Learning and Instruction 52 (2017) 148–160

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Identifying determinants of teachers' judgment (in)accuracy regarding students' school-related motivations using a Bayesian cross-classified multi-level model

Anna-Katharina Praetorius [a], [*], Tobias Koch [b], Annette Scheunpflug [c], Horst Zeinz [d], Markus Dresel [e]

[a] German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany
[b] Center for Methods, Leuphana University of Lüneburg, Germany
[c] Department of Education, University of Bamberg, Germany
[d] Department of Education, University of Münster, Germany
[e] Department of Psychology, University of Augsburg, Germany

ARTICLE INFO

Article history:
Received 24 June 2016
Received in revised form
3 June 2017
Accepted 10 June 2017
Available online 30 June 2017

Keywords:
Judgment accuracy
Teacher judgments
Cross-classification
Motivation
Variance sources

ABSTRACT

Teachers differ considerably in their judgment accuracy of motivational student characteristics. Thus far, only few investigations have focused on explaining these differences. In this study, we investigated to what extent groups of characteristics (i.e., student, information, teacher, and class characteristics) derived from the Realistic Accuracy Model (Funder, 1995) are relevant for explaining differences in teachers' judgment accuracy regarding students' school-related self-concept and autonomous motivation. Data from 1239 students and 341 teachers were analyzed using a Bayesian cross-classified multi-level modeling approach. Our analyses showed that variance in teacher judgments is largely due to variation at the level of judgments and less due to variation in the slope (i.e., the accuracy of teacher judgments). Teachers' judgment accuracy varied to a comparable degree across teachers and classes. Significant determinants for these differences were teachers' subject and students' grade point average.

1. Introduction

Teachers are required to accomplish a wide variety of tasks such as instructing, fostering and educating students as well as counseling students and parents. They need to know their students very well to address these issues effectively, thus calling for a high judgment accuracy regarding different student characteristics. Previous studies have, however, shown that teachers are on average rather inaccurate in judging student characteristics, especially when it comes to judging school-related as well as subject-specific motivational characteristics (e.g., Praetorius, Berner, Zeinz, Scheunpflug, & Dresel, 2013; Spinath, 2005; Urhahne, Chao, Florineth, Luttenberger, & Paechter, 2011). The empirical evidence moreover indicates large differences in teachers' judgment accuracy. So far, these differences have not yet been successfully explained (see also Südkamp, Kaiser, & Möller, 2012). Knowledge about these differences would serve to identify aspects that should be focused when enhancing teachers' judgmental abilities. For advancing the field, it seems pivotal to base respective research on a theoretical model (see Chaplin, 1991; Funder & Colvin, 1997) that is combined with a sound methodological approach.

In the present study, the Realistic Accuracy Model (Funder, 1995) is proposed as a suitable theoretical framework for explaining differences in teacher judgments of students' motivational characteristics. We chose a cross-classified multi-level modeling approach to investigate determinants of teachers' judgment accuracy systematically by disentangling variance sources within students (e.g., gender) from variance sources among classes (e.g., class heterogeneity) and among teachers (e.g., teaching experience). Previous methodological approaches had concentrated on investigating one teacher per class with respect to his or her judgment accuracy, thus confounding class and teacher variance sources. For disentangling

[*] Corresponding author. German Institute for International Educational Research (DIPF), Schlossstrasse 29, 60486 Frankfurt, Germany.
E-mail address: praetorius@dipf.de (A.-K. Praetorius).

http://dx.doi.org/10.1016/j.learninstruc.2017.06.003
0959-4752/© 2017 Elsevier Ltd. All rights reserved.

them, cross-classified designs include data of several teachers judging every student and all teachers judging several students. A necessary precondition for using this cross-classified approach is the focus on student characteristics that can be judged by all teachers teaching a class (e.g., general cognitive abilities or school-related motivation). To our knowledge, this approach has so far not been applied to judgment accuracy. The present paper is thus the first to do so, focusing on school-related motivations of students.

In the following, we first elaborate on the importance of students' school-related motivations, second on the relevance of teachers' abilities to judge respective characteristics, and third on the degree of teachers' judgment accuracy. We also present the Realistic Accuracy Model as an explanation for differences in teachers in their accuracy and finally derive the research questions and approach pursued by our study.

1.1. Conceptualizing students' motivations

Students' school-related (i.e., related to school in general) as well as subject-specific achievement motivation plays a pivotal role for students' task choice, their learning behavior, as well as their performance (Wigfield & Eccles, 2000). Different constructs have been proposed in the literature to describe aspects of students' motivation. In this regard, students' academic self-concept (Marsh, 1990) and students' autonomous motivation (Deci & Ryan, 2000) are considered to be particularly important, representing expectancy as well as value aspects of motivation (see Wigfield & Eccles, 2000).

Self-concepts comprise one's mental cognitive representations of the own abilities in academic domains. From the outset, self-concept research (James, 1892/1999) has focused on domain-specific and more general aspects of self-concepts, which both have shown to be relevant for humans' cognition and behavior. In the academic context, the early work by Shavelson, Hubner, and Stanton (1976) focused on self-concept aspects with different degrees of specificity, including the general, school-related academic self-concept. Recent research such as Brunner et al. (2010) also points to the relevance of both aspects, with the general, school-related academic self-concept accounting for a substantial amount of variance in all subject-specific measures and, thus, considerably influencing subject-specific self-concepts. Regarding effects of academic self-concepts, both school-related and subject-specific aspects are relevant, among others for the quality and persistence of learning efforts as well as achievement (e.g., Dickhäuser & Reinhard, 2006; Guay, Marsh, & Boivin, 2003; Retelsdorf, Köller, & Möller, 2014; Sparfeldt, Schilling, & Rost, 2003)[1].

Although the simultaneous existence of motivational characteristics on a more general academic level and on a more subject-specific level was theorized and analyzed most explicitly for students self-concepts, this rationale also applies to other motivational characteristics, particularly to value-related constructs such as students' autonomous motivation (see e.g., Gottfried, 1985). Autonomous motivation comprises intrinsic motivation (i.e., enjoyment of dealing with school content) and identified motivation (i.e., personal importance of dealing with school content; for an overview, see Deci & Ryan, 2000; see also Reindl, Berner, Scheunpflug, Zeinz, & Dresel, 2015). Prior studies have shown that autonomous motivation exists on both the school-related and the subject-specific level (Bong, 2001); additionally, both aspects are related to learning behavior and achievement (e.g., Fortier, Vallerand, & Guay, 1995).

To sum up, the literature clearly indicates the high relevance of motivational characteristics on a general school-related and a subject-specific level. In the present study, we focus on the general level as this is a necessary precondition for disentangling the various sources of judgment accuracy (see section 1.5).

1.2. Relevance of teachers' judgment accuracy regarding students' motivations

A vast number of studies has shown that students' school-related and subject-specific motivation can be positively influenced through teachers' actions in instructional settings (see e.g., Ames, 1992; Dickhäuser & Stiensmeier-Pelster, 2003; O'Mara, Marsh, Craven, & Debus, 2006; Pintrich, 2003; Skinner & Belmont, 1993). Teaching behaviors that support students' motivation in a comprehensive way seem particularly promising because they address both expectancy and value aspects of student motivation. Students differ in their motivation both on the general school level as well as regarding different subjects, and teachers need to tailor their support, feedback, explanations, and task selection specifically towards their students' motivation for fostering them optimally. Research has shown that such adaptive teaching is related to positive motivational and achievement-related outcomes (for an overview, see Corno, 2008). Tomlinson et al. (2003) argued in this regard that “equality of opportunity becomes a reality only when students receive instruction suited to their varied readiness levels, interests, and learning preferences, thus enabling them to maximize the opportunity for growth” (p. 120). More tangibly, acting in an adaptive way with respect to students' motivations could for example mean helping a student who thinks very low of his or her academic abilities to experience competence by using contingent praise and attributional feedback (see O'Mara et al., 2006). Similarly, students who have a low academic autonomous motivation for school-related activities would benefit from teachers' making them realize the relevance of school content for their own lives (see Hulleman & Harackiewicz, 2009).

To help students develop optimally, teachers need to accurately judge motivational student characteristics: The same behavior (e.g., not engaging in academic class discussions) can have different motivational causes (e.g., low self-concept or low autonomous motivation; see also Givvin, Stipek, Salmon, & MacGyvers, 2001). Teachers thus need to know about their students' motivational characteristics – alongside knowledge about their cognitive abilities and achievements (see the meta-analyses of Machts, Kaiser, Schmidt, & Möller, 2016; Südkamp et al., 2012) – to adapt their teaching behavior appropriately (e.g., Givvin et al., 2001; Karing, 2009; Praetorius et al., 2015; Spinath, 2005; Urhahne & Zhu, 2015). Teachers' abilities to judge their students' motivational characteristics are also deemed important by the Standards for Teacher Competence in Educational Assessment of Students (American Federation of Teachers, the National Council on Measurement in Education, & the National Education Association, 1990) or the Standards of the Interstate New Teachers Assessment and Support Consortium (Council of Chief State School Officers, 2013). Empirical evidence, too, points towards a relation between teachers' judgment accuracy of student achievement as well as student motivation, teaching behavior, and student outcomes (e.g., Behrmann & Souvignier, 2013; Helmke & Schrader, 1987; Praetorius, Scheunpflug, Zeinz, & Dresel, 2015). Finally, recent studies have shown that teachers use their judgments of

[1] In the study of Guay et al. (2003), for example, latent correlations between the school-related academic self-concept and achievement ranged between $0.31 \leq r \leq 0.73$. In a similar but domain-specific study by Retelsdorf et al. (2014), the correlations between reading self-concept and reading achievement ranged between $0.45 \leq r \leq 0.53$.

students' motivational characteristics in the context of teachers' school tracking recommendations (Böhmer, Hörstermann, Gräsel, Krolak-Schwerdt, & Glock, 2015; Pohlmann, 2009), emphasizing once more the relevance of accurate teacher judgments of motivational characteristics.

1.3. Accuracy of teachers' judgments

The common approach to investigating the accuracy of teacher judgments is to compare teachers' judgments to student tests (for achievement) or student self-reports (for motivational characteristics) using intra-individual correlations. Based on such self-other agreements, previous studies have shown that teachers' judgment accuracy differs depending on the student characteristic to be judged. Compared to their judgments of student achievement (e.g., mean correlation $r = 0.63$ in the meta-analysis of Südkamp et al., 2012), teachers are on average rather inaccurate in judging motivational characteristics. For students' school-related and subject-specific academic self-concepts, correlations between teacher judgments and their students' self-reports were found to be small to medium. Spinath (2005) and Urhahne et al. (2011), for example, reported mean correlations of $r = 0.39$ and $r = 0.43$ for the school-related academic self-concept, respectively. Praetorius et al. (2013) found correlations of $r = 0.22$ for the self-concept in the subject of German and $r = 0.27$ for the self-concept in Mathematics. For learning motivation on the school-level (defined by intrinsic motivation as well as learning goals), Urhahne et al. (2011) revealed a correlation of $r = 0.10$ between teachers' judgments and students' self-reports. Spinath (2005) reported a correlation of $r = 0.20$. Teachers' judgments of their students' academic interest (defined as enjoyment in participating in academic instruction) correlated with the respecting students' self-reports at $r = 0.21/.30$ for German and $r = 0.32/.37$ for Mathematics, according to a study by Karing (2009).

Summing up, teachers' judgment accuracy is considerably lower for motivational student characteristics compared to student achievement. This might be explained by the fact that teachers regularly collect information about their students' achievement (Alvidrez & Weinstein, 1999), based on a wide range of different indicators such as tests and verbal behavior. Students' motivational characteristics are way more difficult to judge as they are primarily inner states and do not necessarily need to be consistent with students' achievement (Givvin et al., 2001; Praetorius et al., 2015; Urhahne & Zhu, 2015). Despite these challenges, a series of motivation-related verbal and non-verbal behaviors indicate the current motivational status of students. These can be processed by teachers; this is especially true for students' task choice, persistence, verbalized causal attributions, and expression of emotions (see Berhenke, Miller, Brown, Seifer, & Dickstein, 2011). Teachers need, however, to explicitly focus on these types of information to form a judgment about their students' motivation.

Whereas teachers on average judge students' motivations inaccurately, this is not the case for all teachers. Previous studies have shown substantial variations of teachers in their judgment accuracy. Spinath (2005), for example, presented a study wherein the intra-individual correlations between teacher judgments and students' school-related motivational characteristics ranged between $-0.39 \leq r \leq 0.82$ for students' self-concepts and between $-0.46 \leq r \leq 0.67$ for students' learning motivation. Urhahne and Zhu (2015) reported on similar ranges for positive attitude towards school ($-0.14 \leq r \leq 0.89$). These studies thus indicate large differences in teachers' judgment accuracy regarding student motivational characteristics. Explanations of these differences might be useful for teaching candidates and teachers regarding the enhancement of their judgmental abilities.

1.4. Determinants of teachers' judgment accuracy

One of the most prominent models for explaining the accuracy of individuals' judgments is the Realistic Accuracy Model developed by Funder (1995). According to this model, four steps are necessary for accurate judgments (see also Fig. 1): First, relevant behavior has to be shown (relevance), second, this behavior must be available for the judge (availability), third, the judge has to detect the behavioral information (detection), and fourth, it is necessary that the information is utilized correctly (utilization). Funder (1995) distinguished four different groups of explanatory characteristics that are associated with these steps which he named good trait, good target, good information, and good judge. In the following, evidence regarding teachers' judgment accuracy is summarized for each of them.

Traits differ in how easily they can be judged, thus to what degree relevant behavior is connected to the trait and to what degree this behavior is available to a judge (Funder, 2012). This also holds true for teacher judgments, as can be inferred from the large differences in the average judgment accuracy of teachers for student achievement compared to motivational characteristics (see section 1.3).

Targets differ in the degree to which they show relevant behavior for the trait to be judged and also in the extent to which this behavior is available to a judge. To our knowledge, only a study by Urhahne and Zhu (2015) has so far identified target differences in teachers' judgment accuracy for motivational characteristics; this study indicated that girls are rated more accurately than boys concerning a specific aspect of their well-being (i.e., self-reported social problems in school). Other studies have focused on target differences in achievement-related judgment accuracy, indicating that students with minority status (e.g., ethnicity or disability; Hurwitz, Elliott, & Braden, 2007; Kaiser, Südkamp, & Möller, 2017) as well as younger students (e.g., Kenny & Chekaluk, 1993; Martinéz, Stecher, & Borko, 2009; see, however, Maguin & Loeber, 1996) are judged more accurately.

Additionally, the quantity and quality of *information* that is available to judges may differ (Funder, 1995). A characteristic that has been investigated for teachers' achievement-related judgment accuracy is the school subject: In major subjects, teachers teach their students many hours a week and thus get more information about their students than teachers of minor subjects. Accordingly, higher judgment accuracy regarding achievement has been found for teachers of major compared to minor subjects (Hopkins, George, & Williams, 1985). However, a study of Praetorius et al. (2015) revealed that people who saw a 30-s video of students demonstrated a similar accuracy in judging students' school-related academic self-concepts compared to teachers who judged their own students, contradicting the idea that more information about students is related to higher judgment accuracy.

Judges differ in their ability to detect and utilize relevant student behavior for their judgments. One of the most often investigated teacher characteristics here is teaching experience. However, existing studies found no relation between teaching experience and judgment accuracy, for judging students' self-concepts (Praetorius, Karst, Dickhäuser, & Lipowsky, 2011) but also for students' achievement (e.g., Feinberg & Shapiro, 2003). All other evidence on judge-related characteristics focuses on teachers' achievement-related judgment accuracy. With the exception of cognitive abilities (see Kaiser, Helm, Retelsdorf, Südkamp, & Möller, 2012), none of the tested teacher characteristics – neither perfectionism nor the self-reported ability of perspective-taking (see

Lorenz, 2011) — was related to judgment accuracy.

The groups of characteristics mentioned by Funder (1995) that are assumed to influence judgment accuracy have been developed for personality judgments, the typical situation being that judges judge personality characteristics of individual targets. The situation is different for teachers. Here, usually many students in a certain class are judged by teachers. Thus, the specific class environment might play an important role for explaining differences in judgment accuracy. It therefore seems to be fruitful to focus also on aspects related to the *environment* when explaining teachers' judgment accuracy. Because the environment is important for both targets and judges, environmental aspects are relevant in all four steps of the RAM model. The existing evidence for the relation between environment characteristics and teachers' judgment accuracy focuses exclusively on student achievement: For class size, the findings are inconsistent (e.g., Weinert & Schrader, 1986; Wild & Rost, 1995) whereas the achievement heterogeneity of a class has consistently been found to be positively related to judgment accuracy (e.g., Karing, 2009; Weinert & Schrader, 1986).

Taken together, only few studies were conducted investigating determinants of teachers' judgment accuracy, and results are mixed. By contrast, many studies focus on explaining differences in the level of teacher judgments (e.g., higher achievement judgments for girls, see Hinnant, O'Brien, & Ghazarian, 2009; Tiedemann, 2002). Only very few of the studies on determinants of teachers' judgment accuracy have focused on motivational student characteristics. Our knowledge about determinants of teachers' judgment accuracy regarding motivational characteristics is thus limited.

1.5. The present study

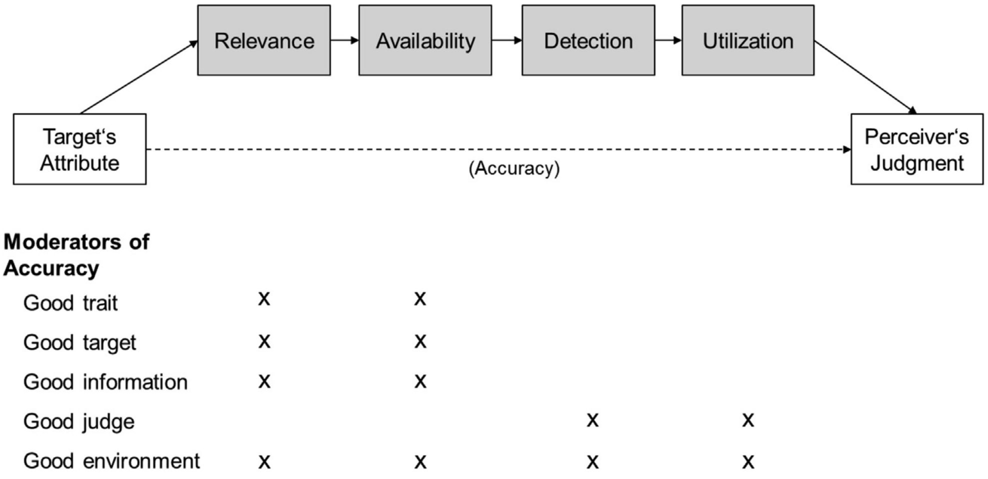
Based on the RAM model (Funder, 1995), the present study aims at explaining differences in teachers' judgment accuracy regarding two school-related motivational student characteristics, namely students' academic self-concept and students' autonomous motivation for school-related activities. Two characteristics instead of one were chosen to check to what extent findings are specific to a certain motivational student characteristic (traits according to the RAM model). The selected characteristics had been focused by studies on teachers' judgment accuracy before (see section 1.4) and represent pivotal aspects of student motivation.

To distinguish the different aspects in the RAM model and to identify specific determinants, certain design features need to be met. Teacher characteristics (judges) and class characteristics (environment) have often been confounded in previous studies (see e.g., Martinéz et al., 2009). To separate student effects (targets) from teacher effects (judges) and classroom effects (environment), an innovative cross-classified multi-level approach was chosen with teachers judging several students and every student being judged by several teachers. In selecting determinants, we ensured that all characteristics and steps in the RAM model were covered.

Targets. Students' age and gender are two of the most relevant characteristics regarding the RAM model, according to prior studies on achievement-related teacher judgments. Female students might show more behavior that is related to their self-concept and their autonomous motivation as it is socially more accepted for girls to show emotions related to their motivation (Hypothesis 1; see Decuir-Gunby & Williams-Johnson, 2014). Motivational characteristics approximate achievement over time and thus can be assumed to map more closely to observable behavior, therefore older students might be judged more accurately than younger students (Hypothesis 2; see e.g., Marsh & Craven, 1991). According to Funder (2012), one of the most important aspects for the availability step is the consistency of target behavior across situations. A useful indicator in the school context for this consistency is students' grade point average (GPA) across different subjects. Students who consistently perform very well or very poorly (i.e., those with a very high or very low GPA) should be easier to judge than students with an average GPA. Thus, we expect GPA to predict teachers' judgment accuracy in a curvilinear way (Hypothesis 3).

Information. Another important aspect for the availability step is the time teachers spend on interacting with the students they are asked to judge (Funder, 2012). In accordance with the RAM model, teachers teaching major subjects should judge their students' self-concepts and autonomous motivation more accurately than teachers teaching minor subjects (Hypothesis 4). We formulated the hypothesis in accordance with theory instead of the mixed evidence regarding teachers' judgment accuracy (Hopkins et al., 1985; Praetorius et al., 2015), as empirical findings might be contradictory due to the methods used.

Judges. Whether or not a teacher detects relevant student behavior depends, among others, on the teacher's capacities to focus attention on individual students. If a teacher focusses mainly on herself or himself due to low confidence in the own teaching competencies (i.e., low self-efficacy), the detection of relevant behavior in the student is unlikely (Hypothesis 5). Whether a teacher can utilize the detected behavior in an appropriate way should depend largely on his or her experience with the range of



Moderators of Accuracy				
Good trait	x	x		
Good target	x	x		
Good information	x	x		
Good judge			x	x
Good environment	x	x	x	x

Fig. 1. The Realistic Accuracy Model and Relevant Moderator Groups (adapted from Funder, 2012, p. 178).

possible behaviors shown by students. The utilization of behavior should thus depend on the teacher's job experience (Hypothesis 6). Again, we formulated this hypothesis in accordance with the theoretical assumptions by Funder rather than prior non-significant findings (e.g., Feinberg & Shapiro, 2003; Praetorius, Greb, Dickhäuser, & Lipowsky, 2011) that may as well have resulted from underpowered designs or suboptimal analytical approaches.

Environment. Possibly, students might not show motivation-related behavior regardless of circumstances. Classroom climate (i.e., the quality of social interactions and relationships between teacher and students, as well as among students, see e.g., Reyes, Brackett, Rivers, White, & Salovey, 2012) constitutes an important classroom characteristic that might influence to what degree relevant behavior for self-concepts and autonomous motivation is shown by students. If students feel accepted by their teacher as well as their classroom peers, they are more likely to show how they think and feel (Hypothesis 7).

2. Method

2.1. Sample

In the study presented here, 15 intermediate track schools ("Realschulen") took part from the German state of Bavaria. We removed one school from the initial sample because for this school, only data from four students was available.

The study from which the data were drawn for the present work encompassed three measurement points (see Reindl et al., 2015). The third measurement point was designed specifically in a cross-classified manner to allow for the analysis of the present study; it was therefore selected for analysis. In total, data from 1239 students (49.5% female, 1 participant did not report gender) from 123 classes (grades 5 through 9) were assessed in this manner. They were, on average, 14.2 years old ($s = 1.3$). Each student was rated by three to five teachers teaching different subjects. Concurrently, each teacher rated several students ($M = 7.0$; range: 2–14). In total, judgment data of 341 teachers (66.7% female; 5.1% missing data) were analyzed. The teachers had been professionally active for 9.7 years on average (range: 1–38 years).

Student data were collected in the classroom. At the same time, teacher questionnaires were distributed and then returned, individually, by the teachers themselves. The language of the questionnaires was German.

2.2. Measurements

Students' academic self-concept was assessed using three items from the well-validated absolute scale of the German SESSKO instrument (Schöne, Dickhäuser, Spinath, & Stiensmeier-Pelster, 2002) using a five-point scale[2]. One example item was "I am for school …" and had to be answered on a scale from 1 (*not talented*) to 5 (*very talented*). The internal consistency of the scale was satisfactory (McDonald's $\omega = 0.77$)[3].

Students' autonomous motivation for school-related activities was measured using three items from the Programme for International Student Assessment (PISA) 2000 survey (Kunter et al., 2003) with a four-point scale ranging from 1 (*disagree*) to 4 (*agree*). One example item was "School is personally important for me". The internal consistency of the scale was satisfactory (McDonald's $\omega = 0.75$).

The teachers were asked to rate their students on every item the students had responded to regarding the academic self-concept as well as the autonomous motivation, regarding all students working on a research project ID with an odd number. A random selection of students per class was used because it would have been too time-consuming to judge the entire class. The internal consistencies were satisfactory (McDonald's $\omega = 0.92$ for the self-concept ratings; $\omega = 0.89$ for the autonomous motivation ratings).

On the student level, we included the following characteristics that might predict judgment accuracy: (a) student gender as a dummy coded variable (1 = female), (b) self-reported grade point average of the grades in the major subjects Mathematics, English as a foreign language (EFL), and German, with higher values indicating better achievement (possible range of grades: 1–6; Math: $M = 3.97$, $s = 0.99$, EFL: $M = 4.01$, $s = 0.83$, German: $M = 4.00$, $s = 0.93$), and (c) student age (in years). On the teacher level, we included (d) job experience (in years), (e) the subject the teacher teaches in the class in which he or she judged the students using a dummy variable (0 = major subject, including 10.86% math teachers, 9.48% English teachers, and 11.24% German teachers; 1 = 68.41% further minor subjects), and (f) teachers' self-efficacy using an eight-item scale with a four-point scale ranging from 1 (*disagree*) to 4 (*agree*) (see Schwarzer & Jersusalem, 1999; e.g., "I know that I am capable of teaching test-relevant content to even the most problematic students", $M = 2.99$, $s = 0.35$, $\omega = 0.73$). On the class level, we included (g) the teacher-reported classroom climate using a four-item scale with a four-point scale ranging from 1 (*disagree*) to 4 (*agree*) (see Clausen, 2002; e.g., "I take time in my instruction for students' personal and social issues", $M = 3.00$, $s = 0.46$, $\omega = 0.66$). For an overview of all items of the scales used see Appendix A.

2.3. Analyses

The data was analyzed using cross-classified multilevel regression models and Bayesian estimation techniques. Teacher judgments of students' self-concepts and autonomous motivation were chosen as the dependent variables. In doing so, we were able to examine the variation in the level of teacher judgment (i.e., intercept of teacher judgment) as well as the variation in judgment accuracy (i.e., regression coefficient of teacher judgment predicted by students' self-ratings) across the different measurement levels. In order to predict the variation in teacher judgment and teacher accuracy, we included explanatory variables[4] into the model. The models were estimated using Markov chain Monte Carlo (MCMC) algorithm and Bayesian estimating techniques, which allow researchers to evaluate the entire posterior density of all model parameters (e.g., fixed and random effects) and compute non-symmetrical credibility intervals (see Gelman & Hill, 2007). All models were estimated using two chains, 30,000 MCMC iterations and a thinning of three (i.e., every 3rd iteration was recorded). The first 10,000 MCMC iterations (out of the entire 30,000) were disregarded and used as burn-in phase. The convergence of the MCMC

[2] The SESSKO instrument takes into account different reference norms a student can compare himself or herself to. Accordingly, it distinguishes four scales: a criterial scale (i.e., comparison with an absolute standard), a social scale (i.e., comparison with other students), an individual scale (i.e., comparison with one's own past achievements), and an absolute scale (i.e., no explicit comparison).

[3] McDonald's ω can be interpreted analogously to Cronbach's α and is more appropriate as it, among others, requires fewer and more realistic statistical assumptions (e.g., Dunn, Baguley, & Brunsden, 2014).

[4] We used a two-step (i.e., factor score regression) approach to account for measurement error influences in the covariates and to avoid a too complex model with many latent variables. In the first step, we computed factor scores for all continuous covariates. In the second step, we used the factor scores as explanatory variables in the cross-classified regression models. Since measurement error influences were rather low, similar results as compared to a classical structural equation modeling approach can be expected.

chains was investigated by visual inspection of the MCMC plots. All models were estimated using the open software program R (R Development Core Team, 2008) rjags (Plummer, 2016), coda (Plummer, Best, Cowles, & Vines, 2006) and mcmcplots (Curtis, 2015). The code for all models and MCMC plots is provided as online supplemental material (Appendix B).

We specified seven cross-classified multilevel regression models for each construct (i.e., self-concept and autonomous motivation). First, we specified two unconditional models (Models 1a and 1b), which were used to examine the variation in teacher judgments (but not in judgment accuracy) across students i and teachers j and classes/schools k. The restricted model (Model 1a) did not include a random effect for classes/schools (i.e., $u_k = 0$), whereas the general model (Model 1b) did:

$$y_{ijk} = \beta_0 + u_i + u_j + u_k + e_{ijk} \quad \text{(Model 1b)} \tag{1}$$

The above Equation (1) states that each rating y_{ijk} belonging to teacher j rating student i in class and school k (combined measurement level) can be decomposed into five components: an overall (or grand)-mean (β_0; overall teacher judgment), a random student effect (i.e., variation of the overall teacher judgment across students; u_i), a random teacher effect (i.e., variation of the overall teacher judgment across teachers; u_j), a random class/school effect (i.e., variation of the overall teacher judgment across classes and schools; u_k), and an error term (e_{ijk}). The error term (ε_{ijk}) captures interaction effects as well as measurement error influences (see e.g., Fielding & Goldstein, 2006). In cross-classified multilevel models, the random effects u_i, u_j, u_k, and e_{ijk} are assumed to be mutually uncorrelated.

To evaluate the amount of variation in the teacher judgments that is attributable to the different levels, we computed different intra-class correlations. The proportion of overall variation in the teacher judgments that is due to teacher characteristics (TC) is:

$$TC = \frac{Var(u_j)}{Var(u_i) + Var(u_j) + Var(u_k) + Var\left(e_{ijk}\right)} = \frac{Var(u_j)}{Var\left(y_{ijk}\right)} \tag{2}$$

and those for students' characteristics (SC) and class/school (CC) characteristics are:

$$SC = \frac{Var(u_i)}{Var\left(y_{ijk}\right)}$$

and

$$CC = \frac{Var(\ u_k)}{Var\left(y_{ijk}\right)} \tag{3}$$

Next, we added students' academic self-concept or autonomous motivation, respectively, as a predictor variable into the model. The continuous predictor variables were centered at the grand mean. Again, we specified two models. Model 2a included random slopes across teachers j, model 2b also included a random slope across classes/schools k.

$$y_{ijk} = \beta_0 + (\beta_1 + \upsilon_{1j} + \upsilon_{1k})s_{c1ik} + u_i + u_j + u_k + e_{ijk} \quad \text{(Model 2b)} \tag{4}$$

The intercept β_0 is the expected rating y_{ijk} for students with an average self-concept or autonomous motivation across students, teachers, and classes/schools. The regression coefficient β_1 between the outcome y_{ijk} and the grand mean centered predictor (s_{c1ik}) represents how well teacher reports can be predicted by students' self-reports, which we will refer to as a measure of judgment accuracy.

Additionally, we included two random effects, that is, a random slopes parameter of teaching accuracy across classes and schools (υ_{1k}), and a random slopes parameter of teaching accuracy (υ_{1j}) across teachers. Note that we did not allow for random slopes with regard to students in Equation (4) because students' self-reports were fixed for each teacher.

Next, we included teacher, class, and student characteristics as well as interaction terms as additional explanatory variables in the model. The interaction terms indicate to what extent the teacher accuracy parameter (β_1) is moderated by teacher, class, or student characteristics. First, we included student characteristics as covariates into the model (Model 3a). Second, we included teacher and class characteristics as additional covariates into the model (Model 3b). Third, we included interaction terms between students' self-reported self-concept or autonomous motivation (i.e., s_{c1ik}) and all covariates (Model 3c). We assumed fixed effects for all additional student variables, teacher variables, and interaction terms, as the corresponding random effects appeared to be low. Furthermore, we fixed the covariances between the random intercepts and random slopes at each measurement level to zero, as they were marginal and non-significant.

In Equation (5) below (model 3c), we included one continuous covariate for teacher characteristics (t_{c1ik}), two continuous covariates for student characteristics (s_{c1ik} and s_{c2ik}), and two interaction terms ($s_{c1ik}*s_{c2ik}$ and $s_{c1ik}*t_{c1jk}$) for reasons of simplicity:

$$y_{ijk} = \beta_0 + (\beta_1 + \upsilon_{1j} + \upsilon_{1k})s_{c1ik} + \beta_2 s_{c2ik} + \beta_3 t_{c1jk} + \beta_4(s_{c1ik}*s_{c2ik}) + \beta_5\left(s_{c1ik}*t_{c1jk}\right) + u_i + u_j + u_k + e_{ijk} \tag{5}$$

The above models were compared by using the Bayesian deviance information criteria (DIC, Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). The DIC is a generalization of classical information criteria (e.g., Akaike information criteria or the Bayesian information criteria) and is also a combined measure of the complexity and the deviance of the model. Models with smaller DIC values should be preferred. Additionally, we computed 95% credibility intervals for all model parameters (i.e., fixed and random effects) using the posterior densities.

3. Results

3.1. Descriptive analyses

Table 1 displays the descriptive statistics for students' academic self-concept, students' autonomous motivation, and the teacher judgments of these two characteristics. In Table 2, the bivariate correlations among these four variables are shown. In line with previous findings, our study suggests that on average, correlations between teacher judgments and students' self-ratings are low to medium-sized.

3.2. Standard deviations and variance decomposition in teacher judgments

The results of the unconditional model (see Equation (1)) and the conditional model using students' self-reports (Equation (4)) are presented in Table 3 for students' academic self-concept and in Table 4 for students' autonomous motivation. A relatively large amount of variance was attributable to variation in the level of teacher judgments, whereas a somewhat smaller amount of variance was due to variation in the accuracy of teacher judgments.

Table 1
Descriptive statistics for the students' self-ratings and teachers' judgments.

Variable	*M*	*s*	Min.	Max.	Skew	Kurtosis	*SE*
Students' academic self-concept	3.53	0.72	1.00	5.00	−0.61	0.95	0.01
Students' autonomous motivation	2.59	0.72	1.00	4.00	−0.26	−0.36	0.01
Teachers' judgments of academic self-concept	3.42	0.83	1.00	4.00	−0.40	−0.04	0.01
Teachers' judgments of autonomous motivation	3.91	0.70	1.00	5.00	−0.63	1.00	0.01

Table 2
Bivariate Correlations between the Student and Teacher Scales Students' academic self-concept.

	Students' academic self-concept	Students' autonomous motivation	Teachers' judgments of autonomous motivation
Students' autonomous motivation	0.33***		
Teachers' judgments of autonomous motivation	0.20***	0.21***	
Teachers' judgments of academic self-concept	0.29***	0.16***	0.59***

Note. The correlations between teacher judgments and student self-reports were calculated intra-individually per teacher. The correlations were Fisher-Z transformed, then averaged, and finally transformed back into a correlation coefficient, which is reported in the table. ***$p < 0.001$.

Table 3
Variance decomposition of the unconditional models (Model 1a and 2a) and the conditional models (Model 1b and 2b) for teacher judgments of students' academic self-concepts.

	Model 1a		Model 1b		Model 2a		Model 2b	
	M [CI]	*SD*	*M [CI]*	*SD*	*M [CI]*	*SD*	*M [CI]*	*SD*
Intercept	3.43* [3.39; 3.48]		3.43* [3.39; 3.48]		3.43* [3.39; 3.47]		3.43* [3.39; 3.48]	
Student self-concept					0.30* [0.25; 0.34]		0.30* [0.25; 0.36]	
Variances Intercept								
Students	0.16* [0.14; 0.18]	0.40	0.15* [0.14; 0.17]	0.39	0.13* [0.11; 0.14]	0.36	0.12* [0.10; 0.14]	0.34
Teacher	0.07* [0.06; 0.09]	0.27	0.07* [0.06; 0.09]	0.26	0.07* [0.06; 0.09]	0.27	0.07* [0.06; 0.09]	0.27
Class/School			0.01* [0.00; 0.01]	0.07			0.00 [0.00; 0.01]	0.05
Interactions/ Residual variance	0.30* [0.29; 0.32]	0.55	0.30* [0.29; 0.32]	0.55	0.30* [29; 0.31]	0.55	0.30* [0.29; 0.31]	0.55
Variances Slope								
Teacher					0.02* [0.00; 0.03]	0.13	0.01* [0.00; 0.03]	0.11
Class/School							0.02* [0.00; 0.05]	0.13
ICC								
Students	0.30* [0.27; 0.32]		0.29* [0.26; 0.32]					
Teachers	0.13* [0.11; 0.16]		0.13* [0.11; 0.16]					
Class/School			0.01* [0.00; 0.03]					
Deviance								
Mean deviance	9571		9570		9474		9480	
Penalty	1031		1031		1038		1030	
Penalized deviance	10602		10601		10512		10511	

Note. *CI* = 95% credibility interval; *SD* = standard deviation. Note that the variances of the random effects were rounded to two decimals. *$p < 0.05$.

Next, we computed the variance coefficients described in section 2.3 (see Equations (2) and (3)). Overall, 28.9% and 27.0% of the total variation in teachers' judgments of students' self-concepts and autonomous motivation were due to student characteristics, respectively; 13.2% and 25.0% were due to teacher characteristics; 0.9% and 2.8% were due to class/school characteristics. Finally, 57.0% and 45.2% were due to interaction effects between teachers and students within classes as well as measurement error. With regard to students' self-concepts, the standard deviations of the random intercepts were considerably larger for teachers (0.27) than for classes/schools (0.05), indicating that the variation of teachers' judgments across classes/schools was marginal. By contrast, the standard deviations of the random slopes were comparable across teachers (0.11) and classes/schools (0.13) (see Model 2b in Table 3). Similar results were found with regard to students' autonomous motivation (see Model 2b in Table 4).

To evaluate the range of teachers' judgment accuracy across teachers and classes/schools, we computed 95% (coverage) intervals. For students' academic self-concept, judgment accuracy ranged from 0.05 to 0.44 across teachers and from 0.02 to 0.47 across classes/schools. For students' autonomous motivation, it ranged from −0.03 to 0.36 across teachers and from −0.01 to 0.33 across classes/schools.

3.3. *Determinants of the teachers' judgment accuracy*

In a next step, we included explanatory variables on the student, teacher, and class level to identify determinants of teachers' judgments as well as their (in)accuracy. The results are provided in Table 5 (for self-concept) and Table 6 (for autonomous motivation).

With regard to students' self-concept, all student characteristics were significant for the *level of teachers' judgments* (see Table 5), indicating higher judgments for girls, younger students, as well as low and high GPA due to a curvilinear effect. Similar results were found with regard to students' autonomous motivation, except for GPA as only high GPA was associated with high judgments (see Table 6). Additionally, both subject and teacher self-efficacy were positively associated with the level of teachers' judgments of students' self-concept (see model 3b in Table 5). With regard to students' autonomous motivation, teaching experience and teacher self-efficacy were positively associated with teacher judgments.

With respect to explaining differences in *teachers' judgment*

Table 4
Variance decomposition of the unconditional models (Model 1a and 2a) and the conditional models (Model 1b and 2b) for teacher judgements of students' autonomous motivation.

	Model 1a		Model 1b		Model 2a		Model 2b	
	M [CI]	*SD*	*M [CI]*	*SD*	*M [CI]*	*SD*	*M [CI]*	*SD*
Intercept	2.83* [2.79; 2.87]		2.83* [2.78; 2.88]		2.83* [2.79; 2.88]		2.83* [2.78; 2.88]	
Student au. motivation					0.17* [0.13; 0.21]		0.17* [0.12; 0.22]	
Variances Intercept								
Students	0.12* [0.10; 0.13]	0.34	0.11* [0.10; 0.12]	0.33	0.10* [0.09; 0.12]	0.32	0.10* [0.09; 0.11]	0.31
Teacher	0.10* [0.08; 0.12]	0.32	0.10* [0.08; 0.12]	0.32	0.10* [0.08; 0.12]	0.32	0.10* [0.08; 0.12]	0.32
Class/School Interactions/			0.01* [0.00; 0.02]	0.10			0.01* [0.00; 0.02]	0.08
Residual variance	0.18* [0.17; 0.19]	0.43	0.18* [0.17; 0.19]	0.43	0.18* [0.17; 0.19]	0.42	0.18* [0.17; 0.19]	0.42
Variances Slope								
Teacher					0.10* [0.08; 0.12]	0.32	0.01* [0.01; 0.02]	0.11
Class/School							0.01* [0.00; 0.03]	0.10
ICC								
Students	0.29		0.27					
Teachers	0.26		0.25					
Class/School			0.03					
Deviance								
Mean deviance	6588		6582		6462		6466	
Penalty	1116		1113		1171		1164	
Penalized deviance	7703		7695		7633		7630	

Note. *CI* = 95% credibility interval; *SD* = standard deviation; au. = autonomous. Note that the variances of the random effects were rounded to two decimals. $^{*}p < 0.05$.

Table 5
Predicting teacher judgments of students' academic self-concept.

	Model 3a	Model 3b	Model 3c
Intercept	3.37* [3.32; 3.41]	3.31* [3.25; 3.37]	3.31* [3.24; 3.37]
Student covariates			
Student self-concept	0.19* [0.13; 0.24]	0.19* [0.13; 0.24]	0.22* [0.14; 0.30]
Gender (female)	0.10* [0.05; 0.16]	0.10* [0.05; 0.15]	0.10* [0.05; 0.16]
Age	−0.04* [−0.06; −0.01]	−0.04* [−0.06; −0.01]	−0.04* [−0.06; −0.01]
GPA linear	0.35* [0.30; 0.40]	0.35* [0.30; 0.40]	0.35* [0.30; 0.40]
GPA quadratic	−0.06* [−0.11; −0.01]	−0.06* [−0.11; −0.01]	−0.07* [−0.12; −0.02]
Teacher and class covariates			
Job experience		−0.00 [−0.01; 0.00]	−0.00 [−0.01; 0.00]
Subject (minor)		0.09* [0.05; 0.15]	0.09* [0.04; 0.14]
Self-efficacy		0.05* [0.00; 0.10]	0.05* [0.00; 0.10]
Classroom climate		0.02 [−0.02; 0.06]	0.02 [−0.02; 0.06]
Interaction terms with self-concept			
Gender (female)			0.05 [−0.04; 0.14]
Age			0.01 [−0.03; 0.05]
GPA linear			0.01 [−0.06; 0.07]
GPA quadratic			0.07* [0.02; 0.13]
Job experience			−0.00 [−0.01; 0.00]
Subject (minor)			−0.04 [−0.10; 0.03]
Self-efficacy			0.04 [−0.01; 0.08]
Classroom climate			−0.03 [−0.06; 0.01]
Deviance			
Mean deviance	9503	9497	9498
Penalty	946.1	944.4	943.8
Penalized deviance	10449	10441	10442

Note. Values in brackets are 95% credibility intervals. $^{*}p < 0.05$.

accuracy, only few significant interaction effects could be found. Students' gender (Hypothesis 1) and their age (Hypothesis 2) were, contrary to our hypotheses, not related to teachers' judgment accuracy for both students' self-concept and their autonomous motivation. In accordance with Hypothesis 3, we found a curvilinear prediction of students' GPA for teachers' judgment accuracy regarding students' academic self-concept, indicating that students with very low and very high GPA are judged more accurately. The hypothesis was, however, only partly supported as the effect did not show for students' autonomous motivation where only a linear effect of students' GPA was found. With regard to the information group, a significant interaction effect for subject and students' self-reported autonomous motivation was found, suggesting higher judgment accuracy for teachers of major subjects; this effect did, however, not occur for students' academic self-concept. Hypothesis 4, expecting higher judgment accuracy for major subject teachers, could therefore only partly be confirmed. Neither a determinant from the judge group (self-efficacy, see Hypothesis 5; and job experience, see Hypothesis 6) nor from the environment group (classroom climate, Hypothesis 7) were found to be related to teachers' judgment accuracy regarding students' academic self-concept or autonomous motivation.

4. Discussion

Thus far, only very few investigations have focused on determinants of the (in)accuracy of teachers' judgments of students' motivational characteristics. In the present study, the RAM model

Table 6
Predicting teacher judgments of students' autonomous motivation.

	Model 3a	Model 3b	Model 3c
Intercept	2.71* [2.66; 2.76]	2.71* [2.65; 2.76]	2.71* [2.65; 2.76]
Student covariates			
Student autonomous motivation	0.11* [0.06; 0.15]	0.10* [0.06; 0.15]	0.14* [0.07; 0.20]
Gender (female)	0.22* [0.17; 0.26]	0.22* [0.17; 0.26]	0.21* [0.17; 0.26]
Age	−0.09* [−0.11; −0.07]	−0.09* [−0.11; −0.07]	−0.09* [−0.12; −0.07]
GPA linear	0.25* [0.22; 0.29]	0.26* [0.22; 0.29]	0.26* [0.22; 0.30]
GPA quadratic	−0.04* [−0.08; −0.01]	−0.04* [−0.08; −0.00]	−0.02 [−0.07; 0.02]
Teacher and class covariates			
Job experience		0.01* [0.00; 0.01]	0.01* [0.00; 0.01]
Subject (minor)		0.01 [−0.04; 0.05]	0.01 [−0.04; 0.06]
Self-efficacy		0.08* [0.02; 0.13]	0.08* [0.02; 0.13]
Classroom climate		0.03 [−0.01; 0.08]	0.03 [−0.01; 0.07]
Interaction terms with autonomous motivation			
Gender (female)			0.04 [−0.03; 0.12]
Age			0.02 [−0.01; 0.05]
GPA linear			0.08* [0.02; 0.14]
GPA quadratic			0.01 [−0.04; 0.07]
Job experience			0.00 [−0.00; 0.00]
Subject (minor)			−0.07* [−0.12; −0.02]
Self-efficacy			0.02 [−0.02; 0.05]
Classroom climate			0.01 [−0.02; 0.04]
Deviance			
Mean deviance	6484	6487	6482
Penalty	1059	1059	1058
Penalized deviance	7543	7546	7540

Note. Values in brackets are 95% credibility intervals. *$p < 0.05$.

developed by Funder (1995) was used to separate relevant groups of determinants of teachers' judgment accuracy (i.e., students, information, teachers, and environment) as well as relevant steps that lead to accurate judgments (i.e., relevance, availability, detection, and utilization of behavior). To investigate teachers' differences with respect to these aspects regarding judgments of students' self-concepts as well as autonomous motivation, we used Bayesian cross-classified multi-level regression analysis. We were thus able to systematically separate different sources of teachers' judgment (in)accuracy.

4.1. Variance sources in teacher judgments

Our analyses showed that variance in teacher judgments of students' self-concepts as well as students' autonomous motivation is mainly due to variation in the level of judgments. Nevertheless, the results also indicated variation in teachers' judgment accuracy across teachers and classes/schools. The variation is, however, considerably smaller than the variation in judgment accuracy reported in prior studies (e.g., Karing, 2009; Spinath, 2005; Urhahne et al., 2011). One possible explanation for this difference is that intra-individual correlations as usually calculated for every teacher in prior studies are not reliable due to the very small sample sizes of 5–30 data points per teacher. Schönbrodt and Perugini (2013) presented a simulation study and showed that using a 95% confidence interval requires between 68 data points (accepting a variation around the actual correlation of ±0.20) and 403 data points (accepting a variation of ±0.10) to get a stable estimate of a medium-sized correlation ($0.30 \leq r \leq 0.50$). Thus, unreliability due to small sample size was not considered in previous studies when calculating the range of intra-individual correlations. This supports researchers' concerns in the field who cautioned that differences between teachers in their judgment accuracy might not be reliable (e.g., Lorenz, 2011; Martinéz et al., 2009). In our study, we opted for a multi-level modeling approach which provides a more appropriate alternative to identifying variation in teacher judgments and their accuracy.

4.2. Determinants of teacher judgments

As could already be expected based on the size of the variance sources, the identified determinants in the present study were mainly related to the level of teacher judgments and less to their judgment accuracy. Regarding determinants for the level of teacher judgments, we found all included student characteristics—gender (female), age, and GPA—to be related to higher levels of judgment. This is in agreement with some prior studies (e.g., Hinnant et al., 2009; Tiedemann, 2002) and indicates that teachers use partly invalid sources for their judgments. Some information and teacher characteristics were related to the level of teacher judgments as well. Among others, teacher judgments tended to be higher for teachers with a higher self-efficacy. This is an interesting finding which might partly explain why some teachers feel more self-efficacious than others: they simply interpret their surrounding in a more positive way (and thus also the students they are interacting with). This finding thus emphasizes that not only the objective classroom events and student behavior might play a role in building teachers' self-efficacy (see Ross, 1998), but also the teachers' interpretation of these events and behaviors.

Regarding teachers' judgment accuracy in a narrower sense, few significant determinants could be identified. Again, GPA turned out to be relevant, but was differently related to judgment accuracy for the two motivational characteristics. For autonomous motivation, students with higher GPA were judged more accurately whereas for academic self-concept, not only students with higher GPA but also those with lower GPA were judged more accurately. The consistency of student behavior (Hypothesis 3) thus indeed seems to facilitate the accuracy of teacher judgments. However, high GPA obviously indicates consistent behavior to a larger degree across student characteristics than does low GPA, as only higher GPA was associated with higher judgment accuracy for both student characteristics. So far, student characteristics were largely neglected in investigating teachers' judgment accuracy. The current findings indicate that some students are easier to judge than others. Accordingly, student characteristics are potentially highly relevant determinants of teachers' judgment accuracy.

Autonomous motivation was judged more accurately by teachers of major subjects. This again is in line with the RAM model (Hypothesis 4) as well as the study by Hopkins et al. (1985), as major subject teachers were better informed about their students than minor subject teachers. For students' self-concepts, however, such a relation could not be identified. This finding is in line with results of Praetorius et al. (2015) who did not identify any differences in teachers' judgment accuracy of their students' self-concepts compared to judgments based on 30-s videos of students. The presented results thus indicate that determinants of judgment accuracy might not be generalizable across motivational characteristics and that the extent to which relevant behavior is shown and available to teachers differs across characteristics. Based on the evidence we have, it is entirely unclear what causes these differences. To find out more, experimental studies that manipulate aspects of relevance and availability may be suited as they allow controlling for other possibly confounding influences.

Interestingly, none of the teacher-related characteristics was found to predict teachers' judgment accuracy – although a quite large teacher sample was realized in the present study. This is surprising as based on the RAM model, one would expect that the detection and correct utilization of relevant behavior differs across teachers. Indeed, the variance decomposition showed this to be the case but the included variables did not capture those differences. One explanation why higher values regarding job experience were not related to higher judgment accuracy might be that teachers usually do not receive feedback on their perceptions of students and their instruction and therefore have only few opportunities to improve their judgment accuracy. Self-efficacy, too, might be too distal to the judgment process to explain differences in judgment accuracy. More proximal characteristics should thus be focused on in the future. In relation to the RAM model step “utilization”, it might be fruitful to focus on causal attributions of teachers which are known to influence to a considerable degree how teachers interpret student behavior (e.g., Wiley, Tankersley, & Simms, 2012). To provide a systematic account of research in the field, future studies on determinants of teachers' judgment accuracy regarding motivational student characteristics would benefit from using Brunswik's lens model (Brunswik, 1956; for an application to teachers' judgment accuracy, see e.g., Marksteiner, Reinhard, Dickhäuser, & Sporer, 2012), to identify the cues used by teachers and the extent to which these cues are valid for the motivational characteristic to be judged. For a better insight into the RAM model step “detection”, it is advisable to separate it from the “utilization” step. To this end, a two-fold approach seems most beneficial, with experimental studies on the one hand and stimulated recall tasks on the other hand (see studies on the validity of observer ratings for measuring teaching quality, e.g., Bell et al., 2015).

Taken together, the evidence for determinants for teachers' judgment accuracy on the teacher level is very limited as so far only determinants related to the students and the quantity of information could be found. This is interesting as differences in judgment accuracy are often interpreted in terms of ability differences in teachers. The current findings, however, indicate that some students seem to be easier to judge than others and that the quantity of information available to teachers can additionally facilitate or impede judgment accuracy. A comparison of teachers who have to judge differently easy-to-judge students and have different quantities of information about their students is thus not fair if these differences are not taken into account.

Our analyses additionally indicate that for explaining judgment accuracy, it might be promising to take a closer look at aspects that are dependent on characteristics of the teacher and concurrently the specific student being judged (e.g., sympathy for a certain student). A few studies have dealt with such interaction effects. Itskowitz, Navon, and Strauss (1988), for example, found overestimation of students' self-concepts for those students the teachers felt attached to, and underestimation especially for the students the teachers described they rather felt detached from. Given the limited evidence on interactive effects of teachers and students, this seems to be a promising direction for future studies in the field of judgment accuracy research.

4.3. Limitations and further directions

The current study contributes to research on teachers' judgment accuracy in important ways as we empirically tested all theoretically relevant sources of inaccuracy in one study with appropriate statistical methods. However, there are some limitations that need to be addressed in future research to enhance evidence even further.

Our study focused exclusively on subject-unspecific aspects of student motivation on the general school level – although it is known that subject-specific aspects are also relevant (e.g., Brunner et al., 2010). This was a prerequisite for disentangling the teacher and the class levels from each other using a cross-classified design. Nevertheless, we believe that our findings extend the literature on determinants of teachers' judgment accuracy considerably: (a) our study is the first that allowed disentangling different judgment components systematically. (b) Moreover, in investigating judgments from teachers teaching different subjects, we were able to draw conclusions regarding relevant determinants beyond a single subject. (c) Additionally, the size of the simple bivariate correlations between teacher judgments and students' actual characteristics in our study was very similar to the ones found in studies using a subject-specific approach (see e.g., Karing, 2009; Marsh & Craven, 1991; Praetorius et al., 2013), underpinning the validity of our approach. Thus, we are convinced that our work can serve as an important starting point for future research. For example, our systematically gained knowledge could in a next step be transferred to research on subject-specific student motivations as well as student achievement.

We used student self-reports to measure motivational characteristics. Self-reports have been criticized extensively because they might be subject to social desirability and other self-protecting biases. However, for measuring intrapsychic characteristics such as motivation, self-reports are based on the most valid information available (see e.g., Paulhus & Vazire, 2007). Therefore, student motivation is commonly measured based on self-reports. Many studies on student motivation indicate that these self-ratings are valid and meaningful. There is, for example, evidence that social desirability and motivational student characteristics are not substantially related (e.g., Al-Hoorie, 2016; Pekrun, Elliot, & Maier, 2009). Studies which indicate such a relation (e.g., Schaffner, Schiefele, & Ulferts, 2013) found that it does not have an impact on the substantial effects of students' motivational characteristics on achievement.

A Bayesian cross-classified multi-level approach was used for data analysis in this study. Latent approaches such as CFA-MTMM models (see Koch et al., 2016) would present a more immediate account of measurement error, however, such approaches have not yet been developed for the analysis of complex cross-classified multi-rater data including multiple groups and interaction variables. Therefore, we opted for the manifest approach.

4.4. Conclusions

We used a theory-based approach for deriving hypotheses about determinants of teachers' judgment accuracy (i.e., Funder's RAM model) combined with a methodologically innovative and appropriate approach (i.e., cross-classified multi-level modeling), and we thus advanced the field in a systematic way. Accordingly, we could identify significant determinants of judgment accuracy for the target group (i.e., students' GPA) as well as for the information group (i.e., subject). Remarkably, these effects differed across the two investigated motivational student characteristics. In contrast, none of the determinants from the judge group or from the environment group were found to be related to teachers' judgment accuracy. An explanation for these findings could be that a large amount of the variation in teacher judgments had to be attributed to differences in level, not in accuracy. Many studies stated that large differences exist between teachers in their judgment accuracy (e.g., Kaiser et al., 2017; Lorenz, 2011; Urhahne & Zhu, 2015), which needs to be reconsidered. More generally, the appropriateness of research on teachers' judgment accuracy is called into question with its current focus on separable single student characteristics instead of student profiles (see also Glock, Krolak-Schwerdt, Klapproth, & Böhmer, 2013), on analyses primarily on the class level, as well as on the mere accuracy of teacher judgments instead of a broader concept of judgmental competence (see also Herppich et al., 2017).

Funding

This research was funded by a grant from the “Stiftung Bildungspakt Bayern” awarded to Annette Scheunpflug.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.learninstruc.2017.06.003.

References

Al-Hoorie, A. H. (2016). Unconscious motivation. Part II: Implicit attitudes and L2 achievement. *Studies in Second Language Learning and Teaching, 6*(4), 619–649. http://dx.doi.org/10.14746/ssllt.2016.6.4.4.

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*, 731–746. http://dx.doi.org/10.1037/0022-0663.91.4.731.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261–271. http://dx.doi.org/10.1037/0022-0663.84.3.261.

Behrmann, L., & Souvignier, E. (2013). The relation between teachers' sensitivity, their instructional activities, and their students' achievement gains in reading. *Zeitschrift für Pädagogische Psychologie, 27*, 283–293. http://dx.doi.org/10.1024/1010-0652/a000112.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., et al. (2015). Improving observational score quality. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems*. San Francisco: John Wiley & Sons, Inc. http://dx.doi.org/10.1002/9781119210856.ch3.

Berhenke, A., Miller, A. L., Brown, E., Seifer, R., & Dickstein, S. (2011). Observed emotional and behavioral indicators of motivation predict school readiness in Head Start graduates. *Early Childhood Research Quarterly, 26*(4), 430–441. http://dx.doi.org/10.1016/j.ecresq.2011.04.001.

Böhmer, I., Hörstermann, T., Gräsel, C., Krolak-Schwerdt, S., & Glock, S. (2015). An analysis of information search in the process of making school tracking decisions: Which judgment rule do teachers apply? [Eine Analyse der Informationssuche bei der Erstellung der Übergangsempfehlung: Welcher Urteilsregel folgen Lehrkräfte?] *Journal for Educational Research Online, 7*, 59–81.

Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task value, and achievement goals. *Journal of Educational Psychology, 93*, 23–34. http://dx.doi.org/10.1037/0022-0663.93.1.23.

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

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality, 59*(2), 143–178. http://dx.doi.org/10.1111/j.1467-6494.1991.tb00772.x.

Clausen, M. (2002). *Unterrichtsqualität – eine Frage der Perspektive?* [*Instructional quality – a matter of perspective?*] Münster, Germany: Waxmann.

Corno, L. Y. N. (2008). On teaching adaptively. *Educational Psychologist, 43*(3), 161–173. http://dx.doi.org/10.1080/00461520802178466.

Council of Chief State School Officers (CCSSO) Interstate Teacher Assessment and Support Consortium. (2013). In *TASC model core teaching standards and learning progressions for teachers 1.0*. Washington, DC: CCSSO Interstate Teacher Assessment and Support Consortium.

Curtis, S. M. (2015). *mcmcplots: Create plots from MCMC output*. R package version 0.4.2 https://CRAN.R-project.org/package=mcmcplots.

Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*, 227–268. http://dx.doi.org/10.1207/s15327965pli1104_01.

Decuir-Gunby, J. T., & Williams-Johnson, M. R. (2014). The influence of culture on emotions: Implications for education. In R. Pekrun, & L. A. Linnenbrink (Eds.), *International handbook of emotions in education* (pp. 539–557). New York, NY: Routledge.

Dickhäuser, O., & Reinhard, M.-A. (2006). Factors underlying expectancies of success and achievement: The influential roles of need for cognition and general or specific self-concepts. *Journal of Personality and Social Psychology, 90*, 490–500. http://dx.doi.org/10.1037/0022-3514.90.3.490.

Dickhäuser, O., & Stiensmeier-Pelster, J. (2003). Wahrgenommene Lehrereinschätzungen und das Fähigkeitsselbstkonzept von Jungen und Mädchen in der Grundschule [Perceived teacher judgments and the academic self-concept of boys and girls in primary schools]. *Psychologie in Erziehung und Unterricht, 50*(2), 182–190.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. http://dx.doi.org/10.1111/bjop.12046.

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18*, 52–65. http://dx.doi.org/10.1521/scpq.18.1.52.20876.

Fielding, A., & Goldstein, H. (2006). *Cross-classified and multiple membership structures in multilevel models: An introduction and review (Research Report RR791)*. Birmingham: University, Department for Education and Skills.

Fortier, M. S., Vallerand, R. J., & Guay, F. (1995). Academic motivation and school performance: Toward a structural model. *Contemporary Educational Psychology, 20*(3), 257–274.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*(4), 652–670.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*(3), 177–182.

Funder, D. C., & Colvin, C. R. (1997). Congruence of others' and self-judgments of personality. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 617–647). San Diego, CA: Academic Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models* (Vol. 1). New York, NY: Cambridge University Press.

Givvin, K. B., Stipek, D. J., Salmon, J. M., & MacGyvers, V. L. (2001). In the eyes of the beholder: Students' and teachers' judgments of students' motivation. *Teaching and Teacher Education, 17*, 321–331. http://dx.doi.org/10.1016/S0742-051X(00)00060-3.

Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2013). Beyond judgment bias: How students' ethnicity and academic profile consistency influence teachers' tracking judgments. *Social Psychology in Education, 16*, 555–573. http://dx.doi.org/10.1007/s11218-013-9227-5.

Gottfried, A. E. (1985). Academic intrinsic motivation in elementary and junior high school students. *Journal of Educational Psychology, 77*(6), 631–645.

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and achievement: Developmental perspective on their causal ordering. *Journal of Educational Psychology, 95*, 124–136. http://dx.doi.org/10.1037/0022-0663.95.1.124.

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgment accuracy on achievement. *Teaching and Teacher Education, 3*, 91–98.

Herppich, S., Praetorius, A. K., Hetmanek, A., Glogger-Frey, I., Ufer, S., Leutner, D., … Südkamp, A. (2017). Ein Arbeitsmodell für die empirische Erforschung der diagnostischen Kompetenz von Lehrkräften. In A. Südkamp, & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen [Judgmental competences of teachers: Theoretical and methodological developments]* (pp. 75–95). Münster: Waxmann.

Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school year. *Journal of Educational Psychology, 101*, 662–670. http://dx.doi.org/10.1037/a0014306.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*, 177–182. http://dx.doi.org/10.1111/j.1745-3984.1985.tb01056.x.

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science, 326*, 1410–1412. http://dx.doi.org/10.1126/science.1177067.

Hurwitz, J. T., Elliott, S. N., & Braden, J. P. (2007). The influence of test familiarity and student disability status upon teachers' judgments of students' test performance. *School Psychology Quarterly, 22*, 115–144. http://dx.doi.org/10.1037/1045-3830.22.2.115.

Itskowitz, R., Navon, R., & Strauss, H. (1988). Teachers' accuracy in evaluating students' self-image: Effect of perceived closeness. *Journal of Educational Psychology, 80*, 337–341. http://dx.doi.org/10.1037/0022-0663.80.3.337.

James, W. (1892/1999). The self. In R. F. Baumeister (Ed.), *The self in social psychology* (pp. 69–77). Philadelphia, PA: Psychology Press. Original: James, W. [1892/1948]. Psychology. Cleveland, OH: World Publishing.

Kaiser, J., Helm, F., Retelsdorf, J., Südkamp, A., & Möller, J. (2012). Zum Zusammenhang von Intelligenz und Urteilsgenauigkeit bei der Beurteilung von Schülerleistungen im Simulierten Klassenraum [On the relation of intelligence and judgement accuracy in the process of assessing student achievement in the simulated classroom]. *Zeitschrift für Pädagogische Psychologie, 26*, 251–261. http://dx.doi.org/10.1024/1010-0652/a000076.

Kaiser, J., Südkamp, A., & Möller, J. (2017). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology*. http://dx.doi.org/10.1037/edu0000156. Advance online publication.

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen [Diagnostic competence of elementary and secondary school teachers in the domains of competence and interests]. *Zeitschrift für Pädagogische Psychologie, 23*, 197–209. http://dx.doi.org/10.1024/1010-0652.23.34.197.

Kenny, D. T., & Chekaluk, E. (1993). Early reading performance: A comparison of teacher-based and test-based assessment. *Journal of Learning Disabilities, 26*, 227–236. http://dx.doi.org/10.1177/002221949302600403.

Koch, T., Schultze, M., Jeon, M., Nussbeck, F., Praetorius, A.-K., & Eid, M. (2016). A cross-classified CFA-MTMM model for structurally different and non-independent interchangeable methods. *Multivariate Behavioral Research, 51*, 67–85. http://dx.doi.org/10.1080/00273171.2015.1101367.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., et al. (2003). *PISA 2000: Dokumentation der Erhebungsinstrumente* [*PISA 2000: Documentation of the survey instruments*]. Berlin, Germany: Max-Planck-Institut für Bildungsforschung.

Lorenz, C. (2011). *Diagnostische Kompetenz von Grundschullehrkräften. Strukturelle Aspekte und Bedingungen* [*Diagnostic competence of primary school teachers. Structural aspects and conditions*]. Bamberg, Germany: University of Bamberg Press.

Machts, N., Kaiser, J., Schmidt, F. T., & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review, 19*, 85–103. http://dx.doi.org/10.1016/j.edurev.2016.06.003.

Maguin, E., & Loeber, R. (1996). How well do ratings of academic performance by mothers and their sons correspond to grades, achievement test scores, and teachers' ratings? *Journal of Behavioral Education, 6*, 405–425. http://dx.doi.org/10.1007/BF02110514.

Marksteiner, T., Reinhard, M.-A., Dickhäuser, O., & Sporer, S. L. (2012). How do teachers perceive cheating students? Beliefs about cues to deception and detection accuracy in the educational field. *European Journal of Psychology of Education, 27*, 329–350. http://dx.doi.org/10.1007/s10212-011-0074-5.

Marsh, H. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review, 2*, 77–172.

Marsh, H. W., & Craven, R. G. (1991). Self-other agreement on multiple dimensions of preadolescent self-concept: Inferences by teachers, mothers, and fathers. *Journal of Educational Psychology, 83*(3), 393–404.

Martinéz, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment, 14*, 78–102. http://dx.doi.org/10.1080/10627190903039429.

National Council on Measurement in Education, Washington, DC.& American Federation of Teachers, Washington, DC. & National Education Association, Washington, DC. (1990). *Standards for teacher competence in educational assessment of students* [Washington, D.C.]: Distributed by ERIC Clearinghouse http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED323186.

O'Mara, A. J., Marsh, H. W., Craven, R. G., & Debus, R. L. (2006). Do self-concept interventions make a difference? A synergistic blend of construct validation and meta-analysis. *Educational Psychologist, 41*(3), 181–206. http://dx.doi.org/10.1207/s15326985ep4103_4.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York: Guilford.

Pekrun, R., Elliot, A. J., & Maier, M. A. (2009). Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of Educational Psychology, 101*, 115–135. http://dx.doi.org/10.10137/a0013383.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*(4), 667–686. http://dx.doi.org/10.1037/0022-0663.95.4.667.

Plummer, M. (2016). *rjags: Bayesian graphical models using MCMC*. R package version 4-6. Retrieved from https://CRAN.R-project.org/package=rjags.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*(1), 7–11.

Pohlmann, S. (2009). *The transition at the end of primary school - The formation of tracking recommendations from the teachers' perspective* [*Der Übergang am Ende der Grundschulzeit – Zur Formation der Übergangsempfehlung aus der Sicht der Lehrkräfte*]. Münster, Germany: Waxmann.

Praetorius, A.-K., Berner, V.-D., Zeinz, H., Scheunpflug, A., & Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging academic self-concepts of students. *Journal of Educational Research, 106*, 64–76. http://dx.doi.org/10.1080/00220671.2012.667010.

Praetorius, A.-K., Drexler, K., Rösch, L., Christophel, E., Heyne, N., Scheunpflug, A., … Dresel, M. (2015). Judging students' self-concepts within 30 seconds? An application of the zero-acquaintance approach to research on teachers' judgment accuracy. *Learning and Individual Differences, 37*, 231–236. http://dx.doi.org/10.1016/j.lindif.2014.11.015.

Praetorius, A.-K., Greb, K., Dickhäuser, O., & Lipowsky, F. (2011). Wie gut schätzen Lehrer die Fähigkeitsselbstkonzepte ihrer Schüler ein? Zur diagnostischen Kompetenz von Lehrkräften [How teacher rate their students: on teachers' diagnostics competence regarding the academic self-concept]. *Psychologie in Erziehung und Unterricht, 58*, 81–91. http://dx.doi.org/10.2378/peu2010.art30d.

R Development Core Team. (2008). *R: A language and environment for statistical computing. R foundation for statistical computing*. Retrieved from http://www.R-project.org.

Reindl, M., Berner, V.-D., Scheunpflug, A., Zeinz, H., & Dresel, M. (2015). Effect of negative peer climate on the development of autonomous motivation in mathematics. *Learning and Individual Differences, 38*, 68–75. http://dx.doi.org/10.1016/j.lindif.2015.01.017.

Retelsdorf, J., Köller, O., & Möller, J. (2014). Reading achievement and reading self-concept – testing the reciprocal effects model. *Learning and Instruction, 29*, 21–30. http://dx.doi.org/10.1016/j.learninstruc.2013.07.004.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 104*(3), 1–13. http://dx.doi.org/10.1037/a0027268.

Ross, J. A. (1998). The antecedents and consequences of teacher efficacy. In J. Brophy (Ed.), *Advances in research on teaching* (pp. 385–400). Greenwich, CT: JAI Press.

Schaffner, E., Schiefele, U., & Ulferts, H. (2013). Reading amount as a mediator of the effects of intrinsic and extrinsic reading motivation on reading comprehension. *Reading Research Quarterly, 48*(4), 369–385. http://dx.doi.org/10.1002/rrq.52.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*, 609–612. http://dx.doi.org/10.1016/j.jrp.2013.05.009.

Schöne, C., Dickhäuser, O., Spinath, B., & Stiensmeier-Pelster, J. (2002). *SESSKO – Skalen zur Erfassung des schulischen Selbstkonzepts* [*Scales to assess the academic self-concept*]. Göttingen, Germany: Hogrefe.

Schwarzer, R., & Jerusalem, M. (1999). *Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*. Berlin: Freie Universität Berlin.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46*(3), 407–441.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571–581.

Sparfeldt, J. R., Schilling, S. R., Rost, D. H., & Müller, C. (2003). Bezugsnormorientierte Selbstkonzepte? Zur Eignung der SESSKO [Reference-oriented self-concepts? Regarding the suitability of SESSKO]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 24*(4), 325–335.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 583–639. http://dx.doi.org/10.1111/1467-9868.00353.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz [Accuracy of teacher judgments of student characteristics and the construct of diagnostic competence]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 19*, 85–95. http://dx.doi.org/10.1024/1010-0652.19.12.85.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743–762. http://dx.doi.org/10.1037/a0027627.

Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in Mathematics, 50*, 49–62. http://dx.doi.org/10.1023/A:1020518104346.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., … Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted, 27*(2–3), 119–145.

Urhahne, D., Chao, S.-H., Florineth, M. L., Luttenberger, S., & Paechter, M. (2011). Academic self-concept, learning motivation, and test anxiety of the underestimated student. *British Journal of Educational Psychology, 81*, 161–177. http://dx.doi.org/10.1348/000709910X504500.

Urhahne, D., & Zhu, M. (2015). Accuracy of teachers' judgments of students' subjective well-being. *Learning and Individual Differences, 43*, 226–232. http://dx.doi.org/10.1016/j.lindif.2015.08.007.

Weinert, F. E., & Schrader, F. W. (1986). Diagnose des Lehrers als Diagnostiker [Diagnosis of the teachers as a diagnostician]. In H. Petillon, J. Wagner, & B. Wolf (Eds.), *Schülergerechte Diagnose. Theoretische und empirische Beiträge zur Pädagogischen Diagnostik [Diagnosis which does justice to the pupil: Theoretical and empirical contributions to diagnostics in teaching]* (pp. 11–29). Weinheim: Beltz.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. http://dx.doi.org/10.1006/ceps.1999.1015.

Wild, K.-P., & Rost, D. (1995). Klassengröße und Genauigkeit von Schülerbeurteilungen [Class size and accuracy of judgments of students]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 27*(1), 78–90.

Wiley, A. L., Tankersley, M., & Simms, A. (2012). Teachers' causal attributions for student problem behavior: Implications for school-based behavioral interventions and research. In B. G. Cook, M. Tankersley, & T. J. Landrum (Eds.), *Classroom behavior, contexts, and interventions* (pp. 279–300). Bingley, UK: Emerald.