2024

Interrater Reliability Between Parents and Teachers Based on the Child's Grade Level in a Rural Community: Behavior Assessment System for Children, Third Edition

Samantha B. Godoy

Interrater Reliability Between Parents and Teachers Based on the Child's Grade Level in a Rural Community: Behavior Assessment System for Children, Third Edition

Samantha B. Godoy

Presented to the Faculty of the

Graduate School of Clinical Psychology

George Fox University

in partial fulfillment

of the requirements for the degree of

Doctor of Psychology

in Clinical Psychology

Newberg, Oregon

Approval Page

Interrater Reliability Between Parents and Teachers Based on the Child's Grade Level: Behavior Assessment System for Children, Third Edition

by

Samantha B. Godoy

has been approved

at the

Graduate School of Clinical Psychology

George Fox University

as a Dissertation for the PsyD degree

Committee Members

Ryan D. Thompson, PsyD, Chair

Elizabeth B. Hamilton, PhD, Member

Amber Nelson, PsyD, Member

February 20, 2024

Abstract

The process of conducting child and adolescent psychoeducational assessments has changed over the past 2 decades (Shapiro & Heick, 2004). In the past, the school psychologists commonly concentrated on behavioral, achievement, and projective assessments and usually did not include systematic multi-rater observation rating scales of behavior. There is now consensus within the professional community that an assessment should meet three criteria including data from multiple methods, data from multiple sources, and data in multiple settings (Alfonso et al., 2020). Multimodal assessment provides a more in-depth perspective and creates less administrator biases, as multiple views of the child's functioning across contexts must be considered. It is not unusual for a child's teacher and parent to have varying views on a child's development, and if only one of these perspectives is considered, an assessment may become unfairly slanted (Alfonso et al, 2020). The current study aimed to understand interrater relationships between teachers and parents based on the grade level of the student. Participants were from an archival database from multiple school districts in rural Oregon. The Behavior Assessment System for Children, Third Edition parent-rating scales and teacher-rating scales were compared between students in elementary school and students in grades above elementary school. Students in elementary school encounter the highest interrater reliability between teachers and parents. There were no significant differences between pre/post COVID data suggesting the pandemic did not interfere with interrater reliability. This information may aid clinicians in understanding interrater relationships and the impact of providing multiple domains when observing children.

Keywords: BASC-3, interrater reliability, rural, externalizing, internalizing, COVID

Table of Contents

Table of Tables

Interrater Reliability Between Parents and Teachers Based on the Child's Grade Level: Behavior Assessment System for Children, Third Edition

Chapter 1

As children grow and learn, they may encounter difficulty learning, concentrating, or getting along with others in their home and school environment. A psychoeducational assessment is often utilized to help those closest to a child understand what specific aspects of that child's development warrant additional support. Through such assessments, a clinician can closely understand the mechanisms behind a child's cognitive processes, learning, and socioemotional development. As such, clinicians in various settings such as hospitals, clinics, and schools are tasked with making complex decisions about a child's welfare from assessment results (Moss & Moss-Racusin, 2021).

The National Center for Education Statistics and Digest of Education Statistics (U.S. Department of Education, 2021) reported that over 7 million children were served under the Individuals with Disabilities Education Act in the 2020–2021 school year. Psychoeducational assessment is an integral part of this process. Arguably, the most valuable part of the psychoeducational assessment is the tailored recommendations unique to a child's diagnosis, that guide the educational and mental health accommodations a child receives. There is also interprofessional utility to these assessments as they can provide diagnostic clarity that is useful to special education staff, counselors, medical providers, and families.

The process of conducting child and adolescent assessments has changed over the past 2 decades (Shapiro & Heick, 2004). In the past, school psychologists commonly concentrated on behavioral, achievement, and projective assessments and usually did not include systematic multi-rater observation rating scales of behavior. There is now consensus within the professional community that an assessment should meet three criteria: data from multiple methods, data from multiple sources, and data in multiple settings (Alfonso et al., 2020). The multimodal assessment provides a more in-depth perspective by integrating multiple views of the child's functioning across situational contexts.

The Behavior Assessment System for Children, Third Edition (BASC-3; Reynolds & Kamphaus, 2015) is a multimethod, multidimensional system used to evaluate emotional and behavioral difficulties in children and adolescents, taking into consideration their history and behavior, not only at home, but also at school. This provides a clinician with the opportunity to minimize referral source bias by asking the child, parents or guardians of the child, and the child's teacher about their unique and individual perspectives related to the child's behaviors and emotional functioning. As noted by Alfonso et al. (2020), this can all be accomplished through the use of one assessment tool, rather than requiring multiple measures to be given.

Emotional and behavioral difficulties may present differently across environmental contexts, based on situational demands. This is why the integrated assessment approach of the BASC-3 components are so desirable, as they incorporate both parent and teacher perspectives. Together, the BASC-3 components offer a comprehensive system for identifying, evaluating, monitoring, and remediating behavioral and emotional problems in children and adolescents. A salient benefit of using the BASC-3 is the ability to use each component individually or in different combinations to gain the best understanding of the child.

Parent Rating Scale/Teacher Rating Scale

The parent rating scale (PRS) and teacher rating scale (TRS) of the BASC-3 are useful in assessing multiple traits of children because they require different respondents in multiple settings. The BASC-3 TRS is a comprehensive measure of both behavioral and adaptive functioning in the school setting. It is designed for use by teachers or others who fill a similar role, such as teacher assistants or preschool caregivers. The forms contain descriptors of behaviors that the respondent rates on a 4-point scale of frequency, ranging from 1 (*never*) to 4 (*almost always*). The TRS takes 10 to 15 min for teachers to complete.

The BASC-3 PRS is a comprehensive measure of a child's behavioral and adaptive functioning in community and home settings. The PRS uses the same four-choice response format as the TRS and takes 10 to 20 min to complete due to having a few more items than the PRS and is broad in range because the number of items increases with the age of the child. The TRS and PRS are almost identical measures that assess for both clinical and adaptive scales, however, there are minor differences which are linked to different contexts. For example, the TRS includes learning problems and study skills while the PRS includes activities of daily living. While informal evaluation and therapeutic interaction also yield important information about a youth's presentation, such clinical data are often somewhat subjective and limited by the unique interaction between particular caregivers or providers and children (Moss & Moss-Racusin, 2021).

Clinical Scales

One of the strengths of the BASC is that behavioral observations are organized into different scales which can guide clinical interpretation and intervention. The Behavioral Symptoms Index measures the overall level of problem behavior and is comprised of six scales, including hyperactivity, aggression, depression, attention problems, atypicality, and withdrawal. The behavioral symptom categories described in the BASC-3 reports are externalizing problems, internalizing problems, and adaptive skills. Externalizing problems consist of a range of behaviors that are directed outward and considered problematic, this scale includes hyperactivity, aggression, and conduct problems. Internalizing problems consist of behaviors that are directed inwardly, and this scale consists of anxiety, depression, and somatization. In addition to the clinical scales, the BASC-3 includes adaptive skills: adaptability, social skills, leadership, study skills, functional communication, and activities of daily living.

Reliability

Before proceeding with a psychological assessment, the assessor needs to understand the reliability and validity of the test instrument they are using. Reliability is the capability of generating reproducible findings and validity is the ability of producing accurate data about the identified domains of functioning. A common concern in psychological assessment of children is that clinicians can over-pathologize a child based on one data point (Moss & Moss-Racusin, 2021). Internal consistency, test-retest reliability, and interrater reliability were all analyzed for the BASC-3 TRS and PRS (Reynolds & Kamphaus, 2015). Each time the BASC-3 is conducted on a child, it looks at interrater reliability among teachers and interrater reliability between parents and teachers. The interrater reliability estimates found on the BASC-3 are generally higher than interrater reliability estimates found on other rating scale instruments (Youngstrom et al., 2000; Reynolds & Kamphaus, 2015). The administrative manual for the BASC-3 notes the following about the relationship between PRS and TRS; “overall, the correlation coefficients between the TRS and PRS scale scores are generally low to moderate, indicating perceived behavioral differences between teachers and parents across settings" (Reynolds & Kamphaus, 2015, p. 280). Further, “among the clinical scales, the correlation coefficients between the TRS and PRS scale scores become higher as age level of the form increases.” (Reynolds & Kamphaus, 2015, p. 280). Lack of interrater reliability could arise for a variety of reasons, such as the child’s behavior changing in different environments, namely school and home. One possibility of the behavioral change between environments may be environmental stressors. It may also be that the expectations of teachers differ from the expectations of parents. The focus of this study is inter-rater reliability, which describes the level of agreement among different raters.

COVID-19 Pandemic

The COVID-19 pandemic has profoundly impacted our school system, catalyzing an unprecedented shift in education worldwide. With the imperative to minimize transmission, schools swiftly adapted by implementing remote learning models, causing a seismic transformation in traditional pedagogical methods. The closure of physical campuses resulted in significant disruptions to academic routines, social interactions, and extracurricular activities, posing challenges for both students and educators. The COVID-19 pandemic has underscored the necessity for reassessing BASC-3 interrater reliability protocols, particularly in virtual or hybrid learning environments, to ensure consistent and accurate behavioral evaluations amidst changing educational dynamics. This study will focus on interrater reliability pre and post COVID-19.

Current Study

It is reported in the BASC-3 manual that the interrater reliability is increased when the students’ ages are increased. All correlations presented in the BASC-3 research are based on clinical and nonclinical cases rather than distinguishing between the two cases. The current study will be based on clinical data from a rural sample. The question arises whether there is a difference between clinical data and nonclinical cases and whether there is a difference in interrater reliability based on geographic area (i.e., rural as opposed to suburban or urban school settings).

Hypothesis 1

It is predicted that exclusively clinical data from a rural sample will not support the assertion in the BASC-3 manual that "among the clinical scales, the correlation coefficients between the TRS and PRS scale scores become higher as age level of the form increases" (Reynolds & Kamphaus, 2015, p. 280). Due to the environmental factors of rural living, it is assumed that teachers of children who are in their earlier years of learning will not only spend far more time with the children in the classroom, but they may also know them from other areas of their life (e.g., church, neighborhood). As the child ages and moves up grade levels, they tend to have more interactions with different teachers and for a shorter time span in each classroom. Teachers of older children and adolescents may not be getting as holistic of a view of the child's behavior and personality due to spending less time with the child. Therefore, it is predicted that students who are younger will have higher interrater reliability between teacher and parents than students who are older.

Hypothesis 2

It is also predicted that there will be higher interrater reliability among parents and teachers when looking at externalizing behavior than when looking at internalizing behaviors. As teachers and parents spend time with the children, it is easier to observe externalizing behaviors in the student. Externalizing behaviors may include acting out in the classroom, throwing objects, screaming, and physical violence. Alternately, internalizing behaviors, like negative self-belief and difficulty concentrating, may be more difficult for the parents and teachers to observe. The hypothesis is that the parents and teachers who spend more time with the student will have higher interrater reliability than parents and teachers who spend less time with the student. There appears to be minimal research on this data, and the current research has nonclinical and clinical populations lumped together. Furthermore, the research that has been conducted focuses on the raters' differences related to specific diagnostic characteristics. This research will be focused on a rural environment with a clinical population. In smaller communities, the teachers are expected to know the children better across the sample. Interrater reliability is likely to be stronger than is reported in the BASC-3 manual due to greater familiarity in a smaller community and the stability of remaining in the same school with the largely the same faculty monitoring the children's behavior over time.

Hypothesis 3

Additionally, it is predicted there will be differences in pre/post-COVID (post-COVID is defined as during the pandemic and after) years due to the unusual challenges parents, teachers, and students faced during this time. Distance learning occurred which was more likely to make it difficult for teachers to become as familiar with their students. It may have also been difficult for parents to objectively view their child's behaviors as they were possibly spending more frequent times with them and were more responsible for their continued attention and learning during distanced learning. Due to the sample being from a rural population, this was likely to be an especially large difference from a suburban or urban setting due to the rural location's tendencies to have higher dual relationships with teachers and students. Interrater reliability pre-COVID is likely to be stronger due to the lack of challenges in building the relationships between teachers and students.

Chapter 2

Methods

Participants

This study utilizes archival data from two rural school district databases located in Oregon, within a behavioral health service group providing assessments for individualized education plans (IEP). The purpose of the assessment is to inform the school district if the child is eligible for an IEP or other school-based accommodations. The participants were referred through the school district and informed consent was previously collected from participant guardians for the purpose of a comprehensive psychoeducational assessment. The data included cases assessed using the BASC-3 from 2016 to 2022. Participants ranged in age from 6 to 18 years and education level from first grade to 12th grade. Other demographic variables, including ethnicity, gender, age, and pre/post COVID will be reported for the final sample.

Table 1

Demographics of the Sample

Item	Category	Frequency	Proportion
Ethnicity	European American	97	0.77
	Latinx	20	0.16
	Biracial	8	0.06
	Unknown	1	0.01
Gender	Male	98	0.78
	Female	27	0.21
	Other	1	0.01
Age	Elementary	48	0.38
	Above elementary	77	0.62

Note. N = 126.

Materials

Demographics

Demographic data were collected from the psycho-educational assessments in student files (see Table 1).

Behavioral Assessment System for Children

The current version of the BASC-3 (Reynolds & Kamphaus, 2015) was published in 2015, it continues to be a multimethod, multidimensional system used to evaluate behavioral and emotional functioning of children aged 2–25 years. The test utilizes multiple respondents in educational settings and home settings to identify clinically significant emotional and behavioral concerns which include externalizing, internalizing, and adaptive functioning.

This study will utilize both the PRS and the TRS which each take 10–20 min to administer. Scores are derived from comparing the individual’s scores to those of age-matched peers. Both the PRS and TRS include questions with a 4-point Likert scale by responding to answers with “N” (*never*), “S” (*sometimes*), “O” (*often*), or “A” (*almost always*). The use of multiple respondents in the BASC-3 allows for the comparison of a child’s behavior across domains and perspectives and reduces risk of misclassification (Stone et al., 2020; Yeguez & Sibley, 2016).

The *T*-scores for multi-rater results for externalizing and internalizing behaviors in the BASC-3 normally range from 30 to 70, with scores higher than 70 indicating more severe externalizing or internalizing behaviors. A *T*-score on the adaptive scales of 40–30 is considered at-risk and below 30 is considered clinically significant.

Table 2

Age Split Descriptive Statistics

Level	Frequency	Proportion
Above Elementary	77	0.62
Elementary	48	0.38

Table 3

Pre/Post COVID Split Descriptive Statistics

Level	Frequency	Proportion
Pre COVID	85	0.68
Post COVID	40	0.32

Procedures

Testing was conducted as a psychoeducational assessment to determine IEP eligibility and provide recommendations for school staff. Doctoral students, supervised by a licensed psychologist, were trained to administer and interpret the assessments. Prior to testing, parental consent was obtained. The BASC-3 PRS and TRS were respectively given to parents and teachers for each student. Due to multiple teachers completing the TRS for the majority of the students, the primary teacher was chosen for children in elementary school and the teacher indicated as “Teacher 1” on their assessment was chosen for middle school and high school.

The archived scores were retrieved from each student’s school file and directly inputted into a digital spreadsheet and coded for the demographic information provided. All information remained confidential and was coded in such a way that individual students are not identifiable.

Chapter 3

Results

Intraclass correlation coefficients (ICC) were calculated for parent and teacher ratings of externalizing and internalizing behaviors for a sample of clinical cases. Further, different ICC values were calculated for the overall sample, for younger students (e.g., elementary, and below) and older students (e.g., elementary, and above). The ICC is a value between 0 and 1, where values below 0.5 indicate poor reliability, between 0.5 and 0.75 moderate reliability, between 0.75 and 0.9 good reliability, and any value above 0.9 indicates excellent reliability (Koo & Lee, 2016). One-sided paired samples *t*-tests were used to assess the differences among these correlations. Consistent with the primary hypothesis, the elementary group is moderately higher in interrater reliability, $t(16) = 2.08$, $p = .027$, $d = 0.504$, 95% CI = [.07, ∞].

Additionally, ICCs were calculated grade level differences (see Table 4) as well as for pre-COVID (prior to March 1,2020) and post-COVID (after March 1, 2020) differences (see Tables 5 and 6). Scales that were not included in the ICC were scales that did not cross the domain of both teacher and parent which included learning problems, study skills, and activities of daily living.

Table 4

Intraclass Correlations

	Scale	Overall ICC	Elementary ICC	Above elementary ICC
1	Hyperactivity	0.60	0.74	0.49
2	Aggression	0.46	0.52	0.36
3	Conduct problems	0.45	0.43	0.46
4	**Externalizing problems**	0.56	0.52	0.48
5	Anxiety	0.15	0.26	0.05
6	Depression	0.27	0.39	0.16
7	Somatization	0.34	0.08	0.45
8	**Internalizing problems**	0.38	0.34	0.41
9	Atypicality	0.23	0.14	0.31
10	Withdrawal	0.26	0.32	0.21
11	Attention problems	0.46	0.60	0.33
12	Behavioral symptoms	0.40	0.55	0.21
13	Adaptability	0.31	0.51	0.14
14	Social skills	0.08	0.05	0.08
15	Leadership	0.13	0.19	0.08
16	Functional communication	0.08	0.09	0.05
17	**Adaptive skills**	0.11	0.12	0.06

Note. Item 4 consists of Items 1–3, Item 8 consists of Items 5–8, and Item 17 consists of Items 13–16.

Table 5

Pre-COVID Intraclass Correlations

	Scale	Overall ICC	Elementary ICC	Above elementary ICC
1	Hyperactivity	0.58	0.74	0.45
2	Aggression	0.51	0.63	0.32
3	Conduct problems	0.47	0.49	0.41
4	**Externalizing problems**	0.58	0.67	0.44
5	Anxiety	0.09	0.24	0.00
6	Depression	0.30	0.44	0.14
7	Somatization	0.37	-0.01	0.50
8	**Internalizing problems**	0.41	0.39	0.42
9	Atypicality	0.27	0.21	0.32
10	Withdrawal	0.18	0.18	0.17
11	Attention problems	0.42	0.48	0.37
12	Behavioral symptoms	0.42	0.61	0.21
13	Adaptability	0.22	0.49	0.00
14	Social skills	0.04	-0.04	0.07
15	Leadership	0.07	0.01	0.07
16	Functional communication	-0.05	-0.01	-0.12
17	**Adaptive skills**	0.08	0.07	0.01

Note. Item 4 consists of Items 1–3, Item 8 consists of Items 5–8, and Item 17 consists of Items 13–16.

Table 6

Post-COVID Intraclass Correlations

	Scale	Overall ICC	Elementary ICC	Above elementary ICC
1	Hyperactivity	0.66	0.74	0.61
2	Aggression	0.36	0.30	0.48
3	Conduct problems	0.40	0.22	0.60
4	**Externalizing problems**	0.53	0.51	0.58
5	Anxiety	0.32	0.32	0.29
6	Depression	0.17	0.10	0.26
7	Somatization	0.21	0.31	0.12
8	**Internalizing problems**	0.27	0.22	0.35
9	Atypicality	0.12	0.06	0.22
10	Withdrawal	0.45	0.56	0.32
11	Attention problems	0.55	0.75	0.22
12	Behavioral symptoms	0.33	0.40	0.26
13	Adaptability	0.56	0.57	0.57
14	Social skills	0.14	0.22	0.11
15	Leadership	0.25	0.39	0.10
16	Functional communication	0.33	0.27	0.40
17	**Adaptive skills**	0.20	0.17	0.22

Note. Item 4 consists of Items 1-3, Item 8 consists of Items 5–8, and Item 17 consists of Items 13–16.

Chapter 4

Discussion

To date, no published study has examined the BASC-3 interrater reliability between parent and teacher in a rural setting. It is important to differentiate between rural and suburban populations because there is a strong environmental difference in the child's upbringing. Rural communities tend to have much smaller populations than urban or suburban communities. This means there may be more overlap in community relationships. The students may see their teachers in more places than just their school setting, such as at the grocery store, place of worship, or community events. Due to the rural population, the teachers may also have multiple roles within the school setting. They may be the student's teacher for English as well as for recreation, allowing them to have more time observing the child.

The data collected indicates there is clinical significance between interrater reliability between teachers and parents of students who are in elementary grades and those who are above elementary grades. A one-sided test indicates that teachers and parents have greater interrater reliability when the student is in elementary school rather than above elementary school. These findings contrast with literature, suggesting that environment (rural vs. suburban) plays a role in the interrater reliability between teachers and parents. This contrasting literature finding suggests potential variations in interrater reliability of BASC-3 assessments between rural and suburban environments, highlighting the importance of considering contextual factors when interpreting behavioral assessments in diverse settings.

Supporting study hypotheses, across domains, externalizing behaviors appear to produce higher interrater reliability which is consistent with literature. This suggests that when observing children, it is more likely a teacher and parent will focus on the outward behaviors rather than the inward (internalizing) behaviors. Interestingly, the elementary group also produced higher rates of interrater reliability in 11 of the 17 scales. This continues to suggest that interrater reliability is higher amongst teachers and parents within the elementary group. As the child ages, therefore, it may become difficult to build a holistic picture of the child's behaviors when they are observed differently at school and at home. As such, it is important to understand how and why the interrater reliability between parents and teachers may differentiate. As children get older and become more independent, they may have less access to community resources and their parents are less involved in their social interactions. Older students may engage in more social activities with their peers rather than spending time with their parents due to a perceived need for independence and less parental support. They may also interact differently around their peers and teachers at school than they do in the household. Thus, providing different behavioral observations between parents and teachers. As children age, they may also become more adept at hiding their behavior problems than younger children, which makes it more difficult for both teachers and parents to discern the child's behavior.

Discussion of the Hypotheses

Hypothesis 1

In a rural population sample, younger students exhibit higher interrater reliability between teacher and parent assessments compared to their older counterparts. This suggests that age differences may influence the consistency of observations across different raters within rural communities. Results confirmed this hypothesis. There was a moderate clinical significance indicating higher interrater reliability between teacher and parents in younger students. Considering this, we also take into account the rural population of the sample. Rural teachers often have a more unique and powerful impact on students' lives (Starrett et al., 2021). They can facilitate cohesive connections that involve supportive relationships with students, thus providing them with a more accurate behavioral observation. In a study by Starrett et al. (2021), it was found that teachers who cultivate supportive relationships with students reported higher levels of accuracy in their behavioral observations, suggesting that these connections play a crucial role in enhancing teachers' ability to perceive and understand students' behaviors.

Hypothesis 2

There will be higher interrater reliability among parents and teachers when looking at externalizing behavior than when looking at internalizing behaviors. Results confirmed this hypothesis for all students. It appears to be easier for parents and teachers to rate a child's externalizing behaviors than their internalizing behaviors. This is noteworthy as there may be some misses in understanding the child's inner thoughts and feelings with just observations. When trying to observe the child's internalizing behaviors it may be helpful to have a conversation with the child to gain further insight into how they view themselves. Research by Eisenberg et al. (2001) underscores the importance of incorporating children's perspectives through direct communication, as it enhances the understanding of their internalizing behaviors and provides valuable insights into their self-perception and emotional experiences. There may be cultural implications here as well when thinking about how we expect students to behave at school and at home.

Hypothesis 3

As the data continued to be broken down and further analyzed, there was a concern the pre/post COVID sample size would not be large enough to deliver adequate results. However, in reviewing acceptable sample sizes, this study provides an appropriate sample size as stated below:

> ICC's of above 0.8, acceptable sample size can be as low as 20 measurements (as either 2 raters and *n* = 10, 4 raters and *n* = 5 or 5 raters and *n* = 4). However, for ICC's of 0.7-0.8 one would need 40 measurements (as either 2 raters and *n* = 20, 4 raters and *n* = 10 or 5 raters and *n* = 8), and ICC's of 0.6-0.7 would require 60 measurements (as either 2 raters and *n* = 30, 3 raters and *n* = 20 or 4 raters and *n* = 15). (Gwet, 2014).

To clarify, this sample consists of two raters, so an example requirement of 60 participants is obtained by having 30 participants in each group, indicating the sample size is acceptable to run an ICC. Overall, there appears to be no significant difference between pre and post COVID differences between grade levels. However, when breaking down the data further, it is interesting to see a difference in functional communication prior to COVID and post COVID. It appears the raters tend to agree more on the student's ability to express ideas and communicate in a way others can understand more post COVID. Prior to COVID, the functional communication scale was the overall least reliable. This may possibly be due to the need to learn how to communicate on screens (distance learning) and not be able to communicate in person. Students may have become more adept at communicating their needs with less distractions or other students around to impact the way they are communicating. When looking at the scale differences, it is interesting to see about half of the scales increased in reliability whereas the other half decreased in reliability. There does not appear to be a pattern when looking at which scales increased and which scales decreased.

Limitations and Future Direction

In this sample using archival data, students were referred by school officials for psychological evaluations due to academic, emotional, and behavioral concerns. As such, this study was not able to include non-clinical controls as an additional comparison group because of the lack of availability. Due to the lack of a non-clinical sample, there is a limitation in understanding how interrater reliability between teachers and parents may be impacted in the general population. The limitation may be that students from a clinical sample may be under more observation as it is and may already be provided with additional interactions from teachers, allowing for stronger interrater reliability. If a non-clinical sample was studied, we may find that there is lower interrater reliability due to less frequent interactions between student and teacher. Future research may indicate that because the clinical sample gets more attention in the community and at school, the teachers may be more aware of the behavioral deficits. Thus, they may over pathologize particular behaviors and emotions and not provide an unprejudiced observation.

Due to the sample being in a rural community in Oregon and not a nationwide sample, we do not know if these findings would be replicable in rural communities outside of Western Oregon. The sample closely matched the ethnic demographics of the area in which the data was collected, but those demographics include very low numbers of Black, Indigenous, and students of color in comparison to many other parts of the country. This is a limitation to the study as the lack of the Black, Indigenous and students of color population does not provide a generalizable view of all rural communities. In addition, the sample size consisted of the majority of participants being male and almost double the amount were above elementary school age. Taking this into consideration, the sample may not accurately capture the interrater reliability of females and elementary aged students.

Further, future researchers may wish to examine the cultural factors of behavior and how those factors influence observations made and completion of the behavior rating scales. Observer biases and gender biases may impact the way the teacher or parent unconsciously observes or rates the student in order to fit their expectation of gender norms. Although gender biases may have declined over the years, gender stereotypes continue, and continue to affect how we subconsciously view others (Tabassum & Nayak, 2021). For example, parents and teachers may view rowdy behaviors in girls as externalizing problems, but the same behaviors in boys may be viewed as normal. Similarly, they may view quiet boys as experiencing significant internalizing problems, but the same behaviors in girls may be viewed as respectful. It may be useful to research how these biases affect the way observers rate the behaviors of children and whether this affects interrater reliability.

It seems intuitive that more interrater reliability is preferable and that it would provide a more helpful picture of a child's functioning. However, one of the reasons examiners gather information from multiple sources is that some lack of interrater reliability is assumed, and the discrepancies give a clearer perception of a child's functioning across settings. As research has been gathered, we understand that children's external and internal behaviors will differ depending on the environment and comfort level of the child. Seeing inconsistencies in interrater reliability is only an additional data point in understanding the child.

Implications

These findings suggest that behavioral observations from multiple areas in a child's life are helpful in providing a holistic view of the child. Where we lack interrater reliability, we may begin to understand how the child may be behaving differently in each setting. This can provide further insight into the child’s inner workings and how they interact socially versus in their home environment. It is important to consider the geographical location of the student when looking at interrater reliability due to the substantial differences within both rural and urban areas.

Conclusions

The age or grade level of the child indicates a significant difference between interrater reliability between teachers and parents on the BASC-3 in a rural setting. Children who are younger in age, or are in elementary school, have higher interrater reliability between teacher and parent. This is likely due to the environmental differences in the rural setting with more overlap in community partnerships and teachers getting to know students in multiple settings. It is also concluded that when children are in lower grade levels, they tend to spend more time in one teacher's classroom, providing a longer timeframe for the teacher to observe the student’s behaviors.

It is confirmed that there is higher interrater reliability between teachers and parents when assessing external behaviors rather than internal behaviors. Due to the easier observability of external behaviors, it is less difficult to accurately identify these behaviors. Furthermore, older children have higher cognitive and social skills, possibly enabling them to mask more of their behavioral problems and emotional dysregulation. The younger child may struggle more to mask their behavior and emotional dysregulation making it easier for any observer to document behavioral patterns. This developmental difference may be why interrater reliability with externalizing and internalizing behaviors between teachers and parents on the BASC-3 may be higher for elementary aged students.

Prior to analyzing the data, there was the belief that there would be a significant difference in interrater reliability between parents and teachers due to the pandemic, distance learning, and decreased interaction with teachers. However, there does not appear to be a statistical relationship between pre/post-COVID and interrater reliability between teachers and parents on the BASC-3. The sample consisted of data from a rural community, where the teachers may have already been provided with ample time to observe the students' behavior prior to COVID. As theorized above, rural populations provide teachers with the ability to observe students in multiple settings, allowing for adequate time to detect behavioral patterns. The environmental location may have been helpful in alleviating a lack of behavioral observations.

References

Alfonso, V. C., Engler, J. R., & Lepore, J. C. C. (2020). Assessing and evaluating young children: Developmental domains and methods. In V. C. Alfonso & G. J. DuPaul (Eds.), *Healthy development in young children: Evidence-based interventions for early education* (pp. 13–44). American Psychological Association. https://doi.org/10.1037/0000197-002

Eisenberg, N., Cumberland, A., Spinrad, T. L., Fabes, R. A., Shepard, S. A., Reiser, M., Murphy, B. C., Losoya, S. H., & Guthrie, I. K. (2001). The relations of regulation and emotionality to children's externalizing and internalizing problem behavior. *Child Development*, *72*(4), 1112–1134. https://doi.org/10.1111/1467-8624.00337

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th Ed). STATAXIS Publishing Company. Advanced Analytics, LLC

Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Moss, N.E., Moss-Racusin, L. (2021). *Practical guide to child and adolescent psychological testing*. Springer. https://doi.org/10.1007/978-3-030-73515-9_1

Shapiro, E. S., & Heick, P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools*, *41*(5), 551–561. https://doi.org/10.1002/pits.10176

Starrett, A., Yow, J., Lotter, C., Irvin, M. J., & Adams, P. (2021). Teachers connecting with rural students and places: A mixed methods analysis. *Teaching and Teacher Education*, *97*, 103231. https://doi.org/10.1016/j.tate.2020.103231

Tabassum, N., & Nayak, B. S. (2021). Gender stereotypes and their impact on women's career progressions from a managerial perspective. *IIM Kozhikode Society & Management Review*, *10*(2), 192–208. https://doi.org/10.1177/2277975220975513

U.S. Department of Education, Office of Special Education Programs, Individuals with Disabilities Education Act (IDEA) database. (2021). *State nonfiscal survey of public elementary and secondary education, 2000-01 through 2019-20 and 2020-21 preliminary*. National Center for Education Statistics, Common Core of Data (CCD). Retrieved February 25, 2022, from https://data.ed.gov/dataset/idea-section-618-data-products.

Yeguez, C. E., & Sibley, M. H. (2016, May 4). Predictors of informant discrepancies between mother and middle school teacher ADHD ratings. *School Mental Health*, *8*(4), 452–460. http://doi.org/10.1007/s12310-016-9192-1

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology*, *68*(6), 1038–1050. https://doi.org/10.1037/0022-006X.68.6.1038