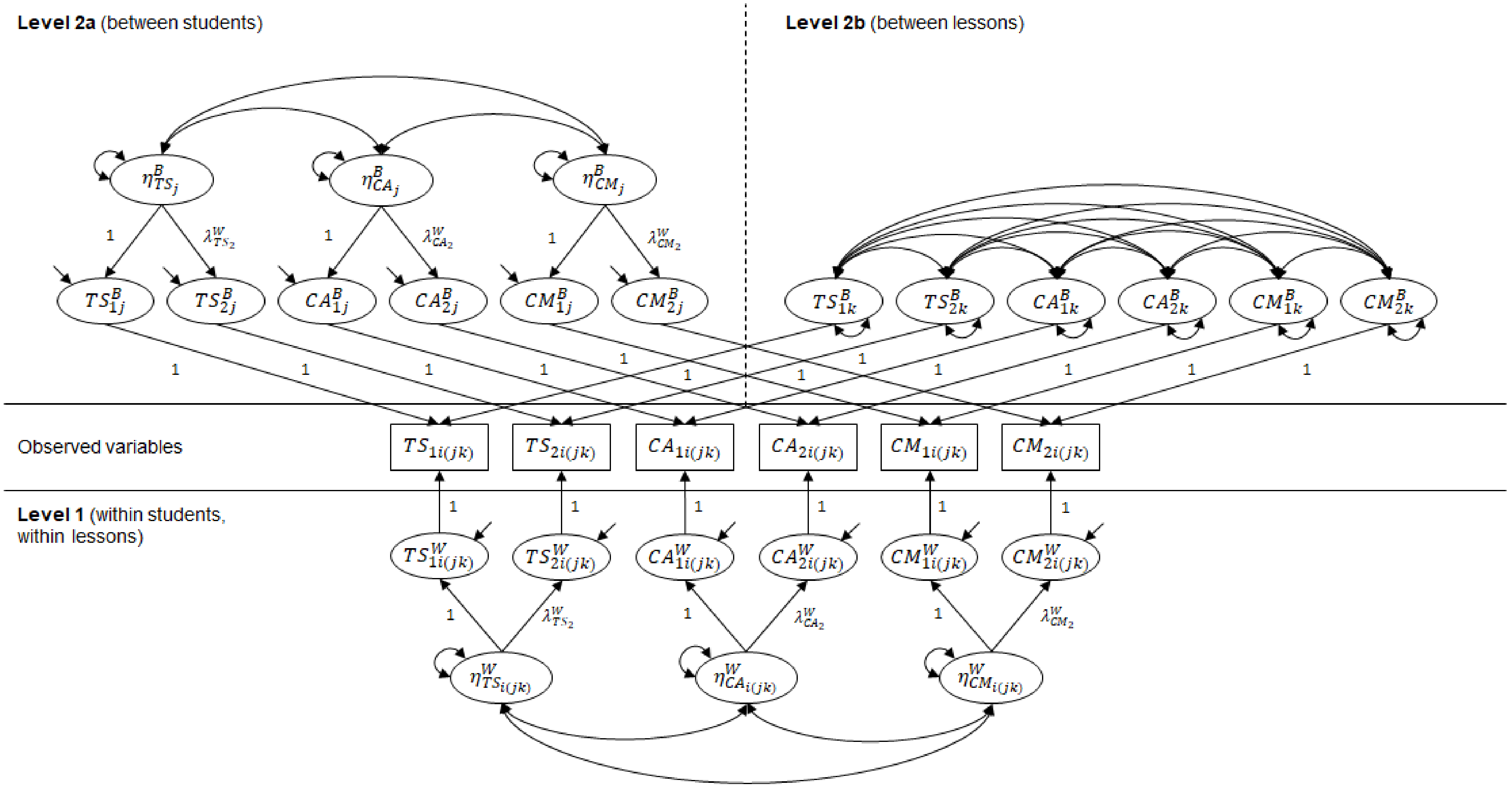
Cross-Classified Models: Factor Correlations Between Aggregated State SPIQ and Trait SPIQ, Grades and Interest at the Pre- and Post-Assessment in Four Subjects

	Aggregated State SPIQ			Trait SPIQ					
Trait Dimension	TS	CA	CM	TS (Pre)	TS (Post)	CA (Pre)	CA (Post)	CM (Pre)	CM (Post)
					Mathematics				
TS (Pre)	.60**	.35**	.22**	--					
TS (Post)	.68**	.42**	.25**	.68**	--				
CA (Pre)	.63**	.52**	.33**	.78**	.60**	--			
CA (Post)	.52**	.62**	.29**	.43**	.63**	.72**	--		
CM (Pre)	.14	.05	.42**	.20**	.15*	.23**	.08	--	
CM (Post)	.21**	.09	.69**	.26**	.22**	.30**	.16**	.67**	--
Grade (Pre)	.22**	.18**	.05	.13*	.14*	.17*	.16*	-.01	.08
INT (Pre)	.39**	.28**	.17**	.29**	.26**	.31**	.18**	.03	.25**
INT (Post)	.37**	.26**	.13*	.19**	.26**	.26**	.23**	-.04	.10
					Physics				
TS (Pre)	.68**	.42**	.10	--					
TS (Post)	.68**	.45**	.30**	.70**	--				
CA (Pre)	.61**	.52**	.23**	.63**	.52**	--			
CA (Post)	.57**	.62**	.21**	.38**	.60**	.60**	--		
CM (Pre)	.26**	.15*	.50**	.14*	.12	.25**	.13	--	
CM (Post)	.27**	.14*	.74**	.20**	.26**	.28**	.19**	.47**	--
Grade (Pre)	.28**	.18*	.06	.12	.12	.21*	.27**	.15	.03
INT (Pre)	.50**	.45**	.22**	.37**	.37**	.33*	.41**	.19**	.19**
INT (Post)	.46**	.39**	.16*	.33**	.41**	.24**	.37**	.11	.16**
					German				
TS (Pre)	.57**	.34**	.11	--					
TS (Post)	.55**	.37**	.18**	.70**	--				
CA (Pre)	.38**	.31**	.08	.71**	.41**	--			
CA (Post)	.46**	.49**	.21**	.53**	.62**	.55**	--		
CM (Pre)	.17**	.10	.46**	.24**	.14*	.27**	.17*	--	
CM (Post)	.18**	.12	.52**	.19**	.26**	.15*	.21**	.53**	--
Grade (Pre)	.11	.29**	.09	.20**	.24**	.18**	.24**	.15*	.10
INT (Pre)	.34**	.27**	.16*	.46**	.40**	.39**	.40**	.28**	.20**
INT (Post)	.38**	.25**	.16*	.43**	.38**	.33**	.40**	.21**	.20**
					English				
TS (Pre)	.53**	.35**	.07	--					
TS (Post)	.43**	.32**	.28**	.62**	--				
CA (Pre)	.40**	.38**	.08	.69**	.50**	--			
CA (Post)	.38**	.46**	.20**	.56**	.68**	.74**	--		
CM (Pre)	.15*	.09	.27**	.26**	.19**	.23**	.14*	--	
CM (Post)	.06	.05	.53**	.15*	.33**	.13*	.22**	.56**	--
Grade (Pre)	.07	-.01	.09	.07	.13	.10	.20**	.01	.03
INT (Pre)	.27**	.17**	.25**	.35**	.36**	.47**	.44**	.06	.12
INT (Post)	.19**	.14*	.33**	.34**	.41**	.42**	.45**	.14*	.14*

Note. TS = Teacher support; CA = Cognitive activation; CM = Classroom management; INT = Academic interest; Pre = Pre-assessment (i.e., week 1); Post = Post-assessment (i.e., week 5). State SPIQ were aggregated via latent aggregation, and all correlations were estimated at the between-students level. * *p* < .05; ** *p* < .01

Figure S10

Hypothesized Cross-Classified Factor Model for State SPIQ within the TBD



Note. TS1 and TS2 = Items measuring perceived teacher support; CA1 and CA2 = Items measuring perceived cognitive activation; CM1 and CM2 = Items measuring perceived classroom management (see Table S1 for item wordings). Subscripted indices *i*, *j*, and *k* denote the measurement points, students, and lessons, respectively. Superscripted letters W and B indicate the within- and between-level, respectively.

Dummy-coded classroom variables served as predictors at the between-students level but are not explicitly represented for better clarity of presentation. The mean structure is not shown in the model. This model assumes cross-level metric invariance between the levels 1 and 2a.

Chapter 4

Students' Personality and the Dynamics Between Lesson-Specific Perceived Instructional Quality and Learning Achievement: An Experience Sampling Approach

This contribution is to be submitted as[1]:

Talić, I., Rauthmann, J. F., Renner, K.-H., Möller, J., & Niepel, C. (to be submitted). Students' personality and the dynamics between lesson-specific perceived instructional quality and learning achievement: An experience sampling approach.

[1] The final version of this contribution published in the journal may differ due to revisions in the course of the peer-review process.

4. Students' personality and the dynamics between lesson-specific perceived instructional quality and learning achievement: An experience sampling approach

Abstract

Students' perceptions of instructional quality (SPIQ) are subjective and time-specific to some extent. Yet they are mostly assessed at one point in time and aggregated across students, thus largely neglecting student- and lesson-specific variance. The present study aimed at shedding light on the dynamics of students' perceptions in specific lessons. Specifically, we examined the role of students' personality traits in state SPIQ and their relation to perceived lesson-specific learning achievement (i.e., self-reported comprehension). Thereby, we distinguished between idiosyncratic and consensual (classroom) SPIQ. To this end, we assessed the three basic dimensions of instructional quality, that is, teacher support, cognitive activation, and classroom management, as state perceptions of all students within classrooms in mathematics' instruction ($N_{\text{observations}}$ = 2,681) across three weeks of 372 German secondary school students' (M_{age} = 15.3 years) daily life. Linear mixed effect models revealed (a) students' personality traits of agreeableness and negative emotionality predicting state SPIQ positively and negatively, respectively, (b) particularly pronounced positive relations between the SPIQ dimension of teacher support and perceived learning achievement, which are (c) stronger for less agreeable students. Differences across idiosyncratic and consensual perceptions could hardly be detected. Thus, the present study sheds light on student characteristics that are related to SPIQ and examined (conditions of) within-student SPIQ–learning achievement relations, while demonstrating a new application for classroom-based state SPIQ that bridges the gap between subjective perceptions and objective instructional behavior.

Keywords: Students’ perceptions of instructional quality; personality; experience sampling; idiosyncratic perceptions; linear mixed effects models

Introduction

Teachers' instructional quality substantially impacts students' achievement (Kunter et al., 2013). Accordingly, students' perceptions of instructional quality (SPIQ) are implemented worldwide in educational large-scale assessments such as the Programme for International Student Assessment (PISA) to provide information on teaching effectiveness at the country level (OECD, 2014). Thus, major efforts have been directed toward assessing the validity of these *perceptions* (e.g., Fauth et al., 2014; Kunter et al., 2007; Ruzek et al., 2022; Wagner et al., 2013; Wisniewski et al., 2020; Wisniewski et al., 2022). In the majority of studies on SPIQ (Praetorius et al., 2018), student ratings are assessed at one time point, aggregated to the class- or school-level, and related to correlates of interest such as achievement (Lüdtke et al., 2009). Such aggregated data neglect (a) individual student characteristics and (b) lesson-specific dynamics, although SPIQ are influenced by both the student raters and time points (Feistauer & Richter, 2017; Wagner et al., 2016). Why different students perceive the same instructional quality differently and the potential differential relations of these differing perceptions to student outcome criteria remain poorly understood. By employing the experience sampling method, the present study demonstrates methodological and content-specific advances compared with cross-sectional, aggregated data with the goal of addressing individual student characteristics and lesson-specific dynamic associations of differential state SPIQ.

To do so, the present study used an experience sampling design that assessed state SPIQ of all students within a classroom attending the same lessons. Assessing perceptions of multiple individuals (i.e., students) in the same situations (i.e., lessons) on the same target (i.e., instructional quality) enables us to transfer insights from situation perception research (Rauthmann et al., 2015) to instructional quality research by conceptualizing perceptions of instructional quality in the classroom as situation perceptions. The SPIQ of multiple students, assessed within shared lessons, allow for bridging the gap between subjective perceptions and the objective instructional quality by differentiating idiosyncratic from consensual perceptions, where consensual, intersubjective perceptions approximate objective instructional quality (Rauthmann et al., 2015). This adds valuable insight into SPIQ research that discussed the subjectiveness of such perceptions (e.g., Wagner et al., 2016; Wisniewski et al., 2022). In addition, the present study sought to shed light on relevant associations of idiosyncratic and consensual components of state SPIQ. Specifically, we examined—for the first time—both the role of students' Big Five personality traits in relation to state SPIQ and the short-term relations between state SPIQ and students' perceived learning achievement (i.e., self-reported comprehension) in the same lessons. While SPIQ–achievement relations are well-documented using higher-level aggregated data (Praetorius et al., 2018), we focus on the level of single lessons to uncover short-term, intraindividual relations between state SPIQ and self-reported comprehension that we used as a subjective learning achievement indicator. In doing so, we differentiated idiosyncratic from consensual perceptions that may produce differential relation patterns. To this end, we used the popular and well-validated framework of Three Basic Dimensions (TBDs; teacher support, cognitive activation, and classroom management; Klieme et al., 2001) to operationalize SPIQ, and adapted it within an experience sampling study to assess lesson-specific perceptions. Ultimately, the present study thus provides insight into the occurrence and relevance of individual SPIQ components, and, in doing so, highlights the role of the *student* in SPIQ.

Differentiating Idiosyncratic and Consensual Components within SPIQ

The TBDs describe SPIQ parsimoniously within the three dimensions of *teacher support* (i.e., providing support during the learning process by, e.g., being sensitive to student needs and avoiding achievement pressure), *cognitive activation* (i.e., encouraging students' thinking and metacognition by, e.g., offering challenging tasks), and *classroom management* (i.e., efficiently using classroom time as learning time by, e.g., imposing clear rules and dealing with disruptions in class effectively; Klieme et al., 2001). The TBDs framework is widely used and empirically well-validated, although most of the validation works use higher-level (e.g., classroom- or school-level) aggregated, habitual (i.e., trait) perceptions (see Praetorius et al., 2018, for an overview), whose results cannot be transferred to the level of the individual (Molenaar, 2004; Murayama et al., 2017). Trait SPIQ that are only assessed at one point in time thus (a) lack the examination of within-student (i.e., intraindividual) variability (in contrast to between-student or interindividual variability) and (b) represent self-reported perceptions whose degree of subjectiveness cannot be estimated without further sources of information (e.g., teachers' self-perceptions; Wisniewski et al., 2022). To overcome (a), the experience sampling method can be implemented where individuals repeatedly report on their momentary (i.e., state) experiences in multiple situations in daily life (Bolger & Laurenceau, 2013). If these state experiences are assessed for multiple individuals regarding the same situation, one can also differentiate (b) idiosyncratic perceptions (that reflect subjectiveness) from overlapping, consensual perceptions (that approximate objectiveness; Rauthmann & Sherman, 2019). Drawing upon the same state SPIQ data as in the present study, Talić et al. (2022) conducted a first validation study on the state perceptions of the TBDs within an experience sampling design and showed a reliable differentiation of the TBDs from lesson to lesson and significant relations to crucial trait outcomes of student motivation and achievement. State SPIQ were reported to entail up to 61 % within-student variance in mathematics lessons across three weeks that remained virtually the same after statistically controlling for shared lesson perceptions of all students in a specific lesson. These shared perceptions were not investigated in more detail. To overcome this drawback, the present investigation uses insights from situation perception research (Rauthmann et al., 2015) for the first time to differentiate idiosyncratic from consensual, shared SPIQ that are usually confounded in the raw SPIQ perception, and to examine students' personality traits and perceived lesson-specific learning achievement in relation to state SPIQ components.[2] Specifically, based on the *experience* of instructional quality (i.e., students' perceptions as reflected in raw scores that are commonly used), one can differentiate between the *construal* of instructional quality (i.e., the unique and idiosyncratic personal reality of instructional quality irrespective of the actual instructional quality) and the *consensus* on instructional quality (i.e., the shared perceptions or social reality of all students within the classroom as indicator of the actual instructional quality; Rauthmann et al., 2015).[3] This distinction enables a comparison of the relative

[2] For more information on the data used in the present study, please see the Methods section below.

[3] Rauthmann et al. (2015) distinguish between liberal and conservative contact, where the former labels consensus as the shared perception among all perspectives of a situation, and the latter labels consensus

importance of subjective SPIQ versus actual instructional quality (as operationalized by the consensual SPIQ) when examining relations to outcome criteria. For instance, if we imagine the example of student A, who interprets (or construes) instructional behavior as cognitively engaging (whereas her classmates do not), her idiosyncratic, construed perception might be more decisive for her increase in lesson comprehension than is the classroom consensus on the degree of cognitive activation (as indicator of actual degree of cognitive activation). Thus, differentiating the three SPIQ components[4] of experience, construal, and consensus taps into the question of the importance of perceptions as subjective versus social realities, emphasizing either the idiosyncratic meaning that individuals give situations (i.e., subjectivist views) or the consensually shared meanings (i.e., objectivist views; Rauthmann & Sherman, 2019).

The Role of Students' Personality in SPIQ

The distinction between subjectivist and objectivist views implies that being exposed to the same situational stimuli (i.e., in our case, instructional quality in the classroom) does not mean that all individuals form the same psychological representation of that situation (Rauthmann & Sherman, 2019). To address the question of why different individuals form different perceptions based on the same situational stimuli, we examine personality traits as possible correlates of differences in information processing and subsequent ratings (Rauthmann & Sherman, 2019). For instance, it might be that student A, whose example we introduced in the previous paragraph, is more open-minded than her classmate student B, and therefore construes innovative teaching methods as more cognitively activating than student B, who prefers more conservative teaching methods. The Big Five personality traits open-mindedness, conscientious-

as a shared perception of only external perspectives. Throughout the remainder of this article, we use the term consensus in the sense of liberal contact.

[4] Throughout this article, we refer to the methodological distinction of SPIQ into experience, construal, and consensus as SPIQ *components*, and to the content-specific distinction (i.e., the framework of TBDs) into teacher support, cognitive activation, and classroom management as SPIQ *dimensions*.

ness, extraversion, agreeableness, and negative emotionality describe human personality in broad and robust domains (John, 2021; McCrae & Costa, 1987). Open-mindedness reflects imaginative, intellectually curious, and flexible characteristics. Conscientiousness describes well-organized, systematic, and disciplined individuals. Extraversion entails sociability, talkativeness, activity, and energy. Agreeableness pertains to the tendency of being cooperative, sympathetic, and trusting. Negative emotionality represents the tendency to experience stress, anxiety, and emotional volatility (Costa & McCrae, 1992). In the educational context, the Big Five traits were mostly examined with regard to academic achievement, where open-mindedness, conscientiousness, and agreeableness were significantly related to academic achievement with conscientiousness as the most important predictor even after controlling for intelligence (e.g., Franzen, Arens, Greiff, van der Westhuizen, et al., 2022; see also meta-analyses from Mammadov, 2022; Poropat, 2009). With regard to SPIQ, *teachers'* personality traits have been considered in some works to be relevant for SPIQ (e.g., Holzberger et al., 2013; Roloff et al., 2020; Toropova et al., 2019). Yet the role of *students'* personality traits in SPIQ has remained unclear, although substantial variance in SPIQ has been shown to be attributable to the student raters (Feistauer & Richter, 2017; Ruzek et al., 2022; Wagner et al., 2016). In a study on perceptions of online learning experiences, conscientiousness has been shown to predict positive evaluations positively, and negative evaluations negatively, while agreeableness and open-mindedness additionally predicted the value of online courses positively (Keller & Karau, 2013), providing some initial insights into the impact of personality traits on learners' perceptions. Yet relations between personality traits and perceptions of instructional quality remain poorly understood.

Lesson-Specific Relations between SPIQ and Learning Achievement

Revisiting the example of students A and B from the previous sections, their differential SPIQ might be differentially related to their learning achievements. If student A perceives a higher degree of cognitive activation than student B, it might be that student A experiences an increase in her learning whereas student B does not. In addition, this relation might be particularly strong for certain students. For instance, a higher degree of cognitive activation might be more strongly related to higher learning achievement for more open-minded students than for less open-minded students, because more open-minded students engage in the cognitively engaging thought process for a longer time than their less open-minded peers.

Of the three TBDs, particularly the dimensions of cognitive activation and classroom management have theoretical and empirical relations to achievement, whereas teacher support is more closely related to student motivation (Praetorius et al., 2018). Cognitive activation enhances student achievement by stimulating higher-order thinking, ultimately resulting in the construction of deep and flexible knowledge (Hardy et al., 2006). Classroom management is related to student achievement by enhancing the time effectively spent on tasks (Seidel & Shavelson, 2007). Empirically, relations to achievement were demonstrated for cognitive activation (Klieme et al., 2001), classroom management (Fauth et al., 2014; Scherer et al., 2016), yet also for teacher support (Fauth et al., 2014; for an overview see Praetorius et al., 2018). Thus, there is evidence for the TBDs' relevance with regard to student achievement. Yet this evidence is based on between-person research designs, where interindividual variance (between students, classes, schools, or countries) is used and individual student deviations from mean perceptions are neglected. To gain initial insight into short-term, near immediate SPIQ–achievement relations that consider within-student variance, we examine within-student relations between lesson-specific SPIQ and perceived learning achievement (as a subjective state achievement indicator) that might reveal new insights, aiding in fostering student achievement in individual lessons.

Further, there is some inconsistency such that postulated relations cannot always be confirmed (Praetorius et al., 2018). One possible reason for inconsistencies in SPIQ–achievement relations could be that these relations might be stronger or weaker for certain groups of students, yielding inconsistent or zero results at the group level (Renner et al., 2020). Therefore, the role of personality traits on the relationship between lesson-specific SPIQ and learning achievement is examined. Such interactive situation perception*trait effects have been found for state behavior (Breil et al., 2019), while other authors have not identified such effects (Abrahams et al., 2021). Thus, extant research on SPIQ–achievement relations is extended such that the within-student level is considered as well with regard to possible interactions with personality traits.

The Present Study

This experience sampling study addresses three distinct yet interrelated research questions that—to the best of our knowledge—have never been examined before. In doing so, we focus on state measures of SPIQ and perceived learning achievement and on trait measures of students' personality in German secondary school students in the domain of mathematics instruction, where the majority of research on the TBDs was conducted (Praetorius et al., 2018). First, students' personality traits are examined as predictors of state SPIQ. Second, short-term, within-student relations between state SPIQ and perceived lesson-specific learning achievement are examined. Third, personality traits are tested as moderators of this short-term, lesson-specific SPIQ–learning achievement relation. In doing this, we apply insights from situation perception research (Rauthmann & Sherman, 2019), enabled by our experience sampling design, to perceptions of instructional quality research with the goal of disentangling purely subjective from consensual components (approximating objectivity) that are confounded within SPIQ raw scores that are typically used in SPIQ research. Thus, the present study considers subjective and intersubjective (consensual) perceptions to gain some insight into different state SPIQ components' relevant differential relations. To provide data for comparing state/trait associations, we additionally provide intercorrelations between all variables at the trait level (i.e., trait SPIQ, personality traits, and math grade and reasoning ability as general achievement indicators) and conduct all analyses using trait variables. We utilize the well-established and parsimonious TBDs framework (Klieme et al., 2001) to assess state SPIQ and perceived learning achievement (i.e., self-reported lesson-specific comprehension; Niepel et al., 2022), that we used as an achievement indicator across three weeks of German secondary school students' daily life, as well as the popular, robust, and parsimonious Big Five framework (Costa & McCrae, 1992) to assess students' enduring personality traits. To control for possible confounding effects, we included students' gender, math grade, and reasoning ability as covariates. Prior research has shown students' gender to influence self-perceived math abilities (Niepel et al., 2019). Gender differences in personality traits have also been reported (Costa et al., 2001). Reasoning ability was related to personality traits (Sutin et al., 2022), while school grades were associated with SPIQ (Jaekel et al., 2021).

In all examined relations, we differentiated between the SPIQ components of *experience* (i.e., raw SPIQ scores provided by the students), *construal* (i.e., the purely idiosyncratic portion within the SPIQ that one respective student does not share with the classmates), and *consensus* (i.e., the intersubjective, overlapping classroom perception that approximates actual, objective instructional quality; Rauthmann et al., 2015) for each of the TBDs of teacher support, cognitive activation, and classroom management.

The present study thus complements and extends existing validation efforts of the TBDs that are mainly limited to the between-person level (Praetorius et al., 2018) by (a) considering intraindividual, within-student variation in addition to interindividual, between-student variation, (b) examining relations between state SPIQ on the one hand and personality traits and perceived lesson-specific learning achievement on the other hand, as well as interactions between state SPIQ and personality traits on perceived lesson-specific learning achievement, respectively, all while (c) disentangling idiosyncratic from consensual perceptions within the raw state SPIQ scores in all analyses. In doing so, we highlight the role of the individual student in SPIQ. To this end, we derived three research questions (RQs):

> *RQ1. How are students' personality traits associated with their state perceptions of instructional quality?*
>
> *RQ2. How are students' state perceptions of instructional quality associated with their perceived learning achievement in the same lesson?*
>
> *RQ3. Do students' personality traits moderate the relationship between state perceptions of instructional quality and perceived learning achievement in the same lesson?*

Methods

Procedure and Participants

In the present study, we used data from the larger intensive longitudinal "DynASCEL" (Dynamics of Academic Self-Concept in Everyday Life) project (Niepel et al., 2022)[5], where a three-week experience sampling study was conducted. Prior to and following the experience sampling phase, a pre- and post-assessment was carried out in paper-and-pencil format that obtained exhaustive student trait variables (e.g., personality traits, SPIQ traits). To address the present RQs, we focused on the experience sampling data on SPIQ and perceived learning achievement in every single mathematics lesson and used trait data from the pre-assessment. We drew on $N = 372$ German secondary school students attending the highest ability track (i.e., the German *Gymnasium*) who participated in the experience sampling part of the study of which $n = 308$ students attended the 9th and $n = 64$ students attended the 10th grade. Our sample consisted of 34.1 % boys (from $n = 301$ students with available gender information). Students reported a mean age of 15.3 years ($SD = 0.68$; range = 13.3-17.4 years; based on $n = 298$ students with available age information) and were nested in 18 classes in six schools from four German states (Rhineland-Palatinate, North Rhine-Westphalia, Baden-Wuerttemberg, and Mecklenburg-Western Pomerania). Based on these 18 classrooms, we assessed students' perceptions of 17 different math teachers' (58.8 % male) instructional quality (i.e., one teacher instructing math to two separate classes). On average, there were 20.6 students in a classroom ($SD = 4.65$; range = 10 – 27). Class constellations were stable across school grades.

In the experience sampling phase, students completed e-diaries on smartphones. Specifically, students responded to a short electronic questionnaire assessing their perceptions of this specific lesson on the application movisensXS (versions 1.3.0-1.3.4;

[5] Data from the project have been and will be used in other manuscripts on different research questions (e.g., Hausen et al., 2022). The intensive longitudinal data examined in the present study (i.e., mathematics state SPIQ and perceived lesson-specific learning achievement) have been used in previous studies addressing different research questions (Niepel et al., 2022; Talić et al., 2022).

movisens GmbH, Karlsruhe, Germany). We preprogrammed the smartphones such that the experience sampling prompts were triggered three minutes prior to the regular ending of every single mathematics lesson during the three-week period according to the class-specific timetables. The number of math lessons thus varied between classes per design (i.e., *M* = 10.11 math lessons; *SD* = 3.39; range = 3 – 16). In total, we obtained 2,681 valid responses (i.e., at least one out of nine items of interest answered; see Measures section below), representing a compliance rate of 70.81 %. Causes of missingness included absences from lessons (e.g., student illness), cancellation of classes, exams or similar events, and technical issues (e.g., empty smartphone batteries).

Students' participation in the study was voluntary. Single items and prompts were skippable. Written parental consent was obtained for participating students and the local ethics review panel of the University of Luxembourg as well as all involved federal education authorities approved of all procedures. This study was not preregistered.

Measures

State Measures

State SPIQ. In the three-week experience sampling phase, state SPIQ in mathematics instruction were assessed within the TBDs of teacher support, cognitive activation, and classroom management using the two-item state scales described by Talić et al. (2022). These were based on the PISA 2012 scales (Mang et al., 2018) and adapted for the use in intensive longitudinal designs. Example items are "*During this lesson, the teacher helped students with their learning"* (for teacher support), "*During this lesson, the teacher gave problems that required us to think for an extended time*" (for cognitive activation), and "*During this lesson, there was noise and disorder*" (for classroom management; negative indicator). Items were responded to on a scale ranging from 0 (*false*) to 5 (*true*) such that higher ratings represented higher perceived instructional quality. Talić et al. (2022) reported two-level reliability coefficients for the state SPIQ scales for the present data in mathematics classes for teacher support at ω = .84 and ω = .97, for cognitive activation at ω = .83 and ω = .97, and for classroom management at ω = .76 and ω = .94 within students and between students, respectively. Further, validity evidence concerning the factor structure and relations to trait SPIQ scales, school grades, and interest were provided, altogether suggesting the applicability of the state SPIQ scales in an experience sampling design (Talić et al., 2022). The German-language items can be found in (Talić et al., 2022).

Perceived Lesson-Specific Learning Achievement. Perceived lesson-specific learning achievement was assessed in each mathematics lesson during the three-week experience sampling phase (i.e., in the same situations as state SPIQ). Three items that were shown to be applicable and meaningful in an experience sampling design (Niepel et al., 2022) were used to assess perceived lesson-specific comprehension and learning progress. Niepel et al. (2022) derived the items from previous research that implemented similar items in e-diaries to assess perceived learning achievement (e.g., Peterson & Miller, 2004; Shernof et al., 2017). The item wordings were "*I was able to follow the last lesson well*", "*I understood a lot in the last lesson*", and "*I learned a lot in the last lesson*" and were responded to on a 6-point Likert scale ranging from 0 (*false*) to 5 (*true*) such that higher scores indicated higher perceived lesson-specific learning achievement. Niepel et al. (2022) reported two-level reliability coefficients of ω = .89 within students and ω = .95 between students.

Trait measures

All examined trait measures were assessed in the preassessment (i.e., prior to the three-week experience sampling phase).

Personality Traits. We assessed personality traits using the German version (Danner et al., 2019) of the Big Five Inventory 2 (Soto & John, 2017). The Big Five personality traits were assessed with 12 items each that showed internal consistencies at α = .84 (for open-mindedness), α = .87 (for conscientiousness), α = .86 (for extraversion), α = .81 (for agreeableness), and α = .88 (for negative emotionality) in a representative German sample (Danner et al., 2019).[6] Items were responded to on a five-

[6] Note that these five broad domain traits can be distinguished into three facets each (i.e., aesthetic sensitivity, intellectual curiosity, and creative imagination for open-mindedness, organization, productiveness, and responsibility for conscientiousness, sociability, assertiveness, and energy level for

point Likert scale ranging from 0 (*disagree completely*) to 4 (*agree completely*) such that higher scores indicate higher trait manifestations.

Trait SPIQ. Trait SPIQ in mathematics instruction were assessed within the TBDs of teacher support, cognitive activation, and classroom management. We used the original full scales implemented in PISA 2012 (Mang et al., 2018), consisting of five (for teacher support and classroom management, each) and nine (for cognitive activation) items that assess general (habitual) perceived instructional quality that is not tied to specific lessons (i.e., aggregated perceptions). Example items are "*The teacher helps students with their learning"* (for teacher support), "*The teacher gives problems that require us to think for an extended time*" (for cognitive activation), and "*There is noise and disorder*" (for classroom management; negative indicator). Items were responded to on a scale ranging from 0 (*never [in no lesson]*) to 5 (*always [in every lesson]*) such that higher ratings represented higher perceived instructional quality. Internal consistencies for the PISA German student sample range between α = .79 to α = .89 (Mang et al., 2018).

Report Card Mathematics Grade. We obtained students' self-reported mathematics grade from their most recent report card. Prior research has shown that self-reported school grades serve as reliable achievement indicators in German student samples that do not underlie systematic reporting biases (Dickhäuser & Plenter, 2005; Sparfeldt et al., 2008). School grades in Germany are assigned on a six-point Likert scale which we recoded such that higher scores represented better achievement, thus ranging from 1 (*insufficient*) to 6 (*very good*).

Reasoning Ability. Students' reasoning ability was assessed using the Intelligenz-Struktur-Test-Screening (IST-Screening; Liepmann et al., 2012), the short version of the well-established Intelligenz-Struktur-Test (IST; Amthauer, 1970;

extraversion, compassion, respectfulness, and trust for agreeableness, and anxiety, depression, and emotional volatility for negative emotionality). Due to the explorative nature of analyses and a magnitude of multiple comparisons, the present study does not examine relations of personality traits at the facet level. For interested readers, intercorrelations between personality traits at the facet level and all other examined variables are provided in Table S1 in the Online Supplementary Material (OSM).

Liepmann et al., 2007). We used version A of the available versions A and B of the test. The IST-Screening measures students' reasoning ability in the three task areas of verbal analogies, number sequences, and figural matrices using 20 items each. The internal consistency of the composite score encompassing all three task areas (i.e., across all 60 items) was reported to be α = .87 by Liepmann et al. (2012). In the present study, we used the composite raw score across all task areas as an indicator of general reasoning ability.

Statistical Analyses

For the statistical analyses, we followed recommendations by a recent experience sampling study that examined the three components experience, construal, and consensus in situation perceptions in an educational context (Abrahams et al., 2021). We conducted all analyses using the statistical software R (R Core Team, 2021). Two-level ω reliability coefficients were calculated with the MplusAutomation package for R (Hallquist & Wiley, 2018). For fitting linear mixed effects models, we used the lme4 package with the optimizer bobyqa to improve convergence (Bates et al., 2015). We used the confint() function to obtain bootstrapped 95 % confidence intervals. To obtain standardized parameters, we used the effectsize package (Ben-Shachar et al., 2020).

To address our research questions, we first disentangled the three SPIQ components experience, construal, and consensus (see also Abrahams et al., 2021; Rauthmann et al., 2015). SPIQ experience is reflected by the raw individual SPIQ scores (as commonly used in previous SPIQ research). SPIQ construal and consensus needed to be calculated based on the raw SPIQ scores. To do this, we first calculated the lesson-specific class means such that for each row in the dataset (i.e., for a specific student in a specific lesson), the respective lesson-specific student SPIQ mean was excluded. In other words, the individual SPIQ did not enter the class mean perception of instructional quality in the same lesson. By doing this, we ensured that the class mean entailed only variance from all other students to avoid an artificial overemphasis of student variance when relating lesson-specific student and class means. SPIQ construal was then obtained by extracting the standardized residual scores from regression analyses, where lesson-specific individual SPIQ (i.e., experience) were regressed on lesson-specific class-mean SPIQ. In other words, variance of the individual SPIQ experience that were not explained by class SPIQ was considered to be purely idiosyncratic (i.e., SPIQ construal). SPIQ consensus reflected consensual perceptions of *all* students within a class in a specific lesson. SPIQ consensus was obtained by extracting the factor scores from factor analyses on individual SPIQ experience and class SPIQ. In other words, variance that was shared across individual and class SPIQ was considered as overlapping (i.e., SPIQ consensus). This procedure was conducted for all three SPIQ dimensions (i.e., the three TBDs teacher support, cognitive activation, and classroom management).

The experience sampling produced data where measurement points (i.e., Level 1) were nested within students (i.e., Level 2) that were, in turn, nested within classes (i.e., Level 3). Clustering in 18 classes at Level 3 was controlled for by adding 17 dummy-coded class-based predictor variables in each model (Hox et al., 2018). To estimate the reliability of our implemented measures, we computed single-level (for traits) and two-level (i.e, within- and between-student, for states) McDonald's ω coefficients (Geldhof et al., 2014). To estimate dependency in the data due to repeated measurements within persons, we calculated intraclass correlation coefficients (ICCs) for the state measures (Aarts et al., 2014). Due to model convergence issues when implementing random slopes, we conducted random intercept models. Predictors at Level 1 were centered within students, while predictors at Level 2 were centered at the grand mean (Enders & Tofighi, 2007). To control for gender, math grade, and reasoning ability, we conducted two sets of models for each RQ that exclude or include these covariates, respectively. Results are presented for both model sets. We reported unstandardized fixed effects coefficients (*b*s) and their bootstrapped 95 % confidence intervals. To estimate the fixed effects' fit to the model, we calculated marginal multiple *R*s (R_m; Nakagawa & Schielzeth, 2013). We calculated standardized regression coefficients as a multilevel model effect size measure (Lorah, 2018). To interpret effect sizes, we draw on the guidelines recommended by Gignac and Szodorai (2016) with coefficients of .10 as relatively small, .20 as typical, and .30 as relatively large. To adjust for multiple testing, we followed the procedure implemented by Abrahams et al. (2021) and used the more conservative level of p < .001 to test for significance. Additionally, findings at the level of p < .05 are highlighted in the tables to inform the readers on any marginal associations due to the explorative nature of analyses, yet these findings are not discussed in the article. The exact model specifications are described at the beginning of the respective results section for enhanced clarity. Data cannot be made available because of data protection concerns. Readers interested in the data can contact the first author.

Results

Preliminary Analyses

Prior to addressing our research questions, we examined the descriptive statistics of our implemented measures (see Table 1). McDonald's ω coefficient was calculated a as single-level reliability estimate for trait measures and as a two-level reliability estimate for state measures. For state measures, within-student [between-student] ω coefficients ranged from $\omega_{\text{within}} = .76$ to $\omega_{\text{within}} = .84$ [$\omega_{\text{between}} = .94$ to $\omega_{\text{between}} = .97$] for SPIQ experience across the TBDs. Perceived lesson-specific learning achievement showed a reliability of $\omega_{\text{within}} = .89$ and $\omega_{\text{between}} = .95$. For trait measures, coefficients ranged between ω = .81 to ω = .93 for trait SPIQ and between ω = .85 to ω = .89 for personality traits.[7] ICC values ranged between ICC = .40 and ICC = .62 across the SPIQ components and perceived lesson-specific learning achievement, indicating a substantial amount of within-student variance in these constructs.[8]

Having disentangled the three SPIQ components experience, construal, and consensus for the TBDs for the first time, we preliminarily examined their intercorrelations at the within- and between-student level (see Table 2). Correlations at the between-student level were higher than at the within-student level. Here, we will focus on the within-student level, where in general, the three components showed close to perfect correlations to one another across dimensions (e.g., teacher support experience and

[7] Descriptive statistics for personality traits at the facet level can be found in Table S2 in the OSM.

[8] Please note that these analyses do not provide entirely new results. Drawing on DynASCEL data, Talić et al. (2022) reported ω and ICC coefficients for state SPIQ, and Niepel et al. (2022) reported ω and ICC coefficients for perceived lesson-specific learning achievement, yet the latter while drawing on a slightly different sample size. To provide all relevant information, we report these coefficients here anew.

teacher support construal), ranging between $r = .88$ to $r = 1$. Thus, the three components experience, construal, and consensus show an extensive overlap in all three state SPIQ dimensions. Teacher support components were moderately related to cognitive activation components (ranges of $r = .43$ to $r = .45$, all ps < .001), while classroom management was uncorrelated with either of the two.

Table 1

Descriptive Statistics

	M	*SD*	ω	ICC	ω_{within}	ω_{between}
State SPIQ Experience						
Teacher Support	3.11	1.37	-	.49	.84	.97
Cognitive Activation	3.06	1.29	-	.44	.83	.97
Classroom Management	3.41	1.37	-	.61	.76	.94
State SPIQ Construal						
Teacher Support	0.00	1.29	-	.47	-	-
Cognitive Activation	0.00	1.26	-	.43	-	-
Classroom Management	0.00	1.20	-	.53	-	-
State SPIQ Consensus						
Teacher Support	0.00	0.94	-	.49	-	-
Cognitive Activation	0.00	0.82	-	.44	-	-
Classroom Management	0.00	0.97	-	.62	-	-
Lesson-Specific Achievement						
Perceived Learning Achievement	3.46	1.17	-	.40	.89	.95
SPIQ Traits						
Teacher Support	2.74	1.32	.93			
Cognitive Activation	2.94	0.72	.81			
Classroom Management	3.25	1.15	.90			
Personality Traits						
Open-Mindedness	2.24	0.63	.85	-	-	-
Conscientiousness	2.36	0.60	.89	-	-	-
Extraversion	2.37	0.61	.89	-	-	-
Agreeableness	2.65	0.57	.82	-	-	-
Negative Emotionality	1.70	0.62	.89	-	-	-
Covariates						
Math Grade	4.36	1.09	-	-	-	-
Reasoning Ability	43.52	5.70	-	-	-	-

Note. Response formats: SPIQ experience [0, 5]; perceived lesson-specific learning achievement [0; 5]; SPIQ traits [0, 5]; personality traits [0, 4]; math grade [1, 6]; reasoning ability [0, 60].

Table 2

Correlations between SPIQ Components within the TBDs and Perceived Learning Achievement

	SPIQ Experience			SPIQ Construal			SPIQ Consensus			Perceived Learning Achievement
	Teacher Support	Cognitive Activation	Classroom Management	Teacher Support	Cognitive Activation	Classroom Management	Teacher Support	Cognitive Activation	Classroom Management	
SPIQ Experience										
Teacher Support	—	**.72**	**.16**	**.94**	**.67**	**.14**	**1**	**.72**	**.16**	**.62**
Cognitive Activation	**.44**	—	.09	**.68**	**.97**	.05	**.72**	**1**	.09	**.47**
Classroom Management	.01	-.02	—	**.13**	.06	**.86**	**.16**	.09	**1**	**.23**
SPIQ Construal										
Teacher Support	**.95**	**.43**	.02	—	**.70**	**.15**	**.94**	**.68**	**.13**	**.59**
Cognitive Activation	**.43**	**.97**	-.02	**.45**	—	.05	**.67**	**.98**	.06	**.44**
Classroom Management	.03	-.01	**.92**	.03	-.01	—	**.14**	.05	**.81**	**.19**
SPIQ Consensus										
Teacher Support	**1**	**.44**	.01	**.95**	**.43**	.03	—	**.72**	**.16**	**.62**
Cognitive Activation	**.44**	**1**	-.02	**.43**	**.97**	-.01	**.44**	—	.09	**.47**
Classroom Management	.00	-.03	**1**	.01	-.02	**.88**	.00	-.03	—	**.23**
Perceived Learning Achievement	**.49**	**.27**	**.07**	**.45**	**.27**	**.07**	**.49**	**.27**	**.07**	—

Note. Correlations below the diagonal represent within-student correlations, and correlations above the diagonal represent between-student correlations. Correlation coefficients printed in **bold** are significant at $p < .05$, and correlation coefficients printed in **bold and gray shading** are significant at $p < .001$.

To gain first insights into the relations between students' personalities and SPIQ, we initially calculated correlations between personality traits and SPIQ traits (i.e., habitual SPIQ not tied to specific lessons; see Table S1 in the OSM). Teacher support was mostly unrelated to the Big Five personality traits with the exception of negative emotionality (r = -.13, p < .05). Cognitive activation showed positive relations to open-mindedness, conscientiousness, extraversion, and agreeableness (ranges of r = .12 to r = .21, ps < .05). Classroom management showed positive relations to conscientiousness and extraversion (r = .13 and r = .15, ps < .05).[1] Thus, trait SPIQ show some relations to personality traits, with the dimension of cognitive activation displaying the most relations as compared to the other two SPIQ dimensions. Across all dimensions, relations to personality traits were positive except for relations with negative emotionality, which were negative in direction.

Students' Personality Traits as Predictors of State SPIQ (RQ 1)

Before we addressed RQ1, we examined the correlations between personality traits and state SPIQ experience, construal, and consensus for the TBDs (see Table 3). First, we noted that SPIQ components' relations to personality traits were similar across components within dimensions (e.g., similar relations between teacher support experience and open-mindedness and teacher support construal and open-mindedness). Second, we detected some differences to the correlations between personality traits and SPIQ traits (see Preliminary Analyses). The trait correlations revealed substantial relations of teacher support to negative emotionality only, of cognitive activation to each personality trait except for negative emotionality, and of classroom management to conscientiousness and extraversion only. In comparison, relations of state SPIQ components revealed that all teacher support components were only related to agreeableness (mean r = .19, p < .001) and negative emotionality (mean r = -.23, p < .001). Cognitive

[1] To test the robustness of these results, we additionally conducted correlations of SPIQ traits using only the two corresponding items from the longer trait scales that were implemented in the state scales. The result pattern was almost identical to that of the long trait scales with the one exception that teacher support was completely unrelated to personality traits.

activation components were related to each personality trait except for extraversion (ranges of r = .10 to r = .16 for open-mindedness, conscientiousness, and agreeableness, and ranges of r = -.19 to r = -.16 for negative emotionality, ps < .05). Classroom management components of experience and consensus were related to each personality trait except extraversion, while classroom management construal was only related to agreeableness and negative emotionality (ranges of r = .10 to r = .19 for open-mindedness, conscientiousness, agreeableness, and ranges of r = -.15 to r = -.10 for negative emotionality, ps < .05). Thus, with regard to state SPIQ components, agreeableness and negative emotionality seemed to be the most crucial personality traits, showing the most relations to SPIQ components, whereas extraversion was unrelated to all state SPIQ component.

To address RQ1, we conducted linear mixed effect models with 17 dummy-coded variables controlling for classroom membership (not displayed in the tables for the sake of brevity), personality traits (and covariates) as simultaneous predictors, and SPIQ components experience, construal, and consensus for the three dimensions teacher support, cognitive activation, and classroom management as outcome variables. Results are presented in Table 4. For all teacher support components, negative emotionality was the only predictor that was significant at p < .001 in the model without covariates (mean b = -0.36) with a mean effect size of β = -.19. In other words, for every unit increase in negative emotionality, an average of 0.36 decrease in experienced, construed, and consensual teacher support is expected. However, after including the covariates gender, math grade, and reasoning ability, this effect no longer reached statistical significance, suggesting a confounding of negative emotionality with (some of) the covariates. Instead, in this model, agreeableness significantly predicted all three teacher support components positively (mean b = 0.34, mean β = .17, p < .001). The models predicting experience, construal, and consensus of cognitive activation and classroom management displayed no statistically significant prediction by personality traits at p < .001. Across models with teacher support as the outcome variable, the average model fit was R_m = 0.38, whereas the models with cognitive activation and classroom management showed an average model fit of R_m = 0.35 although the latter reveal no significant fixed effect. It is important to note that 17 dummy-coded classroom predictor variables partly produced significant fixed effects that inflated estimations for the R_m values.

Table 3

Correlations between Covariates, Personality Traits, SPIQ Components and Perceived Learning Achievement

	SPIQ Experience			SPIQ Construal			SPIQ Consensus			Perceived Learning Achievement
	Teacher Support	Cognitive Activation	Classroom Management	Teacher Support	Cognitive Activation	Classroom Management	Teacher Support	Cognitive Activation	Classroom Management	
Gender	**-.15**	-.09	.03	**-.19**	**-.12**	.00	**-.15**	-.09	.03	**-.23**
Math Grade	**.24**	**.22**	**.11**	**.20**	**.18**	.04	**.23**	**.22**	**.11**	**.40**
Reasoning Ability	.08	.06	.07	**.10**	.05	.04	.08	.06	.07	**.27**
Open-Mindedness	.02	**.10**	**.10**	.02	**.11**	.04	.02	**.10**	**.11**	**.15**
Conscientiousness	.09	**.11**	**.16**	.07	**.11**	.10	.09	**.11**	**.16**	**.15**
Extraversion	.09	.05	.10	.07	.04	.09	.09	.05	.10	**.11**
Agreeableness	**.18**	**.13**	**.18**	**.21**	**.16**	**.15**	**.19**	**.13**	**.19**	**.18**
Negative Emotionality	**-.22**	**-.16**	**-.11**	**-.25**	**-.19**	**-.15**	**-.22**	**-.16**	**-.10**	**-.31**

Note. Gender is coded with 0 = male; 1 = female.

Correlation coefficients printed in **bold** are significant at $p < .05$, and correlation coefficients printed in **bold and gray shading** are significant at $p < .001$.

Table 4

Personality Traits as Predictors of SPIQ Components (RQ1)

	Experience				Construal				Consensus			
Predictors	*b* [95% CI]	β [95% CI]	*t*	R_m	*b* [95% CI]	β [95% CI]	*t*	R_m	*b* [95% CI]	β [95% CI]	*t*	R_m
Outcome: Teacher Support (Model without Covariates)												
Open-Mindedness	-0.09 [-0.27, 0.10]	-.04 [-.12, .04]	-0.93	0.40	-0.09 [-0.28, 0.09]	-.05 [-.14, .05]	-0.96	0.27	-0.06 [-0.18, 0.06]	-.04 [-.12, .04]	-0.93	0.40
Conscientiousness	-0.11 [-0.29, 0.08]	-.05 [-.13, .04]	-1.08		-0.11 [-0.31, 0.07]	-.05 [-.15, .04]	-1.08		-0.07 [-0.19, 0.06]	-.05 [-.13, .04]	-1.08	
Extraversion	-0.02 [-0.19, 0.17]	-.01 [-.09, .07]	-0.21		-0.03 [-0.22, 0.16]	-.01 [-.10, .08]	-0.27		-0.01 [-0.13, 0.12]	-.01 [-.09, .07]	-0.22	
Agreeableness	**0.28** [0.05, 0.47]	**.12** [.03, .21]	2.71		**0.30** [0.08, 0.52]	**.14** [.04, .23]	2.79		**0.19** [0.04, 0.33]	**.12** [.03, .21]	2.72	
Negative Emotionality	**-0.39** [-0.57, -0.19]	**-.18** [-.27, -.09]	-4.03		**-0.41** [-0.63, -0.21]	**-.20** [-.30, -.10]	-4.08		**-0.27** [-0.39, -0.13]	**-.18** [-.27, -.09]	-4.03	
Outcome: Teacher Support (Model with Covariates)												
Open-Mindedness	-0.06 [-0.24, 0.15]	-.03 [-.11, .06]	-0.67	0.45	-0.07 [-0.28, 0.12]	-.03 [-.13, .06]	-0.69	0.34	-0.04 [-0.17, 0.09]	-.03 [-.11, .06]	-0.67	0.44
Conscientiousness	-0.08 [-0.26, 0.11]	-.04 [-.13, .05]	-0.80		-0.09 [-0.31, 0.12]	-.04 [-.14, .06]	-0.84		-0.06 [-0.20, 0.10]	-.04 [-.13, .05]	-0.80	
Extraversion	0.09 [-0.09, 0.26]	.04 [-.04, .13]	0.99		0.09 [-0.09, 0.27]	.04 [-.05, .14]	0.90		0.06 [-0.08, 0.20]	.04 [-.04, .13]	0.99	
Agreeableness	**0.37** [0.16, 0.59]	**.16** [.07, .25]	3.49		**0.40** [0.20, 0.62]	**.18** [.08, .29]	3.58		**0.26** [0.13, 0.40]	**.16** [.07, .26]	3.49	
Negative emotionality	-0.21 [-0.43, 0.00]	-.10 [-.20, .00]	-1.92		**-0.23** [-0.47, -0.01]	**-.11** [-.22, .00]	-1.99		-0.14 [-0.30, 0.00]	-.10 [-.20, .00]	-1.92	
Gender	**-0.42** [-0.67, -0.16]	**-.15** [-.25, -.06]	-3.26		**-0.44** [-0.70, -0.17]	**-.17** [-.27, -.07]	-3.23		**-0.29** [-0.48, -0.12]	**-.15** [-.25, -.06]	-3.25	
Math Grade	**0.16** [0.04, 0.29]	**.13** [.03, .22]	2.61		**0.17** [0.05, 0.29]	**.15** [.04, .25]	2.70		**0.11** [0.03, 0.20]	**.13** [.03, .22]	2.61	
Reasoning Ability	0.00 [-0.02, 0.03]	.02 [-.08, .11]	0.36		0.00 [-0.02, 0.03]	.01 [-.09, .12]	0.23		0.00 [-0.01, 0.02]	.02 [-.08, .11]	0.36	
Outcome: Cognitive Activation (Model without Covariates)												
Open-Mindedness	0.11 [-0.07, 0.29]	.05 [-.03, .14]	1.19	0.30	0.11 [-0.07, 0.31]	.06 [-.03, .15]	1.21	0.20	0.07 [-0.05, 0.17]	.05 [-.03, .14]	1.19	0.29
Conscientiousness	-0.06 [-0.23, 0.15]	-.03 [-.12, .06]	-0.64		-0.05 [-0.25, 0.14]	-.03 [-.12, .07]	-0.54		-0.04 [-0.15, 0.09]	-.03 [-.12, .06]	-0.62	
Extraversion	-0.06 [-0.22, 0.12]	-.03 [-.11, .05]	-0.67		-0.06 [-0.23, 0.12]	-.03 [-.12, .06]	-0.71		-0.04 [-0.15, 0.07]	-.03 [-.11, .06]	-0.66	
Agreeableness	0.17 [-0.03, 0.37]	.08 [-.01, .17]	1.69		0.18 [-0.02, 0.38]	.08 [-.01, .18]	1.73		0.10 [-0.02, 0.23]	.08 [-.01, .17]	1.68	
Negative Emotionality	**-0.26** [-0.45, -0.07]	**-.13** [-.22, -.04]	-2.79		**-0.27** [-0.45, -0.09]	**-.14** [-.23, -.04]	-2.86		**-0.16** [-0.28, -0.05]	**-.13** [-.22, -.04]	-2.79	
Outcome: Cognitive Activation (Model with Covariates)												
Open-Mindedness	0.11 [-0.08, 0.30]	.05 [-.04, .14]	1.13	0.32	0.11 [-0.10, 0.32]	-.03 [-.13, .06]	1.14	0.23	0.07 [-0.06, 0.19]	-.03 [-.11, .06]	1.13	0.32
Conscientiousness	-0.04 [-0.23, 0.16]	-.02 [-.11, .08]	-0.35		-0.03 [-0.23, 0.22]	-.04 [-.14, .06]	-0.28		-0.02 [-0.14, 0.10]	-.04 [-.13, .05]	-0.32	
Extraversion	0.01 [-0.17, 0.20]	.00 [-.09, .09]	0.09		0.01 [-0.18, 0.19]	.04 [-.05, .14]	0.01		0.01 [-0.12, 0.11]	.04 [-.04, .13]	0.10	
Agreeableness	0.20 [-0.02, 0.41]	.09 [-.01, .19]	1.83		0.21 [-0.02, 0.41]	.18 [.08, .29]	1.91		0.12 [-0.02, 0.25]	.16 [.07, .26]	1.83	
Negative Emotionality	-0.16 [-0.38, 0.07]	-.08 [-.18, .03]	-1.45		-0.16 [-0.36, 0.07]	-.11 [-.22, .00]	-1.44		-0.10 [-0.24, 0.04]	-.10 [-.20, .00]	-1.44	
Gender	-0.25 [-0.51, -0.02]	-.10 [-.20, .00]	-1.94		**-0.27** [-0.54, -0.01]	**-.17** [-.27, -.07]	-2.06		-0.16 [-0.32, 0.00]	-.15 [-.25, -.06]	-1.97	

	Experience				Construal				Consensus			
Predictors	*b* [95% CI]	β [95% CI]	*t*	R_m	*b* [95% CI]	β [95% CI]	*t*	R_m	*b* [95% CI]	β [95% CI]	*t*	R_m
Math Grade	0.10 [-0.01, 0.22]	.09 [-.01, .19]	1.71		0.11 [-0.02, 0.24]	.15 [.04, .25]	1.80		0.07 [-0.01, 0.15]	.13 [.03, .22]	1.70	
Reasoning Ability	0.00 [-0.03, 0.02]	-.02 [-.11, .08]	-0.30		0.00 [-0.03, 0.02]	.01 [-.09, .12]	-0.39		0.00 [-0.02, 0.01]	.02 [-.08, .11]	-0.31	
					Outcome: Classroom Management (Model without Covariates)							
Open-Mindedness	0.02 [-0.16, 0.20]	.01 [-.08, .09]	0.16	0.50	0.00 [-0.20, 0.17]	.00 [-.10, .10]	0.03	0.19	0.01 [-0.13, 0.13]	.01 [-.07, .09]	0.19	0.55
Conscientiousness	0.06 [-0.14, 0.23]	.03 [-.06, .12]	0.64		0.07 [-0.11, 0.27]	.04 [-.07, .14]	0.69		0.04 [-0.08, 0.17]	.03 [-.06, .11]	0.64	
Extraversion	0.03 [-0.14, 0.22]	.02 [-.07, .10]	0.38		0.04 [-0.12, 0.22]	.02 [-.07, .12]	0.44		0.02 [-0.10, 0.15]	.02 [-.06, .09]	0.38	
Agreeableness	0.11 [-0.07, 0.31]	.05 [-.04, .14]	1.12		0.11 [-0.10, 0.31]	.06 [-.05, .16]	1.08		0.07 [-0.07, 0.20]	.05 [-.04, .13]	1.12	
Negative Emotionality	-0.10 [-0.29, 0.07]	-.05 [-.14, .04]	-1.10		-0.12 [-0.31, 0.10]	-.06 [-.16, .04]	-1.20		-0.07 [-0.19, 0.05]	-.05 [-.13, .04]	-1.08	
					Outcome: Classroom Management (Model with Covariates)							
Open-Mindedness	0.07 [-0.12, 0.26]	.03 [-.06, .12]	0.68	0.50	0.06 [-0.15, 0.24]	.03 [-.07, .13]	0.55	0.19	0.05 [-0.08, 0.18]	.03 [-.05, .12]	0.71	0.55
Conscientiousness	0.08 [-0.12, 0.32]	.04 [-.06, .13]	0.77		0.09 [-0.13, 0.29]	.05 [-.06, .16]	0.87		0.05 [-0.07, 0.20]	.04 [-.06, .13]	0.78	
Extraversion	0.03 [-0.18, 0.20]	.01 [-.08, .10]	0.31		0.04 [-0.15, 0.24]	.02 [-.08, .13]	0.45		0.02 [-0.11, 0.13]	.01 [-.07, .10]	0.30	
Agreeableness	0.08 [-0.14, 0.30]	.04 [-.06, .13]	0.76		0.08 [-0.14, 0.31]	.04 [-.07, .15]	0.75		0.06 [-0.07, 0.19]	.04 [-.06, .13]	0.76	
Negative Emotionality	-0.09 [-0.31, 0.14]	-.04 [-.15, .06]	-0.83		-0.10 [-0.32, 0.13]	-.05 [-.17, .07]	-0.85		-0.06 [-0.22, 0.08]	-.04 [-.14, .06]	-0.81	
Gender	-0.05 [-0.32, 0.22]	-.02 [-.12, .08]	-0.36		-0.07 [-0.36, 0.20]	-.03 [-.14, .09]	-0.49		-0.03 [-0.20, 0.14]	-.02 [-.11, .08]	-0.37	
Math Grade	-0.03 [-0.16, 0.09]	-.02 [-.12, .08]	-0.42		-0.03 [-0.16, 0.09]	-.03 [-.15, .09]	-0.48		-0.02 [-0.10, 0.05]	-.02 [-.12, .08]	-0.41	
Reasoning Ability	0.00 [-0.02, 0.02]	.00 [-.10, .10]	-0.01		0.00 [-0.02, 0.03]	.01 [-.11, .12]	0.09		0.00 [-0.02, 0.01]	.00 [-.10, .09]	-0.06	

Note. Each model additionally contains 17 dummy-coded predictor variables indicating class membership to control for clustered data at Level 3. For brevity, these fixed effects are not displayed in the table. *b* = unstandardized multilevel regression coefficient; β = standardized multilevel regression coefficient; R_m = marginal multiple R for generalized linear mixed effect models. Gender is coded with 0 = male; 1 = female. Personality traits and covariates were centered at the grand mean. Regression coefficients printed in **bold** are significant at $p < .05$, and regression coefficients printed in **bold and gray shading** are significant at $p < .001$

State SPIQ as Predictors of Perceived Lesson-Specific Learning Achievement (RQ 2)

Before we addressed RQ2, we calculated within- and between-student correlations between the three SPIQ components and perceived lesson-specific learning achievement (see Table 2). Again, correlations at the between-student level were higher than at the within-student level. We focus on the within-student level, where relations to perceived lesson-specific learning achievement were descriptively strongest for teacher support (mean $r = .48$) and cognitive activation (mean $r = .27$) and lowest for classroom management (mean $r = .07$, all ps < .001), with an almost identical pattern across the components of experience, construal, and consensus.

To address RQ2, we conducted linear mixed effect models with 17 dummy-coded variables controlling for classroom membership, three SPIQ dimensions per component (and covariates) as simultaneous predictors, and perceived lesson-specific learning achievement as the outcome variable. Results can be found in Table 5. All components of all SPIQ dimensions significantly and positively predicted perceived lesson-specific learning achievement at $p < .001$ in the models without covariates. There were clear differences in effect sizes. Teacher support showed the largest effect (mean $b = 0.48$, mean $\beta = .32$). Effects of cognitive activation and classroom management were of similar extent (mean $b = 0.08$, mean $\beta = .05$). In the models with covariates, these results remained virtually unchanged. Further, gender predicted perceived lesson-specific learning achievement negatively ($b = -0.36$, $\beta = -.16$), suggesting a negative relation for female students. The math grade predicted perceived lesson-specific learning achievement positively ($b = 0.27$, $\beta = .24$). Reasoning ability did not show any incremental effect on perceived lesson-specific learning achievement above and beyond gender and the math grade. Including the covariates improved the model fit (mean $R_m = 0.44$ without covariates and mean $R_m = 0.55$ with covariates). Concluding, experienced, construed, and consensual teacher support were most decisive for perceived learning achievement in the same lesson.

Table 5

SPIQ Components as Predictors of Perceived Learning Achievement (RQ2)

Predictors	Outcome: Perceived Learning Achievement							
	Model without Covariates				Model with Covariates			
	b [95% CI]	β [95% CI]	*t*	R_m	*b* [95% CI]	β [95% CI]	*t*	R_m
SPIQ Experience								
Teacher Support	**0.42** [0.39, 0.46]	**.33** [.30, .36]	22.65	.45	**0.46** [0.41, 0.50]	**.34** [.31, .37]	21.96	.56
Cognitive Activation	**0.07** [0.03, 0.10]	**.05** [.02, .08]	3.49		**0.07** [0.02, 0.11]	**.05** [.02, .08]	3.30	
Classroom Management	**0.07** [0.04, 0.11]	**.05** [.02, .08]	3.85		**0.09** [0.05, 0.14]	**.06** [.03, .09]	4.45	
Gender	--	--	--		**-0.39** [-0.55, -0.22]	**-.16** [-.24, -.09]	-4.28	
Math Grade	--	--	--		**0.27** [0.18, 0.36]	**.24** [.16, .32]	5.68	
Reasoning Ability	--	--	--		0.01 [-0.01, 0.03]	.07 [-.02, .15]	1.56	
SPIQ Construal								
Teacher Support	**0.40** [0.36, 0.44]	**.30** [.27, .32]	19.77	.43	**0.43** [0.38, 0.47]	**.30** [.27, .33]	18.90	.54
Cognitive Activation	**0.08** [0.04, 0.12]	**.06** [.03, .09]	4.04		**0.08** [0.04, 0.12]	**.06** [.03, .09]	3.72	
Classroom Management	**0.07** [0.03, 0.11]	**.05** [.02, .07]	3.40		**0.08** [0.03, 0.12]	**.05** [.02, .08]	3.52	
Gender	--	--	--		**-0.39** [-0.54, -0.19]	**-.16** [-.24, -.09]	-4.28	
Math grade	--	--	--		**0.27** [0.19, 0.36]	**.24** [.16, .32]	5.67	
Reasoning Ability	--	--	--		0.01 [0.00, 0.03]	.07 [-.02, .15]	1.56	
SPIQ Consensus								
Teacher Support	**0.62** [0.57, 0.67]	**.33** [.30, .36]	22.65	.45	**0.67** [0.60, 0.73]	**.34** [.31, .37]	21.96	.56
Cognitive Activation	**0.10** [0.04, 0.16]	**.05** [.02, .08]	3.49		**0.11** [0.04, 0.17]	**.05** [.02, .08]	3.30	
Classroom Management	**0.11** [0.05, 0.16]	**.05** [.02, .08]	3.84		**0.14** [0.08, 0.20]	**.06** [.04, .09]	4.49	
Gender	--	--	--		**-0.39** [-0.56, -0.20]	**-.16** [-.24, -.09]	-4.29	
Math grade	--	--	--		**0.27** [0.18, 0.35]	**.24** [.16, .32]	5.68	
Reasoning Ability	--	--	--		0.01 [0.00, 0.03]	.07 [-.02, .15]	1.55	

Note. Each model additionally contains 17 dummy-coded predictor variables indicating class membership to control for clustered data at Level 3. For brevity, these fixed effects are not displayed in the Table. *b* = unstandardized multilevel regression coefficient; β = standardized multilevel regression coefficient; R_m = marginal multiple R for generalized linear mixed effect models. Gender is coded with 0 = male; 1 = female. SPIQ components were centered within students. Covariates were centered at the grand mean and added as predictors of random intercepts. Regression coefficients printed in **bold** are significant at $p < .05$, and regression coefficients printed in **bold and gray shading** are significant at $p < .001$

Students' Personality Traits as Moderators of the Association between State SPIQ and Perceived Lesson-Specific Learning Achievement (RQ 3)

Finally, addressing RQ3, we examined personality traits as possible moderators of the link between SPIQ components and perceived lesson-specific learning achievement. We ran a set of preliminary models where we included all possible interaction terms between SPIQ components and personality traits, of which we only used those interaction terms that were significant at $p < .05$ for our final models (for a similar procedure, see Abrahams et al., 2021; Sherman et al., 2015). The elevated alpha level was chosen here for the preliminary models to facilitate the detection of interaction effects that are usually very small (Rauthmann, 2021) and thus might aid in generating new hypotheses. For interpreting moderation effects in the final models, however, we use the criterion of $p < .001$.

In the final models we included 17 dummy-coded variables controlling for classroom membership, three SPIQ dimensions per component (e.g., experience of teacher support, cognitive activation, and classroom management), the respective personality traits and interaction terms between state SPIQ components and personality traits that were significant predictors in the preliminary models (and covariates) as simultaneous predictors, and perceived lesson-specific learning achievement as the outcome variable. Results are displayed in Table 6. Relations between state SPIQ components and perceived lesson-specific learning achievement and between covariates and perceived lesson-specific learning achievement remained virtually the same as those reported for RQ2 and only showed marginal differences in effect sizes. Across all components, the only state SPIQ dimension that interacted with personality traits with regard to perceived lesson-specific learning achievement was teacher support, and the only personality traits that interacted with components of teacher support were agreeableness and negative emotionality. With regard to the preliminary models, of the 45 (3 dimensions per state SPIQ component * 5 personality traits * 3 state SPIQ components) possible interactions, only five interactions reached statistical significance at the $p < .05$ level and were included in the final models. In our test of the final models, only one of those interactions was significant at the $p < .001$ level (see Table 6). Specifically, agreeableness moderated the relation between construed teacher support and perceived lesson-specific learning achievement in the models with and without covariates (*b*s = -0.13, β = -.06 and β = -.05, respectively). In other words, the less agreeable a student is, the stronger is the positive association between construed teacher support and perceived lesson-specific learning achievement. Concerning direct effects, agreeableness showed a positive relation to perceived lesson-specific learning achievement in the model using SPIQ construal and covariates (*b* = 0.31, β = .15, *p* < .001), while negative emotionality showed negative relations to perceived lesson-specific learning achievement in the models using SPIQ experience and consensus (mean *b* = -0.36, mean β = .19, *p* < .001). The average model fit was R_m = 0.52 for the models without covariates and R_m = 0.59 for the models with covariates.

Table 6

Personality Traits as Moderators of the Association between SPIQ Experience, Construal, Consensus and Perceived Learning Achievement (RQ 3)

Predictors	Outcome: Perceived Learning Achievement							
	Model without Covariates				Model with Covariates			
	b [95% CI]	β [95% CI]	*t*	R_m	*b* [95% CI]	β [95% CI]	*t*	R_m
SPIQ Experience								
Teacher Support	**0.45** [0.41, 0.49]	**.34** [.31, .37]	22.53	.54	**0.46** [0.42, 0.50]	**.34** [.31, .37]	22.07	.60
Cognitive Activation	**0.07** [0.03, 0.11]	**.05** [.02, .08]	3.29		**0.06** [0.01, 0.10]	**.04** [.01, .07]	2.73	
Classroom Management	**0.08** [0.03, 0.12]	**.05** [.02, .08]	3.58		**0.08** [0.04, 0.12]	**.05** [.03, .08]	3.85	
Teacher Support x Agreeableness	**-0.08** [-0.14, -0.02]	**-.03** [-.06, -.01]	-2.74		**-0.08** [-0.14, -0.03]	**-.04** [-.06, -.01]	-2.87	
Teacher Support x Negative Emotionality	**0.06** [0.01, 0.12]	**.03** [.00, .05]	2.07		0.06 [0.00, 0.11]	.03 [.00, .05]	1.90	
Agreeableness	0.11 [-0.04, 0.27]	.05 [-.02, .13]	1.34		**0.21** [0.04, 0.37]	**.10** [.02, .18]	2.55	
Negative Emotionality	**-0.45** [-0.59, -0.32]	**-.24** [-.31, -.16]	-6.10		**-0.27** [-0.44, -0.12]	**-.14** [-.22, -.06]	-3.46	
Gender	--	--	--		**-0.37** [-0.56, -0.16]	**-.15** [-.23, -.07]	-3.80	
Math Grade	--	--	--		**0.23** [0.14, 0.32]	**.21** [.13, .29]	4.95	
Reasoning Ability	--	--	--		**0.02** [0.00, 0.04]	**.09** [.01, .17]	2.15	
SPIQ Construal								
Teacher Support	**0.43** [0.37, 0.46]	**.30** [.27, .33]	19.54	.47	**0.43** [0.39, 0.48]	**.30** [.27, .33]	19.05	.57
Cognitive Activation	**0.08** [0.05, 0.13]	**.06** [.03, .09]	4.01		**0.07** [0.03, 0.12]	**.05** [.02, .09]	3.43	
Classroom Management	**0.07** [0.02, 0.11]	**.04** [.02, .07]	3.04		**0.07** [0.02, 0.11]	**.04** [.01, .07]	3.00	
Teacher Support x Agreeableness	**-0.13** [-0.19, -0.07]	**-.05** [-.08, -.03]	-4.34		**-0.13** [-0.19, -0.07]	**-.06** [-.08, -.03]	-4.43	
Agreeableness	**0.25** [0.10, 0.40]	**.12** [.04, .20]	3.07		**0.31** [0.16, 0.46]	**.15** [.08, .23]	3.94	
Gender	--	--	--		**-0.48** [-0.67, -0.29]	**-.20** [-.27, -.12]	-5.16	
Math Grade	--	--	--		**0.25** [0.15, 0.34]	**.23** [.14, .31]	5.32	
Reasoning Ability	--	--	--		0.02 [0.00, 0.03]	.08 [-.01, .16]	1.82	
SPIQ Consensus								

Teacher Support	**0.66** [0.60, 0.72]	**.34** [.31, .37]	22.54	.54	**0.67** [0.61, 0.73]	**.34** [.31, .37]	22.07	.60
Cognitive Activation	**0.11** [0.04, 0.17]	**.05** [.02, .08]	3.29		**0.09** [0.02, 0.16]	**.04** [.01, .07]	2.73	
Classroom Management	**0.11** [0.05, 0.17]	**.05** [.02, .08]	3.63		**0.12** [0.05, 0.19]	**.06** [.03, .08]	3.92	
Teacher Support x Agreeableness	**-0.11** [-0.21, -0.03]	**-.03** [-.06, -.01]	-2.75		**-0.12** [-0.20, -0.04]	**-.04** [-.06, -.01]	-2.87	
Teacher Support x Negative emotionality	**0.08** [0.01, 0.17]	**.03** [.00, .05]	2.08		0.08 [-0.01, 0.16]	.03 [.00, .05]	1.90	
Agreeableness	0.11 [-0.05, 0.25]	.05 [-.02, .13]	1.34		**0.21** [0.05, 0.38]	**.10** [.02, .18]	2.55	
Negative Emotionality	**-0.45** [-0.60, -0.32]	**-.24** [-.31, -.16]	-6.10		**-0.27** [-0.43, -0.11]	**-.14** [-.22, -.06]	-3.46	
Gender	--	--	--		**-0.37** [-0.58, -0.16]	**-.15** [-.23, -.07]	-3.81	
Math Grade	--	--	--		**0.23** [0.13, 0.32]	**.21** [.12, .29]	4.94	
Reasoning Ability	--	--	--		**0.02** [0.00, 0.03]	**.09** [.01, .17]	2.14	

Note. Each model additionally contains 17 dummy-coded predictor variables indicating class membership to control for clustered data at Level 3. For brevity, these fixed effects are not displayed in the table. *b* = unstandardized multilevel regression coefficient; β = standardized multilevel regression coefficient; R_m = marginal multiple R for generalized linear mixed effect models. Gender is coded with 0 = male; 1 = female. SPIQ components were centered within students. Covariates and personality traits were centered at the grand mean and added as predictors of random intercepts. Regression coefficients printed in **bold** are significant at $p < .05$, and regression coefficients printed in **bold and gray shading** are significant at $p < .001$

Discussion

The present study addressed “the perception problem” (Wisniewski et al., 2022) within instructional quality research—differences between perceptions across rating sources—from a different angle. Using an experience sampling design with repeatedly assessed multiple students’ perceptions of the same lesson-specific instructional quality, we performed (a) predictions of state SPIQ and within-student relations and (b) disentangled construed, idiosyncratic perceptions from consensual perceptions that are confounded within raw SPIQ scores. Such analyses are not possible in traditional research designs that assess SPIQ at one point in time and with an unclear target time frame, and aggregate them to higher levels of analyses, thereby considering within-student variation merely as disturbance. We detected substantial effects of students’ personality traits of agreeableness and negative emotionality on state SPIQ. Within-student relations revealed that the dimension of teacher support showed particularly strong positive relations to perceived learning achievement. Additionally, this relation was more pronounced in less agreeable students. Clear differential relations across the three components of SPIQ experience, construal, and consensus could not be detected. Shifting the focus of instructional quality research to individual lessons, within-student relations, and student factors that influence both SPIQ and within-student relations of SPIQ, while always considering the individual relation to the reference group’s perception (i.e., the classroom consensus), essentially shifts the focus of instructional quality research to the student perceiver instead of merely the teachers’ behavior. This ultimately casts a more differentiated picture on instructional quality, classroom interactions, and dynamics in specific lessons.

Experience, Construal, and Consensus of State SPIQ

This study was the first one to differentiate the components of experience, construal, and consensus within SPIQ and the TBDs. An initial examination of intercorrelations revealed large to perfect associations between the different components within the three dimensions of the TBDs framework. In other words, a higher experienced instructional quality (i.e., students’ raw perceptions of instructional quality) is related to higher construed (i.e., students’ idiosyncratic perceptions) and higher consensual (i.e., students’ agreement with classmates’ perceptions) instructional quality. Consistent with this, in all examined relations to personality traits and perceived lesson-specific learning achievement, the three components were considered jointly and showed virtually the same results that only slightly differed in effect sizes in almost all examined relations. Given prior research that discussed the role of the student in SPIQ (e.g., Feistauer & Richter, 2017; Talić et al., 2022; Wagner et al., 2016; Wisniewski et al., 2022), this finding was rather surprising. Some overlap is inherent due to the fact that the components are confounded within one another (i.e., shared variance between individual experience and class mean experience yields consensus, and individual experience variance not explained by class mean experience yields construal). Further, it is important to note that prior research investigated the framework of TBDs as state SPIQ and identified substantial and meaningful within-student variation, where students reliably differentiated between the TBDs from lesson to lesson, even after controlling for shared lesson perceptions (Talić et al., 2022). On this sample-based approach, approximately 53 % of the variance in state SPIQ were attributable to the within-student level, suggesting substantial fluctuations within students. The exact conditions of these fluctuations remained unclear (e.g., fluctuations due to idiosyncratic student characteristics, teacher states, lesson content, or interactions among them; Talić et al., 2022). The present study—following a different, more individual-based approach where construal and consensus perceptions are disentangled from the raw state SPIQ perceptions (i.e., experience) to gain more insight into these state SPIQ fluctuations—found no clear separation of idiosyncratic and consensual SPIQ components within SPIQ experience.

The question of how idiosyncratic SPIQ actually are, remains. Generally, Rauthmann et al. (2015) noted that “most people perceive situations as most other people do,” leaving litte remaining variance after extracting consensual perceptions. In the present study, it might be that variance in SPIQ construal and consensus was too limited to draw reliable conclusions on this question due to limited variance across lessons. Indeed, Talić et al. (2022) reported a maximum of 11 % of variance between lessons (in contrast to a maximum of 54 % of variance between students) on the same dataset. Future research might consider assessing a longer time frame to capture more variability across lessons or compare multiple subjects that might change lesson content more frequently. It is also important to keep in mind that one focus of situation research is the examination of why certain people create certain situations (Rauthmann, 2021). For instance, extraverted people might go to parties or get coffee with their friends because they enjoy the settings (Matz & Harari, 2021). In the present study, however, the situations that were assessed (i.e., lessons in math instruction across three weeks) were not created or deliberately chosen by the students, but constitute a forced environment. The examination of elective subjects might thus offer more insight into idiosyncrasies in SPIQ that go along with a more self-directed choice of attended lessons.

Yet the lack of detecting differential relations across the three components might imply the question of the usefulness of differentiating these components within SPIQ. However, we assert that this differentiation is useful for the examination of SPIQ in shared lessons. First, the advanced insights gained by differentiating different components is theoretically informative. For instance, one could have distinguished between students giving higher ratings for teacher support just because they are more agreeable and thus tend to agree more with the posed item in the questionnaire (i.e., reflecting an effect of agreeableness on *experienced* teacher support) versus students construing instructional behavior as more supportive above and beyond their classmates' perceptions because they are more agreeable (i.e., reflecting an effect of agreeableness on *construed* teacher support) versus students' overlapping in their perceptions with their classmates because they are more agreeable (i.e., reflecting an effect of agreeableness on *consensual* teacher support). Second, although result patterns were largely similar across components in our findings, there are still some noteworthy differences. For instance, we found a significant moderator effect at p < .001 of personality traits on SPIQ–learning achievement relations only for teacher support construal and agreeableness, indicating that it is not the mere rating of instructional behavior as supportive that lowers the positive effect of teacher support on learning achievement, but rather the idiosyncratic construal of more agreeable students. Thus, one could conclude that this might be the portion of the rating which might not necessarily reflect true perceived teacher support but more of an artefact of rating tendencies in agreeable students. Hence, the differentiation in the SPIQ components yields more nuanced insights that offer the elaboration of further hypotheses to be addressed in future studies. At the same time, the current, almost entire lack of differentiation among the three state SPIQ components is also an important indication of how the TBDs framework can be used in future research. Specifically, our results suggest that using the raw state

SPIQ perceptions (i.e., experience) does not change the result pattern with regard to outcome criteria remarkably, thus providing strong validity evidence for implementing state SPIQ perceptions' raw scores in research on the TBDs.

The Role of Students' Personality for Lesson-Specific SPIQ and SPIQ–Learning Achievement Relations

The most crucial of the Big Five personality traits with regard to state SPIQ and particularly teacher support were agreeableness and negative emotionality (RQ1). Agreeableness positively predicted all components of teacher support with small to typical (Gignac & Szodorai, 2016) effect sizes in the models including covariates. Negative emotionality predicted all components of teacher support negatively with typical effect sizes in the models without covariates, whereas these effects do not reach statistical significance at $p < .001$ after including the covariates (see a discussion on the role of covariates below). We could not detect any significant effects at $p < .001$ of personality traits on the other dimensions of cognitive activation and classroom management. Thus, it seems that the dimension of teacher support—which addresses the most affective perceptions of instructional quality (e.g., indicating the teachers' sensitivity for individual student needs), is targeted at the quality of interactions and relationships of agents in the classroom and has a strong link to students' self-determination in the learning process (Praetorius et al., 2018; Ryan & Deci, 2000)—is particularly prone to be influenced by students' personality. The reason that the dimensions of cognitive activation and classroom management might be less prone to personality influences might be their less affective content with clearer physical indications (targeted at task specifics or the learning environment, respectively). The finding that agreeableness was related to higher perceived teacher support is in line with the characteristics of this trait as being cooperative and trusting (Costa & McCrae, 1992) and to prior findings of a positive link between agreeableness and positive course evaluations (Keller & Karau, 2013). Negative emotionality was related to lower perceived teacher support. In other words, students with a higher tendency of experiencing stress and anxiety tended to perceive the same instructional behavior as less supportive. This might indicate a higher need for supportive instructional behavior for those students in order to benefit from it in the classroom. The personality traits of open-mindedness, conscientiousness, and extraversion did not show any relations to SPIQ. Even though we had no clear hypotheses of any relations due to lack of research in this area, this finding is still somewhat surprising. One might have, for instance, assumed open-mindedness to be positively related to cognitive activation due to higher intellectual curiosity, imagination, and divergent thinking that might aid in perceiving instructional behavior as mentally stimulating and challenging. In addition, conscientiousness might have been expected to be associated with classroom management due to the orderliness and dutifulness of conscientious individuals.

Concerning the role of personality traits on the lesson-specific SPIQ–learning achievement relations, we only identified agreeableness as a significant moderator with teacher support construal at $p < .001$ (RQ3), further underpinning the relevance of agreeableness in relation to teacher support. The effect describes a stronger positive relation between construed teacher support and perceived lesson-specific learning achievement for less agreeable students. Agreeable students might construe teacher support such that it does not necessarily reflect true perceived teacher support but rather a tendency to agree from which they cannot benefit in terms of learning achievement gains. In fact, prior research showed the trait of agreeableness to be significantly and positively related to rating leniency, with more agreeable persons providing more favorable ratings even in light of poorer performance (Bernardin et al., 2000; Bernardin et al., 2009; Randall & Sharples, 2012; Yun et al., 2005). The effect sizes of the moderation effect were very small ($\beta = -.05$ and $\beta = -.06$), which is in line with previous research on interaction effects between personality traits and perceived situation characteristics on personality states (Rauthmann, 2021). In the educational context, conscientiousness was found to be the most important Big Five trait in terms of student achievement (Mammadov, 2022). Indeed, conscientiousness showed a significant positive relation to perceived lesson-specific learning achievement at $p < .05$ in the present study ($r = .15$, see Table 3), which albeit did not exceed the relation between the other personality traits and perceived learning achievement. The present study adds the findings that agreeableness and negative emotionality seem to be of higher relevance in perceptions of instructional quality with some interactive effects on a subjective lesson-specific learning achievement indicator, underlining these traits' impact in the educational context. It is important to note that the result pattern of trait SPIQ and personality traits differs in some parts from relations between state SPIQ and personality traits (see Preliminary Analyses). In general, relations between state SPIQ and personality traits were more numerous than between trait SPIQ and personality traits and differed slightly with regard to the personality traits they correlated with. This might suggest some differential relevance of personality traits and momentary, state perceptions of instructional quality versus habitual, time-enduring perceptions. Further research might assess personality states in addition to examine possible mediation effects of personality traits on state SPIQ via personality states (Ching et al., 2014). The present study focused on examinations at the level of individual lessons such that relations between SPIQ trait measures were only of secondary interest and only reported to inform interested readers. For further generation of hypotheses, trait SPIQ relations as well as relations for the 15 personality subfacets of the Big Five traits (Soto & John, 2017) are provided in Table S1 in the OSM.

Short-Term State SPIQ Relations to Perceived Lesson-Specific Learning Achievement

Perceived lesson-specific learning achievement was predominantly positively related to all teacher support components with large effect sizes and to cognitive activation and classroom management components with small effect sizes (RQ2). Given theoretical assumptions and empirical findings on the relation between the TBDs and achievement, this finding is rather unexpected although also relations between teacher support and achievement have been reported (Fauth et al., 2014). Based on between-person research designs, positive relations between cognitive activation and classroom management to student achievement are expected, while teacher support is more closely related to student motivation (Praetorius et al., 2018). The present study identifies the unambiguously strongest relation between teacher support and a lesson-specific, subjective learning achievement indicator at the within-student level. On the one hand, this contrast might reveal differential SPIQ–achievement relations at different levels of analyses due to using interindividual versus intraindividual variance (Molenaar, 2004; Murayama et al., 2017). On the other hand, prior findings that used, for instance, standardized test scores to examine SPIQ–achievement relations can only vaguely be compared to our findings that are based on *perceived* learning achievement (i.e., not reflecting objective achievement). In contrast, the math grade showed substantial relations to teacher support and cognitive activation across all components (see Table S1), demonstrating some differential result patterns for students' perceived learning achievement and their math grade. Specifically, students seem to benefit particularly from teacher support in terms of their perceived learning achievement, whereas the supposed benefit of cognitive activation does not seem to be perceivable by students. Taken together, perceived learning achievement has been shown to be suitable as a daily measure in experience sampling designs, thus maintaining high ecological validity (see Niepel et al., 2022; see also Limitations and Future Research section below) and offering new insights into the dynamics of perceived learning achievement in students' daily life within lessons. Within-student relations between state SPIQ components and perceived lesson-specific learning achievement remained virtually the same after including covariates.

The Role of Gender, Math Grade, and Reasoning Ability

Students' gender, math grade, and reasoning ability were entered in all models as potentially relevant covariates. We report some noteworthy findings that were not the central focus of the present study. First, the effects of negative emotionality on teacher support components did not reach statistical significance after including the covariates, indicating some confounding of these variables. Particularly gender and math grade, that showed relations at $p < .05$ to teacher support, seem to be confounded with negative emotionality. Relations at the $p < .05$ level indicated that female students perceive less teacher support and that students with higher math grades perceive more teacher support. Yet these relations need to be replicated in future studies. Second, gender negatively predicted perceived lesson-specific learning achievement in math with small to typical effect sizes, indicating that female students report lower comprehension in individual math lessons than male students. This finding is in line with prior studies that report on lower self-reported representations of math abilities in female students (even if actual achievement levels are equivalent; Niepel et al., 2019; OECD, 2015). Our findings suggest that this might translate to the level of individual lessons and hints at the need to support girls and young women particularly in mathematics instruction from lesson to lesson. The math grade positively predicted perceived lesson-specific learning achievement with typical to large effect sizes, corroborating the implemented subjective achievement measure. Third, reasoning ability did not play a significant role in any examined relation to SPIQ components or perceived lesson-specific learning achievement above and beyond gender and math grade. One strength of the present study is the inclusion of covariates that have been shown to play a role in the examined relations as well as the presentation of models with and without covariates, thus allowing interested readers to estimate the covariates' effects.

Limitations and Future Research

We note some important limitations. First, throughout the entire article, we focused on students' perceptions of instructional quality. Needless to say, teachers' self-perceptions of instructional quality and lesson-to-lesson variation within these self-perceptions would also be of great interest to study in the future. Students' clustering in classes was controlled for by adding fixed effects of dummy-coded class variables on the one hand, and by focusing on lessons (e.g., students' construal was calculated for specific lessons), which inherently considered classroom clustering. Still, we only addressed students' *perceptions* of teachers' behaviors. The disentanglement of these SPIQ into the three components of experience, construal, and consensus enables an approximation to objective instructional quality by addressing consensual perceptions. That is, if all students within the classroom agree on something, it is *inter*subjective and arguably approximates objectivity. Importantly, this is only an approximation and might still be divergent to teachers' self-perceptions or independent observers' perceptions. Thus, future research should address teachers' state perceptions to gain a more balanced picture of classroom dynamics. It is also recommendable to compute the consensus score based on self- and other ratings (i.e., teachers' and students' perceptions; see Abrahams et al., 2021, for the computation of consensus between self- and other ratings).

To estimate the relevance of the observed effects, it is crucial to discuss effect sizes. For effects that were significant at p < .001, we observed effect sizes that ranged between β = |.05| and β = |.33| with a mean of β = |.17|, reflecting a small to typical effect size (Gignac & Szodorai, 2016). In particularly small effects below .10, the effects' statistical, content, or practical relevance seems questionable at first sight. Yet it is important to keep in mind that even comparatively small effects can have a crucial impact when accumulation over time takes place (Funder & Ozer, 2019; Götz et al., 2022; Matz et al., 2017; Rauthmann et al., 2015). The present study examined dynamics at the level of school lessons (i.e., 45-minute intervals), something experienced by students many thousands of time of during their school career.

Further, it is important to note that our study is correlational and, therefore, cannot imply causality (although we do speak of statistical predictions). For instance, it could also be that perceived lesson-specific learning achievement causally influences the perceptions of instructional quality (e.g., “If I have understood the lesson well, the teacher must have been teaching good”). Thus, experimental research designs including control groups are needed to infer causality. Even though the present study cannot infer causality, it still implemented an intensive longitudinal design where all personality traits and covariates were assessed prior to the experience sampling phase, such that the direction of effects with traits predicting subsequent states seems more plausible than vice versa.

We used perceived lesson-specific learning achievement targeted at the conceptual comprehension of the lesson content as a state achievement indicator. In between-person research designs, usually standardized test scores or school grades are used as achievement indicators (see Arens et al., 2017, for a balanced discussion on different achievement indicators), which are more objective than our perceived state achievement indicator. Yet, in an experience sampling design, the implementation of a standardized test in each lesson is hardly feasible. Further, the positive and substantial relation between math grade and our perceived learning achievement indicator corroborates its validity. In addition, a previous study has demonstrated the empirical distinction of students’ perceived lesson-specific learning achievement versus their perceived lesson-specific math abilities (math self-concept; Niepel et al., 2022), further suggesting its validity in an experience sampling design. In line with our focus on individual perceptions within SPIQ, we thus use perceived lesson-specific learning achievement as a subjective achievement indicator. Future research should, nevertheless, address the question of different indicators of student achievement and their respective implications in an experience sampling design.

Finally, we note that our findings are based on a sample of German secondary school students attending the nineth and 10th grades in schools of the highest ability track, and we only considered math instruction. To test the generalizability of our results, thus, students from other countries, ability tracks, age groups, and subjects are needed.

Implications and Conclusion

Perceptions of instructional quality are omnipresent in daily school life and have wide-ranging implications both for students in terms of student achievement, as well as for teachers in terms of evaluations of their teaching effectiveness even at the country level (OECD, 2014). The present experience sampling study contributed to the understanding of such perceptions within the framework of Three Basic Dimensions (teacher support, cognitive activation, classroom management; Klieme et al., 2001) from the students' perspectives by considering students' personality traits and perceived learning gains in individual lessons. In doing that, we disentangled idiosyncratic from consensual student perceptions that are confounded in the raw perceptions to uncover potentially differential relations across these components and bridge the gap between individual perceptions and actual instructional quality. The present study thus demonstrated new insights to be gained in instructional quality research when examining dynamics at the level of individual lessons in school life and proposes a closer look at the context that students find themselves in.

Online Supplementary Material

Table S1

Correlations between SPIQ Components, Perceived Learning Achievement, Covariates, SPIQ Traits, and Personality Traits and Facets

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
States																		
Experience																		
1. TS	---																	
2. CA	**.72**	---																
3. CM	**.16**	.09	---															
Construal																		
4. TS	**.94**	**.68**	**.13**	---														
5. CA	**.67**	**.97**	.06	**.70**	---													
6. CM	**.14**	.05	**.86**	**.15**	.05	---												
Consensus																		
7. TS	**1.00**	**.72**	**.16**	**.94**	**.67**	**.14**	---											
8. CA	**.72**	**1.00**	.09	**.68**	**.98**	.05	**.72**	---										
9. CM	**.16**	.09	**1.00**	**.13**	.06	**.81**	**.16**	.09	---									
10. P. Ach.	**.62**	**.47**	**.23**	**.59**	**.44**	**.19**	**.62**	**.47**	**.23**	---								
Traits																		
11. Gender	**-.15**	-.09	.03	**-.19**	**-.12**	.00	**-.15**	-.09	.03	**-.23**	---							
12. Grade	**.24**	**.22**	.11	**.20**	**.18**	.04	**.23**	**.22**	**.11**	**.40**	-.06	---						
13. RA	.08	.06	.07	.10	.05	.04	.08	.06	.07	**.27**	-.10	**.38**	---					
14. TS	**.53**	**.31**	.11	**.44**	**.26**	.10	**.53**	**.31**	.11	**.29**	**-.22**	**.12**	-.05	---				
15. CA	**.44**	**.42**	**.32**	**.38**	**.37**	**.24**	**.44**	**.42**	**.33**	**.40**	-.06	**.18**	.05	**.49**	---			
16. CM	.08	.02	**.51**	.09	.02	**.29**	.08	.02	**.54**	**.18**	-.01	-.02	-.08	.02	**.17**	---		
17. O	.02	.10	.10	.02	.11	.04	.02	.10	**.11**	**.15**	**.13**	.03	.09	-.02	**.16**	.04	---	
18. O-AS	-.08	.01	.03	-.06	.01	.00	-.08	.01	.03	-.06	**.34**	-.10	.01	**-.15**	.07	-.01	**.80**	---
19. O-IC	.10	**.17**	**.17**	.10	**.17**	.09	.10	**.17**	**.18**	**.27**	-.07	**.23**	**.16**	**.15**	**.20**	.02	**.72**	**.34**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
20. O-CI	.03	.07	.05	.02	.07	-.02	.03	.07	.06	**.16**	-.03	.01	.04	.00	.08	.10	**.76**	**.39**
21. C	.09	.11	**.16**	.07	.11	.10	.09	.11	**.16**	**.15**	**.20**	**.28**	.00	.02	**.21**	**.13**	**.25**	**.15**
22. C-O	.02	.06	.08	.00	.06	.04	.02	.06	.09	.05	**.20**	**.18**	-.09	-.03	**.13**	.06	.11	.05
23. C-P	**.12**	.10	.08	**.12**	.10	.05	**.12**	.10	.08	**.16**	.08	**.23**	.03	.10	**.17**	**.14**	**.24**	**.14**
24. C-R	**.12**	**.14**	**.30**	.10	**.14**	**.20**	**.12**	**.14**	**.31**	**.22**	**.23**	**.33**	.10	.01	**.22**	**.17**	**.32**	**.22**
25. E	.09	.05	.10	.07	.04	.09	.09	.05	.10	.11	**.14**	-.04	-.10	.00	**.12**	**.15**	**.20**	.10
26. E-S	.06	-.01	.01	.04	-.02	.05	.06	-.01	.00	.03	**.14**	**-.13**	-.10	-.04	.02	**.10**	.06	.03
27. E-A	.01	.00	**.18**	-.01	-.01	**.13**	.01	.00	**.18**	.11	.05	.01	-.02	.01	**.13**	**.19**	**.24**	.10
28. E-EL	**.17**	**.14**	.06	**.13**	**.12**	.04	**.16**	**.14**	.06	**.14**	**.17**	.07	**-.18**	.09	**.20**	.07	**.22**	**.12**
29. A	**.18**	**.13**	**.18**	**.21**	**.16**	**.15**	**.19**	**.13**	**.19**	**.18**	**.24**	**.14**	-.03	.07	**.19**	.05	**.22**	**.19**
30. A-C	**.14**	**.11**	**.16**	**.14**	**.12**	**.14**	**.14**	**.11**	**.16**	**.12**	**.33**	.09	-.03	.06	**.17**	-.01	**.26**	**.22**
31. A-R	**.13**	**.12**	**.19**	**.14**	**.14**	**.12**	**.13**	**.12**	**.20**	**.16**	**.18**	**.18**	-.01	.05	**.17**	**.12**	**.21**	**.14**
32. A-T	**.18**	.09	**.12**	**.23**	**.13**	**.12**	**.19**	.09	**.12**	**.18**	.09	.07	-.05	.07	**.14**	.00	.10	.10
33. NE	**-.22**	**-.16**	**-.11**	**-.25**	**-.19**	**-.15**	**-.22**	**-.16**	-.10	**-.31**	**.26**	**-.17**	.00	**-.13**	-.10	-.09	-.10	.08
34. NE-A	**-.21**	**-.15**	-.06	**-.23**	**-.17**	-.08	**-.21**	**-.15**	-.05	**-.20**	**.30**	-.03	.04	**-.16**	-.04	-.11	-.03	**.14**
35. NE-D	**-.18**	-.11	**-.12**	**-.18**	-.11	**-.14**	**-.18**	-.11	-.11	**-.31**	**.19**	**-.18**	.05	**-.17**	**-.15**	**-.16**	-.11	.07
36. NE-EV	**-.16**	**-.15**	-.10	**-.22**	**-.20**	**-.14**	**-.16**	**-.16**	-.09	**-.26**	**.15**	**-.21**	-.06	-.04	-.08	.04	-.10	-.01

Note. TS = Teacher support; CA = Cognitive activation; CM = Classroom management; P. Ach. = Perceived learning achievement; RA = Reasoning ability; O = Open-mindedness; O-AS = Aesthetic Sensitivity; O-IC = Intellectual Curiosity; O-CI = Creative Imagination; C = Conscientiousness; C-O = Organization; C-P = Productiveness; C-R = Responsibility; E = Extraversion; E-S = Sociability; E-A = Assertiveness; E-EL = Energy Level; A = Agreeableness; A-C = Compassion; A-R = Respectfulness; A-T = Trust; NE = Negative emotionality; NE-A = Anxiety; NE-D = Depression; NE-EV = Emotional Volatility.

Gender is coded with 0 = male; 1 = female.

Correlation coefficients printed in **bold** are significant at $p < .05$, and correlation coefficients printed in **bold and gray shading** are significant at $p < .001$.

Table S1 (Continued)

Correlations between SPIQ Components, Perceived Learning Achievement, Covariates, SPIQ Traits, and Personality Traits and Facets

	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
19. O-IC	---																	
20. O-CI	**.41**	---																
21. C	**.26**	**.20**	---															
22. C-O	**.14**	.08	**.85**	---														
23. C-P	**.27**	**.19**	**.81**	**.48**	---													
24. C-R	**.28**	**.25**	**.79**	**.48**	**.61**	---												
25. E	**.14**	**.23**	**.14**	-.01	**.21**	**.23**	---											
26. E-S	-.02	.10	-.03	**-.14**	.07	.03	**.88**	---										
27. E-A	**.28**	**.20**	**.14**	.01	**.17**	**.24**	**.80**	**.55**	---									
28. E-EL	**.13**	**.27**	**.30**	**.16**	**.32**	**.31**	**.78**	**.56**	**.43**	---								
29. A	**.15**	**.19**	**.36**	**.23**	**.32**	**.40**	**.13**	.10	-.09	**.33**	---							
30. A-C	**.16**	**.22**	**.30**	.17	**.27**	**.36**	**.21**	**.17**	-.01	**.37**	**.89**	---						
31. A-R	**.20**	**.17**	**.47**	**.31**	**.40**	**.50**	.01	-.08	-.07	**.22**	**.83**	**.67**	---					
32. A-T	.02	.09	**.19**	**.15**	**.16**	**.15**	.10	**.13**	**-.13**	**.24**	**.80**	**.54**	**.47**	---				
33. NE	**-.17**	**-.19**	**-.30**	-.05	**-.39**	**-.42**	**-.29**	**-.22**	**-.16**	**-.34**	**-.29**	**-.16**	**-.32**	**-.25**	---			
34. NE-A	-.11	**-.16**	-.10	.07	**-.22**	**-.21**	**-.32**	**-.27**	**-.26**	**-.25**	-.02	.06	-.05	-.05	**.83**	---		
35. NE-D	**-.15**	**-.23**	**-.28**	-.06	**-.38**	**-.36**	**-.41**	**-.33**	**-.26**	**-.44**	**-.24**	**-.17**	**-.22**	**-.20**	**.84**	**.57**	---	
36. NE-EV	**-.17**	-.10	**-.36**	**-.16**	**-.39**	**-.45**	.04	.08	**.14**	**-.14**	**-.45**	**-.28**	**-.52**	**-.35**	**.79**	**.50**	**.45**	---

Note. O-IC = Intellectual Curiosity; O-CI = Creative Imagination; C = Conscientiousness; C-O = Organization; C-P = Productiveness; C-R = Responsibility; E = Extraversion; E-S = Sociability; E-A = Assertiveness; E-EL = Energy Level; A = Agreeableness; A-C = Compassion; A-R = Respectfulness; A-T = Trust; NE = Negative Emotionality; NE-A = Anxiety; NE-D = Depression; NE-EV = Emotional Volatility.

Gender is coded with 0 = male; 1 = female.

Correlation coefficients printed in **bold** are significant at *p* < .05, and correlation coefficients printed in **bold and gray shading** are significant at *p* < .001.

Table S2

Descriptive Statistics of Personality Traits at the Facet Level

	M	*SD*	ω
Open-Mindedness			
Aesthetic Sensitivity	1.96	0.97	.78
Intellectual Curiosity	2.36	0.73	.66
Creative Imagination	2.43	0.75	.81
Conscientiousness			
Organization	2.46	0.95	.88
Productiveness	2.07	0.71	.74
Responsibility	2.53	0.57	.60
Extraversion			
Sociability	2.44	0.88	.84
Assertiveness	2.36	0.71	.74
Energy Level	2.31	0.64	.64
Agreeableness			
Compassion	2.82	0.75	.79
Respectfulness	2.92	0.64	.74
Trust	2.21	0.69	.67
Negative Emotionality			
Anxiety	1.98	0.71	.62
Depression	1.39	0.82	.84
Emotional Volatility	1.73	0.77	.76

Note. Response format: [0, 4].

Chapter 5

Discussion and Outlook

5. Discussion and Outlook

The classroom is a complex and dynamic interactional system where many individuals operate and influence each other (Gardner, 2019). Within each individual, there are stable, trait characteristics (e.g., extraversion) and momentary, state expressions (e.g., momentary enthusiasm) that impact on their behavior in the classroom. In addition, there are numerous external time-varying factors in relation to the classroom setting that further enhances its dynamics (e.g., number of individuals, time of day, week, or school year, lesson content Curby et al., 2011; Praetorius et al., 2014). In this light, the mere assessment of individual self-reports with regard to one domain or at one point in time with the implicit assumption of assessing a 'true' score (Ziegler & Bühner, 2012) seems particularly unreasonable. Therefore, the present dissertation aimed at disentangling confounded variance components within self-reported ratings in three ways in three different research works, while also examining the effect of this disentanglement with regard to crucial related constructs in an educational context.

Central Findings

The first contribution (Chapter 2) presented two latent modeling approaches in specifying hierarchical constructs that entail both domain-specific and domain-general components. These two modeling approaches (i.e., first-order factor (FOF) and nested factor (NF) modeling) were incorporated to specify the construct of test anxiety (TA) in the domains of math and German within the generalized internal/external (GI/E) frame of reference model, that essentially draws on domain-specific processes (Möller et al., 2016). In doing so, the first contribution illustrated a substantial change in result patterns in dependence on the modeling strategy. Particularly dimensional comparisons were prone to the change in modeling strategy. In other words, the NF modeling approach purified the domain-specific construct manifestations (from domain-general construct manifestations; Arens et al., 2021; Brunner et al., 2010; Eid et al., 2017) which then showed more pronounced cross-domain relations. With this, the NF modeling approach might offer one possible explanation for the debated distinction of dimensional contrast and assimilation effects (Möller et al., 2020). The contrasting relation pattern in dependence on the FOF versus NF modeling approach illustrated the importance of considering domain-general manifestations within domain-specific manifestations of hierarchical constructs (e.g., TA), and thus allows for more nuanced insights into the structure and correlates of this construct.

The second contribution (Chapter 3) assessed state students' perceptions of instructional quality (SPIQ) in the Three Basic Dimensions (TBDs; teacher support, cognitive activation, and classroom management; Klieme et al., 2001) to uncover within-student variation from lesson to lesson. SPIQ are traditionally assessed at one point in time and aggregated to higher levels (e.g., class, school, or country; Praetorius et al., 2018), leaving individual student deviations unattended of. In a two-level confirmatory factor analysis model, the second contribution explicitly and simultaneously considered both within- and between-student variation (Dyer et al., 2005; Kim et al., 2016). A substantial proportion of SPIQ variance in the four examined domains of math, physics, German, and English was within-student variance, that exhibited the same cross-level invariant factor structure (Jak & Jorgensen, 2017) as between-student variance and showed similar relations to student trait achievement and motivation. With this, the contribution illustrated a successful assessment of state SPIQ, that vary meaningfully from lesson to lesson. This disentanglement of within-student (time-specific) from between-student (habitual) variation ultimately enables a closer examination of correlates of these different kinds of variation (e.g., lesson-specific dynamics versus long-term relations), and offers empirical support for the multilevel validation of the framework of TBDs.

The third contribution (Chapter 4) examined such lesson-specific dynamics among state SPIQ and perceived learning achievement (i.e., self-reported lesson comprehension) in the light of students' Big Five personality traits using linear mixed effect models. In all examined relations, idiosyncratic state SPIQ were differentiated from consensual (class) state SPIQ to examine potentially differential relations of these components that are usually confounded in the raw SPIQ rating (Rauthmann et al., 2015; Rauthmann & Sherman, 2019). Disentangling these variance components might reveal new insights into the relative importance of one's personal reality (in terms of person-specific or idiosyncratic SPIQ) versus one's social reality (in terms of class consensual SPIQ) with regard to perceived learning achievement in the same lesson. The third contribution illustrated all relations between the raw SPIQ rating (i.e., experience) as well as between the disentangled, idiosyncratic and consensual, respectively, SPIQ and personality traits and perceived learning achievement. These relations hardly differed across SPIQ components, and the components showed a large overlap among each other. In other words, clear idiosyncratic perceptions could not be differentiated from consensual perceptions. Possibly, most of SPIQ are consensual as it is more adaptive to perceive the environment in the way that one's peers do (Rauthmann et al., 2015). Agreeableness was the most important personality trait with regard to state SPIQ, while teacher support was the most important dimension of instructional quality with regard to perceived learning achievement. As such, this contribution offered valuable insights into the person-specificity of SPIQ and provided evidence for the validity of the raw state SPIQ rating.

Further Insights

The present dissertation yielded numerous findings that are reported in detail in the respective contributions. Two higher-level aspects, that are not discussed explicitly or in detail within the contributions are noted here. First, especially in the third contribution, we noted the role of students' gender in educational processes. For instance, gender was related to perceived lesson-specific learning achievement in math, with girls and young women reporting lower comprehension of the lesson content. Further, gender seemed to be confounded with the personality trait of negative emotionality, with girls and young women reporting higher negative emotionality than their male classmates (for details please see Chapter 4), highlighting the need to investigate gender-related dynamics in education. For instance, in the early grades of school, girls and boys display similar achievement levels in math, yet, this changes throughout the course of school, resulting in gender differences in math test scores with boys achieving higher scores than girls (Buchmann et al., 2008). In addition, even if math achievement levels do not differ, girls tend to report lower self-perceived math abilities (Niepel et al., 2019; OECD, 2015). Increasingly more efforts are devoted to addressing gender inequalities in education (UNESCO, 2017). Without placing an emphasis on gender differences, the present dissertation noted the need to do so. Specifically, students' gender should not only be incorporated as control variable, but rather be examined focally in gender-specific analyses and studies with the ultimate goal to understand gender-related educational processes to further approach gender equality. In doing so, not only biological gender should be considered, but also gender role identities (Altstötter-Gleich, 2004) and gender stereotypes (Retelsdorf et al., 2015).

Second, the DynASCEL project illustrates that a three-week experience sampling study (ESM) with up to 42 measurement points per student with a borrowed research smartphone in addition to an exhaustive pre- and post-assessment in paper-and-pencil format was feasible in a sample of secondary school students of the highest ability track with the mean age of 15 years. Importantly, the compliance was acceptable or better with at least about 70 % of the pre-programmed measurement points completed (Rintala et al., 2019; Wen et al., 2017). This is not to be taken for granted, particularly given the combination of the adolescent study participants who undergo various non-school related challenges during their developmental stage on the one hand, and the enhanced participant demands in this intensive longitudinal study on the other hand. Based on high reliabilities, empirical validity evidence, and limited missingness, the DynASCEL data suggest that adolescents – at least those attending the highest-ability track – are eligible participants for such studies. It is important to note, however, that the DynASCEL project was carefully planned and followed state-of-the-art recommendations for intensive longitudinal projects in both the planning and execution phase that assuredly contributed to the data quality. For instance, the sampling protocol was clearly defined and derived in the light of the constructs of interest, applicability in the classroom context and participants (Doherty et al., 2020), a pilot study was conducted and item psychometrics were examined and items, if applicable, adjusted or removed. Participant burden was minimized wherever possible (Fuller-Tyszkiewicz et al., 2013), participation-based, but also completeness-based participant incentives were implemented (Christensen et al., 2003) and on-site study guidance was provided for the first few days of the study for each class. Encouragingly, the DynASCEL project thus illustrates how adolescent study participants provide intensive longitudinal data in a high quality and thereby allow for interesting insights during this crucial developmental stage.

Limitations and Future Research

We note some important limitations of this dissertation and provide recommendations for further research. First, all three contributions of this dissertation are based on (parts of) the same dataset. The first contribution did not make use of the experience sampling data and only used data from the student trait pre-assessment, while the other two contributions focally used the experience sampling data but also data from the trait pre-assessment and, for the second contribution, pre- and post-assessment. Thus, for the replication of results, the analyses should be done on a different dataset of highest-ability track German secondary school students of the ninth and tenth grades with a three-week experience sampling phase in the four core subjects math, physics, German, and English. For the generalization of results, the analyses should be completed on different datasets with students from different grades (i.e., age groups), school tracks, using different subjects (e.g., elective subjects; see Chapter 4) in different cultures.

Second, throughout the three contributions, we cannot speak of causality, as we only obtained correlative relations. Although the terms of predictions in the statistical sense are used, we cannot infer true causal relations. It might be that we did not assess important third variables, or that relations in fact are of the opposite direction or reciprocal. To infer causality, further research should conduct experimental research designs including control groups. This has been done for processes within the GI/E model (e.g., Müller-Kalthoff et al., 2017) and instructional quality (e.g., Hardy et al., 2006).

Third, in the studies on perceptions of instructional quality, we focused on students' perceptions and the role of the student in SPIQ. Yet, of course also the teachers' self-perceptions would be of great interest (Wisniewski et al., 2022). Further research could assess teachers' traits and states besides students' traits and states to gain a more balanced picture of classroom dynamics in perceptions of instructional quality. This would also allow for an estimation of consensus among the students and the teachers, adding a new crucial perspective and thereby approximating the objective situation further (Rauthmann et al., 2015). Further, it would be highly interesting to systematically assess teachers that teach the same class in different school subjects, and teachers teaching the same subject to different classes, allowing for unique examinations of teacher*class interactions (Fauth et al., 2020). Examining the same teacher in the same class in different subjects would also enable an examination of GI/E model-based processes (i.e., social and comparison processes) that have been shown to impact on SPIQ (Arens & Möller, 2016). In teachers' self-perceptions of expertise, empirical support for social and dimensional comparison processes has also been found for different knowledge domains (Paulick et al., 2017) and might translate to different subjects as well.

Fourth, although the dataset was rich and exhaustive with multiple thousands of measurement points (at Level 1), a couple hundreds of students (at Level 2), and 18 classrooms (at Level 3), we partially dealt with limited statistical power depending on the complexity of models and analyses. This showed in the need to fix negative item residual variances (Kline, 2016), the need to conduct two-step estimation procedures where only measurement models were specified in the first step and coefficients thereof were fixed in the second step of specifying the structural models (Anderson & Gerbing, 1988), conducting models with random intercepts only despite presumably varying random slopes (Abrahams et al., 2021), and using dummy-coded predictor variables to control for higher-level clustering (Hox et al., 2018). Experience sampling studies and educational studies with students nested in classrooms yield hierarchical data that needs to be considered in models that can adequately handle dependent observations (Hox et al., 2018). Although we could not always perform the *ideal* analysis (e.g., 3-Level modeling with an explicit specification of the class level, see above), all analyses were state-of-the-art and all dependencies in the data were considered. Nevertheless, further research might focus on collecting larger sample sizes especially with regard to the number of classes. In some of our examined targets (e.g., GI/E model, shared lessons and consensual perceptions) the classrooms were implicitly included as frames of reference, yet, they should be specified explicitly in the models as own level of analysis.

Fifth, ESM designs are able to uncover intraindividual variation (versus interindividual variation) with the aim of "bringing the person back into scientific psychology" as stated in the much-cited article by Molenaar (2004, p. 1). At the same time, it is important to keep in mind that assessing intraindividual variation with the use of ESM does not automatically and implicitly equate to conducting idiographic research. Yet, if analyzed accordingly, ESM studies can bridge the gap between the idiographic/nomothetic divide and distinguish between general laws that always hold (i.e., nomothetic) and individual uniqueness (i.e., idiographic; Renner et al., 2020). The second and third contribution of the present dissertation assessed intraindividual variability, yet, did not perform true idiographic research as still aggregates were used (e.g., aggregated, sample-based intraindividual variability). Throughout the course of the dissertation, analyses became increasingly individual (see the examination of idiosyncratic SPIQ in Chapter 4). Yet, idiosyncracies in SPIQ were still examined at the sample level, across all students. To uncover unique, individual phenomena, future research should thus make use of according statistical procedures (e.g., group iterative multiple model estimation (GIMME); Beltz et al., 2016; Gates & Molenaar, 2012).

Sixth, there are some ways in which the present ESM design might be complemented or elaborated in accordance with the respective research focus. If one were more interested in short-term dynamic processes, additional variables might be assessed as state measures to enable more exhaustive within-person analyses and possible within-person mediations (e.g., personality states; Ching et al., 2014). If one were interested in long-term change of short-term variability, repeated ESM studies across longer time intervals on the same sample (i.e., measurement burst designs; Stawski et al., 2015) could be conducted. If one would like to cross-validate psychological self-reports with other indicators that do not rely on self-report, to highlight possible differential relations between different kinds of indicators (e.g., self-reported test anxiety versus heart rate), or to identify context-specific patterns of psychological experience (e.g., feeling anxious in the classroom but not at home), one could incorporate a multimodal ambulatory assessment (Trull & Ebner-Priemer, 2013). Vast technological progress on devices that are increasingly popular and wide-spread (e.g., smartphones and smartwatches) enables new methods of assessment, including biological stress or well-being indicators, physical activity, GPS sensors, smartphone usage data including specific app usage, sleep times, or social network behavior to just name a few (Harari et al., 2020; Matz & Harari, 2021; Quiroz et al., 2018; Stachl et al., 2020).

Finally, we have demonstrated three examples of variance decomposition in this dissertation. It is important to keep in mind that these examples are not exhaustive in any way as to possible other confounding of variance components. For instance, we could not address student*teacher interactions or student*subject interactions

(Feistauer & Richter, 2017) among many other variance sources. Yet, the present dissertation addressed three widespread examples of confounded ratings, and with this entails important implications.

Implications and Conclusion

The present dissertation demonstrated three examples of how ratings in psychological assessment can be confounded and how these confoundings impact on relations to other constructs of interest. In doing so, we identified (a) that hierarchical constructs that entail distinct domain-specific manifestations with domain-general manifestations at the apex are highly dependent on the modeling strategy. In contrast to FOF models, NF models (Brunner et al., 2010; Eid et al., 2017) disentangle general from domain-specific test anxiety components, and facilitate the detection of dimensional comparison effects in the GI/E model (Möller et al., 2016). Practically, this understanding can aid in handling students' socio-affective experiences more holistically in light of other domains and across multiple domains. Raising this awareness might buffer the detrimental dimensional comparison effects and can, for instance, refine evaluations targeted at reducing test anxiety (von der Embse et al., 2013) in terms of their effectiveness on specific domains or a reduction in general. Further, we uncovered (b) substantial within-student variance from lesson to lesson in three basic dimensions (Klieme et al., 2001) of students' perceptions of instructional quality that is invariant across the within- and between-students levels and shows similar relations to student achievement and motivation across levels. These fluctuations would have been neglected if aggregated, cross-sectional data were used, yet, hold important implications. Considering situation-specific besides habitual perceptions of instructional quality enables a more adaptive view on classroom dynamics and the roles of students and teachers in it. Ultimately, this promotes more flexible views on instructional processes that have the potential to weaken dysfunctional beliefs (e.g., “I have no control over the class. I am a bad teacher”) in favor of more constructive beliefs that are put into perspective (e.g., “Today, I had issues in maintaining a clear classroom structure. Maybe I was too tired today.”). Finally, we illustrated (c) the differentiation between idiosyncratic and consensual students' state perceptions of instructional quality (Rauthmann et al., 2015) and their relations to students‘ personality traits and perceived lesson-specific learning achievement, where we hardly detected any differential relations. Despite substantial lesson-to-lesson variation, idiosyncratic perceptions could not be clearly distinguished from consensual perceptions across all students. On the one hand, this contribution sheds light on personality traits that influence state SPIQ and can provide an enhanced understanding of how SPIQ truly reflect teaching effectiveness (Scherer et al., 2016). On the other hand, state perceptions of teacher support are related to perceived learning achievement in the same lesson, highlighting a distinction between lesson-specific dynamics and long-term relations. Lesson-specific effects can accumulate over time (e.g., Götz et al., 2022) and are thus an important source of information of students' daily experiences. The lack of differentiation of idiosyncratic and consensual SPIQ supports the validity of the use of raw state SPIQ scores and the validity of the TBD framework.

In conclusion, the disentanglement of different variance components within ratings of psychological constructs enables a refinement in both the analyses of constructs' structures and correlates, as well as a relativization of conclusions to these specific variance components. In other words, findings can be more precisely related to a specific domain (instead of a domain-general characteristic), to a specific situation (instead of a habitual style), or to a specific person (instead of more persons' consensus), ultimately enhancing more nuanced and flexible views on psychological constructs and their specificity or generality in terms of domains, situations, and persons.

6. References

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience*, *17*(4), 491–496. https://doi.org/10.1038/nn.3648

Abrahams, L., Rauthmann, J. F., & de Fruyt, F. (2021). Person-situation dynamics in educational contexts: A self- and other-rated experience sampling study of teachers' states, traits, and situations. *European Journal of Personality*, *35*(4), 598–622. https://doi.org/10.1177/08902070211005621

Ahmed, W., Minnaert, A., Kuyper, H., & van der Werf, G. (2012). Reciprocal relationships between math self-concept and math anxiety. *Learning and Individual Differences*, *22*(3), 385–389. https://doi.org/10.1016/j.lindif.2011.12.004

Altstötter-Gleich, C. (2004). Expressivität, Instrumentalität und psychische Gesundheit [Expressivity, instrumentality and psychological health. Validation of a scale assessing gender role self-concept]. *Zeitschrift Für Differentielle und Diagnostische Psychologie*, *25*(3), 123–139. https://doi.org/10.1024/0170-1789.25.3.123

Amthauer, R. (1970). *Intelligenz-Struktur-Test [Intelligence structure test]*. Hogrefe.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411–423. https://doi.org/10.1037/0033-2909.103.3.411

American Psychological Association. (2020). *Publication manual of the American Psychological Association 2020: The official guide to APA style* (7th ed.). American Psychological Association.

Arens, A. K., Becker, M., & Möller, J. (2017). Social and dimensional comparisons in math and verbal test anxiety: Within- and cross-domain relations with achievement and the mediating role of academic self-concept. *Contemporary Educational Psychology*, *51*, 240–252. https://doi.org/10.1016/j.cedpsych.2017.08.005

Arens, A. K., Jansen, M., Preckel, F., Schmidt, I., & Brunner, M. (2021). The structure of academic self-concept: A methodological review and empirical illustration of central models. *Review of Educational Research*, *91*(1), 34–72. https://doi.org/10.3102/0034654320972186

Arens, A. K., Marsh, H. W., Pekrun, R., Lichtenfeld, S., Murayama, K., & vom Hofe, R. (2017). Math self-concept, grades, and achievement test scores: Long-term reciprocal effects across five waves and three achievement tracks. *Journal of Educational Psychology*, *109*(5), 621–634. https://doi.org/10.1037/edu0000163

Arens, A. K., & Möller, J. (2016). Dimensional comparisons in students' perceptions of the learning environment. *Learning and Instruction*, *42*, 22–30. https://doi.org/10.1016/j.learninstruc.2015.11.001

Arens, A. K., Yeung, A. S., Craven, R. G., & Hasselhorn, M. (2011). The twofold multidimensionality of academic self-concept: Domain specificity and separation between competence and affect components. *Journal of Educational Psychology*, *103*(4), 970–981. https://doi.org/10.1037/a0025047

Asparouhov, T., & Muthén, B. (2020). Comparison of models for the analysis of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(2), 275–297. https://doi.org/10.1080/10705511.2019.1626733

Barroso, C., Ganley, C. M., McGraw, A. L., Geer, E. A., Hart, S. A., & Daucourt, M. C. (2021). A meta-analysis of the relation between math anxiety and math achievement. *Psychological Bulletin*, *147*(2), 134-168. https://doi.org/10.1037/bul0000307

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O., & Neubrand, J. (1997). *TIMSS - Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich: Deskriptive Befunde*. *Springer eBook Collection*. VS Verlag fur Sozialwissenschaften GmbH. https://doi.org/10.1007/978-3-322-95096-3

Bellens, K., van Damme, J., van den Noortgate, W., Wendt, H., & Nilsen, T. (2019). Instructional quality: catalyst or pitfall in educational systems' aim for high achievement and equity? An answer based on multilevel SEM analyses of TIMSS 2015 data in Flanders (Belgium), Germany, and Norway. *Large-Scale Assessments in Education*, *7*(1). https://doi.org/10.1186/s40536-019-0069-2

Beltz, A. M., Wright, A. G. C., Sprague, B. N., & Molenaar, P. C. M. (2016). Bridging the nomothetic and idiographic approaches to the analysis of clinical data. *Assessment*, *23*(4), 447–458. https://doi.org/10.1177/1073191116648209

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Bernardin, H. J., Cooke, D. K., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *The Journal of Applied Psychology*, *85*(2), 232–236. https://doi.org/10.1037/0021-9010.85.2.232

Bernardin, H. J., Tyler, C. L., & Villanova, P. (2009). Rating level and accuracy as a function of rater personality. *International Journal of Selection and Assessment*, *17*(3), 300–310. https://doi.org/10.1111/j.1468-2389.2009.00472.x

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass.

Bolger, N., & Laurenceau, J.-P. (Eds.). (2013). *Methodology in the Social Sciences Ser. Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Breil, S. M., Geukes, K., Wilson, R. E., Nestler, S., Vazire, S., & Back, M. D. (2019). Zooming into real-life extraversion – How personality and situation shape sociability in social interactions. *Collabra: Psychology*, *5*(1), Article 7. https://doi.org/10.1525/collabra.170

Brunner, M., Keller, U., Dierendonck, C., Reichert, M., Ugen, S., Fischbach, A., & Martin, R. (2010). The structure of academic self-concepts revisited: The nested Marsh/Shavelson model. *Journal of Educational Psychology*, *102*(4), 964–981. https://doi.org/10.1037/a0019644

Brunner, M., Keller, U., Hornung, C., Reichert, M., & Martin, R. (2009). The cross-cultural generalizability of a new structural model of academic self-concepts. *Learning and Individual Differences*, *19*(4), 387–403. https://doi.org/10.1016/j.lindif.2008.11.008

Buchmann, C., DiPrete, T. A., & McDaniel, A. (2008). Gender inequalities in education. *Annual Review of Sociology*, *34*(1), 319–337. https://doi.org/10.1146/annurev.soc.34.040507.134719

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295. https://doi.org/10.1006/ceps.2001.1094

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology*, *97*(2), 268–274. https://doi.org/10.1037/0022-0663.97.2.268

Ching, C. M., Church, A. T., Katigbak, M. S., Reyes, J. A. S., Tanaka-Matsumi, J., Takaoka, S., Zhang, H., Shen, J., Arias, R. M., Rincon, B. C., & Ortiz, F. A. (2014). The manifestation of traits in everyday behavior and affect: A five-culture study. *Journal of Research in Personality*, *48*, 1–16. https://doi.org/10.1016/j.jrp.2013.10.002

Christensen, T. C., Barrett, L. F., Bliss-Moreau, E., & Lebo, K. (2003). A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, *4*(1), 53–78. https://doi.org/10.1023/A:1023609306024

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, *4*(1), 5–13.

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322–331. https://doi.org/10.1037/0022-3514.81.2.322

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., Hamre, B., & Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade. *The Elementary School Journal*, *112*(1), 16–37. https://doi.org/10.1086/660682

Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., Soto, C. J., & John, O. P. (2019). Die deutsche Version des Big Five Inventory 2 (BFI-2). *Diagnostica*, *65*(3), 121–132. https://doi.org/10.1026/0012-1924/a000218

Devine, A., Fawcett, K., Szücs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, *8*(33), 1–9. https://doi.org/10.1186/1744-9081-8-33

Dickhäuser, O., & Plenter, I. (2005). „Letztes Halbjahr stand ich zwei“: Zur Akkuratheit selbst berichteter Noten [On the accuracy of self-reported school marks]. *Zeitschrift Für Pädagogische Psychologie*, *19*(4), 219–224. https://doi.org/10.1024/1010-0652.19.4.219

Doherty, K., Balaskas, A., & Doherty, G. (2020). The design of ecological momentary assessment technologies. *Interacting with Computers*, *32*(3), 257–278. https://doi.org/10.1093/iwcomp/iwaa019

Dörendahl, J., Scherer, R., Greiff, S., Martin, R., & Niepel, C. (2021). Dimensional comparisons in the formation of domain-specific achievement goals. *Motivation Science*, *7*(3), 306–318. https://doi.org/10.1037/mot0000203

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*, *16*(1), 149–167. https://doi.org/10.1016/j.leaqua.2004.09.009

Eid, M. (2020). Multi-faceted constructs in abnormal psychology: Implications of the bifactor S-1 model for individual clinical assessment. *Journal of Abnormal Child Psychology 48*(7), 895–900. https://doi.org/10.1007/s10802-020-00624-9

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff, J., Lüdtke, O., Polikoff, M. S., Klusmann, U., & Trautwein, U. (2020). Don't blame the teacher? The need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology*, *112*(6), 1284–1302. https://doi.org/10.1037/edu0000416

Feistauer, D., & Richter, T. (2017). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education*, *42*(8), 1263–1279. https://doi.org/10.1080/02602938.2016.1261083

Franzen, P., Arens, A. K., Greiff, S., & Niepel, C. (2022). Student profiles of self-concept and interest in four domains: A latent transition analysis. *Learning and Individual Differences*, *95*, 102139. https://doi.org/10.1016/j.lindif.2022.102139

Franzen, P., Arens, A. K., Greiff, S., van der Westhuizen, L., Fischbach, A., Wollschläger, R., & Niepel, C. (2022). Developing and validating a short-form questionnaire for the assessment of seven facets of conscientiousness in large-scale assessments. *Journal of Personality Assessment*, *104*(6), 759–773. https://doi.org/10.1080/00223891.2021.1998083

Fuller-Tyszkiewicz, M., Skouteris, H., Richardson, B., Blore, J., Holmes, M., & Mills, J. (2013). Does the burden of the experience sampling method undermine data quality in state body image research? *Body Image*, *10*(4), 607–613. https://doi.org/10.1016/j.bodyim.2013.06.003

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gardner, R. (2019). Classroom interaction research: The state of the art. *Research on Language and Social Interaction*, *52*(3), 212–226. https://doi.org/10.1080/08351813.2019.1631037

Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, *63*(1), 310–319. https://doi.org/10.1016/j.neuroimage.2012.06.026

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. https://doi.org/10.1037/a0032138

Gentry, M., Gable, R. K., & Rizza, M. G. (2002). Students' perceptions of classroom activities: Are there grade-level and gender differences? *Journal of Educational Psychology*, *94*(3), 539–544. https://doi.org/10.1037/0022-0663.94.3.539

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Goetz, T., Keller, M. M., Lüdtke, O., Nett, U. E., & Lipnevich, A. A. (2020). The dynamics of real-time classroom emotions: Appraisals mediate the relation between students' perceptions of teaching and their emotions. *Journal of Educational Psychology*, *112*(6), 1243–1260. https://doi.org/10.1037/edu0000415

Goetz, T., Lüdtke, O., Nett, U. E., Keller, M. M., & Lipnevich, A. A. (2013). Characteristics of teaching and students' emotions in the classroom: Investigating differences across domains. *Contemporary Educational Psychology*, *38*(4), 383–394. https://doi.org/10.1016/j.cedpsych.2013.08.001

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., & Preckel, F. (2014). "My questionnaire is too long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology*, *39*(3), 188–205. https://doi.org/10.1016/j.cedpsych.2014.04.002

Gogol, K., Brunner, M., Martin, R., Preckel, F., & Goetz, T. (2017). Affect and motivation within and between school subjects: Development and validation of an integrative structural model of academic self-concept, interest, and anxiety. *Contemporary Educational Psychology*, *49*, 46–65. https://doi.org/10.1016/j.cedpsych.2016.11.003

Gogol, K., Brunner, M., Preckel, F., Goetz, T., & Martin, R. (2016). Developmental dynamics of general and school-subject-specific components of academic self-concept, academic interest, and academic anxiety. *Frontiers in Psychology*, *7*(356), 1–15. https://doi.org/10.3389/fpsyg.2016.00356

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52*(11), 1182–1186. https://doi.org/10.1037//0003-066x.52.11.1182

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, *33*(5), 313–317. https://doi.org/10.1027/1015-5759/a000450

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*(4), 407–434. https://doi.org/10.1207/s15327906mbr2804_2

Hallquist, M. N., & Wiley, J. F. (2018). Mplusautomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Harari, G. M., Vaid, S. S., Müller, S. R., Stachl, C., Marrero, Z., Schoedel, R., Bühner, M., & Gosling, S. D. (2020). Personality sensing for theory development and assessment in the digital age. *European Journal of Personality*, *34*(5), 649–669. https://doi.org/10.1002/per.2273

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking.". *Journal of Educational Psychology*, *98*(2), 307–326. https://doi.org/10.1037/0022-0663.98.2.307

Hausen, J. E., Möller, J., Greiff, S., & Niepel, C. (2022). Students' personality and state academic self-concept: Predicting differences in mean level and within-person variability in everyday school life. *Journal of Educational Psychology*, *114*(6), 1394–1411. https://doi.org/10.1037/edu0000760

Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2020). Giving G a meaning: An application of the bifactor-(S-1) approach to realize a more symptom-oriented modeling of the Beck depression inventory–II. *Assessment*, *27*(7), 1429-1447. https://doi.org/10.1177/1073191118803738

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47–77. https://doi.org/10.3102/00346543058001047

Hock, M., Renner, K.-H., Bergner-Köther, R., & Laux, L. (submitted). Separating states and traits in anxiety and depression.

Hodapp, V. (1991). Das Prüfungsängstlichkeitsinventar TAI-G: Eine erweiterte und modifizierte Version mit vier Komponenten [The Test Anxiety Inventory TAI-G: An expanded and modified version with four components]. *Zeitschrift Für Pädagogische Psychologie*, *5*(2), 121–130.

Hodapp, V., Rohrmann, S., & Ringeisen, T. (2011). *Prüfungsangstfragebogen: PAF*. Hogrefe.

Holzberger, D., Philipp, A., & Kunter, M. (2013). How teachers' self-efficacy is related to instructional quality: A longitudinal analysis. *Journal of Educational Psychology*, *105*(3), 774–786. https://doi.org/10.1037/a0032198

Hox, J. J., Moerbeek, M., & van Schoot, R. de. (2018). *Multilevel analysis: Techniques and applications* (Third edition). *ProQuest Ebook Central*. Routledge Taylor & Francis. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5046900

Hoyle, R. H., & Gottfredson, N. C. (2015). Sample size considerations in prevention research applications of multilevel modeling and structural equation modeling. *Prevention Science*, *16*(7), 987–996. https://doi.org/10.1007/s11121-014-0489-8

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, F. L. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *The Journal of Experimental Education*, *84*(1), 175–196. https://doi.org/10.1080/00220973.2014.952397

Jaekel, A.-K., Göllner, R., & Trautwein, U. (2021). How students' perceptions of teaching quality in one subject are impacted by the grades they receive in another subject: Dimensional comparisons in student evaluations of teaching quality. *Journal of Educational Psychology*, *113*(4), 770–783. https://doi.org/10.1037/edu0000488

Jak, S., & Jorgensen, T. D. (2017). Relating measurement invariance, cross-level invariance, and multilevel reliability. *Frontiers in Psychology*, *8*, 1640. https://doi.org/10.3389/fpsyg.2017.01640

Janis, R. A., Burlingame, G. M., & Olsen, J. A. (2016). Evaluating factor structures of measures in group research: Looking between and within. *Group Dynamics: Theory, Research, and Practice*, *20*(3), 165–180. https://doi.org/10.1037/gdn0000043

John, O. P. (2021). History, measurement, and conceptual elaboration of the Big-Five trait taxonomy: The paradigm matures. In O. P. John, Robins, & R. W. (Eds.), *Handbook of personality: Theory and research* (pp. 35–82). The Guilford Press.

Jonassen, D. (2011). Supporting problem solving in PBL. *Interdisciplinary Journal of Problem-Based Learning*, *5*(2). https://doi.org/10.7771/1541-5015.1256

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2. ed.). *Advanced quantitative techniques in the social sciences: Vol. 10*. Sage. https://doi.org/10.4135/9781452226576

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). The Guilford Press.

Keller, H., & Karau, S. J. (2013). The importance of personality in students' perceptions of the online learning experience. *Computers in Human Behavior*, *29*(6), 2494–2500. https://doi.org/10.1016/j.chb.2013.06.007

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research*, *51*(6), 881–898. https://doi.org/10.1080/00273171.2016.1228042

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: "Aufgabenkultur" und Unterrichtsgestaltung [Mathematicsematics instruction at secondary level. Task culture and instructional design]. In Bundesministerium für Bildung und Forschung (Ed.), *TIMMS - Impulse für Schule und Unterricht* (pp. 43–57). Bundesministerium für Bildung und Forschung.

Kline, R. B. (2016). *Principles and practice of structural equation modeling (4th ed.).* The Guilford Press.

Klumb, P., Elfering, A., & Herre, C. (2009). Ambulatory assessment in industrial/organizational psychology. *European Psychologist*, *14*(2), 120–131. https://doi.org/10.1027/1016-9040.14.2.120

Krapp, A. (2002). An educational-psychological theory of interest and its relation to SDT. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 405–427). University of Rochester Press.

Kunter, M., & Baumert, J. (2007). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction*, *9*(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, *105*(3), 805–820. https://doi.org/10.1037/a0032583

Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV: A multicriteria analysis. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Mathematics Teacher Education: Vol. 8. Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project* (pp. 97–124). Springer. https://doi.org/10.1007/978-1-4614-5149-5_6

Lachowicz, M. J., Preacher, K. J., & Kelley, K. (2018). A novel measure of effect size for mediation analysis. *Psychological Methods*, *23*(2), 244–261. https://doi.org/10.1037/met0000165

Lazarides, R., & Ittel, A. (2012). Instructional quality and attitudes toward mathematics: Do self-concept and interest differ across students' patterns of perceived instructional quality in mathematics classrooms? *Child Development Research*, *2012*, 1–11. https://doi.org/10.1155/2012/813920

Liebert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, *20*(3), 975–978. https://doi.org/10.2466/pr0.1967.20.3.975

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000R [Intelligence structure test 2000R]*. Hogrefe.

Liepmann, D., Beauducel, A., Brocke, B., & Nettelnstroth, W. (2012). *Intelligenz-Struktur-Test-Screening [Intelligence structure test-screening]*. Hogrefe.

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education*, *6*(1). https://doi.org/10.1186/s40536-018-0061-2

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 modeling. *Contemporary Educational Psychology*, *34*(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Mammadov, S. (2022). Big Five personality traits and academic performance: A meta-analysis. *Journal of Personality*, *90*(2), 222–255. https://doi.org/10.1111/jopy.12663

Mang, J., Ustjanzew, N., Schiepe-Tiska, A., Prenzel, M., Sälzer, C., Müller, K., & Gonzaléz Rodríguez, E. (2018). *PISA 2012 Skalenhandbuch: Dokumentation der Erhebungsinstrumente*. Waxmann.

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal*, *23*(1), 129–149. https://doi.org/10.3102/00028312023001129

Marsh, H. W. (1988). The content specificity of math and English anxieties: The high school and beyond study. *Anxiety Research*, *1*(2), 137–149. https://doi.org/10.1080/10615808808248226

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (1st ed., pp. 319–383). Springer. https://doi.org/10.1007/1-4020-5742-3_9

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science*, *1*(2), 133–163. https://doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*(2), 106–124. https://doi.org/10.1080/00461520.2012.670488

Marsh, H. W., Relich, J. D., & Smith, I. D. (1983). Self-concept: The construct validity of interpretations based upon the SDQ. *Journal of Personality and Social Psychology*, *45*(1), 173–187. https://doi.org/10.1037/0022-3514.45.1.173

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality*, *74*(2), 403–456. https://doi.org/10.1111/j.1467-6494.2005.00380.x

Matz, S. C., Gladstone, J. J., & Stillwell, D. (2017). In a world of big data, small effects can still matter: A reply to Boyce, Daly, Hounkpatin, and Wood (2017). *Psychological Science*, *28*(4), 547–550. https://doi.org/10.1177/0956797617697445

Matz, S. C., & Harari, G. M. (2021). Personality-place transactions: Mapping the relationships between Big Five personality traits, states, and daily places. *Journal of Personality and Social Psychology*, *120*(5), 1367–1385. https://doi.org/10.1037/pspp0000297

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90. https://doi.org/10.1037//0022-3514.52.1.81

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review*, *120*(3), 544–560. https://doi.org/10.1037/a0032459

Möller, J., Müller-Kalthoff, H., Helm, F., Nagy, & Marsh, H. W. (2016). The generalized internal/external frame of reference model: An extension to dimensional comparison theory. *Frontline Learning Research*, *4*(2), 1–11. https://doi.org/10.14786/flr.v4i2.169

Möller, J., Zitzmann, S., Helm, F., Machts, N., & Wolff, F. (2020). A meta-analysis of relations between achievement and self-concept. *Review of Educational Research*, *90*(3), 376–419. https://doi.org/10.3102/0034654320919354

Morris, L. W., Davis, M. A., & Hutchings, C. H. (1981). Cognitive and emotional components of anxiety: Literature review and a revised worry–emotionality scale. *Journal of Educational Psychology*, *73*(4), 541–555. https://doi.org/10.1037/0022-0663.73.4.541

Movisens GmbH. (2017). movisensXS (Version 1.3.0-1.3.4) [Mobile App]. https://www.movisens.com

Müller-Kalthoff, H., Jansen, M., Schiefer, I. M., Helm, F., Nagy, N., & Möller, J. (2017). A double-edged sword? On the benefit, detriment, and net effect of dimensional comparison on self-concept. *Journal of Educational Psychology*, *109*(7), 1029–1047. https://doi.org/10.1037/edu0000171

Murayama, K., Goetz, T., Malmberg, L.-E., Pekrun, R., Tanaka, A., & Martin, A. J. (2017). Within-person analysis in educational psychology: Importance and illustration. In D.W. Putwain & K. Smart (Ed.), *British Journal of Educational Psychology Monograph Series II: Psychological aspects of education - Current trends: The role of competence beliefs in teaching and learning* (pp. 71–87). Wiley.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide (8th ed.)*. Muthén & Muthén.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Niepel, C., Brunner, M., & Preckel, F. (2014). The longitudinal interplay of students' academic self-concepts and achievements within and across domains: Replicating and extending the reciprocal internal/external frame of reference model. *Journal of Educational Psychology*, *106*(4), 1170–1191. https://doi.org/10.1037/a0036307

Niepel, C., Marsh, H. W., Guo, J., Pekrun, R., & Möller, J. (2022). Revealing dynamic relations between mathematics self-concept and perceived achievement from lesson to lesson: An experience-sampling study. *Journal of Educational Psychology*, *114*(6), 1380–1393. https://doi.org/10.1037/edu0000716

Niepel, C., Stadler, M., & Greiff, S. (2019). Seeing is believing: Gender diversity in STEM is related to mathematics self-concept. *Journal of Educational Psychology*, *111*(6), 1119–1130. https://doi.org/10.1037/edu0000340

OECD. (2014). *PISA 2012: Technical report*. OECD Publishing.

OECD. (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence. PISA*. OECD. https://doi.org/10.1787/9789264229945-en

Paulick, I., Großschedl, J., Harms, U., & Möller, J. (2017). How teachers perceive their expertise: The role of dimensional and social comparisons. *Contemporary Educational Psychology*, *51*, 114–122.

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-Scale Assessments in Education*, *2*(1), 1–17. https://doi.org/10.1186/s40536-014-0005-4

Peterson, S. E., & Miller, J. A. (2004). Quality of college students' experiences during cooperative learning. *Social Psychology of Education*, *7*(2), 161–183. https://doi.org/10.1023/B:SPOE.0000018522.39515.19

Podsakoff, N. P., Spoelma, T. M., Chawla, N., & Gabriel, A. S. (2019). What predicts within-person variance in applied psychology constructs? An empirical examination. *The Journal of Applied Psychology*, *104*(6), 727–754. https://doi.org/10.1037/apl0000374

Pohlmann, B., Möller, J., & Streblow, L. (2005). Bedingungen leistungsbezogenen Verhaltens im Sportunterricht. *Zeitschrift Für Sportpsychologie*, *12*(4), 127–134. https://doi.org/10.1026/1612-5010.12.4.127

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322–338. https://doi.org/10.1037/a0014996

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Quiroz, J. C., Geangu, E., & Yong, M. H. (2018). Emotion recognition using smart watch sensor data: Mixed-design study. *JMIR Mental Health*, *5*(3), e10153. https://doi.org/10.2196/10153

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Randall, R., & Sharples, D. (2012). The impact of rater agreeableness and rating context on the evaluation of poor performance. *Journal of Occupational and Organizational Psychology*, *85*(1), 42–59. https://doi.org/10.1348/2044-8325.002002

Rauthmann, J. F. (2021). Capturing interactions, correlations, fits, and transactions: A person-environment relations model. In J. F. Rauthmann (Ed.), *The handbook of personality dynamics and processes* (pp. 427–522). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-813995-0.00018-2

Rauthmann, J. F., & Sherman, R. (2019). Toward a research agenda for the study of situation perceptions: A variance componential framework. *Personality and Social Psychology Review*, *23*(3), 238–266. https://doi.org/10.1177/1088868318765600

Rauthmann, J. F., Sherman, R. A., Nave, C. S., & Funder, D. C. (2015). Personality-driven situation experience, contact, and construal: How people's personality traits predict characteristics of their situations in daily life. *Journal of Research in Personality*, *55*, 98–111. https://doi.org/10.1016/j.jrp.2015.02.003

Renner, K.-H., Hock, M., Bergner-Köther, R., & Laux, L. (2018). Differentiating anxiety and depression: The State-Trait Anxiety-Depression Inventory. *Cognition & Emotion*, *32*(7), 1409–1423. https://doi.org/10.1080/02699931.2016.1266306

Renner, K.-H., Klee, S., & Oertzen, T. von (2020). Bringing back the person into behavioural personality science using big data. *European Journal of Personality*, *34*(5), 670–686. https://doi.org/10.1002/per.2303

Retelsdorf, J., Schwartz, K., & Asbrock, F. (2015). "Michael can't read!" Teachers' gender stereotypes and boys' reading self-concept. *Journal of Educational Psychology*, *107*(1), 186–194. https://doi.org/10.1037/a0037107

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, *31*(2), 226–235. https://doi.org/10.1037/pas0000662

Roloff, J., Klusmann, U., Lüdtke, O., & Trautwein, U. (2020). The predictive validity of teachers' personality, cognitive and academic abilities at the end of high school on instructional quality in Germany: A longitudinal study. *AERA Open*, *6*(1). https://doi.org/10.1177/2332858419897884

Rost, D. H., & Sparfeldt, J. R. (2002). Facetten des schulischen Selbstkonzepts. *Diagnostica*, *48*(3), 130–140. https://doi.org/10.1026//0012-1924.48.3.130

Ruzek, E., Aldrup, K., & Lüdtke, O. (2022). Assessing the effects of student perceptions of instructional quality: A cross-subject within-student design. *Contemporary Educational Psychology*, 102085. https://doi.org/10.1016/j.cedpsych.2022.102085

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 583–601. https://doi.org/10.1080/10705510903203466

Scherer, R., & Gustafsson, J.-E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: An application of multilevel bifactor structural equation modeling. *Frontiers in Psychology*, *6*, 1550. https://doi.org/10.3389/fpsyg.2015.01550

Scherer, R., & Nilsen, T. (2016). The relations among school climate, instructional quality, and achievement motivation in mathematics. In T. Nilsen & J.-E. Gustafsson (Eds.), *IEA Research for Education. Teacher quality, instructional quality and student outcomes* (Vol. 2, pp. 51–80). Springer International Publishing. https://doi.org/10.1007/978-3-319-41252-8_3

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology*, *7*, 110. https://doi.org/10.3389/fpsyg.2016.00110

Schilling, S. R., Sparfeldt, J. R., & John, M. (2005). Besser in Mathe, besorgter in Deutsch? – Beziehungen zwischen Schulleistungen, Selbstkonzepten und Prüfungsängsten im Rahmen des I/E-Modells. In S. R. Schilling, J. R. Sparfeldt, & C. Pruisken (Eds.), *Aktuelle Aspekte pädagogisch-psychologischer Forschung. Detlef H. Rost zum 60. Geburtstag* (pp. 159–178). Waxmann.

Schmitz, B. (2006). Advantages of studying processes in educational research. *Learning and Instruction*, *16*(5), 433–449. https://doi.org/10.1016/j.learninstruc.2006.09.004

Schneider, R., Sparfeldt, J. R., Niepel, C., Buch, S. R., & Rost, D. H. (2022). Measurement invariance of test anxiety across four school subjects. *European Journal of Psychological Assessment*, *38*(5), 356–364. https://doi.org/10.1027/1015-5759/a000676

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499. https://doi.org/10.3102/0034654307310317

Sherman, R. A., Rauthmann, J. F., Brown, N. A., Serfass, D. G., & Jones, A. B. (2015). The independent effects of personality and situations on real-time expressions of behavior and emotion. *Journal of Personality and Social Psychology*, *109*(5), 872–888. https://doi.org/10.1037/pspp0000036

Shernof, D. J., Ruzek, E. A., Sannella, A. J., Schorr, R. Y., Sanchez-Wall, L., & Bressler, D. M. (2017). Student engagement as a general factor of classroom experience: Associations with student practices and educational outcomes in a university gateway course. *Frontiers in Psychology*, *8*, 994. https://doi.org/10.3389/fpsyg.2017.00994

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Sparfeldt, J. R., Buch, S. R., Rost, D. H., & Lehmann, G. (2008). Akkuratesse selbstberichteter Zensuren [The accuracy of self-reported grades in school]. *Psychologie in Erziehung Und Unterricht*, *55*, 68–75.

Sparfeldt, J. R., Rost, D. H., Baumeister, U. M., & Christ, O. (2013). Test anxiety in written and oral examinations. *Learning and Individual Differences*, *24*, 198–203. https://doi.org/10.1016/j.lindif.2012.12.010

Sparfeldt, J. R., Schilling, S. R., Rost, D. H., Stelzl, I., & Peipert, D. (2005). Leistungsängstlichkeit: Facetten, Fächer, Fachfacetten? Zur Trennbarkeit nach Angstfacette und Inhaltsbereich [Test anxiety: The relevance of anxiety facets as well as school subjects]. *Zeitschrift Für Pädagogische Psychologie [German Journal of Educational Psychology]*, *19*(4), 225–236. https://doi.org/10.1024/1010-0652.19.4.225

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality*, *34*(5), 613–631. https://doi.org/10.1002/per.2257

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520. https://doi.org/10.3102/1076998616646200

Stawski, R. S., MacDonald, S. W. S., & Sliwinski, M. J. (2015). Measurement burst design. In S. K. Whitbourne (Ed.), *The encyclopedia of adulthood and aging* (pp. 1–5). John Wiley & Sons Inc. https://doi.org/10.1002/9781118521373.wbeaa313

Steinmayr, R., Crede, J., McElvany, N., & Wirthwein, L. (2016). Subjective well-being, test anxiety, academic achievement: Testing for reciprocal effects. *Frontiers in Psychology*, *6*(1994), 1–13. https://doi.org/10.3389/fpsyg.2015.01994

Sticca, F., Goetz, T., Bieg, M., Hall, N. C., Eberle, F., & Haag, L. (2017). Examining the accuracy of students' self-reported academic grades from a correlational and a discrepancy perspective: Evidence from a longitudinal study. *PloS One*, *12*(11), e0187367. https://doi.org/10.1371/journal.pone.0187367

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study. Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States*. U.S. Government Printing Office.

Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, *24*(3), 236–243. https://doi.org/10.1207/S15324796ABM2403_09

Streblow, L. (2004). *Bezugsrahmen und Selbstkonzeptgenese. Pädagogische Psychologie und Entwicklungspsychologie: Vol. 42*. Waxmann.

Sutin, A. R., Stephan, Y., Luchetti, M., Strickhouser, J. E., Aschwanden, D., & Terracciano, A. (2022). The association between five factor model personality traits and verbal and numeric reasoning. *Aging, Neuropsychology and Cognition*, *29*(2), 297–317. https://doi.org/10.1080/13825585.2021.1872481

Talić, I., Scherer, R., Marsh, H. W., Greiff, S., Möller, J., & Niepel, C. (2022). Uncovering everyday dynamics in students' perceptions of instructional quality with experience sampling. *Learning and Instruction*, *81*, 101594. https://doi.org/10.1016/j.learninstruc.2022.101594

Toropova, A., Johansson, S., & Myrberg, E. (2019). The role of teacher characteristics for student achievement in mathematics and student perceptions of instructional quality. *Education Inquiry*, *10*(4), 275–299. https://doi.org/10.1080/20004508.2019.1591844

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, *9*, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science*, *23*(6), 466–470. https://doi.org/10.1177/0963721414550706

UNESCO. (2017). *Cracking the code: Girls' and women's education in science, technology, engineering and mathematics (STEM)*. UNESCO.

van der Westhuizen, L., Arens, A. K., Greiff, S., Fischbach, A., & Niepel, C. (2022). The generalized internal/external frame of reference model with academic self-concepts, interests, and anxieties in students from different language backgrounds. *Contemporary Educational Psychology*, *68*, 102037. https://doi.org/10.1016/j.cedpsych.2021.102037

von der Embse, N., Barterian, J., & Segool, N. (2013). Test anxiety interventions for children and adolescents: A systematic review of treatment studies from 2000-2010. *Psychology in the Schools*, *50*(1), 57–71. https://doi.org/10.1002/pits.21660 von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, *28*, 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, *108*(5), 705–721. https://doi.org/10.1037/edu0000075

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, *63*(3), 249. https://doi.org/10.2307/1170546

Wen, C. K. F., Schneider, S., Stone, A. A., & Spruijt-Metz, D. (2017). Compliance with mobile ecological momentary assessment protocols in children and adolescents: A systematic review and meta-analysis. *Journal of Medical Internet Research*, *19*(4), e132. https://doi.org/10.2196/jmir.6641

Wisniewski, B., Röhl, S., & Fauth, B. (2022). The perception problem: a comparison of teachers' self-perceptions and students' perceptions of instructional quality. *Learning Environments Research*, *25*(3), 775–802. https://doi.org/10.1007/s10984-021-09397-4

Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. (2020). Obtaining secondary students' perceptions of instructional quality: Two-level structure and measurement invariance. *Learning and Instruction*, *66*, 101303. https://doi.org/10.1016/j.learninstruc.2020.101303

Wolff, F., & Möller, J. (2021). Dimensional comparison theory: Minimal intervention affects strength of dimensional comparison effects. *The Journal of Experimental Education*, *89*(4), 625–642. https://doi.org/10.1080/00220973.2020.1843128

Wolff, F., Sticca, F., Niepel, C., Götz, T., van Damme, J., & Möller, J. (2020). The reciprocal 2I/E model: An investigation of mutual relations between achievement and self-concept levels and changes in the math and verbal domain across three countries. *Journal of Educational Psychology*, Advance online publication. https://doi.org/10.1037/edu0000632

Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment*, *13*(2), 97–107. https://doi.org/10.1111/j.0965-075X.2005.00304.x

Zeidner, M. (1998). *Test anxiety: The state of the art. Perspectives on individual differences*. Springer. http://lib.myilibrary.com/detail.asp?id=20699

Zeidner, M. (2020). Test anxiety. In B. J. Carducci, C. S. Nave, A. Di Fabio, D. H. Saklofske, & C. Stough (Eds.), *The Wiley encyclopedia of personality and individual differences: Personality processes and individual differences, Volume III* (pp. 445–449). John Wiley & Sons.

Ziegler, M., & Bühner, M. (2012). *Grundlagen der Psychologischen Diagnostik. SpringerLink Bücher*. VS Verl. für Sozialwiss. https://doi.org/10.1007/978-3-531-93423-5

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice*, *12*(2), 127–140. https://doi.org/10.1037/1089-2699.12.2.127