DOI: 10.1002/jclp.22840

WILEY

RESEARCH ARTICLE

Interrater agreement and discrepancy when assessing problem behaviors, social-emotional skills, and developmental status of kindergarten children

Sebastian Bergold[1] | Hanna Christiansen[2] | Ricarda Steinmayr[1]

[1]Department of Psychology, TU Dortmund University, Dortmund, Germany

[2]Department of Psychology, Philipps-University Marburg, Marburg, Germany

Correspondence
Sebastian Bergold, Department of Psychology, TU Dortmund University, Emil-Figge-Straße 50, D-44227 Dortmund, Germany.
Email: sebastian.bergold@tu-dortmund.de

Abstract

Objective: The present study examined parent-teacher agreement and discrepancy when assessing kindergarten children's behavioral and emotional problems, social-emotional skills, and developmental status.

Method: Parents and teachers of overall $n = 922$ kindergarten children ($M_{age} = 3.99$; 449 girls) rated the children using the Conners Early Childhood, the Strengths and Difficulties Questionnaire, and the Questionnaire for Assessing Preschool Children's Behavior.

Results: Agreement was moderate for problem behaviors and social-emotional skills and substantial for developmental status. Agreement was stronger for externalizing than for internalizing problems. Agreement on the clinical relevance of problem behaviors and of social-emotional skills was stronger for children with a clinical diagnosis than for those without. Parents tended to report more problems, but also greater social-emotional skills and developmental status, than teachers.

Conclusions: The findings corroborate the importance of situational specificity for understanding interrater agreement and discrepancy. Future teacher questionnaires should more specifically assess children's functioning in kindergarten.

KEYWORDS
externalizing and internalizing behavior, interrater discrepancy, multiple informants, parent-teacher agreement, preschool children, situational specificity

1 | INTRODUCTION

About 15% of children between 2 and 5 years of age have a mental disorder (see Cree et al., 2018; McCue Horwitz et al., 2012, for review). To intervene at the earliest possible time and thus prevent disorder chronicity, diagnostics must be carried out as early as possible. However, making a valid clinical diagnosis at this young age is especially challenging, as these children are not yet able to make reliable self-assessments (Angold & Egger, 2007). Therefore, external assessments from various perspectives assume an especially prominent role when clinically assessing preschool children's psychopathology.

As the diagnostic criteria usually demand problem behaviors to be present across multiple settings to diagnose a mental disorder (American Psychiatric Association, 2013; World Health Organization, 2016), assessments by caregivers from different settings are required. For preschool children, assessments are often made by their parents and kindergarten teachers. However, an important question is the extent to which parents and teachers agree on their assessment of the child's problems and skills. To date, numerous studies have been devoted to parent-teacher agreement and discrepancy when assessing children in middle childhood and adolescence (e.g., Achenbach, McConaughy, & Howell, 1987; De Los Reyes et al., 2015; Stone, Otten, Engels, Vermulst, & Janssens, 2010). However, due to the long-lasting lack of clinical rating instruments tailored to young children (Angold & Egger, 2007), studies on parent-teacher agreement and discrepancy when assessing preschool children are few. Furthermore, results vary from study to study at least to some extent (Achenbach et al., 1987; De Los Reyes et al., 2015). Whether this is due to differences between samples, questionnaires, or problem behaviors cannot unequivocally be answered unless one sample is questioned concerning various child characteristics with different questionnaires.

The present study helps fill this gap by examining agreement and discrepancy between parents' and teachers' assessments of kindergarten children's behavioral and emotional problems, social-emotional skills, and developmental status. In addition, we examined type of problem behavior, clinical diagnostic status, and the teacher's familiarity with the child as potential moderators of agreement and discrepancy.

2 | OVERALL PARENT-TEACHER AGREEMENT AND DISCREPANCY: FINDINGS ON OLDER CHILDREN AND ADOLESCENTS

Many studies comparing clinical assessments of children or adolescents made by different raters (parents, teachers, other caregivers, the children themselves) have been published. The seminal meta-analysis by Achenbach et al. on interrater agreement dating back to 1987 included 119 studies already. They found that the mean correlation between parent and teacher assessments was modest ($r = 0.27$). Similar results were reported by Rescorla et al. (2014) and by De Los Reyes et al. (2015) in their recent meta-analysis based on 341 studies. With regard to social skills, the meta-analysis by Renk and Phares (2004) detected overall agreement of $r = 0.38$. Such results reveal mediocre agreement between parents and teachers.

Studies addressing absolute agreement are less numerous. Rescorla et al. (2014) found that across the 21 countries they covered, parents tended to report more problems than teachers, even though in some countries, there were no notable differences between the two rating groups (see also Kennerley et al., 2018; Narad et al., 2015; Youngstrom, Loeber, & Stouthamer-Loeber, 2000).

Besides comparing raw scores, comparing standard values might be especially interesting for clinical practice. When making a psychiatric diagnosis, what is important is whether raters from different settings agree on the clinical relevance

of a child's problems. However, surprisingly few studies on parent-teacher discrepancies have investigated interrater agreement on whether or not the child's problems are clinically relevant. Results have revealed rather weak agreement, ranging from $\kappa \approx 0.10$ to $\kappa \approx 0.30$ (Deng, Liu, & Roosa, 2004; Rescorla et al., 2014; Zahner & Daskalakis, 1998).

3 | REASONS FOR WEAK AGREEMENT AND STRONG DISCREPANCY BETWEEN RATERS

For a long time, weak interrater agreement has either been attributed to measurement error or to the inability of a certain rater group to validly assess children's behavior problems (Achenbach et al., 1987; De Los Reyes, 2011). However, at the latest, since the work by Achenbach et al. (1987), the attention of researchers in the field has been drawn to the assumption that different raters validly provide different information, and that weak interrater agreement is primarily a consequence of situational specificity. In particular, children's behaviors can vary across settings. This variation might additionally be strengthened by the raters themselves because they usually interact with the child, thereby evoking different behaviors (e.g., Achenbach et al., 1987; De Los Reyes, 2011; De Los Reyes & Kazdin, 2005). If the behavior in school or kindergarten differs from behavior at home, weak agreement and notable differences between parents and teachers would be the logical consequence and cannot be considered the mere result of an allegedly invalid assessment. In line with this hypothesis, De Los Reyes, Henry, Tolan, and Wakschlag (2009) showed in their laboratory study that interrater discrepancies, when assessing preschoolers' disruptive behavior, were related to actual cross-situational differences in children's behavior. Furthermore, interrater agreement is markedly stronger between raters coming from the same context (e.g., mother-father, teacher-teacher; $r = 0.60$) than between raters originating from different contexts (e.g., parent-teacher; $r = 0.28$; Achenbach et al., 1987; see also De Los Reyes et al., 2015). Furthermore, interrater discrepancies remain stable over time (De Los Reyes, 2011), and parent and teacher assessments both predict child outcomes such as later symptoms or a clinical diagnosis independently from each other (Kerr, Lunkenheimer, & Olson, 2007; Korsch & Petermann, 2014). Thus, different rater groups seem to contribute valid and partly unique information on children's problem behaviors.

At the same time, however, situational specificity is probably not the sole reason for interrater discrepancies. For instance, Kennerley et al. (2018) investigated parent and teacher ratings of attention-deficit/hyperactivity symptoms and compared those ratings with observations made by a clinician at school. The teacher ratings, but not the parent ratings, of the total symptoms correlated with the respective observer ratings. Initially, this finding seems to support the notion of situational specificity. However, the correlation between teacher and observer ratings was only weak ($r = 0.26$). Further, when investigating the subscales, the teacher rating of hyperactivity/impulsivity correlated only slightly higher with the respective observer rating than did the respective parent rating ($r = 0.35$ vs. $r = 0.31$) and there was not even any difference in correlations when considering inattention. Thus, informant characteristics seem to play a role in the emergence of interrater discrepancies.

De Los Reyes and Kazdin (2005) elaborated this argument in their Attribution Bias Context (ABC) Model. According to their model, interrater discrepancies are (at least partly) the result of rater differences in three characteristics: (a) Raters can differ in what they attribute to be the cause of the problematic behavior (disposition vs. context). For example, the child might feel his or her behavior is caused by contextual factors, whereas parents or teachers might perceive it to be caused by child dispositions (fundamental attribution bias). (b) Perspectives of raters involved in the diagnostic process could differ. Due to their different attributions of problem behavior, children and teachers or parents might harbor different perspectives on whether and which behavior warrants treatment and therefore retrieve different information from their memories, resulting in different responses. As another example, parental stress might lower parents' threshold for perceiving a behavior as problematic. (c) Raters can differ in how far their goals are consistent with the goals of the diagnostic process (i.e., collecting information on the child's negative behaviors, deciding on treatment, planning treatment). For example, some protagonists might not want treatment and therefore refuse to report problematic behavior, whereas others might strongly wish

treatment and therefore report or even overestimate problematic behavior. According to this model, discrepancies between parents and teachers might be the result of different perspectives on whether and which problem behavior warrants treatment, leading to different information being retrieved from memory and to different response tendencies and thresholds and possibly also to different goals within the diagnostic process, all factors that might be exacerbated by actual cross-situational differences in the child's behavior.

4 | WHY PUT SPECIAL FOCUS ON PRESCHOOL CHILDREN?

Studies focusing on preschool children are scarce. To illustrate, when examining child's age as an agreement moderator, Achenbach et al. (1987) stated that the number of studies with children under 6 years of age was too small to enable a mean correlation for that age group. The same still seems to apply today: In their recent meta-analysis, De Los Reyes et al. (2015) condensed the studies focusing on young children to just one group defined as children of 10 years or younger. This suggests that there are still very few studies on young children, hampering fine-grained comparisons between very young age groups.

In fact, a child's age might be important for interrater agreement. For example, Achenbach et al.'s (1987) meta-analysis revealed that interrater agreement was stronger for children from 6 to 11 years of age ($r = 0.51$) than for 12- to 19-year-olds ($r = 0.41$). De Los Reyes et al. (2015) observed similar results. Already Achenbach et al. (1987) conjectured (in line with the notion of situational specificity) that younger children's behaviors might be easier to discern or more consistent across situations and that raters would thus obtain stronger agreement (see also De Los Reyes & Kazdin, 2005). Following this line of argument, interrater agreement for preschoolers might be even stronger than for schoolchildren, because preschool children's problems are by nature especially externalized, and, therefore, presumably more apparent (Grietens et al., 2004). Furthermore, younger children require more intensive care from their caregivers than do older children, which might increase caregivers' agreement when assessing the children's problems and skills (see also De Los Reyes & Kazdin, 2005; De Los Reyes et al., 2015; Renk & Phares, 2004).

5 | PREVIOUS WORK FOCUSING ON PRESCHOOL CHILDREN

Against the expectation that agreement for preschoolers should be stronger than for older children, the few existing studies focusing on preschool children reported levels of parent-teacher agreement resembling those for schoolchildren and adolescents (Berg-Nielsen, Solheim, Belsky, & Wichstrom, 2012; Dinnebeil et al., 2013; Grietens et al., 2004). As with older children, parents tended to report more problems than did teachers (Berg-Nielsen et al., 2012; Dinnebeil et al., 2013; Grietens et al., 2004; Korsch & Petermann, 2014; Kuschel, Heinrichs, Bertram, Naumann, & Hahlweg, 2007; Winsler & Wallace, 2002). With regard to social skills or prosocial behavior, overall agreement was modest, too, and parents tended to rate their children more positively than did teachers (Dinnebeil et al., 2013; Schönmoser, Schmitt, Lorenz, & Relikowski, 2018; Winsler & Wallace, 2002). However, there is nothing known about interrater agreement when assessing young children's developmental status.

5.1 | Moderation effects on parent-teacher agreement and discrepancy

There is some evidence in support of factors increasing or decreasing parent-teacher agreement. Potential moderators can be subdivided into child characteristics such as type and severity of problem behavior, gender, or age; parent characteristics such as psychopathology, stress, or socioeconomic status; family characteristics such as family status or number of siblings; and teacher-related characteristics such as familiarity with the child or teacher-child relationship (for an overview, see Berg-Nielsen et al., 2012; De Los Reyes & Kazdin, 2005). In the following, we focus on three of the most important potential moderators.

5.1.1 | Type of problem behavior

The type of problem behavior has been identified as a consistent moderator of interrater agreement for older children. In the meta-analysis by Achenbach et al. (1987), agreement (across different rater pairs) was stronger for externalizing ($r = 0.41$) than for internalizing problems ($r = 0.32$). For parent-teacher agreement specifically, De Los Reyes et al. (2015) detected correlations of $r = 0.28$ for externalizing and $r = 0.21$ for internalizing problems. Some studies focusing on preschool children have confirmed these findings so far (e.g., Grietens et al., 2004; Winsler & Wallace, 2002). Interrater discrepancies were found to be stronger for externalizing than for internalizing problems (e.g., Grietens et al., 2004; Winsler & Wallace, 2002; but cf. Berg-Nielsen et al., 2012). In Rogge, Koglin, and Petermann (2018), parents reported more conduct problems and hyperactivity (and prosocial behavior) than did teachers, whereas teachers reported more emotional and peer problems.

5.1.2 | Clinical diagnostic status

The clinical diagnostic status (whether or not a child has an existing clinical diagnosis) has attracted relatively little attention so far, although it might be important for interrater agreement and discrepancy, too. One could hypothesize that for children with an existing diagnosis, agreement might be stronger and discrepancy weaker, because the problem behavior should be more severe and therefore more obvious and consistent across settings. On the other hand, one might also expect weaker agreement and stronger discrepancy for those children because they would be more likely given extreme scores by one rater, which would make weak agreement and strong discrepancy more probable, as Narad et al. (2015) suggested. Very few studies have examined preschoolers' clinical diagnostic status (or problem severity) as a moderator. Korsch and Petermann (2014) directly compared a clinical and a nonclinical sample of preschoolers and found no significant differences in agreement between the two groups. Kuschel et al. (2007) and Berg-Nielsen et al. (2012), however, identified markedly weaker agreement and stronger discrepancy, respectively, for deviant than nondeviant children.

5.1.3 | Teacher's familiarity with the child

How long teachers have known the child whose behavior they are rating might also moderate agreement or discrepancy. Answering this question would be especially relevant with regard to the situational specificity of problem behavior. If interrater discrepancies were really mainly due to cross-situational variability in the child's behavior, the teacher's familiarity with the child should not raise interrater agreement. Instead, it should either have no impact at all, or even lower agreement. The latter is because the longer a teacher has known the child, the better he or she should be able to judge its behavior. In turn, if a child's behavior really differs between school and home, agreement should lessen as the teacher gains familiarity with the child. However, if familiarity enhances agreement, that would be indicative of one rater group's aligning their assessments to the judgments of the other rater group, which would mean emerging discrepancies because one rater has provided the more valid assessment than the other (and not because the children behaved differently in different contexts).

Despite its importance, this moderation has seldom been addressed. Zahner and Daskalakis (1998) found in their sample of 6- to 11-year-olds that agreement on internalizing problems, did, in fact, increase with teachers' familiarity with the child. The same was true for agreement on externalizing problems, although the reverse pattern occurred in a clinical subsample. Berg-Nielsen et al. (2012), on the other hand, identified no such moderation via teacher's familiarity.

In summary, our knowledge about interrater agreement and discrepancy when assessing preschool children relies on very few studies, and some of the findings are contradictory. This contradiction can have several reasons, for example cultural differences (Rescorla et al., 2014). Still another might be a methodological issue, as described below.

6 | MEASUREMENT INVARIANCE AS PREREQUISITE FOR INTERPRETABILITY

The value of previous studies notwithstanding, most of their results must be interpreted with caution because raw scores were often compared without scrutinizing measurement invariance beforehand. But in fact—at least when comparing raw scores—certain levels of measurement invariance between the rater groups must hold to enable meaningful comparisons between raters (comparisons of standard values do not require measurement invariance as long as there are separate norms for the rater groups, because group-specific standardizations per se imply the assumption of non-invariance). More precisely, whenever relations between variables (i.e., interrater agreement) are interpreted, (at least partial) metric invariance (equal factor loadings of corresponding indicators) across the rater groups must be given; and whenever mean differences (i.e., interrater discrepancy) are interpreted, (at least partial) scalar invariance (equal factor loadings and intercepts) across the rater groups must be given (e.g., Brown, 2006).

If measurement invariance does not hold (and is not controlled for), results of agreement or discrepancy, respectively, cannot be meaningfully interpreted because the items function differently in the rater groups. For example, agreement might be artificially diminished or discrepancy artificially increased because items differ between the groups in their meaning or because the answer scale does not work equally for the groups. However, attention has been paid only recently to the issue of measurement invariance when examining interrater agreement and discrepancy (e.g., Narad et al., 2015; Rogge et al., 2018). Therefore, for most of the studies reviewed above, it remains unclear as to what extent their results reflect methodological artifacts or "real" agreement or discrepancy, respectively. To our knowledge, only one study (Rogge et al., 2018) focusing on preschoolers has tested for measurement invariance. However, this valuable study relied on just one instrument, meaning ratings of the same behavior could not be compared or investigated across different instruments, and it did not include an assessment of the children's developmental status. Nor was agreement between raters evaluated, but rather discrepancy only; moderation effects were not addressed either. The present study deals with all these factors.

7 | THIS STUDY

To sum up, although clinical diagnosis in this very age group demands particular attention, few studies have focused on preschool children. Furthermore, agreement on clinical relevance has seldom been examined in the few existing studies on preschoolers, and samples were sometimes small and more or less preselected. In addition, assessments of the child's developmental status or of positive child attributes such as social-emotional skills or prosocial behavior have rarely been considered. The issue of measurement invariance has also often been neglected.

In the present study, we investigated parent-teacher agreement and discrepancy while assessing a wide range of behavioral and emotional problems, but also socio-emotional skills and developmental status. To this end, we analyzed a large and relatively unselected sample of kindergarten children. We based our analyses on both raw scores and standard values informing about raters' views on the clinical relevance of the children's problems. Before analyzing raw scores, we tested the instruments' model fit and measurement invariance across rater groups to ensure our assessments' comparability.

Meta-analyses (mostly based on studies with older children and adolescents) revealed that agreement tends to be stronger for middle-aged children than adolescents. Theoretical considerations (see above) suggest that (a) parent-teacher agreement should be even stronger for preschool children and (b) agreement should be stronger for externalizing than for internalizing problems. Further, (c) agreement might be stronger for children with a clinical diagnosis than for children without. As situational specificity is usually assumed to be a main reason for differences between parent and teacher judgments, we expected that (d) teacher's familiarity would not strengthen agreement (or might even weaken it). With regard to discrepancy, we expected that (e) parents would report more problem

behaviors than teachers. Moderation effects on interrater discrepancy were examined on an exploratory basis. As there has been no study on interrater agreement on developmental status to date, we examined agreement and discrepancy considering developmental status on an exploratory basis, too.

8 | METHOD

8.1 | Participants

Parents and kindergarten teachers (henceforth: teachers) of $n = 922$ kindergartners (449 girls, 467 boys, 6 without information on gender) from urban and suburban areas in seven federal states in Germany provided their assessments. Our sample is a convenience sample from different kindergartens and clinics for child and adolescent psychiatry and psychotherapy recruited between autumn 2013 and summer 2015. The data were collected by five study centers located in Marburg, Dortmund, Koblenz, Saarbrücken, and Rostock. Whereas the study centers in Dortmund and Koblenz approached kindergartens only, the study centers in Saarbrücken and Rostock recruited in clinics only. The study center in Marburg collected data in kindergartens and in a psychiatric clinic as well. Information on the study was distributed via e-mail lists and personal contacts. Only those institutions that responded were included. As we relied on lists, no exact response rate can be determined, although many more institutions were approached than responded. Parents were not approached individually, except for the patient sample.

Children's age ranged from 2 to 6 years ($M = 3.99$, $SD = 1.25$). Eighty-one children (32 girls and 59 boys) were recruited in clinics and had received a diagnosis, although no information on the diagnosis' type or frequency was available for data-protection reasons. All diagnoses were ICD-10 based and made by trained senior staff at the clinics. Teachers were not systematically informed about the children's diagnostic status.

In 490 cases, mothers provided the parents' assessment; in 39 cases, fathers did so; in 54 cases, parents collaborated. For the remaining cases, we had no information as to which parent had made the assessment. On average, mothers were 35.11-years old ($SD = 5.88$) and fathers were 38.82-years old ($SD = 5.73$). We asked the parents to indicate their highest educational level in order to estimate their socioeconomic status (educational level is considered as the broadest and most stable single proxy for socioeconomic status; Sirin, 2005). Parents with a higher educational level were overrepresented: 43.3% of the mothers and 50.7% of the fathers had either a university or a university of applied sciences degree. The Abitur (i.e., the highest school-leaving qualification in Germany) was the highest educational level for 10.1% of the mothers and 3.7% of the fathers; 23.3% of the mothers and 18.9% of the fathers had a school-leaving qualification lower than the Abitur; 1.8% of the mothers and 1.8% of the fathers had no school-leaving qualification.

Of the teachers, 97.6% were female. They were on average 41.27-years old ($SD = 11.33$) and they knew the rated child for $M = 20.45$ months ($SD = 12.69$).

8.2 | Measures

Depending on the local conditions in the study centers, different combinations of questionnaires were given to the parents and teachers. As Table 1 shows, 555 parents and 450 teachers were given all three questionnaires (described in more detail below). However, in other cases, parents and teachers were given only one or two instruments. In some cases, assessments were collected from just one informant (i.e., parent *or* teacher). In 437 cases, the child was rated on all three instruments by the parents *and* the teacher. In five cases, the same two instruments (i.e., Conners Early Childhood [EC] and Strengths and Difficulties Questionnaire [SDQ]) were filled out by parents and teachers, and in 25 cases, one (the same) instrument (the Conners EC) was filled out by parents and teachers. To evaluate model fit, we used all the data available from the respective instrument. To assess measurement invariance and compare parents' and teachers' assessments, however, we used only those cases in

which the child had been rated by his or her parents *and* the teacher, so that parents' and teachers' assessments could be directly compared.

8.2.1 | Conners Early Childhood

n = 667 teachers and *n* = 795 parents rated the children on the German version (Harbarth, Steinmayr, Neidhardt, & Christiansen, 2017) of the Conners EC (Conners, 2009). The Conners EC allows a DSM (American Psychiatric Association, 2013)-based assessment of the behavioral and emotional problems of children aged 2 to 6 years via a teacher and a parent questionnaire (Harbarth et al., 2017). The behavioral and emotional problems are assessed on four empirical behavior scales (Inattention/Hyperactivity, Defiant/Aggressive Behaviors, Social Functioning/Atypical Behaviors, and Anxiety) and two theoretically driven scales (Mood and Affect and Physical Symptoms). However, as the latter two were not verifiable for the teachers in the present sample (see also Harbarth et al., 2017), we excluded them from our analyses.

Teacher and parent versions of the problem behavior scales differ to some extent in the number of items and their content. When analyzing interrater agreement and discrepancy in raw scores, we only relied on similar items, that is, 16 items for Inattention/Hyperactivity ($\alpha = 0.95$ for teachers, $\alpha = 0.94$ for parents), 15 items for Defiant/Aggressive Behaviors ($\alpha = 0.92/0.83$), 22 items for Social Functioning/Atypical Behaviors ($\alpha = 0.85/0.71$), and 13 items for Anxiety ($\alpha = 0.85/0.80$). On all items, parents or teachers were asked how often the child exhibited the respective behavior in the last month. Items were answered on a 4-point scale (0 = not true at all [never, seldom], 3 = very applicable [very often, very frequently]).

Furthermore, with the Conners EC, a child's developmental status can be assessed on five theoretically driven scales (Adaptive Skills, Communication, Motor Skills, Play, and Pre-Academic/Cognitive Skills). The overlap between parent and teacher versions used to analyze raw scores was 12 items for Adaptive Skills ($\alpha = 0.91/0.87$), 17 items for Communication ($\alpha = 0.93/0.91$), 17 items for Motor Skills ($\alpha = 0.93/0.91$), five items for Play ($\alpha = 0.85/0.83$), and 19 items for Pre-Academic/Cognitive Skills ($\alpha = 0.95/0.93$). On all items, informants were asked whether the child does the respective task without help. All items were answered on a 3-point scale, with 0 = no (never or rarely), 1 = sometimes, and 2 = yes (always or almost always).

TABLE 1 Contingency table of teachers and parents having filled out different combinations of instruments

	Parents								
	All	Two instruments			One instrument			No	
Teachers	instruments	CEC+SDQ	CEC+VBV	SDQ+VBV	CEC	SDQ	VBV	instrument	Σ
All instruments	437	0	3	1	0	0	0	9	450
Two instruments									
CEC+SDQ	3	5	0	0	2	0	0	0	10
CEC+VBV	63	0	0	1	0	0	0	1	65
SDQ+VBV	2	0	0	0	0	0	0	1	3
One instrument									
CEC	0	3	0	0	25	0	0	114	142
SDQ	0	0	0	0	0	0	0	0	0
VBV	7	0	0	0	0	0	0	0	7
No instrument	43	0	0	0	202	0	0	0	245
Σ	555	8	3	2	229	0	0	125	922

Abbreviations: CEC, Conners Early Childhood; SDQ, Strengths and Difficulties Questionnaire; VBV, Verhaltensbeurteilungsbogen für Vorschulkinder 3–6 (Questionnaire for Assessing Preschool Children's Behavior 3–6).

8.2.2 | Strengths and Difficulties Questionnaire

n = 463 teachers and *n* = 565 parents rated the children on the German version (Klasen, Woerner, Rothenberger, & Goodman, 2003) of the SDQ (Goodman, 1997). The SDQ assesses behavioral and emotional problems as well as prosocial behavior of children and adolescents aged 4 to 16 years (although it may also be applied to 2- and 3-year-olds). It comprises the Emotional Problems (α = 0.70/0.73), Conduct Problems (α = 0.69/0.61), Hyperactivity (α = 0.86/0.85), Peer Problems (α = 0.75/0.65), and Prosocial Behavior (α = 0.82/0.72) scales, each being represented by five items formulated as statements answered on a 3-point scale, with 0 = not true, 1 = somewhat true, and 2 = certainly true regarding the last 6 months.

8.2.3 | Questionnaire for Assessing Preschool Children's Behavior

n = 525 teachers and *n* = 560 parents rated the children on the Verhaltensbeurteilungsbogen für Vorschulkinder 3–6 (VBV; Questionnaire for Assessing Preschool Children's Behavior 3–6; Döpfner, Berner, Fleischmann, & Schmidt, 1993), a common German questionnaire for screening behavioral and emotional problems and social-emotional skills of 3- to 6-year-olds. The VBV consists of the Social-Emotional Skills, Oppositional-Aggressive Behaviors, Hyperactivity versus Playing Endurance, and Emotional Problems scales. The overlap between the parent and teacher versions was five items for Social-Emotional Skills (α = 0.71/0.59), 18 items for Oppositional-Aggressive Behaviors (α = 0.92/0.93), seven items for Hyperactivity versus Playing Endurance (α = 0.87/0.86), and seven items for Emotional Problems (α = 0.74/0.72). On all items, informants were asked how often the child had shown the respective behavior in the last 4 weeks. All items were answered on a 5-point scale (0 = never, 4 = several times daily).

8.3 | Analyses

8.3.1 | Model fit

To test whether the theoretical structure of the (reduced) instruments was tenable in our study sample, we evaluated all the instruments' model fit in the overall sample, in the teacher sample, and in the parent sample. To evaluate model fit, we used the χ^2 value, the comparative fit index (CFI), the Tucker-Lewis Index (TLI), the root mean square error of approximation (RMSEA), and— in the case of continuous data—the standardized root mean square residual (SRMR). Given that the χ^2 value is sensitive to larger sample sizes, we placed stronger focus on the other fit indices.

Usually, a CFI (TLI) ≥0.95 is considered indicative of a good or at least satisfactory fit (Schermelleh-Engel, Moosbrugger, & Müller, 2003). Still, in some cases (e.g., for questionnaire data) even a CFI (TLI) ≥0.90 can indicate an acceptable fit (Marsh, Hau, & Wen, 2004). With regard to the RMSEA, values ≤0.05, ≤0.08, and ≤0.10 indicate a good, an acceptable, and a mediocre fit, respectively; for the SRMR, values <0.05 and <0.10 indicate a good and an acceptable fit, respectively (Schermelleh-Engel et al., 2003).

For continuous data, the robust maximum likelihood (MLR) estimator was used. For categorical data (Conners developmental milestones and SDQ), we applied the mean and variance adjusted weighted least squares (WLSMV) estimator.

8.3.2 | Measurement invariance

Before further analyzing the data, we tested for metric and scalar invariance of all scales by adding constraints stepwise for the respective model parameters (factor loadings, intercepts/thresholds). We employed the χ^2 difference test (with Satorra-Bentler correction when analyzing models with exclusively continuous data and the derivatives from the respective less restrictive model when analyzing models containing categorical data)

and—because of the strictness of the χ^2 difference test in larger samples—especially the changes in CFI and RMSEA to identify non-invariance. Both ΔCFI and ΔRMSEA have proven to be sensitive to all types of non-invariance and to be relatively unaffected by sample size and model complexity (Chen, 2007; Cheung & Rensvold, 2002). A ΔCFI ≥ −0.01 in combination with a ΔRMSEA ≥ 0.015 was taken as indicative of non-invariance (Chen, 2007; Cheung & Rensvold, 2002).

8.3.3 | Interrater agreement

We computed Pearson correlations and intraclass correlations (ICC) to examine interrater agreement. Unlike the Pearson correlations, ICCs consider the variance between two raters on the individual level, and rating variance across all children (e.g., McDonald et al., 2016). The ICCs were derived from a two-way random effects model (absolute agreement, single measure; McGraw & Wong, 1996; Shrout & Fleiss, 1979). Both agreement measures were computed for our overall sample and separately by clinical diagnostic status in order to investigate potential moderation effects of clinical diagnostic status on interrater discrepancies. To investigate moderation effects from a teacher's familiarity with the child (a continuous variable), we conducted moderated regression analyses predicting the parent assessment from *z*-standardized teacher assessments, *z*-standardized familiarity, and their interaction term.

In addition, we investigated interrater agreement on the child's mental status (clinically relevant vs. clinically irrelevant) and whether clinical diagnostic status and familiarity would moderate agreement. For this purpose, we relied on the complete parent and teacher versions and converted the raw scores into standard values based on the instruments' norming samples.[1] The cut-off values between "clinically relevant" and "clinically irrelevant" were chosen according to the instrument manuals. Symptoms were defined as clinically relevant if $T \geq 65$ (Conners EC; equals a percentile rank ≥93), percentile rank ≥90 (SDQ), and Stanine ≥8 (VBV; equals a percentile rank ≥93).

We then calculated Cohen's κ. We used Shrout's (1998) guidelines to interpret κ values in clinical research (0 to 0.10: virtually no agreement; 0.11 to 0.40: slight; 0.41 to 0.60: fair; 0.61 to 0.80: moderate; 0.81 to 1: substantial). Furthermore, we conducted receiver operating characteristic (ROC) analyses with the parents' assessments as the reference value (note, however, that the choice of either group's assessments as the reference is arbitrary). That is, we computed the sensitivity, specificity, positive predictive power (PPP), and negative predictive power (NPP) of the teacher ratings. In addition, we referred to the area under the curve (AUC) as an effect size. Finally, we conducted binary logistic regression analyses with interrater agreement on mental status (0 = no agreement, 1 = agreement) as dependent variable and diagnostic status (0 = no diagnosis, 1 = diagnosis) and teacher's familiarity with the child (in months) as independent variables to investigate whether diagnostic status or familiarity would moderate agreement.

8.3.4 | Interrater discrepancy

Differences in latent means between groups were derived either from the scalar invariance models from a multi-group analysis of mean and covariance structures (MG-MACS; Conners developmental milestones, SDQ) or from a multiple indicators-multiple causes (MIMIC) model, wherein scalar non-invariance was controlled for (Conners behavior scales, VBV), depending on results from the scalar invariance tests (see below and Supporting Information Material S2). Moderation of interrater discrepancy by diagnostic status was tested by computing discrepancies separately according to diagnostic status. Moderation by a teacher's familiarity with the child was tested by means of regression analyses with discrepancies at the individual level as the dependent variable and familiarity as the independent variable.

These analyses were performed with Mplus 6.12 and with SPSS 25. All estimations in Mplus terminated normally and there were no convergence problems. Missing data were handled with the full information maximum likelihood approach for analyses conducted in Mplus. For analyses with manifest variables in SPSS, we followed the

instructions in the instruments' manuals on how to handle missing values as we were aiming to investigate interrater agreement and discrepancy as they occur in clinical practice (i.e., scales with more missing values than permissible according to the manuals were excluded from the analyses).

9 | RESULTS

9.1 | Preliminary analyses

First, we inspected multivariate normality of the continuous variables. Mardia's test revealed clear non-normality for both the Conners behavior scales and the VBV. The distributions were rather exponential in nature following a probability density function, which, however, is expected for clinical ratings. In the case of non-normality, the MLR estimator (in contrast to the ML estimator) provides an unbiased estimation of model fit. The Mahalanobis distance (calculated separately according to diagnostic status) revealed no extreme multivariate outliers.

Second, we inspected the model fit of the Conners EC, the SDQ, and the VBV, and tested for their metric and scalar invariance. The fit of all models was satisfactory. Detailed results of the model tests are found in the Supporting Information Material S1. Full scalar invariance was given for the Conners developmental milestones and the SDQ. We noted a considerable lack of scalar invariance for the Conners behavior scales and for the VBV. To control for this scalar non-invariance, we set up MIMIC models in which this lack could be controlled for, making reasonable interpretations of mean differences possible. Detailed results of these invariance tests and information on the MIMIC models are found under Supporting Information Material S2.

Descriptive statistics and intercorrelations of the scales are found in Tables S3 and S4 in the Supporting Information Material S3. The means in problem behavior scales were relatively low for children without a clinical diagnosis, but higher for children with one. We observed the opposite pattern in conjunction with developmental status and social skills/prosocial behavior. Table S4 shows that the correlational patterns fell in line with what would be theoretically likely (i.e., high correlations between comparable symptom scales from different instruments, lower correlations between scales targeting different symptoms, negative correlations between problem scales and developmental milestones as well as scales assessing social skills/prosocial behavior).

Below, we turn to the results of our main analyses, that is, (a) correlations between parents' and teachers' ratings (raw scores), (b) agreement between parents and teachers on the clinical relevance of the children's problems, and (c) mean differences between parents' and teachers' ratings.

9.2 | Correlations between parents' and teachers' ratings

We had predicted in Hypothesis 1 that parent-teacher agreement would be stronger for preschool children than the agreement reported in previous studies with older children, which found correlations of $r \approx 0.30$. Pearson correlations and ICCs are shown in Tables 2 and 3, respectively. When averaged across the mean correlations for both externalizing and internalizing problems (see below), parent-teacher agreement was $r = 0.37$ (adjusted for differences in sample size). Thus, agreement was only slightly stronger than for older children. Agreement for social-emotional skills was $r = 0.38$ and ICC = 0.05. Agreement for prosocial behavior was $r = 0.38$ and ICC = 0.34, replicating meta-analytical findings exactly (Renk & Phares, 2004). Thus, taken together, support for Hypothesis 1 was very limited. Interestingly, agreement on developmental milestones was generally stronger than on problem behaviors and social-emotional skills or prosocial behavior ($0.53 \leq r \leq 0.84$; $0.51 \leq \text{ICC} \leq 0.82$).

In Hypothesis 2, we had predicted that agreement would be stronger for externalizing than for internalizing problems. Pearson correlations and ICCs indicated stronger agreement for externalizing problems (hyperactivity: $0.52 \leq r \leq 0.58$, $0.37 \leq \text{ICC} \leq 0.57$; oppositional/aggressive behavior: $0.37 \leq r \leq 0.45$; $0.35 \leq \text{ICC} \leq 0.44$; peer problems: $r = 0.43$, ICC = 0.43) than for internalizing problems ($0.21 \leq r \leq 0.33$; $0.20 \leq \text{ICC} \leq 0.33$). When averaged across symptom scales (and adjusted for differences in sample size), overall agreement for externalizing problems was $r = 0.47$ and overall agreement for

internalizing problems was $r = 0.26$. This difference was statistically significant ($z = 3.79$, $p < 0.001$), thus supporting Hypothesis 2.

In Hypothesis 3, we had predicted that agreement would be stronger for children with a clinical diagnosis than for those without. When the Pearson correlations were evaluated separately by diagnostic status, only three moderation effects emerged. The correlation was higher for children with a clinical diagnosis than for those without one on the Conners EC Inattention/Hyperactivity ($r = 0.69$ vs. $r = 0.41$, $z = 2.62$, $p = 0.009$), the developmental milestone Play ($r = 0.70$ vs. $r = 0.47$, $z = 2.27$, $p = 0.023$), and SDQ Hyperactivity ($r = 0.72$ vs. $r = 0.50$, $z = 2.25$, $p = 0.024$). With regard to the ICCs, we found only one moderation effect: The ICC was significantly higher for children with a clinical diagnosis than for those without one on the Conners EC Inattention/Hyperactivity scale (ICC = 0.69 vs. ICC = 0.40), given that both 95% confidence intervals did not overlap. No other moderation effect occurred, revealing weak support for Hypothesis 3.

In Hypothesis 4, we had predicted that teacher's familiarity with the child would not strengthen interrater agreement. Indeed, parent-teacher agreement was moderated by familiarity only occasionally. Only four out of 18 possible moderation effects were statistically significant. Parent-teacher agreement was somewhat stronger on Conners EC Inattention/Hyperactivity ($b_{\text{familiarity}} = -0.18$, *n.s.*, $b_{\text{interaction}} = 0.73$, $p = 0.035$), on Conners EC Anxiety

TABLE 2 Pearson correlations between parents' and teachers' assessments

	Entire sample		By clinical diagnostic (CD) status			
Scale	***N***	***r***	$n_{\text{CD/no CD}}$	r_{CD}	$r_{\text{no CD}}$	$z_{\text{CD-no CD}}$
Conners behavior scales						
Inattention/hyperactivity	532	.52***	47/485	.69***	.41***	2.62**
Defiant/aggressive behaviors	532	.37***	45/487	.33*	.35***	−0.14
Social functioning/atypical behavior	529	.45***	47/482	.44**	.39***	0.38
Anxiety	534	.21***	47/487	.24	.19***	0.33
Conners developmental milestones						
Adaptive skills	529	.81***	47/482	.81***	.83***	−0.39
Communication	526	.75***	46/480	.80***	.73***	1.07
Motor skills	512	.82***	46/466	.81***	.82***	−0.19
Play	527	.53***	47/480	.70***	.47***	2.27*
Pre-academic/cognitive skills	506	.84***	46/460	.87***	.83***	0.91
SDQ						
Emotional problems	447	.33***	47/399	.07	.35***	−1.86
Conduct problems	447	.45***	47/399	.35*	.41***	−0.44
Hyperactivity	446	.58***	47/399	.72***	.50***	2.25*
Peer problems	446	.43***	47/399	.29*	.43***	−1.02
Prosocial behavior	446	.38***	47/399	.51***	.33***	1.38
VBV						
Social-emotional skills	489	.38***	50/439	.44**	.33***	0.84
Oppositional-aggressive behavior	471	.38***	50/421	.29*	.35***	−0.43
Hyperactivity vs. playing endurance	482	.54***	48/434	.64***	.46***	1.67
Emotional problems	493	.25***	52/441	.05	.28***	−1.58

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

TABLE 3 Intraclass correlations between parents' and teachers' assessments

	Entire sample		By clinical diagnostic (CD) status			
Scale	ICC	CI	ICC_{CD}	CI_{CD}	$ICC_{no\ CD}$	$CI_{no\ CD}$
Conners behavior scales						
Inattention/hyperactivity	.51***	.44–.57	.69***	.51–.81	.40***	.32–.48
Defiant/aggressive behaviors	.35***	.25–.43	.31*	.04–.54	.32***	.23–.41
Social functioning/atypical behavior	.45***	.38–.52	.45**	.18–.65	.38***	.30–.45
Anxiety	.20***	.11–.29	.21	−.06–.46	.17***	.08–.26
Conners developmental milestones						
Adaptive skills	.82***	.79–.84	.79***	.66–.88	.82***	.79–.85
Communication	.71***	.61–.78	.77***	.61–.87	.68***	.58–.76
Motor skills	.81***	.77–.84	.77***	.62–.87	.81***	.78–.84
Play	.51***	.42–.58	.70***	.52–.82	.45***	.36–.53
Pre-academic/cognitive skills	.79***	.58–.88	.80***	.36–.92	.79***	.59–.87
SDQ						
Emotional problems	.33***	.24–.41	.07	−.22–.35	.35***	.26–.43
Conduct problems	.44***	.36–.51	.34**	.07–.57	.40***	.31–.48
Hyperactivity	.57***	.51–.63	.73***	.56–.84	.50***	.42–.57
Peer problems	.43***	.35–.50	.26*	−.02–.51	.43***	.35–.51
Prosocial behavior	.34***	.24–.43	.40***	.08–.63	.31***	.21–.40
VBV						
Social-emotional skills	.05***	−.04–.16	.12*	−.08–.34	.04***	−.04–.13
Oppositional-aggressive behavior	.36***	.28–.44	.28*	.00–.51	.33***	.24–.41
Hyperactivity vs. playing endurance	.37***	.19–.51	.36***	−.01–.62	.32***	.16–.44
Emotional problems	.21***	.12–.30	.04	−.22–.31	.24***	.15–.34

Abbreviations: CI, 95% confidence interval; ICC, intraclass correlation derived from a two-way random effects model (absolute agreement, single measure).
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

($b_{familiarity} = -0.06$, *n.s.*, $b_{interaction} = 0.44$, $p = 0.023$), and on SDQ Hyperactivity ($b_{familiarity} < 0.01$, *n.s.*, $b_{interaction} = 0.18$, $p = 0.047$) if teachers had known the child for longer. On Conners EC Adaptive Skills, agreement was slightly weaker if teachers had known the child for longer ($b_{familiarity} = -0.48$, $p < 0.001$, $b_{interaction} = -0.40$, $p = 0.015$). Each of these moderation effects was small ($0.08 \leq \beta \leq 0.10$; $0.004 \leq \Delta R^2 \leq 0.010$). Thus, in short these findings supported Hypothesis 4.

9.3 | Interrater agreement on clinical relevance

Thus far, we have investigated interrater agreement in terms of raw scores only. Yet in clinical practice, the most common key issue is whether parents and teachers agree in their appraisal of the child's mental status. Therefore, we also investigated whether both teachers and parents perceived the child's problem behaviors, social-emotional skills, and developmental status as clinically relevant.

Agreement on clinical relevance is illustrated in Table 4. Although all κ values were statistically significant, agreement was not substantial for any of the scales according to Shrout (1998). With regard to problem behaviors,

TABLE 4 Interrater agreement on child's mental status (clinically relevant vs. not clinically relevant)

	% clinically relevant			Receiver operating characteristic analysis					
Scale	Parents	Teachers	κ	Sensitivity	Specificity	PPP	NPP	AUC	95% CI_{AUC}
Conners behavior scales									
Inattention/ hyperactivity	9.14	9.92	.46***	.53	.94	.49	.95	.74***	.65–.83
Defiant/aggressive behaviors	8.14	7.36	.21***	.26	.94	.29	.94	.60*	.50–.70
Social functioning/ atypical behavior	7.42	8.01	.44***	.50	.95	.46	.96	.73***	.63–.83
Anxiety	7.95	7.16	.12**	.18	.94	.19	.93	.56	.46–.66
Conners developmental milestones									
Adaptive skills	9.33	13.69	.61***	.81	.93	.55	.98	.87***	.80–.94
Communication	11.01	10.55	.67***	.69	.97	.72	.96	.83***	.75–.91
Motor skills	9.25	13.10	.52***	.69	.93	.49	.97	.81***	.73–.89
Play	8.63	9.41	.33***	.41	.94	.38	.94	.67***	.58–.77
Pre-academic/ cognitive skills	9.59	12.04	.62***	.74	.95	.59	.97	.85***	.77–.92
SDQ									
Emotional problems	11.87	11.87	.25***	.34	.91	.34	.91	.63**	.53–.72
Conduct problems	7.58	12.12	.15**	.30	.89	.19	.94	.60	.48–.71
Hyperactivity	10.13	11.90	.45***	.55	.93	.47	.95	.74***	.64–.84
Peer problems	15.95	17.72	.29***	.43	.87	.39	.89	.65***	.57–.73
Prosocial behavior	8.86	6.58	.24***	.26	.95	.35	.93	.61*	.50–.72
VBV									
Social-emotional skills	12.35	10.90	.32***	.37	.93	.42	.91	.65**	.56–.74
Oppositional-aggressive behavior	9.51	9.28	.17***	.24	.92	.25	.92	.58	.48–.68
Hyperactivity vs. playing endurance	11.20	10.67	.48***	.52	.95	.55	.94	.74***	.64–.83
Emotional problems	11.76	12.83	.23***	.34	.90	.31	.91	.62**	.52–.72

Note: Parents' assessments were used as reference value for the ROC analyses.
Abbreviations: AUC, area under the curve; CI, confidence interval; NPP, negative predictive power; PPP, positive predictive power.
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

agreement was rather weak (oppositional/aggressive behavior: $0.15 \leq \kappa \leq 0.21$; peer problems: $\kappa = 0.29$; internalizing problems: $0.12 \leq \kappa \leq 0.25$). Only for hyperactivity ($0.45 \leq \kappa \leq 0.48$) and social functioning/atypical behavior ($\kappa = 0.44$), agreement was fair. The results from the ROC analyses mirrored the results from the κ statistic (see Table 4). The AUC was small and partly not above chance for oppositional/aggressive behavior ($0.58 \leq AUC \leq 0.60$), peer problems (AUC = 0.65), and internalizing problems ($0.56 \leq AUC \leq 0.63$). Agreement was moderate at least for hyperactivity and social functioning/atypical behavior. For example, in 52 to 55% of all cases in which parents rated

their child as exhibiting clinically relevant hyperactivity, teachers also rated the child as hyperactive (sensitivity). In 93 to 95% of all cases in which parents rated their child as non-hyperactive, the teachers did too (specificity). In 47 to 55% of all cases in which teachers rated the child as hyperactive, the parents agreed (PPP). And in 94 to 95% of all cases in which teachers rated the child as non-hyperactive, parents did too (NPP). AUC = 0.74 ($p < .001$) for all hyperactivity scales and AUC = 0.73 ($p < 0.001$) for social functioning/atypical behavior. Agreement was also weak for social-emotional skills ($\kappa = 0.32$; AUC = 0.65, $p = 0.001$) and prosocial behavior ($\kappa = 0.24$; AUC = 0.61, $p = 0.040$). Again, these findings only yield limited support for Hypothesis 1. The evidence that agreement on clinical relevance was stronger for internalizing than for externalizing problems (Hypothesis 2) was only partial. The level of agreement on clinical relevance depended instead on the specific problem behavior in question.

Agreement tended to be stronger on the developmental milestones, however. We observed at least moderate agreement on three of the five developmental milestones (adaptive skills: $\kappa = 0.61$; communication: $\kappa = 0.67$; pre-academic/cognitive skills: $\kappa = 0.62$). For motor skills, $\kappa = 0.52$, and for play, $\kappa = 0.33$. Sensitivity, specificity, PPP, and NPP were relatively high for the developmental milestones, resulting in quite high AUC values ranging from 0.81 to 0.87 (with the exception of play, AUC = 0.67), all of those exceeding chance ($p < 0.001$).

Finally, we investigated whether a child's clinical diagnostic status and teacher's familiarity with that child would moderate agreement on clinical relevance, again testing Hypotheses 3 and 4. Binary logistic regression analyses showed that clinical diagnostic status was a quite strong and consistent moderator for problem behaviors and social-emotional skills. Agreement was stronger for children with a clinical diagnosis than for those without one on all Conners behavior scales, four of the five SDQ scales, and on three of the four VBV scales (Table 5). Agreement was about two to seven times more likely for children with a clinical diagnosis than for those without. This pattern appeared strongest for social functioning/atypical behavior (odds ratio [OR] = 7.38). It was also clear and consistent across instruments for oppositional/aggressive behavior ($4.63 \leq OR \leq 6.07$), hyperactivity ($2.72 \leq OR \leq 4.36$), and internalizing problems ($1.94 \leq OR \leq 4.42$). For social-emotional competencies, OR = 2.60. Hypothesis 3 was thus well supported. By contrast, we detected just one such moderation effect for the developmental milestones (play: OR = 3.18).

With regard to teacher's familiarity with the child, only a few, relatively weak moderation effects appeared in conjunction with problem behaviors (Conners EC Social Functioning/Atypical Behavior: OR = 1.06; VBV Emotional Problems: OR = 1.03; SDQ Peer Problems: 1.02), again supporting Hypothesis 4. Instead, we noted consistent and stronger moderation effects for the developmental milestones. Each additional month of familiarity made agreement between teachers and parents 1.07 to 1.12 times more likely.

9.4 | Mean differences between parents' and teachers' ratings

According to Hypothesis 5, we had expected that parents would report more problems than teachers. Table 6 shows the differences between the latent means of parents' and teachers' assessments (positive d values indicate greater means for parents). Parents reported significantly higher values than did teachers on eight of the 11 problem behavior scales. The effect sizes were small for seven of these eight scales ($0.19 \leq d \leq 0.40$). The difference was medium-sized ($d = 0.51$) only on VBV Oppositional-Aggressive Behavior. When averaged across the problem behavior scales (and adjusted for sample size), the parent-teacher discrepancy was $d = 0.23$ for externalizing problems and $d = 0.27$ for internalizing problems. We thus found no strong evidence of larger discrepancies for externalizing or internalizing problems. Taken together, Hypothesis 5 was supported. Parents also rated their children higher than did teachers on prosocial behavior ($d = 0.35$).

With regard to developmental status, parents rated their children's skills higher than did teachers in pre-academic/cognitive skills ($d = 0.17$) and on developmental milestones required for positive social interaction (communication: $d = 0.29$, play: $d = 0.31$). There was virtually no difference on adaptive skills and motor skills.

On an exploratory basis, we investigated whether the direction or magnitude of the mean differences would differ according to the child's clinical diagnostic status and according to a teacher's familiarity with the child. Although there were some descriptive differences in d values between children with and children without a

TABLE 5 Logistic regression analysis for the prediction of interrater agreement (0 = no agreement, 1 = agreement) from clinical diagnostic status (0 = no diagnosis, 1 = diagnosis) and teacher's familiarity with the child (months)

	Clinical diagnostic status			Teacher's familiarity with child				
Scale	*B*	*SE*	OR	*B*	*SE*	OR	C&S-R^2	N-R^2
Conners behavior scales								
Inattention/hyperactivity	1.47**	0.44	4.36	0.01	0.01	1.01	.02	.05
Defiant/aggressive behaviors	1.80***	0.41	6.07	0.02	0.01	1.02	.04	.07
Social functioning/atypical behavior	2.00***	0.45	7.38	0.06**	0.02	1.06	.05	.12
Anxiety	1.49***	0.41	4.42	−0.01	0.01	1.00	.03	.05
Conners developmental milestones								
Adaptive skills	0.73	0.59	2.08	0.09***	0.02	1.10	.06	.13
Communication	0.98	0.60	2.66	0.07**	0.02	1.07	.04	.09
Motor skills	0.60	0.60	1.82	0.11***	0.02	1.12	.08	.17
Play	1.16*	0.49	3.18	0.09***	0.02	1.09	.07	.14
Pre-academic/cognitive skills	−0.01	0.78	0.99	0.10***	0.02	1.11	.06	.14
SDQ								
Emotional problems	1.33**	0.40	3.78	0.01	0.01	1.01	.03	.05
Conduct problems	1.53***	0.40	4.63	−0.02	0.01	0.99	.04	.08
Hyperactivity	1.00*	0.47	2.72	0.03	0.01	1.03	.02	.04
Peer problems	0.82*	0.40	2.27	0.02*	0.01	1.02	.02	.03
Prosocial behavior	0.21	0.57	1.24	0.02	0.01	1.02	<0.01	.01
VBV								
Social-emotional skills	0.96*	0.43	2.60	0.05***	0.01	1.06	.05	.08
Oppositional-aggressive behavior	1.73***	0.40	5.66	0.02	0.01	1.02	.04	.07
Hyperactivity vs. playing endurance	1.39**	0.46	4.01	0.02	0.02	1.02	.02	.05
Emotional problems	0.66	0.42	1.94	0.03*	0.01	1.03	.02	.03

Abbreviations: C&S R^2, Cox & Snell R^2; N-R^2, Nagelkerke's R^2; OR, odds ratio.
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

diagnosis, none of those differences were statistically significant (see Table 6). Thus, child's diagnostic status did not moderate interrater discrepancy. With regard to teachers' familiarity, only two moderation effects appeared, both referring to developmental status. The discrepancy between parent and teacher ratings of adaptive skills ($b = 0.02$, $p = 0.028$) and communication ($b = 0.05$, $p = 0.003$) became slightly larger if teachers had known the child longer.

10 | DISCUSSION

In the present study, we investigated parent-teacher agreement and discrepancy when assessing kindergarten children's behavioral and emotional problems, social-emotional skills, and developmental status. We relied on a

TABLE 6 Differences in latent means of parents' and teachers' assessments for the entire sample and by clinical diagnostic status

Scale	Entire sample *d*	By clinical diagnostic (CD) status		
		d_{CD}	$d_{no\ CD}$	$z_{CD\text{-}no\ CD}$
Conners behavior scales				
Inattention/hyperactivity	0.19**	0.11	0.22**	−0.14
Defiant/aggressive behaviors	0.38***	0.47	0.38***	0.81
Social functioning/atypical behavior	0.27**	0.35	0.27**	0.70
Anxiety	0.40***	0.36	0.42***	0.31
Conners developmental milestones				
Adaptive skills	0.05	0.09	0.04	0.16
Communication	0.29**	0.25	0.30**	−0.16
Motor skills	0.07	0.19	0.06	0.41
Play	0.31***	0.16	0.34***	−0.57
Pre-academic/cognitive skills	0.17*	0.33	0.16	0.53
SDQ				
Emotional problems	0.04	-	0.01	-
Conduct problems	0.26**	-	0.25**	-
Hyperactivity	0.11	-	0.10	-
Peer problems	−0.06	-	−0.03	-
Prosocial behavior	0.35***	-	0.29**	-
VBV				
Social-emotional skills	0.17	0.40	0.15	1.03
Oppositional-aggressive behavior	0.51***	0.56*	0.52***	0.84
Hyperactivity vs. playing endurance	0.22**	0.30	0.22**	0.65
Emotional problems	0.35***	0.29	0.37***	−0.34

Note: Positive *d* values indicate greater means for parents. Results for the Conners behavior scales and for the VBV were derived from the scalar invariance MIMIC models, results for the Conners developmental milestones and for the SDQ were derived from MG-MACS analyses. Differences in latent means on the SDQ could not be computed for the CD group because some of the SDQ items had at least one empty cell in their bivariate frequency table so that these items could not be used for the CD group.

*$p < 0.05$.

**$p < 0.01$.

***$p < 0.001$.

rather unspecific sample of parents and teachers of overall 922 kindergarten children aged between 2 and 6 years, finding evidence of an adequate model fit and measurement invariance before analyzing the data. Therefore, the current findings do not reflect a lack of measurement invariance but rather "genuine" agreement and discrepancy.

10.1 | Interrater agreement and discrepancy on kindergarten children's problem behaviors and social-emotional skills

Assuming that younger children's behavior is both more consistent across situations and easier to observe, we had expected that parent-teacher agreement would be stronger for preschool children than the agreement reported in previous studies focusing on older children and adolescents (e.g., Achenbach et al., 1987; De Los Reyes et al., 2015; Stone et al., 2010). In the present study, overall parent-teacher agreement on children's

problem behaviors (raw scores) was somewhat stronger than in studies with older children, but not by much ($r = 0.37$ vs. $r \approx 0.30$). The level of agreement for social-emotional skills detected in this study even replicated exactly results obtained in the Renk and Phares (2004) meta-analysis. Agreement on clinical relevance was stronger than for older children only in terms of hyperactivity and social functioning/atypical behavior, but not regarding other problem behaviors or social-emotional skills. Therefore, agreement for very young children seems—at best—to be only slightly stronger than for older children. This suggests that the behavior of kindergarten children is not much more consistent across situations or easier to capture than is the behavior of schoolchildren. Studies have shown that self-regulation development with regard to emotions and behavior increases sharply from age three on (e.g., Carlson, 2005; Josephs, 1994). Therefore, preschool children might already understand how they should behave away from home (e.g., in kindergarten), and possess enough self-regulatory skill to do so. Their behavior might therefore be more context-specific than previously thought, resulting in levels of interrater agreement comparable with that of older children.

In accordance with our expectations, parents tended to rate their children's problems higher than did teachers. As noted above, preschool children might already have learned to behave better in kindergarten or in other contexts outside the home. Therefore, they might exhibit more problematic behavior at home than in kindergarten, so that their parents would report more problems than the child's teachers.

We observed quite consistent results for the same problem behavior across instruments. This was even true although the target children rated on one instrument were not always the same children rated on the other instruments. All in all, this finding suggests that differences in interrater agreement between studies are not that much a question of the instruments being employed (as long as measurement invariance has been established) but maybe rather of sample characteristics. Like any other sample, ours might also have its peculiarities. We challenge future researchers to try to replicate our findings.

10.2 | Moderation effects on interrater agreement and discrepancy

As we expected from previous research (e.g., Achenbach et al., 1987; Stone et al., 2010), we observed stronger agreement (raw scores) for externalizing than for internalizing problems. Since externalizing behaviors should be more obvious than internalizing problems irrespective of child's age, it makes sense that it was the type of problem behavior that proved to be a moderator at preschool age also (see also Grietens et al., 2004; Winsler & Wallace, 2002). However, with regard to clinical relevance, agreement was only slightly stronger for externalizing than for internalizing problems. It instead seemed to depend on the particular problem behavior.

Clinical diagnostic status was a weak moderator of agreement (raw scores): Only three moderation effects became apparent, and in all, agreement was stronger for deviant than nondeviant children. However, if the child had a diagnosis, agreement on the clinical relevance of problem behaviors and social-emotional skills was two to six times stronger than if the child had no diagnosis. This finding is not in line with previous studies and even contradicts findings by Berg-Nielsen et al. (2012) and Kuschel et al. (2007). Nevertheless, our finding makes sense because in children with a diagnosis, problem behaviors and the lack of social skills might be more obvious across settings, thus increasing agreement.

If deficits in agreement are really due to situational specificity, teacher's familiarity with the child should not moderate agreement. Indeed, we found teacher's familiarity with the child was no clear moderator of agreement either in terms of raw scores or clinical relevance. This finding corroborates the importance of situational specificity of children's behaviors in terms of interrater agreement.

10.3 | Interrater agreement and discrepancy on developmental status

To the best of our knowledge, ours is the first study to investigate interrater agreement and discrepancy when assessing children's developmental status. The present study now provides initial evidence that agreement on

developmental status is markedly stronger than agreement on problem behavior and social skills—and becomes even stronger once teachers have known the children longer. It seems that as teachers get to know the child better, they become better at assessing the child's skills in different domains of development. Situational (un-)specificity might explain why we found such strong agreement in conjunction with most developmental milestones. Competence as reflected in motor skills or adaptive skills either exists already or it does not, irrespective of the context (home or kindergarten). We should observe greater cross-situational consistency of behaviors related to these skills than for problem behaviors or social-emotional skills. In turn, this should strengthen agreement and reduce discrepancy between parents and teachers. The cross-situational consistency of such skills is also why we believe that the moderation of agreement on developmental milestones via the teacher's familiarity with the child we observed does not contradict the role of situational specificity in general (as it would have had we detected a moderation of the agreement on problem behaviors, which vary more strongly than skills do).

Parents rated their children somewhat higher than did teachers on three of the five developmental milestones (communication skills, play, and in pre-academic/cognitive skills). At first glance, our finding that parents rated their children's behaviors as more problematic (see above) while rating their children's developmental status in some aspects as more advanced (and for teachers vice versa) might appear contradictory. However, parents and teachers might evaluate a child's behavioral and emotional problems in light of the child's developmental status. If its developmental status is perceived as being comparatively advanced, certain behavioral problems might be judged as being more severe because they lag behind what one would expect for that level of development. If developmental status is perceived as being delayed, the same behavior problems might be rated as minor because they are thought to suffice for that child's current level of development.

10.4 | Implications for diagnosing children

Meanwhile, researchers in this field agree on the importance of the situational specificity of children's emotional and behavioral problems (see Section 1). Thus both parents and teachers can be considered valuable sources capable of providing valid information on context-specific behaviors of the child. Using information from multiple perspectives might help us identify main problem areas and identify etiological factors of the disorder (e.g., factors originating at home vs. at kindergarten/school) in the individual case (Rescorla et al., 2014). For example, if parents are reporting more problem behaviors than teachers, causes for that behavior might be found at home. By contrast, if teachers are reporting more problem behaviors than parents, those causes might be found at kindergarten/school. As a consequence, we need to take a hard look at the information we obtain from teachers in order to maximize the benefits from interrater discrepancies. Traditionally, teachers' questionnaires are either identical to or strongly resemble questionnaires filled out by parents. However, as home and school contexts differ in so many ways, future teacher questionnaires should be designed to more specifically assess how the child functions at school.

Generally speaking, instruments should devote more attention to behavior in concrete situations. When parents and teachers are asked to make their assessments, they tend to perceive disordered behaviors and emotions as a disposition, as proposed by the ABC Model (De Los Reyes & Kazdin, 2005). Indeed, current rating instruments are still constructed to assess a disorder as a trait. In particular, they focus on the overall frequencies of behavior, ignoring its cross-situational variability (Hartley, Zakriski, & Wright, 2011; Rescorla et al., 2014). Thus, new measures should more fully and explicitly incorporate the behavior context (as exemplified by Hartley et al., 2011). This might also help integrate information from parents, teachers, and the child itself to create a coherent picture of the child's problems, thus reducing informant and setting error variance without sacrificing important information.

10.5 | Limitations and future directions

Although the present study examined some moderator effects on interrater agreement and discrepancy, we did not investigate other possible moderators, especially those associated with parents or teachers (e.g., parental

psychopathology or socioeconomic status, immigration background, ethnicity). Higher-educated parents were overrepresented in our sample. Parents who participated in this study might have been more engaged in general and may have sought more interaction with teachers, factors that could have resulted in over-rated interrater agreement. On the other hand, socioeconomic status has not tended to moderate interrater agreement and discrepancy (De Los Reyes & Kazdin, 2005). Furthermore, families with an immigration background or from ethnic minorities might have been underrepresented, although we have no data on that. Another important issue would have been to examine interrater agreement on officially diagnosed symptoms. Unfortunately, we had no information on the exact ICD diagnoses the deviant children had been given. Future studies focusing on preschool children should put emphasis on a larger set of moderators also including parent and teacher variables and, if possible, actual diagnoses when investigating clinical samples. Another point is that teacher assessments are probably influenced by the reference group effect (e.g., Dinnebeil et al., 2013; Schönmoser et al., 2018). We did not investigate whether that was the case in this study, as our goal was to examine interrater agreement and discrepancy at an absolute level, that is, as they appear in clinical practice. However, future studies might determine the strength of the reference group effect on kindergarten teachers' assessments. Further, as already noted, as not all instruments were administered to parents or teachers, our sample sizes for analyses were smaller than our overall sample. However, differences in how our instruments were administered were largely due to local conditions at the study centers. As these conditions were independent from participant characteristics, the missing data were not systematically absent (rather by design), meaning that their absence does not have the potential to substantially distort our findings. Finally, although we had little information on whether it was the mother or father (or both) who had completed the questionnaires, that issue is fairly negligible in terms of parent-teacher agreement and discrepancy (Achenbach et al., 1987; Grietens et al., 2004; Kuschel et al., 2007).

11 | CONCLUSIONS

The present study provides insights into parent-teacher agreement and discrepancy when assessing emotional and behavioral problems, social-emotional skills, and the developmental status of kindergarten children. We identified moderate agreement, which was only slightly stronger than for older children, and notable discrepancy in conjunction with behavior problems and social-emotional skills. However, agreement on developmental status in some areas was strong, and discrepancy was virtually nonexistent. Our findings corroborate the importance of situational specificity for understanding both agreement and discrepancy between the parents' and teachers' appraisal of the child. Both agreement and discrepancy deliver valid information that may be used for diagnosis, for identifying etiological factors, and planning interventions. To improve diagnosis, future instruments for teachers should be designed as to more specifically assess children's functioning at kindergarten or school, respectively.

ETHICAL APPROVAL

All procedures performed in the study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

INFORMED CONSENT

Informed consent was obtained from all individual participants included in the study.

CONFLICT OF INTERESTS

The authors declare that there are no conflicts of interests.

ENDNOTE

[1]For the SDQ, there is currently no German standardized version for teacher assessments. Therefore, we used the Spanish norms of parent and teacher assessments for 3- to 4-year-olds as well as the British norms of parent and teacher assessments for 5- to 10-year-olds, retrieved from http://www.sdqinfo.com. We decided to use the norms from those two countries (and not from one of those countries only) because for Great Britain, there are currently neither any parent norms provided for 4-year-olds nor teacher norms for 3- to 4-year-olds. For Spain, on the other hand, there are no norms provided for children older than 4 years of age. Therefore, using the norms from those two countries allowed us to cover nearly the entire age range of our study sample (with the exception of $n = 128$ 2-year-olds and $n = 33$ children without age specified, both of which were excluded from our analysis). No norms from another country (or combination thereof) provided at the SDQ website would have enabled better coverage. In any case, the use of norms from two different countries should be unproblematic for the purpose of the present study as long as both teacher and parent norms for a certain age (or gender) group stem from the same country.

ORCID

Sebastian Bergold http://orcid.org/0000-0002-6424-9134

REFERENCES

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213–232. https://doi.org/10.1037/0033-2909.101.2.213

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5)*. Washington, D. C.: American Psychiatric Association Publishing.

Angold, A., & Egger, H. L. (2007). Preschool psychopathology: Lessons for the lifespan. *Journal of Child Psychology and Psychiatry*, *48*, 961–966. https://doi.org/10.1111/j.1469-7610.2007.01832.x

Berg-Nielsen, T. S., Solheim, E., Belsky, J., & Wichstrom, L. (2012). Preschoolers' psychosocial problems: In the eyes of the beholder? Adding teacher characteristics as determinants of discrepant parent-teacher reports. *Child Psychiatry & Human Development*, *43*, 393–413. https://doi.org/10.1007/s10578-011-0271-0

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*, *28*, 595–616. https://doi.org/10.1207/s15326942dn2802_3

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Conners, C. K. (2009). *Conners Early Childhood (Conners EC)*. Toronto, Canada: Multi-Health Systems.

Cree, R. A., Bitsko, R. H., Robinson, L. R., Holbrook, J. R., Danielson, M. L., Smith, C., … Peacock, G. (2018). Health care, family, and community factors associated with mental, behavioral, and developmental disorders and poverty among children Aged 2–8 Years—United States, 2016. *MMWR. Morbidity and Mortality Weekly Report*, *67*, 1377–1383. https://doi.org/10.15585/mmwr.mm6750a1

De Los Reyes, A. (2011). Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *40*, 1–9. https://doi.org/10.1080/15374416.2011.533405

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *141*, 858–900. https://doi.org/10.1037/a0038498

De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology*, *37*, 637–652. https://doi.org/10.1007/s10802-009-9307-3

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*, 483–509. https://doi.org/10.1037/0033-2909.131.4.483

Deng, S., Liu, X., & Roosa, M. W. (2004). Agreement between parent and teacher reports on behavioral problems among Chinese children. *Journal of Developmental & Behavioral Pediatrics*, *25*, 407–414. https://doi.org/10.1097/00004703-200412000-00004

Dinnebeil, L. A., Sawyer, B. E., Logan, J., Dynia, J. M., Cancio, E., & Justice, L. M. (2013). Influences on the congruence between parents' and teachers' ratings of young children's social skills and problem behaviors. *Early Childhood Research Quarterly*, *28*, 144–152. https://doi.org/10.1016/j.ecresq.2012.03.001

Döpfner, M., Berner, W., Fleischmann, T., & Schmidt, M. (1993). *Verhaltensbeurteilungsbogen für Vorschulkinder (VBV 3-6) [Questionnaire for Assessing Preschool Children's Behavior]*. Weinheim, Germany: Beltz.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Grietens, H., Onghena, P., Prinzie, P., Gadeyne, E., Van Assche, V., Ghesquière, P., & Hellinckx, W. (2004). Comparison of mothers', fathers', and teachers' reports on problem behavior in 5- to 6-year-old children. *Journal of Psychopathology and Behavioral Assessment*, *26*, 137–146. https://doi.org/10.1023/B:JOBA.0000013661.14995.59

Harbarth, S., Steinmayr, R., Neidhardt, E., & Christiansen, H. (2017). *Conners Skalen zu Aufmerksamkeit, Verhalten und Entwicklungsmeilensteinen im Vorschulalter (Conners EC) [Conners Scales for Attention, Behavior, and Developmental Milestones in Preschool Age]*. Göttingen, Germany: Hogrefe.

Hartley, A. G., Zakriski, A. L., & Wright, J. C. (2011). Probing the depths of informant discrepancies: Contextual influences on divergence and convergence. *Journal of Clinical Child & Adolescent Psychology*, *40*, 54–66. https://doi.org/10.1080/15374416.2011.533404

Josephs, I. E. (1994). Display rule behavior and understanding in preschool children. *Journal of Nonverbal Behavior*, *18*, 301–326. https://doi.org/10.1007/BF02172291

Kennerley, S., Jaquiery, B., Hatch, B., Healey, M., Wheeler, B. J., & Healey, D. (2018). Informant discrepancies in the assessment of attention-deficit/hyperactivity disorder. *Journal of Psychoeducational Assessment*, *36*, 136–147. https://doi.org/10.1177/0734282916670797

Kerr, D. C. R., Lunkenheimer, E. S., & Olson, S. L. (2007). Assessment of child problem behaviors by multiple informants: A longitudinal study from preschool to school entry. *Journal of Child Psychology and Psychiatry*, *48*, 967–975. https://doi.org/10.1111/j.1469-7610.2007.01776.x

Klasen, H., Woerner, W., Rothenberger, A., & Goodman, R. (2003). [German version of the Strength and Difficulties Questionnaire (SDQ-German)--overview and evaluation of initial validation and normative results]. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, *52*, 491–502. http://hdl.handle.net/20.500.11780/2700

Korsch, F., & Petermann, F. (2014). Agreement between parents and teachers on preschool children's behavior in a clinical sample with externalizing behavioral problems. *Child Psychiatry & Human Development*, *45*, 617–627. https://doi.org/10.1007/s10578-013-0430-6

Kuschel, A., Heinrichs, N., Bertram, H., Naumann, S., & Hahlweg, K. (2007). Wie gut stimmen Eltern und Erzieherinnen in der Beurteilung von Verhaltensproblemen bei Kindergartenkindern überein? [How well do parents' and teachers' reports agree on behaviour problems in pre-schoolaged children?]. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, *35*, 51–58. https://doi.org/10.1024/1422-4917.35.1.51

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*, 320–341. https://doi.org/10.1207/s15328007sem1103_2

McCue Horwitz, S., Hurlburt, M. S., Heneghan, A., Zhang, J., Rolls-Reutz, J., Fisher, E., … Stein, R. E. K. (2012). Mental health problems in young children investigated by U.S. child welfare agencies. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*, 572–581. https://doi.org/10.1016/j.jaac.2012.03.006

McDonald, C. A., Lopata, C., Donnelly, J. P., Thomeer, M. L., Rodgers, J. D., & Jordan, A. K. (2016). Informant discrepancies in externalizing and internalizing symptoms and adaptive skills of high-functioning children with autism spectrum disorder. *School Psychology Quarterly*, *31*, 467–477. https://doi.org/10.1037/spq0000150

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. https://doi.org/10.1037/1082-989X.1.4.390

Narad, M. E., Garner, A. A., Peugh, J. L., Tamm, L., Antonini, T. N., Kingery, K. M., … Epstein, J. N. (2015). Parent-teacher agreement on ADHD symptoms across development. *Psychological Assessment*, *27*, 239–248. https://doi.org/10.1037/a0037864

Renk, K., & Phares, V. (2004). Cross-informant ratings of social competence in children and adolescents. *Clinical Psychology Review*, *24*, 239–254. https://doi.org/10.1016/j.cpr.2004.01.004

Rescorla, L. A., Bochicchio, L., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., … Verhulst, F. C. (2014). Parent–teacher agreement on children's problems in 21 societies. *Journal of Clinical Child & Adolescent Psychology*, *43*, 627–642. https://doi.org/10.1080/15374416.2014.900719

Rogge, J., Koglin, U., & Petermann, F. (2018). Do they rate in the same way?: Testing of measurement invariance across parent and teacher SDQ ratings. *European Journal of Psychological Assessment*, *34*, 69–78. https://doi.org/10.1027/1015-5759/a000445

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Schönmoser, C., Schmitt, M., Lorenz, C., & Relikowski, I. (2018). Prosoziales Verhalten von Kindergartenkindern—Ein Vergleich der Eltern- und Erzieherperspektive. *Zeitschrift für Erziehungswissenschaft*, *21*, 317–337. https://doi.org/10.1007/s11618-017-0783-x

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*, 301–317. https://doi.org/10.1191/096228098672090967

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Eesearch*, *75*, 417–453. https://doi.org/10.3102/00346543075003417

Stone, L. L., Otten, R., Engels, R. C. M. E., Vermulst, A. A., & Janssens, J. M. A. M. (2010). Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-year-olds: A review. *Clinical Child and Family Psychology Review*, *13*, 254–274. https://doi.org/10.1007/s10567-010-0071-2

Winsler, A., & Wallace, G. L. (2002). Behavior problems and social skills in preschool children: Parent-teacher agreement and relations with classroom observations. *Early Education & Development*, *13*, 41–58. https://doi.org/10.1207/s15566935eed1301_3

World Health Organization. (2016). *International statistical classification of diseases and related health problems* (5th ed.). Geneva, Switzerland: Word Health Organization.

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology*, *68*, 1038–1050. https://doi.org/10.1037//0022-006X.68.6.1038

Zahner, G. E. P., & Daskalakis, C. (1998). Modelling sources of informant variance in parent and teacher ratings of child psychopathology. *International Journal of Methods in Psychiatric Research*, *7*, 3–16. https://doi.org/10.1002/mpr.30

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Bergold S, Christiansen H, & Steinmayr R. Interrater agreement and discrepancy when assessing problem behaviors, social-emotional skills, and developmental status of kindergarten children. *J. Clin. Psychol.* 2019;75:2210–2232. https://doi.org/10.1002/jclp.22840