Catching sight of children with internalizing symptoms in upper elementary classrooms

Zee, M.; Moritz Rudasill, K.

DOI
10.1016/j.jsp.2021.05.002

Publication date
2021

Document Version
Final published version

Published in
Journal of School Psychology

Journal of School Psychology 87 (2021) 1–17

Contents lists available at ScienceDirect

Journal of School Psychology

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

Catching sight of children with internalizing symptoms in upper elementary classrooms

Marjolein Zee [a,*], Kathleen Moritz Rudasill [b]

[a] Research Institute of Child Development and Education, University of Amsterdam, the Netherlands
[b] Virginia Commonwealth University, USA

ARTICLE INFO

Action Editor: Nate von der Embse

Keywords:
Internalizing behavior
Teacher self-efficacy
Student–teacher relationships
Background characteristics
Measurement bias

ABSTRACT

Teachers play a crucial role in the assessment of children's internalizing symptoms but may not always succeed in accurately identifying such symptoms in class. Using a multilevel structural equation modeling (MSEM) approach, this study aimed to explore teacher and child characteristics that may explain measurement bias in teachers' ratings of internalizing symptoms at the between- and within-teacher level. Upper elementary school teachers (N = 92, 74.9% female) filled out the Strengths and Difficulties Questionnaire, Student-Teacher Relationship Scale, and Student-Specific Teacher Self-Efficacy Scale for randomly selected children (N = 690, 50.5% girls, Grades 3–6) from their classrooms. Participating teachers and children also responded to several background questions. Multilevel SEMs suggested that teachers' self-efficacy beliefs toward, relationship experiences with, and externalizing symptom ratings of individual children affected their ratings of these children's internalizing symptoms at the within-teacher level. Specifically, given equal levels of internalizing behavior, teachers were likely to systematically under-identify symptoms of anxiety and over-identify bullying for children with more externalizing behavior and conflictual relationships, or in circumstances where teachers had lower self-efficacy. Children with high levels of closeness received systematically higher ratings on somatic complaints and lower ratings on solitary behavior and peer problems. At the between-teacher level, less experienced teachers were more likely to over-identify symptoms of worries than were more experienced teachers, given equal levels of internalizing symptoms. As such, these findings extend the limited body of evidence on children's internalizing symptoms in upper elementary school.

1. Introduction

Reports of children with internalizing problems, or inner-directed and overcontrolled behaviors including symptoms related to anxiety, somatic complaints, or withdrawal, are becoming increasingly common in upper elementary school (Reijneveld et al., 2006). During this period where puberty, social (media) pressures, and emotional changes occur simultaneously (Goldstein et al., 2015), prevalence rates of internalizing symptoms are likely to increase from 10% to almost 30% (Tandon et al., 2009). These problems frequently continue into adulthood (Costello et al., 2011) and may eventually lead to serious social, psychological, and academic difficulties in later life (Fergusson et al., 2006; Valdez et al., 2011). To counter such difficulties, accurate and early identification of

* Corresponding author at: Research Institute of Child Development and Education, University of Amsterdam, P.O. Box 15776, NL-1001, NG, Amsterdam, the Netherlands.
E-mail addresses: m.zee@uva.nl (M. Zee), kmrudasill@vcu.edu (K.M. Rudasill).

https://doi.org/10.1016/j.jsp.2021.05.002
Received 1 May 2020; Received in revised form 26 February 2021; Accepted 22 May 2021
Available online 17 June 2021

children's internalizing symptoms in upper elementary school is needed.

Teachers frequently are a primary information source for the identification and assessment of internalizing symptoms. Due to their extensive contact with children in class, teachers have the unique opportunity to observe the behaviors and subtle cues of children with internalizing symptoms across the school day and for longer periods of time (Keiley et al., 2003). Moreover, teachers who have extensive experience with a large variety of child behaviors, needs, and (dis)abilities may have an implicit normative database against which they can interpret and assess the behavior of children with internalizing symptoms (Abikoff et al., 1993).

Despite their value and utility, one persistent limitation in teacher ratings of internalizing symptoms is that there are no gold standards against which the accuracy (*validity*) and precision (*reliability*) of such ratings can be judged (Kraemer et al., 2003). Several scholars have indicated, for instance, that between 13% and 50% of the variance in teachers' ratings of child behavior tend to be attributable to discrepancies among teachers (Johnson et al., 2016; Peters et al., 2014; Splett et al., 2018, 2020). Little is known about the potential sources of such discrepancies, although it is likely that they reflect as much about teachers themselves as about the child to be rated and teachers' experiences with this child in the classroom (De Los Reyes & Kazdin, 2005; Konold & Pianta, 2007). As teachers' ratings may ultimately affect their decisions concerning educational placement (e.g., Proctor & Prevatt, 2003), it is important to advance insight into the processes that influence teachers' ratings of internalizing child symptoms.

To address this need, we utilized a multilevel structural equation modeling (MSEM) approach to exploring and explaining measurement bias in teachers' ratings of internalizing symptoms at the between- and within-teacher level. This method offers various advantages over common multilevel regression models in which broadband ratings are regressed on various teacher and child characteristics to explain between-teacher level variance in internalizing concerns (e.g., Mashburn et al., 2006; Splett et al., 2018). First, MSEM allows for an exploration of systematic differences between teacher ratings of internalizing behavior that are not attributable to *true* differences in the symptoms that such ratings are intended to measure (Jak et al., 2014). Thus, rather than focusing on broadband factors of internalizing behavior, we can investigate the degree to which the consistency of responses on specific items is equal within and across teachers. This is a vital step in determining which of the specific symptoms are the most valid and precise indicators of children's internalizing behavior, and which of those are influenced by other sources (Bauer et al., 2013). Second, MSEM allows for the inclusion of violators at the within-teacher (i.e., predictors of variance within teachers) and between-teacher level (i.e., predictors of variance across teachers) that explain potential non-invariance in teacher ratings of internalizing symptoms (Jak et al., 2014). Thus, this study may represent a small step toward understanding the sources that may influence the accuracy and precision of teachers' ratings of internalizing symptoms in upper elementary school.

1.1. Theoretical background

Thus far, the majority of research on rater bias has, due to a general lack of coherent frameworks, mainly relied on prior empirical findings to explain potential discrepancies in teachers' ratings of children's internalizing symptoms. However, De Los Reyes and Kazdin (2005) proposed the Attribution Bias Context (ABC) Model for understanding informant discrepancies in the assessment of child psychopathology that may also allow for a conceptualization of why there are idiosyncrasies in teacher ratings of internalizing symptoms. The ABC Model proposes that differences in raters' beliefs and attributions about the causes of problematic child behavior, and in their perspectives of such behaviors, may result in discrepancies that cannot be attributed to true differences in problem behavior (De Los Reyes & Kazdin, 2005).

Based on the basic tenets of the ABC Model, there are two possible routes to explain why there may be differences within and across teachers in the accuracy of their ratings of internalizing symptoms. First, it is likely that teachers start the process of rating specific children in their classrooms with different attributions about whether or which of the child's behavior might be problematic. Such attributions may, in part, be dependent upon various teacher characteristics. For instance, increased exposure to internalizing symptoms due to growth in teaching experience may affect teachers' attributions such that they may perceive such symptoms as more troublesome than less experienced teachers (Kokkinos et al., 2004).

Second, as a consequence of these attributions, discrepancies are likely to arise in the way information about a specific child is stored and recalled when teachers rate the severity of this child's internalizing symptoms (De Los Reyes & Kazdin, 2005; Splett et al., 2020). Conceivably, this process is a function of the experiences that teachers might have with a particular child in class, as well as characteristics and perceived behaviors of this child (Kraemer et al., 2003). Based on their prior beliefs, teachers may, for example, recall particular events from memory (e.g., being unable to "get through" to the child) that support their views of a child's internalizing symptoms and may disregard events that fail to conform to their beliefs (e.g., a long, positive conversation with the child). Following these two routes, we consider both teacher factors (i.e., gender, teaching experience, self-efficacy, affective relationship experiences) and child factors (i.e., gender, ethnicity, externalizing behavior) that may explain measurement bias in internalizing symptoms ratings at the between- and within-teacher level.

1.2. Between-teacher predictors of rater discrepancies in children's internalizing symptoms

Following the ABC Model's first route to explaining rater bias, several teacher characteristics can be proposed that cause differences among teachers' ratings of children's internalizing symptoms. Of these characteristics, the greatest amount of attention in the literature has probably been given to teachers' years of teaching experience. A handful of prior studies has suggested that teachers with little teaching experience first have to be alerted to the subtle signs of internalizing symptoms before they are actually able to perceive such symptoms as significant (Borg, 1998). This is perhaps not surprising: Children with internalizing symptoms, by virtue of their passively compliant attitude, do not often disturb teachers' lessons, challenge their authority, or evoke frustration in them (Rubin &

Coplan, 2004). Accordingly, children with internalizing symptoms are believed to be less likely to attract the attention of novice teachers than are children with externalizing behavior, thereby increasing the likelihood of these teachers' underreporting of internalizing symptoms (e.g., Tandon et al., 2009).

Empirical findings from several correlational studies largely substantiate the assumption that relatively inexperienced teachers tend to appraise disruptive behaviors as more serious and troublesome than internalizing symptoms and feel more helpless and anxious when confronted with disruptive behaviors (Kokkinos et al., 2004, 2005; Shen et al., 2009). Yet, evidence from more rigorous multilevel studies on differences across teachers in their rating of children's social-emotional behavior seems less conclusive. In a diverse sample of 711 children from 210 pre-K classes, Mashburn et al. (2006) evaluated several teacher factors associated with their ratings of children's problem behaviors and identified teaching experience as a significant positive predictor. However, both Peters et al. (2014) and Splett et al. (2018) could not establish such associations.

Next to teaching experience, teacher gender has been pinpointed as a common characteristic that might influence teachers' judgments about children's behavior. Findings from several studies support the premise that female teachers are more likely to accurately identify children who had recently been labelled as seriously emotionally disturbed (Ritter, 1989) and to be more sensitive to these children's needs than male colleagues (Hopf & Hatzichristou, 1999). In contrast, Splett et al. (2018) noted that male teachers are likely to provide higher ratings of children's behavioral and emotional risks than female teachers, but rate their male students lower than expected. Results from both Caldarella et al. (2009) and Kokkinos et al. (2005) even suggested there may be no effects of teachers' gender on their ratings of the seriousness of internalizing symptoms in elementary school. Thus, the extent to which teaching experience and teacher gender may actually influence (the accuracy of) teachers' internalizing symptoms ratings has yet to be determined.

1.3. Within-teacher predictors of rater discrepancies in children's internalizing symptoms

The ABC Model also proposes that discrepancies among teachers' ratings of internalizing child symptoms stem from the way that information about a specific child's behavior is stored and subsequently accessed from memory (De Los Reyes & Kazdin, 2005). Based on Pianta's conceptual model of student–teacher relationships, such information is likely to be stored in mental representational models, which reflect the set of feelings and beliefs about teachers' relationships with individual children and guide their perspectives about these children's behaviors (Pianta et al., 2003). Among the most relevant beliefs and experiences investigated so far are teachers' self-efficacy in relation to and affective relationships with individual children.

Teachers' self-efficacy, or beliefs in their capabilities to organize and perform the actions required to produce given attainments, has been increasingly considered one of the central determinants of teachers' thought processes, affective states, and actions (Bandura, 1997). Results from Zee and Koomen's (2016) review suggested that highly self-efficacious teachers are likely to perceive difficult children as less challenging, take more adequate approaches to improving these children's behaviors, and to be well-attuned to students' signals, needs, and expectations. As such, it is possible that teachers' self-efficacy may affect how they assess the internalizing symptoms of individual children in class as well.

Thus far, a handful of multilevel studies has included investigations of teachers' self-efficacy in relation to students with internalizing symptoms. With only one exception (i.e., Peters et al., 2014), these studies all suggest that teachers' self-efficacy beliefs shape their perspectives of child behavior (Mashburn et al., 2006; McLean et al., 2019). In this literature, however, teachers' self-efficacy beliefs have been entered as teacher-level predictors to account for variance in teachers' ratings of children's internalizing symptoms at the between-teacher level only. Such an approach leaves open the possibility of exploring how teachers' self-efficacy toward *particular children* leads to idiosyncrasies in their own ratings of these children's internalizing symptoms. For the study of teacher-perceived internalizing symptoms, such a student-specific approach is important as teachers are likely to develop differentiated sets of beliefs about their abilities to deal with individual children, depending on these students' behaviors in class (e.g., Zee, Koomen, et al., 2016). As far as we know, there has been only one study of teachers' self-efficacy in relation to individual children with a variety of social-emotional behaviors (i.e., Zee, de Jong, & Koomen, 2016). The study's results, involving 69 Dutch elementary school teachers and 526 upper elementary students, indicated that teachers felt less efficacious in providing adequate instruction and emotional support to individual children with internalizing symptoms. Hence, more positive, student-specific self-efficacy beliefs may help teachers to be more sensitive to the cues of children with internalizing symptoms and to identify these symptoms within an individual child.

Following the ABC Model, the affective quality of relationships between teachers and children can also be considered a within-teacher level factor influencing the way teachers store and retrieve information about specific children when rating children's internalizing symptoms. This quality is commonly characterized by the degree of closeness (i.e., warmth, trust, and open communication) and conflict (i.e., negativity, tension, and hostility) in teacher–child relationships (e.g., Hamre & Pianta, 2001). To date, a relatively large body of empirical findings on links between internalizing behavior and relationship quality has been assembled, yet the results of these studies are somewhat mixed. For instance, internalizing symptoms have repeatedly been associated with elementary teachers' reports of closeness with children across time, both positively (Roorda et al., 2014) and negatively (Rudasill, 2011; Valiente et al., 2012). Yet in other longitudinal research, evidence for the role of internalizing student behavior for teacher-perceived closeness could not be established (Jerome et al., 2009; Mejia & Hoglund, 2016; Zee & Koomen, 2017). Results regarding conflict seem to be more consistent. Findings from Murray and Murray (2004) and Jerome et al. (2009) suggested that teachers are likely to report higher levels of conflict with students who display internalizing symptoms. Longitudinal research conducted among a sample of 175 Kindergartners (Roorda et al., 2014) substantiates these findings, indicating that internalizing behaviors are associated with higher levels of teacher-perceived conflict over time. Based on these results, it could be that poor-quality relationships with children who have

internalizing behaviors may place teachers at a disadvantage when it comes to recognizing and evaluating their symptoms. However, the specific roles of closeness and conflict in the identification and assessment of internalizing symptoms has not been explored.

1.4. Child predictors of rater discrepancies in children's internalizing symptoms

Another issue that complicates the accuracy of teacher ratings is that teachers may be influenced by specific characteristics of children with internalizing symptoms, rather than having unbiased or consistent perceptions of all children with these symptoms (e.g., De Los Reyes & Kazdin, 2005; Konold & Pianta, 2007). Hence, not surprisingly, the ABC Model has also drawn from empirical studies suggesting that teachers make judgments about and form expectations for a child with internalizing behavior based on information about the child's background, resulting in within-teacher differences in their ratings of children's behavior.

Several studies have explored gender biases in teachers' ratings of internalizing symptoms. However, evidence in this area is largely inconclusive, with some studies indicating higher ratings of behavioral and emotional risk for boys than for girls (Kokkinos et al., 2004; Mashburn et al., 2006; Peters et al., 2014; Splett et al., 2018, 2020) and others showing no gender differences in internalizing symptom ratings (McLean et al., 2019; Shen et al., 2009). Furthermore, multi-sample confirmatory factor results from a large-scale Dutch study using the Strengths and Difficulties Questionnaire (Goodman, 2001) provided evidence for measurement invariance of this instrument, indicating that elementary teachers are likely to perceive children's emotional symptoms and peer problems similarly across gender (Zwirs et al., 2011).

Teacher biases in their ratings of internalizing symptoms have also been evident with students from minoritized backgrounds. Again, however, evidence in this area has been scarce and at times contradictory. A small number of empirical studies have suggested that patterns of discrepant teacher ratings of internalizing symptoms may be affected by ethnicity-related beliefs (e.g., Dulin, 2001; Loo & Rapport, 1998). For instance, Chang and Sue (2003) noted that teachers were more likely to rate overcontrolled behaviors such as shyness and worries as more typical for Asian children than for White or African-American children. Furthermore, in studies predicting between-teacher level variance of teacher-rated child behaviors, Peters et al. (2014) and Splett et al. (2020) found evidence for the idea that Hispanic children tended to be rated as having less internalizing concerns than their White peers. Results from other studies, however, seem to contradict these findings, documenting no mean differences in teacher ratings of children's internalizing symptoms (Mashburn et al., 2006; McLean et al., 2019; Sonuga-Barke et al., 1993) or measurement nonequivalence across ethnicity (e.g., Crijnen et al., 2000; Zwirs et al., 2011).

Last, it is important to note that ratings of internalizing symptoms are frequently comorbid with externalizing symptoms (e.g., aggression, hyperactivity), with correlations ranging between .45 and .63 (Achenbach, 1991). This may have important implications for how teachers perceive children's internalizing symptoms, as such perceptions are most likely constructed from information conveyed by experienced events and specific child behaviors in the classroom. Indeed, a plethora of studies have documented potential biases in teacher ratings of externalizing behaviors. Prior research across various countries has consistently shown that teachers are likely to (a) be less tolerant of children with externalizing behavior as compared to other types of problems (Merrell, 2008; Shen et al., 2009), (b) perceive these behaviors as more serious (e.g., Kokkinos et al., 2004, 2005), and (c) refer children with externalizing problems more often for special services and interventions than children with internalizing symptoms (Caldarella et al., 2008). Hence, the co-occurrence between internalizing and externalizing symptoms within a child may possibly negatively affect teachers' ratings of the child's internalizing symptoms.

1.5. Present study

Drawing upon the basic tenets of the ABC Model, we aimed to explore and explain measurement bias in teachers' ratings of internalizing symptoms at the between- and within-teacher level. To this end, we used a multilevel SEM approach, which is well suited to detect violations of measurement invariance across teachers' ratings of internalizing symptoms in multilevel data (i.e., cluster bias; Jak et al., 2014). Generally, the occurrence of cluster bias indicates that teachers might perceive *specific items* regarding children's internalizing symptoms differently, despite providing similar ratings. Such an item-level approach is particularly relevant for practitioners, as it may provide clues about which of the internalizing symptoms are most difficult to detect and may therefore be the least valid indicators of teacher-rated internalizing behavior. To explore the processes that may influence teacher-rated internalizing symptoms, we examined two questions:

1. What proportion of the variance in teachers' ratings of internalizing symptoms can be attributed to differences between and within teachers?
2. What teacher characteristics (i.e., teacher gender, teaching experience, student-specific self-efficacy, teacher–child closeness and conflict) and child characteristics (i.e., gender, ethnicity, externalizing behavior) explain measurement bias in teachers' ratings of internalizing symptoms at the between- and within-teacher level?

2. Method

2.1. Participants

Data for the present study were collected as a part of a larger research project on teachers' efficacy and skills managing diversity in their classrooms. In this project, behavioral rating scales and teacher surveys on their beliefs and feelings toward particular children in

class were collected, as well as child reports of their classroom experiences and assessments of their achievement. The sample for this study was created using data from two part-projects that were collected in two consecutive school years (Winter/Spring 2013–2014 and 2014–2015) and in which whole classrooms participated. All participating children in this study were enrolled in general education-based upper elementary classrooms; institutional Ethics Review Board approvals were obtained prior to initiating study procedures.

In both part-projects, we used a three-stage stratified sampling procedure that first selected participating schools from a random pool of approximately 600 elementary schools across the Netherlands and through personal contacts and social networking sites (i.e., Facebook, LinkedIn). Of all schools that were initially contacted, 40 schools were willing to take part in this study. These schools were located in both urban and rural areas across the Netherlands and school sizes ranged from 67 to 374 students. Non-participation was mainly due to schools' already full agendas or participation in other research projects.

Second, after schools agreed to participate, approximately 250 teachers, all of whom taught in the upper elementary grades (Grades 3–6), received information letters about the nature and purposes of the study. On average, two teachers per participating school (range = 1–8) made an informed and voluntary decision to participate in this study. This resulted in a sample of 92 teachers (response rate = 36.8%), of whom 13 (14.1%) taught in Grade 3, 26 (28.3%) in Grade 4, 30 (32.6%) in Grade 5, and 23 (25.0%) in Grade 6. These teachers (74.9% female) ranged from 20 to 63 years of age ($M = 40.03$, $SD = 12.54$) and their years of professional teaching experience ranged from six months to 44 years ($M = 15.83$, $SD = 11.94$). These demographic characteristics are comparable to those of the larger population of Dutch teachers, who have a mean age of 43.25 years (range = 19–67 years) and are typically female (84%; DUO, 2014).

Last, a subsample of children was selected as tertiary sampling unit. Eight children per classroom were randomly sampled by the first author from the total pool of participating 2102 children (23 children per classroom on average). Due to absence or illness, however, in 18 classrooms less than eight children took part ($M = 7.51$, $SD = 2.82$, range = 4–8 children). This resulted in a final sample of 690 children who attended Grades 3 ($n = 66$), 4 ($n = 206$), 5 ($n = 202$), and 6 ($n = 216$), and were primarily of Dutch nationality (72.0%). Other ethnicities were Turkish (7.4%), Moroccan (6.1%), Surinamese or Antillean (1.4%), or other (13.1%). The sampled children ranged from seven years and seven months to 13 years and two months of age ($M = 10.66$, $SD = 1.13$); the gender composition was evenly distributed with 341 boys (49.4%) and 347 girls (50.3%). Two children did not provide information about their gender. Teacher reports indicated that participating children's parents included a diverse range of educational attainment: 12.8% of the parents had only completed elementary education, 40.3% of the parents had finished high school and/or vocational education, and 35.2% of the parents had finished higher education. For 10.0% of the children, information about their parents' educational attainment was not available. Last, the degree of internalizing child behavior, as identified by participating teachers, varied considerably: On a 5-point scale, 14.8% of children were rated as having no internalizing behavior difficulties, 75.2% of participating children were rated as having only mild levels of internalizing behavior difficulties (mean scores ≤ 3.0), and the remaining sample (10.0%) was rated as having medium to severe internalizing difficulties. This is roughly similar to the percentages of children with elevated scores on internalizing behavior in a Dutch normative sample (Reijneveld et al., 2006).

2.2. Procedure

During recruitment, either school principals or participating teachers distributed information letters and consent forms to parents of all children from teachers' classrooms. On average, parental consent rates per classroom ranged between 44% and 100%. From all parental consents received (total parental consent rate = 95%), we randomly selected eight students from each teacher's classroom; this was to reduce teacher burden and distribute study resources across a broader number of classrooms. For these students, teachers filled out questionnaires regarding their relationships with (i.e., Student-Teacher Relationship Scale) and self-efficacy beliefs toward (i.e., Student-Specific Teacher Sense of Efficacy Scale) these students, as well as these children's externalizing and internalizing behaviors (i.e., Strengths and Difficulties Questionnaire). Additionally, teachers responded to some general questions regarding their own background characteristics. The total survey took approximately 1 hr for teachers to complete. The overall teacher participation rate was 93%.

During a planned school visit in the Winter/Spring of the school years 2013–2014 and 2014–2015, participating children were asked to respond to several questions about their age, gender, and ethnicity, as well as several other questionnaires about their classroom experiences that were part of the larger research project, but beyond the scope of this study. The total survey took approximately 30–45 min to complete and teachers were asked to leave the classroom to facilitate children's free and honest answering. A research facilitator was present in the classroom to explain the procedure, answer students' questions, and discourage response acquiescence and inconsiderate answering. Completed student reports were available for 94% of the sample. Nonparticipation was mainly due to absence or illness at the time of data collection. After participation, schools and teachers were provided with school reports containing a conceptual overview of the study's results.

2.3. Instruments

2.3.1. Children's internalizing and externalizing behaviors

Teachers were asked to complete the authorized Dutch version of the Strengths and Difficulties Questionnaire (SDQ; Van Widenfelt et al., 2003) to evaluate children's social-emotional behaviors. The SDQ is a brief 25-item behavioral screening questionnaire that measures students' adjustment and psychopathology in the classroom. Initially, this scale was recommended for clinical use, but is currently used widely in schools as well (Stone et al., 2010). The original scale consists of positive and negative student attributes that together represent five factors reflecting children's strengths (i.e., Prosocial Behavior) and difficulties (i.e., Emotional Symptoms,

Conduct Problems, Hyperactivity-Inattention, Peer Problems). In this study, however, we used the more general Internalizing and Externalizing subscales proposed by Goodman et al. (2010), which have been suggested to be more appropriate when evaluating social-emotional child behavior in low-risk samples such as the one for this study.

The Internalizing Behavior subscale (8 items) includes all items from the Emotional Symptoms factor (e.g., "Often complains of headaches, stomach-aches or sickness", "Many fears, easily scared") as well as three items from the Peer Problems factor (i.e., "Rather solitary, tends to play alone", "Gets on better with adults than with other children", "Picked on or bullied by other children"). The Externalizing Behavior dimension (10 items) combines the subscales of Hyperactivity-Inattention and Conduct Problems, with items such as "Restless, hyperactive, cannot sit still for long" and "Often has temper tantrums or hot tempers." All items were rated on a 5-point Likert scale, ranging from 1 (*not true*) to 5 (*certainly true*).

The psychometric properties of the three-factor SDQ model, including Externalizing Behavior, Internalizing Behavior, and Pro-social Behavior, have been demonstrated to be especially suited for use in non-risk samples (Dickey & Blumberg, 2004; Goodman et al., 2010; Van Leeuwen et al., 2006). Specifically, the internal consistency, validity, and mean inter-informant product-moment correlations were considered acceptable (Muris et al., 2004; Van Widenfelt et al., 2003; Zee, de Jong, & Koomen, 2016). Furthermore, several scholars (e.g., Goodman & Scott, 1999; Muris et al., 2004) have shown that the SDQ, despite its brevity, is at least as good in distinguishing among community samples and at-risk samples and is equivalent to other, much longer scales such as the Child Behavior Checklist, Child Depression Inventory, and Revised Children's Manifest Anxiety Scale, with correlations between .43 and .74. In the present study, Cronbach's alphas were .81 for Internalizing Behavior and .87 for Externalizing Behavior.

2.3.2. Teachers' perceptions of the student–teacher relationship quality

Teachers' views of the quality of their relationships with individual children were measured using a short form of the authorized translated Dutch version of the Student-Teacher Relationship Scale (STRS; Koomen et al., 2012). This scale is intended to measure the degree of teacher-perceived Closeness, Conflict, and Dependency in the student–teacher relationship by using a 5-point Likert scale (1 = *definitely does not apply*; 5 = *definitely applies*). In this study, we only made use of the Closeness and Conflict dimensions of the STRS. The Closeness dimension (5 items) considers the extent to which teachers perceive the student–teacher relationship to be warm, open, and secure, with items such as "I share an affectionate and warm relationship with this child". The Conflict dimension (5 items) focuses on negative aspects of the student–teacher relationship, including tension, anger, and mistrust in the relationship. An example item is "This child and I always seem to be struggling". In previous studies, the psychometric properties of this short form of Dutch STRS have been demonstrated to be adequate, with internal consistencies ranging from .86 to .93, and factor loadings between .62 and .87 for Closeness and between .70 and .93 for Conflict (Zee et al., 2013; Zee & Koomen, 2017). In the present study, alpha coefficients were .85 for Closeness and .89 for Conflict, respectively.

2.3.3. Teachers' self-efficacy in relation to individual children

Teachers rated their self-efficacy in relation to individual students using a short, 16-item version of the Student-Specific Teacher Self-Efficacy Scale (Student-Specific TSES; Zee, Koomen, et al., 2016). Unlike the original TSES, the student-specific version reflects teachers' self-referent beliefs about their ability to deal with a specific child, rather than the classroom as a whole. Moderate levels of correspondence between classroom-level and student-specific TSE suggest that both instruments tap into different aspects of the self-efficacy belief system (Zee et al., 2018; Zee, Koomen, et al., 2016).

Generally, the short Student-Specific TSES consists of four teaching domains of four items each: Instructional Strategies, Student Engagement, Behavior Management, and Emotional Support. Given the high intercorrelations between the four factors in prior studies (e.g., Zee, Koomen, et al., 2016), we combined these dimensions into one total score. Example items are "How well can you respond to difficult questions from this student?" (Instructional Strategies), "How well can you prevent this student from negatively affecting the classroom atmosphere?" (Behavior Management), "To what extent can you motivate this student for his/her schoolwork?" (Student Engagement), and "How well can you establish a safe and secure environment for this student?" (Emotional Support). The 16 items were rated by teachers on a seven-point Likert scale, ranging from 1 (*nothing*) to 7 (*a great deal*). Support for the reliability and construct validity of the Student-Specific TSES has been provided by Zee and colleagues (Zee et al., 2018; Zee, Koomen, et al., 2016), with reliabilities ranging between .85 and .94 and factor loadings ranging between .60 and .91. Cronbach's alpha in the present study was excellent, $\alpha = .95$.

2.3.4. Demographic background variables

Teachers provided information about their gender and years of teaching experience. Gender was dummy coded, such that male teachers were assigned a value of 0 and female teachers a value of 1. Children themselves were asked to report on their ethnicity and gender. Children's gender was dummy coded (0 = boys; 1 = girls). Ethnicity was based on children's reports of their mother's country of birth. Children could choose among different options, including the Netherlands, Turkey, Morocco, and Surinam. Given the small proportions of ethnic groups other than Dutch in the sample (7.4% Turkish, 6.1% Moroccan, 1.4% Surinamese/Antillean, 13.1% other), ethnic minority children were treated as one group and contrasted with the Dutch majority group.[1] This variable was also dummy coded (0 = ethnic minority children; 1 = ethnically Dutch children).

[1] Preliminary analyses of variance were performed to determine whether these minority groups differed with regard to internalizing symptoms. The results showed no significant differences ($p > .05$). Therefore, ethnic minorities were treated as one group and contrasted with the Dutch majority group in the main analyses.

2.4. Data analysis

We used multilevel confirmatory factor analysis (MCFA) to explore various teacher and child factors that may influence how teachers identify and assess internalizing symptoms among children in class. With this analytic technique, model fit and parameter estimate biases can be avoided by decomposing the total sample covariance matrix into a pooled within-group (Σ_{WITHIN}) and a between-group ($\Sigma_{BETWEEN}$) covariance matrix (Muthén, 1994). In addition, MCFA is well suited to detect violations of measurement invariance across clusters in multilevel data (Jak et al., 2014). This technique is particularly useful when collecting the same construct from qualitatively different groups or individuals operating in distinct contexts as it aims to take differences in response processes into account (Muthén & Asparouhov, 2013). Generally, cluster bias would suggest that the internalizing behaviors of children who actually display similar levels of these symptoms in class may be rated differently depending on the teacher by whom they are taught. Thus, in this study, the presence of cluster bias would indicate that the internalizing symptom items do not measure the same construct across teachers and that part of the variance in teachers' judgments of internalizing symptoms may be attributed to teacher and/or child characteristics at the between-teacher level (i.e., *Level 2 violators*) or the within-teacher level (i.e., *Level 1 violators*; Jak et al., 2014). Because internalizing behavior comprises a variety of symptoms (e.g., depressive feelings, anxiety, solitary behavior), some of which are easier to detect than others, we investigated bias with respect to each indicator separately.

2.4.1. Modeling procedure

We followed five analytical steps based on the guidelines proposed by Jak et al. (2014). In Step 1 we calculated the intraclass correlation coefficients (ICC) for each of the model's indicators and tested whether the between-teacher level variance and covariance deviated significantly from zero. To this end, we fitted a Null Model ($\Sigma_{BETWEEN} = 0$, Σ_{WITHIN} = free) and an Independence Model ($\Sigma_{BETWEEN}$ = diagonal, Σ_{WITHIN} = free) to the data (Jak et al., 2014; Muthén, 1994). Generally, poor fit of these models are indicative of meaningful between-teacher level variance and covariance (Hox, 2002).

In Step 2 and Step 3 we evaluated the factor structure of the Internalizing Symptoms scale at the within-teacher level using the sample pooled-within covariance matrix (Hox, 2002; Muthén, 1994). This means that we simultaneously specified a measurement model at the within-teacher level and a saturated model at the between-teacher level; we also investigated potential sources of model misspecification. We then used this measurement model to investigate bias regarding Level 1 violators (i.e., teachers' relationship with and self-efficacy beliefs toward individual children as well as children's gender, ethnicity, and externalizing behavior). These violators were separately added as covariates to the model (i.e., they reflect the association with each of the indicators of the measurement model and are represented by a β-coefficient). Only Level 1 violators that were significantly related to the model's indicators were retained in the final model (Jak et al., 2014).

In Step 4 we used the final measurement model established in Step 2 to investigate cluster bias. We started with a fully constrained model in which all factor loadings were constrained to be equal across the between- and within-teacher level and residual variances at the between-teacher level were fixed at zero. To test whether strong factorial invariance held across clusters, we sequentially allowed the between-teacher level residual variances to be freely estimated. Generally, residual variances greater than zero are indicative of cluster bias in their corresponding indicators (Jak et al., 2014). Subsequently, we evaluated whether factor loadings could be considered equal across clusters. Unequal factor loadings indicate that the internalizing symptoms factor at the between-teacher level cannot merely be assumed to be the within-teacher level factor's aggregate. Finally, in Step 5 we used the final model established in Step 4, with all violators at the between- and within-teacher level as covariates (i.e., they reflect the association with each of the indicators of the measurement model and are represented by a β-coefficient). Only statistically significant between-level violators were retained in the model.

2.4.2. Model goodness-of-fit

Multilevel CFAs were fitted in M*plus* 7.11 using robust maximum likelihood estimation (MLR; Muthén & Muthén, 1998–2012). Although this estimation method can also be used with categorical data, we treated all Likert scale items as continuous. Additionally, missing data were treated using full information maximum likelihood estimation (FIML).[2] Overall goodness-of-fit of the models was evaluated by the mean-adjusted χ^2-test, with non-significant chi squares indicating satisfactory fit. Approximate fit was determined using the Root Mean Square Error of Approximation (RMSEA), with values below .05 reflecting close fit and values below .08 signifying a satisfactory fit (Browne & Cudeck, 1993; Hu & Bentler, 1999; Kline, 2011), and the comparative fit index (CFI), with values $\geq$.95 indicating close fit and values $\geq$.90 indicating acceptable fit (Bentler, 1992). The model's modification indices, residual correlations, and Standardized Root Mean Square Residual (SRMR) were used to evaluate component fit. Values $\leq$.08 indicate good model fit (Kline, 2011). To compare alternative models, we employed the (Satorra–Bentler scaled) chi-square difference test (TRd; Satorra & Bentler, 2010), with non-significant chi-squares indicating equivalent fit, and the CFI-difference, with CFI changes $\geq$.02 being indicative of model nonequivalence (Cheung & Rensvold, 2002).

[2] Randomly missing data patterns were evident in teachers' reports of relationship quality (6.1%), children's ethnicity (3.3%), and teaching experience (4.5%). Missingness in internalizing symptom items ranged from 0% to 23.8% ($M = 6.4\%$). *t*-tests indicated that the sample with missing data did not significantly differ from the sample without any missing data in mean scores on teachers' student-specific self-efficacy, relationship experiences, reports of internalizing and externalizing behavior, and background characteristics ($p > .05$).

3. Results

3.1. Descriptive statistics

Means, standard deviations, and zero-order correlations are displayed in Table 1. Both teachers' Gender ($r = -.17$, $p < .001$) and Teaching Experience ($r = .14$, $p < .001$) were significantly associated with their ratings of children's Internalizing Symptoms. Moreover, statistically significant positive associations of children's Externalizing Behavior ($r = .39$, $p < .001$) and teacher-perceived Conflict ($r = .27$, $p < .001$), and negative associations of teachers' Student-Specific Self-Efficacy ($r = -.34$, $p < .001$) and Closeness ($r = -.15$, $p < .001$) were found. The correlations of children's Gender and Ethnicity with Internalizing Symptoms did not deviate significantly from zero.

3.2. Multilevel confirmatory factor analysis of internalizing symptoms

3.2.1. Step 1: evaluating between-teacher level variance and covariance

Table 2 lists the content and intraclass correlations (ICCs) for the Internalizing Symptom items. The ICCs ranged between .06 (Item 6; "Picked on or bullied by other children") and .27 (Item 7; "Gets along better with adults than with other children"), with a mean ICC of .16. Fit indices of the Null Model, χ^2 (36) = 355.62, RMSEA = .113, CFI = .703, $SRMR_{WITHIN}$ = .050, $SRMR_{BETWEEN}$ = .576, and the Independence Model, χ^2 (28) = 141.88, RMSEA = .077, CFI = .894, $SRMR_{WITHIN}$ = .052, $SRMR_{BETWEEN}$ = .576, indicated that there is meaningful between-teacher level variance and covariance.[3] Hence, these clustering effects were substantial enough to use MCFA.

3.2.2. Step 2: evaluating internalizing symptoms at the within-teacher level

Using the sample pooled-within covariance matrix, the overall fit of the one-factor model was not satisfactory, χ^2 (20) = 171.91, $p < .001$, RMSEA = .105, 90% CI [.091–.120], CFI = .866, SRMR = 0.060. The model's modification indices suggested significant model improvement by adding a correlation between the residuals of Item 5 ("Nervous or clingy in new situations, easily loses confidence") and Item 8 ("Many fears, easily scared"), which showed a considerable conceptual overlap. Adding this residual correlation resulted in an adequate model fit, χ^2 (19) = 80.34, $p < .001$, RMSEA = .068, 90% CI [.053–.084], CFI = .946, SRMR = .044.

To justify the appropriateness of the one-factor Internalizing Symptoms model, we additionally examined an alternative two-factor model, including an Emotional Symptoms factor and a Peer Problems factor (see Goodman, 2001). This model fitted the data slightly worse than the one-factor model, χ^2 (19) = 130.85, $p < .001$, RMSEA = .092, 90% CI [.078–.108], CFI = .902, SRMR = .049. For this reason, we retained the one-factor model as our final model.

3.2.3. Step 3: exploring bias regarding within-teacher level violators

In Step 3, to avoid convergence problems and increase statistical power, we fitted restrictive factor models (RFAs) for each of the within-teacher level violators (Gender, Ethnicity, Externalizing Behavior, Self-Efficacy, Closeness, and Conflict) separately. In these models, the Level 2 covariance matrix still remained saturated. Model fit indices of these models are displayed in Table 3. The overall model fit of all baseline RFAs was reasonable, with RMSEA and SRMS values <.08 and CFI values >.90. Regarding children's background characteristics, modification indices suggested further model improvement by adding direct effects of teacher-perceptions of Externalizing Behavior on Item 6 ("Picked on or bullied by other children"; $\beta = .20$, $p < .001$) and Item 8 ("Many fears, easily scared"; $\beta = -.14$, $p < .01$), suggesting that these two items were biased with respect to Externalizing Behavior. Hence, given equal levels of Internalizing Symptoms, teachers are likely to rate items regarding bullying systematically higher and items regarding anxiety systematically lower for children who display higher levels of Externalizing Behavior than for children without such behaviors. Both Gender and Ethnicity were not directly associated with any of the indicators of Internalizing Symptoms.

With respect to teachers' beliefs and affective experiences, the model's modification indices suggested several direct effects (i.e., bias) of these Level-1 violators on the indicators. Similar to Externalizing Behavior, small to moderate associations of Conflict and Student-Specific Self-Efficacy with Item 6 ("Picked on or bullied by other children"; $\beta_{conflict} = .16$, $p < .01$; $\beta_{efficacy} = -.12$, $p < .05$) and Item 8 ("Many fears, easily scared"; $\beta_{conflict} = -.11$, $p < .01$; $\beta_{efficacy} = .14$, $p < .001$) were found. Additionally, teachers' perceptions of Closeness were positively associated with Item 1 regarding psychosomatic complaints ($\beta = .12$, $p < .05$) and negatively associated with Item 2 regarding solitary behavior ($\beta = -.21$, $p < .001$) and Item 7 regarding peer problems ($\beta = -.13$, $p < .05$). Thus, for equal degrees of Internalizing Symptoms, children with conflictual student–teacher relationships or lower self-efficacious teachers obtained systematically higher ratings for items regarding bullying and lower ratings for items regarding anxiety. Children with high levels of Closeness in the relationship received systematically higher ratings on items regarding somatic complaints and lower ratings on items regarding solitary behavior and peer problems than children with less relational Closeness.

Correlations among the within-teacher level covariates and Internalizing Symptoms indicated that children's Externalizing Behavior, but not their Gender and Ethnicity, was directly and positively associated with Internalizing Symptoms ($r = .35$, $p < .001$). Additionally, we found direct negative correlations of teachers' Student-Specific Self-Efficacy ($r = -.38$, $p < .001$) and teacher–child Closeness ($r = -.19$, $p < .001$) with the Internalizing Symptoms factor. In contrast, Conflict was positively associated with this common

[3] To ensure that the reliability of the Internalizing Symptoms scale at the between-teacher level was sufficient, we also calculated the ICC2 for this factor (Morin et al., 2014), which can be calculated by the following formula: $\tau^2_x / \tau^2_x + (\sigma^2_x / n_j)$. In this study the ICC2 value of the Internalizing Symptoms factor was 0.70, which can be considered good (Cicchetti, 1994).

Table 1
Means, standard deviations, and zero-order correlations.

	1. Teacher gender	2. Teaching experience	3. Child ethnicity	4. Child gender	5. Externalizing behavior	6. Teacher self-efficacy	7. Closeness	8. Conflict	9. Internalizing behavior
Inter-individual factors:									
1. Teacher gender	1.00								
2. Teaching experience	−.31**	1.00							
Intra-individual factors:									
3. Child ethnicity	−.06	−.01	1.00						
4. Child gender	.03	−.01	.01	1.00					
5. Child externalizing behavior	−.09*	−.01	.02	−.25**	1.00				
6. Teacher self-efficacy (student-specific)	.00	.12**	.00	.20**	−.67**	1.00			
7. Teacher-perceived closeness	.07	.05	−.11**	.27**	−.29**	.42**	1.00		
8. Teacher-perceived conflict	−.05	−.04	.00	−.17**	.69**	−.60**	−.42**	1.00	
Internalizing symptoms									
9. Child internalizing symptoms	−.17**	.14**	.05	.03	.39**	−.34**	−.15**	.27**	1.00
Descriptive statistics:									
M		15.83			1.91	5.75	3.90	1.55	1.99
SD		11.94			0.79	0.86	0.83	0.86	0.80

Note. Gender: 0 = boy/male, 1 = girl/female. Ethnicity: 0 = minoritized background, 1 = ethnically Dutch background.
* $p < .05$.
** $p < .01$.

Table 2
Items and intra-class correlations of the internalizing symptoms items.

Item	ICC
1. Often complains of headaches, stomach-aches or sickness	.16
2. Rather solitary, prefers to play alone	.11
3. Many worries or often seems worried	.14
4. Often unhappy, depressed or tearful	.20
5. Nervous or clingy in new situations, easily loses confidence	.18
6. Picked on or bullied by other children	.06
7. Gets along better with adults than with other children	.27
8. Many fears, easily scared	.19

factor ($r = .32$, $p < .001$).

3.2.4. Step 4: exploring cluster bias in teachers' ratings of internalizing symptoms

To explore cluster bias, we fitted the final one-factor model at both the between teacher and within-teacher levels. In this model, all factor loadings were constrained to be equal across both levels and there was no residual variance at the between-teacher level (Jak et al., 2014). This model did not fit the data satisfactorily, χ^2 (54) = 227.17, $p < .001$, RMSEA = .068, CFI = .839, $SRMR_{within} = .050$, $SRMR_{between} = .368$. We therefore freed the residual variances with the highest modification indices at the between-teacher level one by one. This resulted in a well-fitting model in which only the residual variances of Item 4 and Item 6 (see Table 2) were constrained to be zero, χ^2 (48) = 95.00, $p < .001$, RMSEA = .038, CFI = .956, $SRMR_{within} = .040$, $SRMR_{between} = .110$. This means that cluster bias was present in all items except for Item 4 and Item 6. Further improvements of fit could be made by freeing the factor loading of Item 3 ("Many worries or often seems worried") across levels, which had a higher factor loading and was therefore more indicative of Internalizing Symptoms at the between-teacher level. The overall fit of this final model was adequate, χ^2 (47) = 87.56, $p < .001$, RMSEA = .035, CFI = .962, $SRMR_{within} = .047$, $SRMR_{between} = .106$.

Table 3
Model fit statistics for restrictive factor models evaluating bias regarding level 1 violators.

	TRd (df)	CFI	RMSEA	$SRMR_{within}$	$SRMR_{between}$	TRd (df)	ΔCFI
Child Gender:							
Baseline RFA Model	77.81 (26)***	.941	.055	.049	.039		
Model with direct effects of violators							
Child Ethnicity:							
Baseline RFA Model	81.79 (26)***	.951	.056	.044	.022		
Model with direct effects of violators							
Child Externalizing Behavior:							
Baseline RFA Model	110.55 (26)***	.932	.069	.050	.027		
Model with direct effects of violators	72.28 (24)***	.961	.054	.040	.023	23.27 (2)***	.029
Teacher Self-Efficacy:							
Baseline RFA Model	87.24 (26)***	.948	.058	.047	.026		
Model with direct effects of violators	65.43 (24)***	.965	.050	.042	.024	13.98 (2)***	.017
Closeness							
Baseline RFA Model	95.26 (26)***	.940	.062	.056	.033		
Model with direct effects of violators	60.11 (23)***	.968	.048	.038	.022	28.02 (3)***	.028
Conflict							
Baseline RFA Model	85.92 (26)***	.948	.058	.049	.025		
Model with direct effects of violators	67.67 (24)***	.962	.051	.042	.024	13.89 (2)**	.014

Note. TRd = Sattora-Bentler Scaled Chi-Square Difference, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Square Residual, RFA = Restrictive Factor Analysis. Violators in the Baseline RFA Model were added as covariates.
*** $p < .001$.
** $p < .01$.

3.2.5. Step 5: exploring cluster bias regarding between-teacher level violators

The presence of cluster bias in all Internalizing Symptom items except for Item 4 (unhappy feelings) and Item 6 (bullied by others) allowed us to explore whether there were inter-individual factors other than Internalizing Symptoms that explained differences in the items of this factor. To this end, we first took the final model from Step 4 and conducted an RFA with teachers' Gender and Teaching Experience as covariates at the between-teacher level. This model fitted the data well, χ^2 (61) = 109.57, $p < .001$, RMSEA = .034, CFI = .957, $SRMR_{within}$ = .047, $SRMR_{between}$ = .107. The model's modification indices suggested bias with respect to Teaching Experience in Item 3 only ("Many worries or often seems worried"; $\beta = -.29$, $p < .001$). Hence, given equal levels of Internalizing Symptoms, less experienced teachers tended to give higher ratings on this item than more experienced teachers. The correlations of Internalizing Symptoms with Teacher Gender ($r = -.32$, $p < .05$) and Teaching Experience ($r = .31$, $p < .01$) were both statistically significant. A graphical representation of this full model is displayed in Fig. 1.

4. Discussion

Upper elementary teachers play a crucial role in the assessment of children's internalizing symptoms but may not always succeed in accurately identifying such symptoms in class (e.g., Konold & Pianta, 2007). Not only do teachers frequently lack the time or experience to catch sight of the subtle cues of children with internalizing symptoms, they may also be affected by their own classroom experiences or their students' characteristics (Rudasill & Kalutskaya, 2014). Consequently, children with internalizing symptoms may go unnoticed by teachers and frequently miss out on the emotional and cognitive supports they need to participate in all aspects of school life (Kokkinos et al., 2004).

Based on the ABC Model, in this study we set out to explore and explain measurement bias in teachers' ratings of internalizing symptoms at the between- and within-teacher level. Our intermediate models suggested that teachers' self-efficacy beliefs toward, relationship experiences with, and externalizing symptom ratings of individual children in their classes affected their assessments of these children's internalizing symptoms at the within-teacher level. Additionally, teachers' years of teaching experience, but not their gender, explained differences in teachers' average ratings of children's internalizing symptoms. Although these findings cannot be fully generalized to scales other than the SDQ, they extend the relatively limited body of evidence on children's internalizing symptoms in upper elementary school.

4.1. Differences in teacher ratings of children's internalizing symptoms

Generally, the results of this study suggested that teachers' SDQ ratings of children's internalizing symptoms in class were likely to vary both within and across teachers. Largely consistent with prior research (e.g., McLean et al., 2019; Splett et al., 2018, 2020), intraclass correlations indicated that the variability at the within-teacher level was substantially larger than the between-teacher level variability, which ranged from 6%–27%. As such, this study is one of the first to reveal that teachers' assessment of internalizing symptoms may be primarily based on the specific characteristics, behaviors, and needs that children bring to class, rather than basic knowledge that teachers may have acquired through teacher training. This premise is largely consistent with research taking a dyadic

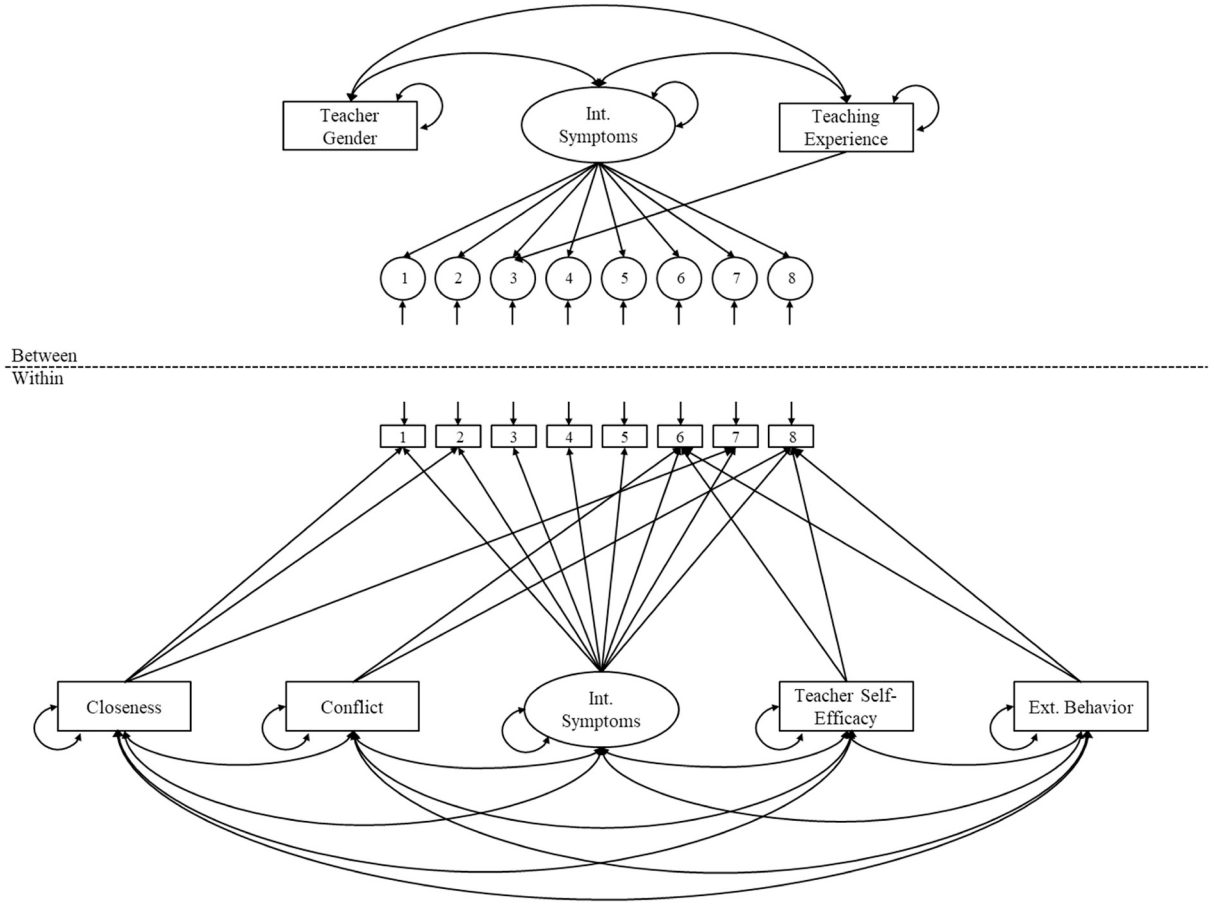

Fig. 1. Final multilevel confirmatory factor model of teachers' ratings of children's internalizing symptoms.
Note. Model fit: χ^2 (125) = 244.75, $p < .001$, RMSEA = 0.037, CFI = 0.953, $SRMR_{within}$ = 0.040, $SRMR_{between}$ = 0.257. To avoid confusion, parameter estimates are not displayed in this model.

approach to the study of internalizing symptoms, suggesting that teachers' views of these children are dependent upon reciprocal exchanges of information about expectations, behaviors, and personal features (Pianta et al., 2003).

Furthermore, to ensure the precision of teachers' ratings of children's internalizing symptoms at the between-teacher level was sufficient, we calculated the reliability of the aggregated SDQ-scale. This ICC2-coefficient (Morin et al., 2014) suggested that the average of teachers' internalizing symptoms ratings for the seven children in their classrooms was reliable (Cicchetti, 1994). Although we cannot provide any answers about the *accuracy* of teacher ratings due to a lack of external criteria against which these ratings were judged (Splett et al., 2018), our findings at least seem to suggest that teachers are able to provide precise judgments for students' internalizing symptoms in their classrooms.

4.2. Cluster bias in teacher ratings of children's internalizing symptoms

To further evaluate potential differences in teacher-perceived internalizing symptoms ratings, we tested for violations of measurement invariance across teachers. This method is different both from multilevel regression as well as other factor analytic methods, such as Bauer et al.'s (2013) tripartite model, which is designed to generate integrated scores from item-level data across multiple informants instead of one type of informant (i.e., teachers). Various researchers (e.g., Jak et al., 2014; Muthén & Asparouhov, 2013) have stressed the importance of using this technique, as it attempts to account for differences in response processes that arise from personal and contextual characteristics, yet still allows for group comparisons on similar latent variables.

Our results suggested the presence of cluster bias in six of eight internalizing symptom items of the SDQ. Only two indicators about the extent to which the child is often unhappy, depressed, or tearful (Item 4) or is picked on or bullied by other children (Item 6) appeared to be invariant across teachers. Two reasons may explain this finding. First, the absence of cluster bias in Item 6, but not Item 4, may be due to the relatively smaller amount of variance in this indicator of internalizing symptoms across teachers. Indeed, the ICC suggests that only 6% of the variance in this item occurred between teachers. Another interpretation is that these two indicators refer

to more easily observed symptoms of internalizing behavior than is the case with more subtle, inwardly-directed symptoms, such as worries, fears, and feelings of anxiety. To some extent, this corroborates prior evidence from Goodman et al. (2010), particularly in that the factor loading of Item 4 (.92) was substantially higher than other Emotional Problems items (range = .64–.84). Overall, the absence of cluster bias for Items 4 and 6 suggested that these symptoms may have been the most valid indicators of teacher-rated internalizing behavior and that the other items were likely influenced by the idiosyncratic views of participating teachers (Bauer et al., 2013; Von der Embse et al., 2019). However, research using scales other than the SDQ are needed to further evaluate these conclusions.

4.3. Between-teacher predictors of rater discrepancies in children's internalizing symptoms

On a conceptual level, cluster bias in the majority of SDQ symptom items suggests that teachers' internalizing symptom ratings did not, in large part, reflect the same construct in every participating classroom. Based on the ABC Model's tenet that discrepancies across teacher ratings arise as a result of differences in teacher characteristics, we considered teachers' gender and teaching experience as between-teacher level factors causing differences in internalizing symptoms ratings. Yet, our findings only partly lent credence to this hypothesis. Specifically, despite a direct negative association with the internalizing symptoms factor, teachers' gender did not explain cluster bias in distinct symptom items at the between-teacher level. Although female teachers' ratings of internalizing symptoms appeared to be more negative than those of their male colleagues, there were no differences between male and female teachers in how they assessed and identified different symptoms. This is unlike prior studies from Splett et al. (2018) and Hopf and Hatzichristou (1999), who reported that female teachers' ratings of internalizing symptoms, including shy/isolated behavior, unhappy feelings, and psychosomatic complaints, are likely to be more positive than their male counterparts. However, both studies combined these symptoms to form a broadband factor of internalizing behavior instead of focusing on separate symptom indicators and explored gender differences through observed rather than latent scores. These differences in approach and design between prior research and our investigation may explain why teachers' gender did not serve as a source of between-teacher variance in teachers' ratings of internalizing symptoms. More research is needed to further explore gender differences in teachers' ratings of internalizing symptoms in class.

Teaching experience only accounted for differences between teachers' average SDQ ratings of the extent to which the child has many worries or often seems worried (Item 3). Conceivably, teachers with less teaching experience in this sample were more likely to identify such symptoms in upper elementary children than were more experienced educators. Generally, this further supports findings of Splett et al. (2018) and Mashburn et al. (2006), suggesting that teachers with both more general teaching experience and professional development experience in behavior screening rate the risk of having social-emotional problem behavior as lower than less experienced teachers. Possibly, teachers may become better judges of the subtle signs of worried children when they accumulate teaching experience and, as such, are less likely than novice teachers to perceive worrisome child behavior as problematic (Kokkinos et al., 2004, 2005). However, even after accounting for differences regarding teaching experience, cluster bias remained in the majority of SDQ-indicators of internalizing symptoms. Hence, not all differences across the teachers were caused by differences in teachers' average ratings of these symptoms in different classes. This highlights the necessity of exploring other factors at the between-teacher level that cause differences in the internalizing symptoms items. Possible examples are teachers' ethnicity, qualifications, and knowledge regarding internalizing symptoms, and interpersonal classroom climate (e.g., Kokkinos et al., 2005; Lau et al., 2004; Zee, de Jong, & Koomen, 2016).

4.4. Within-teacher predictors of rater discrepancies in children's internalizing symptoms

Based on the ABC Model, we also aimed to explain discrepancies among teachers' ratings of internalizing child symptoms by considering the way that information about a specific child's behavior is stored and subsequently accessed from memory (De Los Reyes & Kazdin, 2005). Following prior theory and research on mental relationship representations (Pianta, 1999), we considered teachers' affective relationships with and self-efficacy in relation to individual children as potential violators of measurement invariance at the within-teacher level. Thus far, previous empirical research has almost exclusively included findings on *direct* associations among children's internalizing symptoms and teacher-perceived relationship quality (e.g., Rudasill, 2011; Zee & Roorda, 2018). However, the specific role of teacher–child relationships in the identification and assessment of internalizing symptoms has, to our knowledge, never been explored before. In our study, teachers' perceptions of relational closeness and conflict violated their judgments of individual children's internalizing behavior, both as direct covariates and predictors of distinct symptoms.

Our intermediate RFA models indicated, first, that teachers' perceptions of closeness in teacher–child relationships were likely to cause variation in their assessments of symptoms related to psychosomatic complaints (Item 1), social withdrawal (Item 2), and peer problems (Item 7). Thus, with equal levels of internalizing symptoms, children with whom the teacher experienced high levels of closeness received higher scores on headaches, stomach-aches, or sickness, and lower scores on solitary or age-inappropriate social behavior. According to extended attachment theory, teachers, similar to responsive parents, may provide children with a secure base from which they can explore their environment and a secure haven that helps children maintain proximity to their teachers in times of stress or need (Birch & Ladd, 1998; Hamre & Pianta, 2001). It is possible that when teachers experience high levels of warmth and affection in relation to a child with internalizing symptoms, this may result in a greater sense of knowing that this child experiences psychosomatic problems and thus needs the teacher as a source of emotional support when faced with novel or stressful situations (Pianta, 1999; Zee & Roorda, 2018). Such knowledge may help teachers to more accurately identify the subtle symptoms of children with more internalizing behaviors and may explain why closeness serves as a source of bias for Items 1, 2, and 7 of the internalizing symptoms factor.

In contrast to closeness, conflict only resulted in differences in SDQ Item 6 (picked on or bullied by other children) and Item 8 (many fears, easily scared). Hence, given equal levels of internalizing symptoms, teachers may rate items regarding bullying systematically higher and items regarding anxiety systematically lower for children with whom they experience conflict. However, it is important to note that the strength of these direct associations decreased substantially in the final model in which children's externalizing behaviors were included as well. An explanation is that children's externalizing behaviors may serve as a strong and direct proxy for teachers' experiences of conflict in their relationships with children. For instance, Hamre et al. (2008) used a large sample of preschoolers and teachers and found that more than 50% of the variance in teachers' perceptions of conflict was explained by their ratings of externalizing behavior. Furthermore, other longitudinal studies have shown that children's disruptive behavior may promote vicious cycles of disharmonious relationships and escalating problem behaviors (e.g., Crockett et al., 2018; Doumen et al., 2008; Roorda et al., 2014). In our own sample, the correlation between externalizing behavior and conflict was .69. This high correlation also suggests that the effects of teacher-reported conflict and externalizing behavior are better interpreted in combination with each other, rather than separately (Maassen & Bakker, 2001).

Also interesting is the finding that teachers' self-efficacy in relation to individual children was both a moderate negative predictor of internalizing symptoms and a potential source of measurement bias in individual indicators, such that children with highly self-efficacious teachers obtained systematically lower scores on Item 6 (picked on or bullied by other children) and Item 8 (many fears, easily scared). These results are similar to our findings regarding children's externalizing behavior and teacher–child conflict, which also caused bias in these particular items, but in the opposite direction. Moreover, it complements results of prior empirical research (e.g., Zee, de Jong, & Koomen, 2016), indicating that behaviors in the internalizing spectrum may be directly associated with teachers' student-specific self-efficacy beliefs. However, whereas these past studies have mainly focused on broadband factors of internalizing behavior, our results additionally suggest that *specific aspects* of these overcontrolled behaviors are affected by such self-efficacy beliefs. It is possible that particular symptoms, such as anxious behavior, may reflect a poorer fit with teachers' expectations for appropriate behavior in upper elementary classrooms than other symptoms, such as being bullied, depressive thoughts, or psychosomatic complaints. Such symptoms may be beyond children's control (cf. Chang & Davis, 2009) and are therefore less likely to result in teachers' overreporting of such symptoms. Hence, these findings underline the relevance of investigating teachers' self-efficacy as a source of bias.

Overall, the intermediate models presented in this study highlight the relevance of children's behavior and teachers' beliefs and affective experiences as sources of bias for teachers' ratings of distinct internalizing symptoms, including psychosomatic complaints, social withdrawal, anxiety, and victimization. Yet, the fact that measurement bias in other symptoms (e.g., depressive thoughts or nervous or clingy behavior) could not be explained by these between- and within-teacher level factors warrants further investigation of other factors biasing teachers' assessment of internalizing symptoms.

4.5. Child predictors of rater discrepancies in children's internalizing symptoms

Drawing upon the propositions of the ABC Model, teachers may be influenced by specific characteristics of children with internalizing symptoms rather than having consistent perceptions of all children with these symptoms. Yet, the child factors included in our models only partly accounted for measurement bias. Specifically, children's gender and ethnicity failed to emerge as significant covariates of either the latent internalizing symptoms factor or any of the factor's indicators. The fact that these demographic characteristics were unlikely to cause differences in teachers' ratings of individual children's internalizing symptoms suggests a universality of patterns of children's gender and ethnicity, at least in the context of this study.

Findings from our intermediate models seem to contradict several prior studies in which patterns of discrepant teacher ratings in relation to children's gender (e.g., Mashburn et al., 2006; Splett et al., 2018) and ethnicity (Peters et al., 2014; Splett et al., 2020) have been found. Although we failed to replicate these findings in the current study, this difference may be due to variance in sample characteristics and instrumentation. For instance, instead of including different ethnic groups, we treated ethnic minority children as one group and contrasted these children with the Dutch majority group. Furthermore, we used the relatively brief SDQ to explore teacher ratings of children's internalizing symptoms whereas the majority of prior studies employed universal mental health screening instruments. Yet, in line with Abidin and Robinson's (2002) suggestions, a more optimistic interpretation is that teachers' ratings of internalizing symptoms, as a result of increased awareness, may be increasingly grounded in children's actual behavior in class rather than children's demographic characteristics. This interpretation lends further credence to findings of Zwirs et al. (2011) whose factor analytic models of the SDQ indicated that Dutch upper elementary teachers tended to perceive children's emotional symptoms and peer problems similarly across gender and ethnicity.

In contrast to children's demographic characteristics, their externalizing behaviors resulted in differences in teachers' ratings of these children's internalizing symptoms. Specifically, even if two children in the same classroom may have an equal chance to be picked on or bullied by other children (Item 6) or to be easily scared (Item 8), their teacher appeared to be more likely to rate the one child with externalizing behaviors higher on these items than the other child without such problematic behaviors. From a methodological standpoint, these results speak to the presence of measurement bias regarding externalizing behavior in these two items. From a conceptual viewpoint, however, internalizing symptoms are frequently comorbid with externalizing behavior (Achenbach et al., 1987). Indeed, in our models, the correlation between externalizing behavior and the internalizing symptoms factor was 0.35.

To some degree, our results support those of a meta-analytic study on predictors of bullying and victimization (Cook et al., 2010). In this study, links of both internalizing and especially externalizing behavior with bullying and victimization had effect sizes approaching a medium effect ($r = 0.20$). This suggests that the typical (bully) victim can be characterized as a child who has comorbid externalizing and internalizing problems. Given the moderate positive correlation between internalizing and externalizing behavior in

our sample, teachers may have used individual children's under-controlled actions, including aggressive, disruptive, or noncompliant behaviors, as a lens through which they interpreted other behaviors that may be more internalizing in nature, including (social) anxiety and victimization. However, teachers' ratings of other symptoms that may be more difficult to observe, including unhappy feelings and psychosomatic complaints, seem less likely to be affected by externalizing behavior.

4.6. Limitations

This study's findings need to be interpreted in the context of several limitations. First, it should be noted that measurement bias was tested with the SDQ only. Therefore, caution is warranted when generalizing the present study's findings beyond the scale tested. However, several studies (e.g., Goodman & Scott, 1999; Muris et al., 2004) have indicated that the SDQ, despite its brevity, is at least as good in distinguishing among community samples and at-risk samples and is highly equivalent to other, much longer scales such as the Child Behavior Checklist and Child Depression Inventory, with correlations between the internalizing subscales ranging from .62–.74. It is possible that our results would have been different if we had used a different instrument. In addition, some loadings of the internalizing symptoms factors at the between- and within-teacher level appeared to be relatively weak. These weak factor loadings might have negatively influenced the results of our study. Furthermore, it should be noted that we combined several symptoms to form a broadband factor of internalizing behavior. Despite the fact that the one-factor model had a better fit to the data than a two-factor model of internalizing symptoms, we could not properly investigate whether the between- and within-teacher level factors of interest acted as direct covariates of various latent internalizing symptoms factors, including depressed behavior, psychosomatic symptoms, or withdrawn behavior. Therefore, it is recommended that future researchers also use empirically-based multifactorial syndrome scales other than the SDQ in any attempt to replicate the results. One example is the Teachers' Report Form (Achenbach, 1991) that includes various subscales (e.g., anxious/depressed, somatic complaints).

Second, despite the potential advantages of MSEM, its success heavily depends on data quality and computational problems such as model nonconvergence and unstable estimates, which are likely to arise when MSEMs become increasingly complex (Lee et al., 2018). In our sample, the number of clusters ($N = 92$) was modest and ICC values of some items were below .10. To ensure enough statistical power and avoid computational problems, we therefore ran MSEMS for each of the within-teacher level factors separately. Although this computationally efficient approach helped us to get a clearer picture of which particular items may be biased regarding the Level 1 violators and to avoid potential multicollinearity among them, results from the intermediate models may not provide the complete picture. Therefore, this study could be extended through investigations using a full-model approach and larger samples, which would allow for a deeper understanding of which particular teacher and child factors explain measurement bias in teachers' ratings of internalizing symptoms.

Third, although the method of cluster bias detection offers a small step toward understanding why teachers' ratings of children's internalizing symptoms may be inaccurate, this study does not provide any answers about the actual validity of teacher ratings due to a lack of external criteria of true behavior against which these ratings can be judged. Hence, teacher-perceived ratings would have been strengthened by using external raters, including school psychologists or children themselves, or observations of children's internalizing behavior.

Fourth, caution is warranted when generalizing the results of this study to other populations and settings. Specifically, this study included a non-risk sample of children who were taught by relatively experienced and primarily white female teachers in classes across the Netherlands. These teachers, by virtue of their experience and non-risk student population, may have felt better prepared to deal with children displaying internalizing symptoms in different situations. With this in mind, it is important to extend the generalizability of the current study's results by including a sample of teachers and children from a wider range of backgrounds.

Last, it should be noted that teachers completed the SDQ for seven randomly selected children from their classrooms on average. This may possibly raise questions of selection bias. However, Snijders and Bosker (1999) have previously suggested that inclusion of all children from participating teachers' classrooms is needless when the cluster size of the sample is sufficient, as is the case in the present study. Furthermore, the ICC2 coefficient of .70 indicates that teachers' aggregated internalizing symptoms scores were fairly reliable.

4.7. Implications for research and practice

Despite these limitations, findings from this study have implications for researchers, school psychologists, and other school personnel. First, our findings speak to the importance of investigating idiosyncrasies in teachers' ratings of internalizing symptoms by suggesting that such ratings may have limited use for making comparisons across teachers whose background characteristics, and experiences with and perceptions of children in their classrooms, are different. Although research using scales other than the SDQ are needed to further evaluate this conclusion, our findings stress the need to improve the effectiveness of screening by training teachers to identify the subtle signs of emotional distress. Such efforts may be particularly important as biases within and across teachers due to differences in experience, beliefs, or perceptions of children may lead to under- or overidentification of problem behavior. Following the ABC Model, this may ultimately complicate practical decisions about whether and which of the child's symptoms should be supported in or outside class (De Los Reyes & Kazdin, 2005).

Second, one major assumption in prior research (e.g., Mashburn et al., 2006) on predictors of discrepancies among teachers' ratings of internalizing concerns is that children who are rated similarly on the internalizing symptoms factor will also receive similar ratings by their teachers, regardless of the characteristics of teachers or children. Even though the teachers in our sample could judge the internalizing symptoms of their own students in class in a relatively precise way, our results also show that measurement bias may occur with a variety of internalizing symptoms. Thus, our results point to the potential benefits of evaluating teacher ratings at the

symptom level, next to investigating and making decisions about assessment, classification, and treatment at the construct level.

Third, children's externalizing behavior, but not their gender and ethnicity, was related to teachers' assessments of these children's internalizing symptoms. On a positive note, this suggests that teachers, potentially due to increased awareness, make no distinction between boys and girls or minority or ethnically Dutch children when assessing their internalizing symptoms (cf. Abidin & Robinson, 2002). Yet, to give preservice teachers the greatest chance of success as well as sustain highly experienced teachers as they meet new challenges in class, teacher training and professional development programs should continue to provide teachers with the information needed to understand the implications of social-emotional concerns on students' educational outcomes and respond in productive and meaningful ways to their students in class.

Teaching experience, teachers' beliefs, and affective experiences with individual children also explained differences in teachers' ratings. Hence, the ways teachers appraise their relationships with and self-efficacy beliefs toward individual children may play a vital role in teachers' ability to accurately assess children's internalizing symptoms (Zee, de Jong, & Koomen, 2016). Teachers must be made aware that these beliefs and experiences may have serious implications for their assessments of children's behavior in class, especially since such feelings and beliefs may also be closely related to teachers' behaviors and actions in class (e.g., Bosman et al., 2019).

In conclusion, these findings can inform practices used by school psychologists to both assess students' internalizing problems and assist teachers with students who present problematic behaviors. Although using a different, longer measure may provide a more comprehensive assessment of students' internalizing symptomology, this study provides a more nuanced understanding of teachers' reports of students' internalizing symptoms when using items from the SDQ. Best practices dictate that school psychologists employ multiple measures to obtain the most accurate picture of students' mental health; however, the use of teacher reports on the SDQ may be a useful way to accurately and quickly assess students' internalizing symptoms and assist school psychologists as they work with and support teachers.

Acknowledgment

This research was supported by the Netherlands Organization for Scientific Research (NWO-PROO) and the Van der Gaag grant from the Royal Netherlands Academy of Arts and Sciences (KNAW).

References

Abidin, R. R., & Robinson, L. L. (2002). Stress, biases, or professionalism: What drives teachers' referral judgments of students with challenging behaviors? *Journal of Emotional and Behavioral Disorders, 10*, 204–212, 0.1177/10634266020100040201.

Abikoff, H., Courtney, M., Pelham, W. E., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology, 21*, 519–533. https://doi.org/10.1007/BF00916317.

Achenbach, T. M. (1991). *Manual for the Teacher's report form and 1991 TRF profile*. University of Vermont, Department of Psychiatry.

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.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W.H. Freeman.

Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods, 18*, 475–493. https://doi.org/10.1037/a0032475.

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the bulletin. *Psychological Bulletin, 112*, 400–404. https://doi.org/10.1037/0033-2909.112.3.400.

Birch, S. H., & Ladd, G. W. (1998). Children's interpersonal behaviors and he teacher–child relationship. *Developmental Psychology, 34*, 934–946. https://doi.org/10.1037/0012-1649.34.5.934.

Borg, M. G. (1998). Secondary school teachers' perception of pupils' undesirable behaviors. *British Journal of Educational Psychology, 68*, 67–79.

Bosman, R. J., Zee, M., & Koomen, H. M. (2019). Do teachers have different mental representations of relationships with children in cases of hyperactivity versus conduct problems? *School Psychology Review, 48*, 333–347. https://doi.org/10.17105/SPR-2018-0086.V48-4.

Browne, M. W., & Cudeck, R. (1993). *Alternative ways of assessing model fit. 154* p. 136). Sage Focus Editions.

Caldarella, P., Shatzer, R. H., Richardson, M. J., Shen, J., Zhang, N., & Zhang, C. (2009). The impact of gender on Chinese elementary school teachers' perceptions of student behavior problems. *New Horizons in Education, 57*, 17–31.

Caldarella, P., Young, E. L., Richardson, M. J., Young, B. J., & Young, K. R. (2008). Validation of the systematic screening for behavior disorders in middle and junior high school. *Journal of Emotional and Behavioral Disorders, 16*, 105–117.

Chang, D. F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology, 71*, 235–242.

Chang, M. L., & Davis, H. A. (2009). Understanding the role of teacher appraisals in shaping the dynamics of their relationships with students: Deconstructing teachers' judgments of disruptive behavior/students. In P. A. Schutz, & M. Zembylas (Eds.), *Advances in teacher emotion research* (pp. 95–127). Springer.

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

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Cook, C. R., Williams, K. R., Guerra, N. G., Kim, T. E., & Sadek, S. (2010). Predictors of bullying and victimization in childhood and adolescence: A meta-analytic investigation. *School Psychology Quarterly, 25*, 65–83. https://doi.org/10.1037/a0020149.

Costello, E. J., Copeland, W., & Angold, A. (2011). Trends in psychopathology across the adolescent years: What changes when children become adolescents, and when adolescents become adults? *Journal of Child Psychology and Psychiatry, 52*, 1015–1025.

Crijnen, A. A., Bengi-Arslan, L., & Verhulst, F. C. (2000). Teacher-reported problem behaviour in Turkish immigrant and Dutch children: A cross-cultural comparison. *Acta Psychiatrica Scandinavica, 102*, 439–444. https://doi.org/10.1034/j.1600-0447.2000.102006439.x.

Crockett, L. J., Wasserman, A. M., Rudasill, K. M., Hoffman, L., & Kalutskaya, I. (2018). Temperamental anger and effortful control, teacher–child conflict, and externalizing behavior across the elementary school years. *Child Development, 89*, 2176–2195. https://doi.org/10.1111/cdev.12910.

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.

Dickey, W. C., & Blumberg, S. J. (2004). Revisiting the factor structure of the strengths and difficulties questionnaire: United States, 2001. *Journal of the American Academy of Child and Adolescent Psychiatry, 43*, 1159–1167. https://doi.org/10.1097/01.chi.0000132808.36708.a9.

Doumen, S., Verschueren, K., Buyse, E., Germeijs, V., Luyckx, K., & Soenens, B. (2008). Reciprocal relations between teacher–child conflict and aggressive behavior in kindergarten: A three-wave longitudinal study. *Journal of Clinical Child and Adolescent Psychology, 37*, 588–599. https://doi.org/10.1080/15374410802148079.
Dulin, J. M. (2001). Teacher ratings of early elementary students' social–emotional behavior. *Dissertation Abstracts International, 61*, 3469.
DUO. (2014). *Leeftijd van personeel in het primair onderwijs. [Age of employees in primary education]*. Retrieved from http://www.onderwijsincijfers.nl/kengetallen/primair-onderwijs/personeelpo/leeftijd-personeel.
Fergusson, D. M., Horwood, L. J., & Boden, J. M. (2006). Structure of internalising symptoms in early adulthood. *The British Journal of Psychiatry, 189*, 540–546. https://doi.org/10.1192/bjp.bp.106.022384.
Goldstein, S. E., Boxer, P., & Rudolph, E. (2015). Middle school transition stress: Links with academic performance, motivation, and school experiences. *Contemporary School Psychology, 19*, 21–29. https://doi.org/10.1007/s40688-014-0044-4.
Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to use broader internalising and externalising subscales instead of the hypothesized five subscales on the strengths and difficulties questionnaire (SDQ): Data from British parents, teachers and children. *Journal of Abnormal Child Psychology, 38*, 1179–1191. https://doi.org/10.1007/s10802-010-9434-x.
Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1337–1345. https://doi.org/10.1097/00004583-200111000-00015.
Goodman, R., & Scott, S. (1999). Comparing the strengths and difficulties questionnaire and the child behavior checklist: Is small beautiful? *Journal of Abnormal Child Psychology, 27*, 17–24. https://doi.org/10.1023/A:1022658222914.
Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72*, 625–638.
Hamre, B. K., Pianta, R. C., Downer, J. T., & Mashburn, A. J. (2008). Teachers' perceptions of conflict with young students: Looking beyond problem behaviors. *Social Development, 17*, 115–136. https://doi.org/10.1111/j.1467-9507.2007.00418.x.
Hopf, D., & Hatzichristou, C. (1999). Teacher gender-related influences in Greek schools. *British Journal of Educational Psychology, 69*, 1–18. https://doi.org/10.1348/000709999157527.
Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118.
Jak, S., Oort, F. J., & Dolan, C. V. (2014). Using two-level factor analysis to test for cluster bias in ordinal data. *Multivariate Behavioral Research, 49*, 544–553. https://doi.org/10.1080/00273171.2014.947353.
Jerome, E. M., Hamre, B. K., & Pianta, R. C. (2009). Teacher–child relationships from kindergarten to sixth grade: Early childhood predictors of teacher-perceived conflict and closeness. *Social Development, 18*, 915–945. https://doi.org/10.1111/j.1467-9507.2008.00508.x.
Johnson, A. H., Miller, F. G., Chafouleas, S. M., Welsh, M. E., Chris Riley-Tillman, T., & Fabiano, G. (2016). Evaluating the technical adequacy of DBR-SIS in tri-annual behavioral screening: A multisite investigation. *Journal of School Psychology, 54*, 39–57. https://doi.org/10.1016/j.jsp.2015.10.001.
Keiley, M. K., Lofthouse, N., Bates, J. E., Dodge, K. A., & Pettit, G. S. (2003). Differential risks of covarying and pure components in mother and teacher reports of externalizing and internalizing behavior across ages 5 to 14. *Journal of Abnormal Child Psychology, 31*, 267–283. https://doi.org/10.1023/A:1023277413027.
Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). Guilford Press.
Kokkinos, C. M., Panayiotou, G., & Davazoglou, A. M. (2004). Perceived seriousness of pupils' undesirable behaviours: The student teachers' perspective. *Educational Psychology, 24*, 109–120. https://doi.org/10.1080/0144341032000146458.
Kokkinos, C. M., Panayiotou, G., & Davazoglou, A. M. (2005). Teacher appraisals of student behaviors. *Psychology in the Schools, 42*, 79–89. https://doi.org/10.1002/pits.20031.
Konold, T. R., & Pianta, R. C. (2007). The influence of informants on ratings of children's behavioral functioning: A latent variable approach. *Journal of Psychoeducational Assessment, 25*, 222–236. https://doi.org/10.1177/0734282906297784.
Koomen, H. M. Y., Verschueren, K., Van Schooten, E., Jak, S., & Pianta, R. C. (2012). Validating the student-teacher relationship scale: Testing factor structure and measurement invariance across child gender and age in a Dutch sample. *Journal of School Psychology, 50*, 215–234. https://doi.org/10.1016/j.jsp.2011.09.001.
Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry, 160*, 1566–1577. https://doi.org/10.1176/appi.ajp.160.9.1566.
Lau, A. S., Garland, A. F., Yeh, M., Mccabe, K. M., Wood, P. A., & Hough, R. L. (2004). Race/ethnicity and inter-informant agreement in assessing adolescent psychopathology. *Journal of Emotional and Behavioral Disorders, 12*, 145–156. https://doi.org/10.1177/10634266040120030201.
Lee, J., Shapiro, V. B., Kim, B. E., & Yoo, J. P. (2018). Multilevel structural equation modeling for social work researchers: An introduction and application to healthy youth development. *Journal of the Society for Social Work and Research, 9*, 689–719.
Loo, S. K., & Rapport, M. D. (1998). Ethnic variations in children's problem behaviors: A cross-sectional, developmental study of Hawaii school children. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 39*, 567–575. https://doi.org/10.1017/S0021963098002261.
Maassen, G. H., & Bakker, A. B. (2001). Suppressor variables in path models: Definitions and interpretations. *Sociological Methods & Research, 30*, 241–270. https://doi.org/10.1177/0049124101030002004.
Mashburn, A. J., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2006). Teacher and classroom characteristics associated with teachers' ratings of prekindergartners' relationships and behaviors. *Journal of Psychoeducational Assessment, 24*, 367–380. https://doi.org/10.1177/0734282906290594.
McLean, D., Eklund, K., Kilgus, S. P., & Burns, M. K. (2019). Influence of teacher burnout and self-efficacy on teacher-related variance in social-emotional and behavioral screening scores. *School Psychology, 34*, 503–511. https://doi.org/10.1037/spq0000304.
Mejia, T. M., & Hoglund, W. L. (2016). Do children's adjustment problems contribute to teacher–child relationship quality? Support for a child-driven model. *Early Childhood Research Quarterly, 34*, 13–26. https://doi.org/10.1016/j.ecresq.2015.08.003.
Merrell, K. W. (2008). *Behavioral, social, and emotional assessment of children and adolescents* (3rd ed.). Lawrence Erlbaum Associates.
Morin, A. J., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education, 82*, 143–167. https://doi.org/10.1080/00220973.2013.769412.
Muris, P., Meesters, C., Eijkelenboom, A., & Vincken, M. (2004). The self-report version of the strengths and difficulties questionnaire: Its psychometric properties in 8- to 13-year-old non-clinical children. *British Journal of Clinical Psychology, 43*, 437–448.
Murray, C., & Murray, K. M. (2004). Child level correlates of teacher–student relationships: An examination of demographic characteristics, academic orientations, and behavioral orientations. *Psychology in the Schools, 41*, 751–762. https://doi.org/10.1002/pits.20015.
Muthén, B., & Asparouhov, T. (2013). *New methods for the study of measurement invariance with many groups*. Technical report (Retrieved from http://www.statmodel.com).
Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*, 376–398. https://doi.org/10.1177/0049124194022003006.
Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (6th ed.).
Peters, C. D., Kranzler, J. H., Algina, J., Smith, S. W., & Daunic, A. P. (2014). Understanding disproportionate representation in special education by examining group differences in behavior ratings. *Psychology in the Schools, 51*, 452–465. https://doi.org/10.1002/pits.21761.
Pianta, R. C. (1999). Assessing child-teacher relationships. In R. C. Pianta (Ed.), *Enhancing relationships between children and teachers* (pp. 85–104). Washington, DC: American Psychological Association. https://doi.org/10.1037/10314-005.
Pianta, R. C., Hamre, B., & Stuhlman, M. (2003). Relationships between teachers and children. In W. M. Reynolds, & G. E. Miller (Eds.), *vol. 7. Handbook of psychology: Educational psychology* (pp. 199–234). Wiley.
Proctor, B., & Prevatt, F. (2003). Agreement among four models used for diagnosing learning disabilities. *Journal of Learning Disabilities, 36*, 459–466. https://doi.org/10.1177/0022219403036005070l.

Reijneveld, S. A., Vogels, A. G., Hoekstra, F., & Crone, M. R. (2006). Use of the pediatric symptom checklist for the detection of psychosocial problems in preventive child healthcare. *BMC Public Health, 6*, 197. https://doi.org/10.1186/1471-2458-6-197.

Ritter, D. R. (1989). Teachers' perceptions of problem behavior in general and special education. *Exceptional Children, 55*, 559–564.

Roorda, D. L., Verschueren, K., Vancraeyveldt, C., Van Craeyevelt, S., & Colpin, H. (2014). Teacher–child relationships and behavioral adjustment: Transactional links for preschool boys at risk. *Journal of School Psychology, 52*, 495–510. https://doi.org/10.1016/j.jsp.2014.06.004.

Rubin, K. H., & Coplan, R. J. (2004). Paying attention to and not neglecting social withdrawal and social isolation. *Merrill-Palmer Quarterly, 50*, 506–534. https://doi.org/10.1353/mpq.2004.0036.

Rudasill, K. M. (2011). Child temperament, teacher– Child interactions, and teacher– Child relationships: A longitudinal investigation from first to third grade. *Early Childhood Research Quarterly, 26*, 147–156. https://doi.org/10.1016/j.ecresq.2010.07.002.

Rudasill, K. M., & Kalutskaya, I. (2014). Being shy at school. *Sex Roles, 70*, 267–273. https://doi.org/10.1007/s11199-014-0345-0.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*, 243–248. https://doi.org/10.1007/s11336-009-9135-y.

Shen, J., Zhang, A., Zhang, C., Caldarella, P., Richardson, M. J., & Shatzer, R. H. (2009). Chinese elementary school teachers' perceptions of students' classroom behavior problems. *Educational Psychology, 29*, 187–202.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Sage Publishers.

Sonuga-Barke, E. J., Minocha, K., Taylor, E. A., & Sandberg, S. (1993). Inter-ethnic bias in teachers' ratings of childhood hyperactivity. *British Journal of Developmental Psychology, 11*, 187–200. https://doi.org/10.1111/j.2044-835X.1993.tb00597.x.

Splett, J. W., Raborn, A., Brann, K., Smith-Millman, M. K., Halliday, C., & Weist, M. D. (2020). Between-teacher variance of students' teacher-rated risk for emotional, behavioral, and adaptive functioning. *Journal of School Psychology, 80*, 37–53. https://doi.org/10.1016/j.jsp.2020.04.001.

Splett, J. W., Smith-Millman, M., Raborn, A., Brann, K. L., Flaspohler, P. D., & Maras, M. A. (2018). Student, teacher, and classroom predictors of between-teacher variance of students' teacher-rated behavior. *School Psychology Quarterly, 33*, 460–468. https://doi.org/10.1037/spq0000241.

Stone, L. L., Otten, R., Engels, R. C., Vermulst, A. A., & Janssens, J. 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/s10567-010-0071-2v.

Tandon, M., Cardeli, E., & Luby, J. (2009). Internalizing disorders in early childhood: A review of depressive and anxiety disorders. *Child and Adolescent Psychiatric Clinics of North America, 18*, 593–610. https://doi.org/10.1016/j.chc.2009.03.00.

Valdez, C. R., Lambert, S. F., & Ialongo, N. S. (2011). Identifying patterns of early risk for mental health and academic problems in adolescence: A longitudinal study of urban youth. *Child Psychiatry & Human Development, 42*, 521–538. https://doi.org/10.1007/s10578-011-0230-9.

Valiente, C., Swanson, J., & Lemery-Chalfant, K. (2012). Kindergartners' temperament, classroom engagement, and student–teacher relationship: Moderation by effortful control. *Social Development, 21*, 558–576. https://doi.org/10.1111/j.1467-9507.2011.00640.x.

Van Leeuwen, K., Meerschaert, T., Bosmans, G., De Medts, L., & Braet, C. (2006). The strengths and difficulties questionnaire in a community sample of young children in Flanders. *European Journal of Psychological Assessment, 22*, 189–197. https://doi.org/10.1027/1015-5759.22.3.189.

Van Widenfelt, B. M., Goedhart, A. W., Treffers, P. D. A., & Goodman, R. (2003). Dutch version of the strengths and difficulties questionnaire (SDQ). *European Child & Adolescent Psychiatry, 12*, 281–289. https://doi.org/10.1007/s00787-003-0341-3.

Von der Embse, N., Kim, E. S., Kilgus, S., Dedrick, R., & Sanchez, A. (2019). Multi-informant universal screening: Evaluation of rater, item, and construct variance using a trifactor model. *Journal of School Psychology, 77*, 52–66. https://doi.org/10.1016/j.jsp.2019.09.005.

Zee, M., de Jong, P. F., & Koomen, H. M. Y. (2016). Teachers' self-efficacy in relation to individual students with a variety of social-emotional behaviors: A multilevel investigation. *Journal of Educational Psychology, 108*, 1013–1027. https://doi.org/10.1037/edu0000106.

Zee, M., & Koomen, H. M. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: A synthesis of 40 years of research. *Review of Educational Research, 86*, 981–1015. https://doi.org/10.3102/0034654315626801.

Zee, M., Koomen, H. M., & de Jong, P. F. (2018). How different levels of conceptualization and measurement affect the relationship between teacher self-efficacy and students' academic achievement. *Contemporary Educational Psychology, 55*, 189–200. https://doi.org/10.1016/j.cedpsych.2018.09.006.

Zee, M., Koomen, H. M., Jellesma, F. C., Geerlings, J., & de Jong, P. F. (2016). Inter-and intra-individual differences in teachers' self-efficacy: A multilevel factor exploration. *Journal of School Psychology, 55*, 39–56. https://doi.org/10.1016/j.jsp.2015.12.003.

Zee, M., & Koomen, H. M. Y. (2017). Similarities and dissimilarities between teachers' and students' relationship views in upper elementary school: The role of personal teacher and student attributes. *Journal of School Psychology, 64*, 43–60. https://doi.org/10.1016/j.jsp.2017.04.007.

Zee, M., Koomen, H. M. Y., & van der Veen, I. (2013). Student–teacher relationship quality and academic adjustment in upper elementary school: The role of student personality. *Journal of School Psychology, 51*, 517–533. https://doi.org/10.1016/j.jsp.2013.05.003.

Zee, M., & Roorda, D. L. (2018). Student–teacher relationships in elementary school: The unique role of shyness, anxiety, and emotional problems. *Learning and Individual Differences, 67*, 156–166. https://doi.org/10.1016/j.lindif.2018.08.006.

Zwirs, B., Burger, H., Schulpen, T., Vermulst, A. A., HiraSing, R. A., & Buitelaar, J. (2011). Teacher ratings of children's behavior problems and functional impairment across gender and ethnicity: Construct equivalence of the strengths and difficulties questionnaire. *Journal of Cross-Cultural Psychology, 42*, 466–481. https://doi.org/10.1177/0022022110362752.