Considering Equity of Evidence: Examining Teachers' Justifications for DBR Scale Scores

Jessica B. Koslouski, Ph.D.[1], Kristabel Stark, Ph.D.[2], Sandra M. Chafouleas, Ph.D.[1], T. Chris Riley-Tillman[3]

[1]University of Connecticut, Storrs, CT, USA

[2]University of Maryland, College Park, MD, USA

[3]University of Missouri, Columbia, MO, USA

Author Note

Published in *School Mental Health* in 2023.

The authors want to acknowledge the additional PIs on the grant (Fabiano, Welsh) and express sincere appreciation to the research team members and teachers who participated in this study.

Preparation of this article was supported by a grant from the Institute for Education Sciences, U.S. Department of Education (R324B060014). Opinions expressed herein do not necessarily reflect the position of the U.S. Department of Education, and such endorsements should not be inferred.

Correspondence concerning this article should be addressed to Jessica B. Koslouski, Ph.D., University of Connecticut, Neag School of Education, 249 Glenbrook Road, Unit 3604, Storrs, CT 06269. Email: jessica.koslouski@uconn.edu

Abstract

Social, emotional, and behavioral (SEB) instruments are currently used in schools to screen, refer, and progress monitor students. Although many of these instruments have demonstrated strong technical adequacy, there has been far less examination of their consequential validity—that is, positive or negative intended and unintended consequences of measure use. A stated purpose of SEB instruments is to facilitate equitable assessment practices; therefore, examining consequential validity is needed. In this study, we examined the unintended negative consequences of one existing instrument: the Direct Behavior Ratings-Single Item Scales (DBR-SIS). We investigated unintended negative consequences by examining variation in the types of evidence teachers use to justify different students' DBR-SIS scores. Participants included twenty-eight teachers (13 elementary, 15 secondary) who watched standardized video clips in which student actors engaged in a variety of behaviors. Using a verbal protocol procedure, we had participants rate the behavior of four focal students using the DBR-SIS while explaining how they arrived at each focal student's score. Using conventional content analysis, we found that teachers' justifications often did not align with definitions provided on the instrument. In these cases, teachers justified scores using labels they applied to students, references to classroom experience, personal expectations for student behavior, instructor redirection, comparisons to other students, and misapplied definitions. Justifications were not consistently applied across students by race and gender; Black students were generally described more harshly than White students. We discuss the potential social consequences of these results and implications for teachers' professional learning.

Keywords: consequential validity, unintended consequences, behavior assessment, implicit bias, qualitative

Considering Equity of Evidence: Examining Teachers' Justifications for Direct Behavior Ratings Scale Scores

Nearly half of students have experienced a potentially traumatic event (Bethell et al., 2017) and approximately one-sixth of students demonstrate behavior or emotional patterns aligned with diagnosable mental health disorders (Danielson et al., 2021). In addition,14% of students qualify for special education services (National Center for Education Statistics [NCES], 2021), and approximately 5% of these students' primary disability category is emotional and behavioral disability (U.S. Department of Education, 2020). Students' experiences of trauma, mental health disorders, and disabilities may all have distinct effects on their abilities to make academic progress (IDEA, 2004; Larson et al., 2017). To disentangle various potential barriers to academic progress, schools rely on tools to identify, address, and progress monitor students' academic and social, emotional, and behavioral (SEB) needs, recognizing that SEB needs often affect students' academic progress.

The Potential and Promise of SEB Instruments

Students' social and emotional experiences in school influence their academic progress (Camacho-Morales et al., 2021; Immordino-Yang et al., 2018; Pekrun et al., 2017). As such, data regarding a students' SEB needs may shed light on the nature of academic challenges and support the identification of appropriate supports. Gathering SEB data is of particular importance for students whose social-emotional histories may function as a barrier to academic progress, such as students who have experienced childhood trauma. SEB tools, which are most often completed by teachers, aim to provide systematic data about students' SEB needs to guide equitable decision making (e.g., Briesch et al., 2021). The J.E.D.I Collaborative (2022) defines equity as "allocating resources to ensure everyone has access to the same resources and opportunities. Equity recognizes that advantages and barriers—the ‘isms’—exist.” (p. 1). Schools use SEB instruments to make decisions about whether to initiate, modify, or discontinue student supports, and the level (i.e., individual, class-wide, school-wide) at which to provide such supports (Johnson et al., 2016). Thus, equity in this decision-making is of paramount importance.

There are several SEB instruments currently used by schools, many of which have been developed over the past decade (Kim et al., 2021). A recent systematic review, for example, identified 29 SEB measures used across 109 studies (Brann et al., 2021). These tools have been developed for universal screening, to inform intervention selection for students exhibiting challenges, and to progress monitor those receiving intervention. Thus, data garnered from these instruments can have powerful consequences for a students’ educational experiences, including referral to special education or the discontinuation of services (Chafouleas et al., 2021; McKeon, 2019). There has also been growing interest in using SEB instruments to universally screen students. Universal screening is intended to proactively and equitably identify students’ needs, and to provide schools with data with which to respond with differentiated supports. As use of these instruments expands, it is even more crucial to ensure that such instruments are well validated and offer opportunities to guide equitable decision making (AERA et al., 2014).

Existing Evidence on SEB Instrument Validation

Instrument validity is a multifaceted concept. According to the Standards for Educational and Psychological Testing (AERA et al., 2014), validity is defined as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (p. 11). Thus, validity includes an examination of both the technical adequacy of the instrument as well as the consequences of the instrument’s use.

In a recent review examining the psychometric properties of SEB instruments, Brann and colleagues (2021) found that the majority of instruments demonstrated technical adequacy (e.g., reliability, content validity, construct validity). However, few studies have considered the consequential validity of these instruments (Cizek et al., 2010). Consequential validity evaluates the social consequences, or short- and long-term intended and unintended consequences of measure use (Messick, 1995, 1998). Importantly, intended and unintended consequences can be either positive or negative. Intended positive consequences are typically the desired goal. Unintended positive consequences could be an unexpected, favorable result. From an ethical standpoint, it is expected that intended negative consequences are avoided from the outset of measure development. Therefore, negative unintended consequences are most concerning. As Messick (1998) explains, "Ideally, there should be no adverse consequences associated with bias in scoring and interpretation, with unfairness in test use, or with negative effects on teaching and learning." (p. 11). In addition, in examining consequential validity—and negative unintended consequences specifically—it is important to examine any differential impact of an instrument on groups of students (Kane, 2013). Differential impact due to bias in scoring and interpretation or unfairness in test use would suggest that the instruments are not producing systematic data with which schools can make decisions, which is likely to have negative effects on teaching and learning.

As schools continue to grapple with increasing rates of SEB concerns (Patrick et al., 2020), it is important that they are able to use standardized data collection systems to identify students' needs and equitably align resources to meet these needs. However, like any measure, SEB ratings can be influenced by the rater's own identity, interpretation of the instrument, and beliefs about the student being rated. Teachers can be just as biased as any other rater; teacher biases regarding race and/or gender may result in disproportionately positive or negative ratings of students (Bryan et al., 2012; Tenenbaum & Ruck, 2017). For example, Kozlowski (2015) found that White and Asian students were more likely to be perceived as trying hard by their teachers even when the student did not rate themselves as hard-working. Weathers (2019) found that teachers in an urban district were more likely to rate Black and Asian students as at-risk of internalizing behaviors if they shared the same race. As such, given widespread interest and use of SEB instruments, we need to better understand if and how bias appears in teachers' use of these instruments. More specifically, we need to better understand whether SEB instruments are producing standardized data to be used by schools or simply reinforcing existing patterns of inequity. If the latter is the case, steps need to be taken to mitigate these unintended negative consequences while strengthening intended positive consequences.

The Present Study

The present study provides a preliminary exploration of the consequential validity of an existing SEB instrument: the Direct Behavior Ratings-Single Item Scales (DBR-SIS; Chafouleas et al.,, 2010). Combining features of behavior rating scales and systematic direct observation, the DBR-SIS (Chafouleas et al., 2010) has teachers rate student behavior at the time and place of the behavior (Christ et al., 2009). It addresses three important aspects of student social behavior: Academic Engagement, Respectfulness, and Disruptive Behavior. Using direct observation, the DBR-SIS uses proportion of time (i.e., 0%–100% of observation period) as its scale. As this format may be more objective than Likert scales traditionally used on SEB instruments (e.g., never, rarely, sometimes, often), the DBR-SIS may hold potential to reduce teacher-as-rater effects. The DBR-SIS is also a brief SEB instrument that can be implemented efficiently, and thus, shows promise of usability for school-wide models.

Previous investigations of the DBR-SIS have demonstrated evidence of inter-rater reliability in a diverse range of general education classrooms (e.g., Briesch et al., 2013; Chafouleas et al., 2010; Johnson et al., 2016). The DBR-SIS has demonstrated concurrent validity with the Behavioral and Emotional Screening System (BESS; Kamphaus & Reynolds, 2007), Student Risk Screening System (SRSS; Drummond, 1994), and systematic direct observation (SDO; Briesch et al., 2013; Johnson et al., 2016; Kilgus et al., 2014; Smith et al., 2018). DBR-SIS ratings are also sensitive to change, suggesting utility of the DBR-SIS as a progress monitoring instrument (Smith et al., 2018). Lastly, evidence suggests that teachers find the DBR-SIS to be acceptable and usable (Smith et al., 2018). However, the DBR-SIS has not yet been evaluated for consequential validity.

In this study, we specifically explore the potential for unintended negative consequences related to teacher use of the measure. To do so, we examine whether variation exists in the types of evidence teachers use to justify the scores they give to students on the DBR-SIS. We used a verbal protocol procedure (Ericsson & Simon, 1993) to gather these data. Teachers watched video clips of a mock classroom in which student actors enacted behaviors that ranged in duration and severity. Teachers then completed the DBR-SIS for four focal students, verbalizing their justifications for each score. Although participants used the DBR-SIS, results may shed light on teachers' decision making while using SEB instruments more broadly.

We examined teachers' justifications of the scores they gave to various students as a way to investigate any unintended negative consequences that may be present as schools use SEB instruments. We examined the extent to which teachers' justifications aligned with the definitions provided on the instrument. We also examined whether teachers' justifications varied by student race and gender when students were enacting standardized behaviors. Understanding how teachers used the instrument informs the types of professional learning potentially needed to facilitate intended and equitable use. The research questions for this study were: (1) How do teachers justify their DBR-SIS ratings? (2) Do these justifications vary by student race and gender?

Method

Participants

A total of 28 teachers (13 elementary, 15 secondary) were recruited from Northeast United States schools. Teachers were recruited from schools participating in larger studies of implementation of the DBR-SIS as a screening and progress-monitoring tool. Teachers were mostly White (24 teachers), female (23 teachers), and well-educated (24 teachers had at least a master's degree). Twenty-five had more than five years of teaching experience; nine had more than 20 years of experience. Teachers worked in a variety of settings (12 urban, 9 suburban, and 11 rural). Participant demographics are shown in Table 1. Although homogeneous in relation to teacher race and sex, these demographics reflect current U.S. teachers (NCES, 2022a). Teachers received a $100 gift card for participating in the study.

[Insert Table 1 here]

Materials

DBR-SIS

The DBR-SIS (Chafouleas et al., 2010) is a behavioral rating scale that combines features of behavior rating scales and systematic direct observation by having teachers rate student behavior at the time and place that the behavior occurs (Christ et al., 2009). The DBR-SIS can be used to evaluate student behavior, guide decisions related to behavior supports, communicate between stakeholders, and progress monitor. It addresses three important aspects of student social behavior: Academic Engagement, Disruptive Behavior, and Respectfulness (Christ et al., 2009). The DBR-SIS measure includes a definition of each target behavior and example behaviors at the top of the measure. Academic Engagement is defined as *actively or passively participating in the classroom activity*. Provided examples include writing, raising hand, answering a question, talking about a lesson, listening to the teacher, reading silently, and looking at instructional materials. Disruptive Behavior is defined as *student action that interrupts regular school or classroom activity*. Examples provided on the instrument include out of seat, fidgeting, playing with objects, acting aggressively, and talking or yelling about things that are unrelated to classroom instruction. Respectfulness is defined as *compliant and polite behavior in response to adult direction and/or interactions with peers and adults*. Examples include following teacher direction, pro-social interactions with peers, positive responses to adult request, and verbal or physical disruption with a negative tone/connotation. Raters indicate the proportion of the observation period (0-100%; recoded as 0-10) that the student exhibited each behavior. For each target behavior, scores closer to 10 indicate higher levels of academic engagement, respectfulness, and disruptive behavior. The DBR-SIS rating form includes a reminder to raters that a lower score for Disruptive Behavior is more desirable. See Supplemental Materials for the full measure.

Video Clips

In order to understand teachers' use of the DBR-SIS instrument, we standardized the behaviors observed by teachers by pre-recording video clips of classroom instruction. In each video clip, a teacher (female for elementary clips, male for secondary clips) led a classroom lesson to a group of approximately 12 students. To provide variation, we elected to have one female and one male teacher. These were previous or current classroom teachers hired to plan and teach a recorded 60-minute lesson. The videos were filmed in an intermediate school classroom with students' desks arranged in groups of six. The elementary teacher taught a reading lesson and the secondary teacher taught a science lesson. Each lesson included large group instruction, partner or small group activities, and independent work. Video clips included "filler students" as well as four focal students. Prior to recording the video clips, researchers provided the students, child actors recruited from local theater schools, with a script that specified how they should behave. The scripts included displaying academically engaged behavior (e.g., following along with materials), as well as disrespectful (e.g., talking back to teacher) and disruptive (e.g., fidgeting) behaviors. During recording, study staff also used off-camera cue cards to prompt students to display specific behaviors (e.g., whisper to others or tap pencils). The teachers leading the lessons were unaware of the study's purpose and the various conditions of student behavior. The teachers only knew student actors would behave in a variety of ways and were instructed not to respond to student behavior.

In order to examine whether there were differences in scoring by race and gender, we specifically recruited student actors with four distinct identities: a Black girl, a Black boy, a White girl, and a White boy. The research team created two sets of video clips, one with upper elementary aged children and one with children in late middle school or early high school. After filming, each video clip was edited to ensure counterbalancing of behaviors across the four focal students and across clips. Because it is impossible to fully counterbalance all behaviors (e.g., yawns, whether or not another student responded to scripted behavior) across students and clips, we prioritized the counterbalancing of disruptive behaviors. When editing the video clips, we confirmed the counterbalancing of behaviors by calculating the number of minutes each student was disruptive in each clip. As the DBR-SIS prompts users to rate the proportion of the observation period the student exhibited the behavior, this follow-up analysis ensured that disruptive behaviors were appropriately counterbalanced. Each final videoclip was 10 minutes long.

Across the video clips, all four students exhibited expected classroom behavior (e.g., following directions) in one clip and low to high intensity disruption for short to long durations in other clips. Low intensity behaviors included talking out of turn, fidgeting, and humming to oneself while working. Medium intensity behaviors included whispering with peers, making faces at others, repeatedly tapping pencils, or gently poking others. High intensity behaviors included talking back to the teacher, yelling at peers, pushing objects or people, and throwing objects. Each student actor was instructed on what level of behavioral intensity to exhibit in the video and for how long (short duration = 1minute, medium duration = 5 minutes, medium-high duration = 6 minutes, long duration = 8 minutes). The counterbalancing of student behaviors across clips is shown in Table 2.

[Insert Table 2 here]

Procedure

Teachers participated in this study individually. All teachers reviewed an online training module about using the DBR-SIS in the two weeks prior to participation (http://dbrtraining.education.uconn.edu). The training reviews the definition and examples of each construct and provides guided practice opportunities with explanations for recommended scores. This training session is consistent with what is recommended for DBR-SIS use and has been shown to be effective in increasing rater accuracy (Chafouleas et al., 2012). Study sessions were conducted in a university lab setting using a computer and audio recorder. During the study session, teachers were told that they would be rating the behavior of four focal students.

Teachers were shown the set of video clips most consistent with their teaching placement (i.e., elementary, secondary). Teachers were asked to continuously talk aloud about what they observed in the four focal students; if needed, the interviewer prompted them to "keep talking" (Ericsson & Simon, 1993). Teachers were intentionally asked to attend to multiple students simultaneously as this mirrors the responsibilities of teachers in classrooms. After each clip, they were asked to continue talking aloud, describing their thinking and ratings as they completed the DBR-SIS form for each focal student. After the teacher rated the four students, the interviewer asked the teacher to elaborate on any scores that did not exemplify perfect student behavior (i.e., 10 for academically engaged and respectful, 0 for disruptive). The presentation of video clips was counterbalanced across participants to eliminate any ordering effects (Campbell & Stanley, 1963; Shadish et al., 2002). Each teacher watched all six video clips in the set (i.e., elementary set or secondary set). Each audio recorded session lasted 100-130 minutes and was transcribed verbatim. Approximately sixty minutes of the session were participants' verbalizations while watching the videos; the remaining 40-70 minutes were their verbalizations and explanations while scoring the focal students on the DBR-SIS.

Analysis

DBR-SIS Score Justifications

Because we were interested in how teachers scored students varying in race and gender on a standardized SEB instrument, we focused our analysis for this study on the second portion of the interview when teachers were describing their thinking while rating students. Although teachers remarked about student behavior while watching the videos, we were particularly interested in how they justified their scores while using the instrument as this more directly relates to the potential consequences (i.e., positive or negative) of instrument use. Because teachers referred to the students using names and pronouns, coders were aware of gender attributions. Names that are used across racial groups were chosen as pseudonyms for the child actors so as to not reveal student race. Coders, who did not watch the videos, were not aware of the race of the focal students while analyzing the transcripts. All coding was completed using NVivo software (QSR International, 2020).

To begin, the first two authors and two research assistants identified each piece of evidence the teacher used to justify their score (e.g., "He was listening to the teacher, he was writing, he was looking at the materials" = three pieces of evidence). There were 6,844 pieces of evidence across the 28 participants. We then used conventional content analysis (Hsieh & Shannon, 2005) to identify whether each piece of evidence that was aligned or not aligned with the definitions provided on the instrument (definitions provided above in measure description). We considered statements that were provided in definitions on the instrument, synonyms of those definitions (e.g., on task), and direct opposites (e.g., off task) to be aligned with the instrument. All other evidence was considered to not be aligned with the instrument. This initial coding was divided amongst the research team, and each piece of evidence was identified and assessed as aligned or not aligned with the instrument by one coder. Prior to this process, research assistants received 60 minutes of training on how to identify evidence and code evidence as aligned or not aligned with the instrument. They received weekly supervision, and all of their coding was reviewed by the first author. The first two authors met weekly to ensure consistent coding procedures and to discuss any questions related to coding.

Next, because we were interested in how raters may have deviated from the instrument, we focused on evidence that was not aligned with the definitions provided for each target behavior. After removing all evidence that was aligned with those definitions, the first two authors inductively coded the evidence that was not aligned (*n* = 2,418 pieces of evidence; 35% of evidence/justifications). As inductive codes were developed, the first two authors reached consensus on a definition for each code and created a codebook (Table 3). The first two authors independently coded the same 50% of data. The coders had 98% agreement in their coding. Given this high level of agreement, the first two authors then each coded half of the remaining data, discussing any borderline cases or uncertainties in weekly meetings. As the codebook was finalized, all data were reviewed once more to ensure accurate coding was applied.

[Insert Table 3 here]

Once all of the evidence was coded, we used matrices (Miles et al., 2020) to look for patterns in how each type of evidence that was not aligned with the instrument was applied by teachers for students of varied race and gender. We examined the number of teachers using each type of evidence and patterns of application to the various students. We also performed further inductive content analysis of the contents of each type of evidence (e.g., whether it was used to justify more or less favorable ratings). Because primary and secondary teachers viewed different video clips (i.e., with elementary or secondary students), we first assessed results of elementary and secondary teachers separately. As our final step, we compared the findings of the elementary and secondary teachers to determine if consistent or divergent patterns emerged.

Throughout coding, we considered our own positionalities as they related to the data and research questions. Our work was informed by our shared values regarding the nature of equitable outcomes for students, including the need to provide teachers with supports to promote such outcomes. The first and second authors have both worked as teachers in primary and secondary settings. As such, our analysis was informed by our own knowledge and experience completing SEB instruments and participating in meetings where teachers shared about student behavior using a variety of evidence, including instruments, formal observation protocols, and their own speculations or conclusions about student behavior. While this insight helped us to design the study and interpret data, we were also careful to consider how participants' experiences were distinct from our own. We held weekly meetings throughout the project to process and debrief our interpretations of teacher justifications within the data set. As a team of White female coders, we took steps to reduce the influence of our own implicit biases during the analysis. For example, none of the coders were aware of teacher or student racial demographics while coding. In addition, we used line-by-lining coding with a detailed codebook that allowed for teachers’ justifications to be matched with a definition and example.

DBR-SIS Scores

After completing the qualitative analysis, we examined whether there were quantitative differences in the scores teachers gave to students on the DBR-SIS. Because student actors were instructed to behave differently in each clip (see Table 2), we averaged the scores given to students when their behavior was scripted to be of the same duration and intensity (i.e., no maladaptive behavior, medium duration – medium intensity, short duration – high intensity, medium-high duration, low intensity). Within each of these conditions, we calculated the average Academically Engaged, Respectful, and Disruptive score given by the participants (13 elementary, 15 secondary) to each student. Finally, we calculated the average Academically Engaged, Respectful, and Disruptive score given to each student across the four conditions.

Results

Table 4 shows the average DBR-SIS scores given to each student by condition. As each condition included scripted behaviors for a specified duration, we would expect DBR-SIS scores to be similar across students, within conditions. At the elementary level, we see differences for specific behavioral intensities. For example, when displaying medium intensity behaviors for a medium duration, girls were scored as more respectful and less disruptive than boys. However, when averaged across conditions, elementary students were all scored within one point of each other. We see greater differences at the secondary level. In some cases, the Black boy or Black girl were scored several points above or below the White boy and White girl. When scores were averaged across conditions, the Black students—and the Black boy in particular—were rated more harshly. Across various conditions, there are quantitative differences in scores by student race, gender, or the intersection of race and gender.

Examining teachers' justifications for the scores they gave provides insight into how these differences may have arisen. In addition to referencing the definitions provided on the DBR-SIS instrument, teachers used six types of evidence to justify student scores: (1) labels they applied to students, (2) references to classroom experience, (3) personal expectations for student behavior, (4) instructor redirection, (5) comparisons to other students, and (6) misapplied definitions. Across these six types of evidence, teachers overwhelmingly commented on students' misbehavior rather than what students did well. For example, rather than state what the student did that *was* respectful, one teacher explained, "He really didn't do anything other than maybe tapping his pencil a little bit and not paying attention to the teacher, so [for] *Respectful*, I will give him a 90%."

Labels Applied to Students

The majority of teachers (12 elementary, 13 secondary) justified at least one score by labeling a student rather than comparing the student's behavior to definitions on the instrument. Although behaviors were standardized across students, participants labeled the students differently. All students received labels with both positive and negative connotations. At the elementary level, labels were most commonly given to the White boy (11 teachers); 4-5 teachers labeled each of the remaining students. The White boy was most commonly labeled as *silly* (8 teachers), but also *an instigator* (3 teachers), *smart*, *a nuisance*, *awful*, *a pain in the butt*, *not a nice kid*, and *out of control* (1 teacher each). The Black boy was labeled as *easily distracted*, *sneaky*, *silly*, *well behaved*, and *interesting* (1 teacher each). The White girl was labeled as *bored* (2 teachers), *easily distracted*, *frustrated*, *out of control*, *a bit defiant*, and *a train wreck* (1 teacher each). Finally, the Black girl was labeled as *interesting* (3 teachers), *a model student*, *polite*, *very quiet*, *a puzzle, tired, spacey, rude, an instigator,* and *a disaster* (1 teacher each).

At the secondary level, labels were most commonly applied to the Black boy (10 teachers); 7 teachers labeled each of the remaining students. The Black boy was labeled as *goofy*, *silly*, *distracted*, *busy*, *annoying*, *an itch*, *rude*, *distracting*, *mischievous*, *evil*, *horrible*, *ridiculous*, *crazy*, *a nuisance*, and *an instigator* (1 teacher each). The White boy was labeled as *a model citizen*, *quiet*, *a listener*, *not a rioter*, *passive*, *busy*, *playful*, *rude*, *mischievous*, *a bad boy, a challenge,* and *a ringleader* (1 teacher each). The White girl was labeled as *quiet* (4 teachers), *an honor student, a good girl, Cinderella, a nuisance, defiant, sassy,* and *restless* (1 teacher each). The Black girl was labeled as *a good student, social, busy, a booger, sassy, obnoxious, disinterested, a mess, part of the crew,* and *a ringleader* (1 teacher each).

References to Classroom Experience

Rather than justifying scores using definitions provided on the instrument, teachers often referenced their own classroom experience. Nine of the 13 elementary teachers used their classroom experience to interpret inattentive student behavior. Teachers provided positive interpretations of the White girl's (6 teachers) and White boy's (3 teachers) inattention. For example, one teacher explained, "Teaching as many years as I have, I know that sometimes even the blank stared kids are listening to you, they just might not be looking at the paper when [they] are auditory listeners." However, no teachers offered positive interpretations of the Black boy's inattentive behavior. Instead, four teachers expressed that his inattentive behavior was evidence of not being academically engaged. For example, one teacher conveyed, "Even though once in a while he seemed to be engaged, I don't think, even when he seemed to be engaged, he was not actually actively engaged." She later stated that he "pretended to be engaged." Another teacher expressed, "When you're putting more focus on your materials than the person who's talking, that's not being academically engaged." Teachers offered alternative explanations for the Black girl's inattention. For example, one teacher stated, "something was either on her mind or something like that" and another stated that "these lessons are too long for [her]."

At the secondary level, 14 of the 15 teachers drew on their classroom experience to justify scores. Teachers most frequently drew on their own experience to offer positive interpretations for the White boy's behavior (11 teachers). One shared, "He is a listener. Can he fold airplanes and listen at the same time? Yes." Another stated, "Maybe he needs to hold something; some students need to do that." Another teacher offered a favorable interpretation of him questioning the teacher. She explained, "although he did get argumentative, like 'why do I need to know this?' That shows that he was actually engaged in what they were learning, and he wanted to know why he had to learn it and why it was important." Only two teachers had less favorable interpretations of the White boy's behavior. For example, one explained, "I would say he was doing things he knew he'd get away with, flaunting rules within the classroom, just wasn't getting caught." One teacher also provided a positive interpretation of the White girl's behavior, explaining, "I think the only reason she was off track was maybe just an academic reason that she was just like, I really can't follow what's going on. It's not that she's trying to be off task." When noting that the Black girl put her head on the desk, teachers offered both positive ("She really just had her head down for that portion. She could have just been listening though") and negative ("she was just not able to learn cause her head was down") interpretations for this behavior. Teachers rarely drew on their classroom experience when interpreting the behavior of the Black boy.

Personal Expectations for Student Behavior

Although the DBR-SIS provides definitions for each of the constructs, nine elementary and 11 secondary teachers drew on their own expectations or definitions of academically engaged, respectful, or disruptive when assessing student behavior. Most often, teachers described inattentive or disruptive behavior as disrespectful. For example, one teacher expressed, "I think they are less respectful just by not paying attention to me, by not paying attention while I am teaching the lesson or not looking or doing things that I have asked them to do." At the secondary level, teachers also imposed expectations for eye contact (e.g., "I would have liked to see as a teacher more eye contact, more acknowledgment that he was focused on what the lesson was"), participation ("I would have liked to seen her look up and raise her hand"), and seating position ("I think it's a sign of respect when you're in a lesson, you don't sit on top of your chair.").

Instructor Redirection

Whereas the DBR-SIS is an instrument intended to be based only on student behavior, the majority of teachers (13 elementary, 14 secondary) justified some scores based on their observation of the students' instructor as well. Participants justified scores based on whether student behavior prompted instructor redirection. For example, teachers explained ratings by sharing, “the teacher had to speak with him numerous times,” and “she had to be redirected by the teacher.”

Comparisons to Other Students

Similarly, the majority of teachers (13 elementary, 13 secondary) used observations of other students as evidence for their DBR-SIS ratings. Teachers justified scores by stating that a student’s behavior was better, similar to, or worse than their peers. Patterns emerged in how students were compared. The secondary White girl was never described as behaving worse than her peers, and the secondary Black boy was never described as behaving better than his peers. The elementary White boy, elementary Black girl, and secondary Black girl were more commonly described as behaving *worse* than their peers. The elementary White girl, elementary Black boy, and secondary White boy were more commonly described as behaving *better* than their peers.

Misapplied Definitions

Lastly, despite the DBR-SIS including definitions for each construct on the rating form, most of the teachers (12 elementary, 12 secondary) misapplied definitions in at least one case. Fidgeting was used as evidence of disrespectful behavior, rather than as evidence of disruptive behavior, and pouting and eye rolling were used as justifications for disruptive behavior rather than as evidence of disrespectful behavior. In addition, eight elementary teachers and two secondary teachers stated that being academically disengaged (i.e., distracted) was disruptive. Whereas the definition of Disruptive is behavior that disturbs *others*, some teachers identified certain students as disrupting themselves. Three elementary teachers evaluated the Black boy as “disturbing himself.” One teacher also evaluated the White girl as disturbing herself. At the secondary level, two teachers identified the Black girl as “only disruptive to herself” and one identified the White boy as “disrupting his own academic engagement – his own opportunity to learn.”

Discussion

This study sought to understand how teachers make sense of and justify their scores while rating students using a well-established SEB instrument, the DBR-SIS. We found that teachers often used evidence outside of the construct definitions provided on the instrument. Teachers justified scores with evidence based on labels they applied to students, references to classroom experience, personal expectations for student behavior, instructor redirection, comparisons to other students, and misapplied definitions. We also found that DBR-SIS scores and justifications for those scores were not provided consistently across students of varied race and gender. Despite the fact that all students in our simulated classrooms engaged in standardized behaviors, teachers’ scores, descriptions of behaviors, and justifications for DBR-SIS scores varied widely. Had these been actual students in a school, it is likely that students may have been over- or under-identified for support (e.g., classroom-based support from the teacher or additional school-based supports) based on teachers’ appraisals and ratings of student behavior. Although we do not measure consequences for students directly, our study demonstrates that teachers’ use of the DBR-SIS may have unintended, adverse, and inequitable consequences. As such, we offer this study as a means to consider how school practices, such as SEB assessment, that aim to mitigate inequities (Chafouleas et al., 2022) necessitate critical examination.

Despite two of the three constructs being positively worded (i.e., Academic Engagement, Respectful), teachers much more commonly highlighted student misbehavior rather than what students did well. Thus, an unintended social consequence of these instruments may be the promotion and reinforcement of deficit thinking about students, whereby teachers remember and emphasize infrequent misbehavior over student success (i.e., negativity bias; Rozin & Royzman, 2001). Teachers' application of labels to students rather than behaviors also raises concerns about the potential for enduring judgments about students. Whether these labels were positive (e.g., smart) or negative (e.g., a troublemaker), they may affect teachers' interpretations of students' future behaviors, making it more difficult for teachers to modify their thinking about particular students. As teachers' expectations for students greatly influence their academic achievement (e.g., through the opportunities and scaffolding that are subsequently provided; de Boer et al., 2018), these labels hold particular weight. Negative labels may also increase punitive responses to students while reducing academic expectations, rigor, and scaffolding. Meanwhile, positive labels (e.g., hardworking) may cause teachers to overlook signs that a student may need supplemental support.

The use of instructor redirection and comparisons to other students raises several concerns about the validity of data garnered by the instrument. It is likely that instructors redirect students at different rates; this may reflect their biases or their knowledge of and relationships with students (Starck et al., 2020). Therefore, justifying student behavior ratings based on instructor redirection challenges the construct validity of the instrument as it measures the instructor's professional practices rather than the student's behavior. In addition, justifying student scores based on comparison to other students means that the bar is always changing as each lesson and classroom will have different students and behavior. SEB instruments are intended to be used to make decisions about the provision of services (e.g., refer, continue, discontinue). However, scores based on comparisons to peers in the classroom compromise the capacity of schools to accurately assess students across classrooms and over time.

We found that DBR-SIS scores awarded to students and the justifications for those scores varied based on both student gender and race, and that Black students were generally described more harshly than White students. For example, when teachers provided evidence based on their professional experience, the elementary Black boy was evaluated much more harshly than his peers whereas judgements of the secondary White boy were much more forgiving than those applied to his peers. When teachers compared students to their peers, the secondary Black boy was never described as behaving better than his peers and the secondary White girl was never described as behaving worse than her peers. Although teachers sometimes tempered negative interpretations for behavior, no teachers interpreted the behavior of the secondary Black boy with any softness or nuance. This aligns with research that finds that Black students are disciplined more harshly for the same behaviors as White students (Anderson & Ritter, 2017) and that Black students as young as preschool are suspended and expelled at disproportionate rates (Gilliam, 2016). Our results extend prior research, demonstrating that teachers' harsher judgments of Black students' behavior start with fairly low intensity classroom behaviors such as inattention.

Because our results demonstrate that implicit biases against students of color may manifest in teachers' use of SEB instruments, we recommend that school leaders critically consider requiring staff who use these instruments to engage in anti-racist professional learning opportunities in addition to training on the instrument itself. In 2020, the National Association of School Psychologists (NASP) acknowledged the role of school psychologists in sustaining systematic racism (NASP, 2020), and developed a library of professional learning resources to support educators looking to disrupt inequities in their schools (i.e., NASP, 2021). School leaders should also recognize that combating biases takes prolonged engagement, and a stand-alone training will not be sufficient. Leonard & Woodland (2022) offer an example of how the development of professional learning communities (PLCs) helped educators in an urban school district promote anti-racist work in sustainable ways.

Our results also suggest a need for an approach to reducing bias that considers students' intersectional identities. For example, in their brief on the disciplinary experiences of Black girls, Crenshaw and colleagues (2015) note that the effects of punitive discipline on Black girls are often distinct from the effects on Black boys and may require different types of resources to disrupt. We urge school leaders to evaluate the implicit and explicit ways that the intersectionality of a students' race and gender may impact teachers' interpretations of their behaviors in the classroom, and to provide training to teachers to raise critical awareness of the nature of their interactions with students. Furthermore, because a growing proportion of youth express feeling displaced or misplaced within a binary gender framework, it is also crucial that schools provide opportunities for teachers to learn how to create inclusive environments and proactively recognize behaviors that may be associated with discrimination based on gender and sexuality (Miller, 2018).

Limitations and Future Directions

It is important to consider the limitations of this study. First, teachers might use the instrument differently when rating their own students. "Talking aloud" while viewing the video clips may have resulted in different types of evidence than if teachers were observing students silently. Because teachers were also viewing video clips of a mock classroom with child actors with whom they had no existing relationship, their evidence and score justifications may have been less nuanced than they would be in the context of a relationship. Second, we recognize that neither race nor gender are binary constructs, and that there is high variability within commonly used race and gender constructs. For the purposes of this study, we chose to use binary identity categories (White or Black; boys or girls) in order to consider trends in the data within our sample size. Given the growing diversity of the student population (NCES, 2022b), future research could examine how teachers use the instrument with students of a more diverse range of racial, ethnic, and gender identities. In addition, future research could examine how other aspects of students' identities, such as sexual orientation, religion, home language, and social class impact teachers' ratings of students. Although we discuss race and gender as distinct identity markers in much of this paper, we recognize that each students' identity is an intersection of many identity characteristics, and that a student's intersectional identity may at times protect them from bias and at other times put them at risk of bias. Third, due to the demographic composition of our sample, we did not examine whether aspects of the teachers' identities shaped their ratings. In future research, scholars could purposively sample to have a large and diverse enough participant pool to examine whether teachers' identities are associated with the ways in which they score particular students.

Although we used scripts, cue cards, and video editing to maximize the counterbalancing of behaviors, not every single behavior (e.g., yawns, the presence or absence of a filler student reacting) could be counterbalanced. However, as the DBR-SIS uses duration as its scale, we confirmed that each of the focal students displayed disruptive behavior for the number of minutes planned. It should also be noted that we specifically analyzed teachers' justifications that were not aligned with the definitions provided on the instrument. Future research should also investigate if and how justifications aligned with the instrument are applied (e.g., do teachers reference body language aligned with the instrument [looking at the teacher] in equitable patterns or are there differences by race and gender?).

Finally, although our study points toward the directions for professional learning necessary to enable equitable and effective use of this instrument, we were not able to test whether additional training would have such intended effects. Future research could consider using videos such as those in this study with teacher candidates to practice equitable scoring, discuss discrepancies in ratings, and examine whether candidates improve reliability over time. Researchers should also test the efficacy of in-service professional learning opportunities in which educators have opportunities to practice and receive feedback to promote positive social consequences.

Conclusion

The implications of our results add to a small but growing literature on unintended negative consequences of SEB instruments. Our results suggest that teachers' scores on SEB instruments may be biased towards or against certain students. This may reinforce deficit thinking about particular students and potentially result in systematic over- and under-identification of students in need of services. As a variety of key groups (i.e., parents, teachers, administrators; Briesch et al., 2020) support the use of universal screening, the use of SEB instruments is likely to increase. Therefore, it is critical that these issues be addressed through continued research and professional learning.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, K. P., & Ritter, G. W. (2017). Disparate use of exclusionary discipline: Evidence on inequities in school discipline from a U.S. state. *Education Policy Analysis Archives, 25*(49). http://doi.org/10.14507/epaa.25.2787

Bethell, C. D., Davis, M. B., Gombojav, N., Stumbo, S., & Powers, K. (2017). *Issue brief: Adverse childhood experiences among US children*. http://www.cahmi.org/wp-content/uploads/2018/05/aces_fact_sheet.pdf

Brann, K. L., Daniels, B., Chafouleas, S. M. & DiOrio, C. A. (2021). Usability of social, emotional, and behavioral assessments in schools: A systematic review from 2009 to 2019, *School Psychology Review*, *51*(1), 6-24. https://doi.org/10.1080/2372966X.2020.1836518.

Briesch, A. M., Chafouleas, S. M., Dineen, J. N., McCoach, D. B., & Donaldson, A. (2021). School building administrator reports of screening practices across academic, behavioral, and health domains. *Journal of Positive Behavior Interventions*, 1-12. https://doi.org/10.1177/10983007211003335

Briesch, A. M., Cintron, D. W., Dineen, J. N., Chafouleas, S. M., McCoach, D. B., & Auerbach, E. (2020). Comparing stakeholders' knowledge and beliefs about identifying and supporting students' social, emotional, and behavioral health in schools. *School Mental Health, 12*, 222-238. https://doi.org/10.1007/s12310-019-09355-9

Briesch, A. M., Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C., & Christ, T. J. (2013). The influence of alternative scale formats on the generalizability of data obtained from Direct Behavior Rating Single-Item Scales (DBR-SIS). *Assessment for Effective Intervention, 38*(2), 127–133. https://doi.org/10.1177/1534508412441966

Bryan, J., Day-Vines, N. L., Griffin, D., & Moore-Thomas, C. (2012). The disproportionality dilemma: Patterns of teacher referrals to school counselors for disruptive behavior. *Journal of Counseling & Development*, *90*(2), 177-190. https://doi.org/10.1111/j.1556-6676.2012.00023.x

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Houghton Mifflin.

Camacho-Morles, J., Slemp, G. R., Pekrun, R., Loderer, K., Hou, H., & Oades, L. G. (2021). Activity achievement emotions and academic performance: A meta-analysis. *Educational Psychology Review*, *33*(3), 1051-1095. https://doi.org/10.1007/s10648-020-09585-3

Chafouleas, S. M., Briesch, A. M., Lane, K. L., & Oakes, W. P. (2022). Improving educations' use of data-driven problem-solving to reduce disciplinary infractions for students with emotional disturbance. In P. Fenning & M. Johnson (Eds.), *Discipline disparities among students with disabilities: Creating equitable environments* (pp. 108-123). Teachers College Press.

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., Christ, T. J., Black, A. C., & Kilgus, S. P. (2010). An investigation of the generalizability and dependability of direct behavior rating single item scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. *Journal of School Psychology*, *48*(3), 219–246. https://doi.org/10.1016/j.jsp.2010.02.001

Chafouleas, S. M., Johnson, A. H., Riley-Tillman, T. C., & Iovino, E. A. (2021). *School-based behavioral assessment: Informing prevention and intervention* (2nd ed.). The Guilford Press.

Chafouleas, S., Kilgus, S., Riley-Tillman, T., Jaffery, R., & Harrison, S. (2012). Preliminary evaluation of various training components on accuracy of Direct Behavior Ratings. *Journal of School Psychology, 50*, 317-334. https://doi.org/10.1016/j.jsp.2011.11.007

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34*(4), 201–213. https://doi.org/10.1177/1534508409340390

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement,* 70(5), 732-743. https://doi.org/10.1177/0013164410379323

Crenshaw, K., Ocen, P., & Nanda, J. (2015). *Black girls matter: Pushed out, overpoliced, and underprotected*. Accessed November 1, 2022 from http://static1.squarespace.com/static/53f20d90e4b0b80451158d8c/t/54dcc1ece4b001c03e323448/1423753708557/AAPF_BlackGirlsMatterReport.pdf

Danielson, Bitsko, R. H., Holbrook, J. R., Charania, S. N., Claussen, A. H., McKeown, R. E., Cuffe, S. P., Owens, J. S., Evans, S. W., Kubicek, L., & Flory, K. (2020). Community-based prevalence of externalizing and internalizing disorders among school-aged children and adolescents in four geographically dispersed school districts in the United States. *Child Psychiatry and Human Development, 52*(3), 500–514. https://doi.org/10.1007/s10578-020-01027-z de Boer, H., Timmermans, A. C., & van der Werf, M. P. C. (2018). The effects of teacher expectation interventions on teachers' expectations and student achievement: Narrative review and meta-analysis. *Educational Research and Evaluation, 24*(3-5), 180-200. https://doi.org/10.1080/13803611.2018.1550834

Drummond, T. (1994). The Student Risk Screening Scale (SRSS). Josephine County Mental Health Program. https://doi.org/10.1037/t27737-000

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data (Rev. ed.). The MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

Gilliam, W. (2016). Early childhood expulsions and suspensions undermine our nation's most promising agent of opportunity and social justice. Robert Wood Johnson Foundation. http://themoriahgroup.com/wp-content/uploads/2018/05/early-childhood-expulsions-and-suspensions.pdf

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qual Health Res*, 15(9), 1277-1288. https://doi.org/10.1177/1049732305276687

Individuals With Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Immordino-Yang, M. H., Darling-Hammond, L., & Krone, C. (2018). *The brain basis for integrated social, emotional, and academic development: How emotions and social relationships drive learning*. https://www.aspeninstitute.org/wp-content/uploads/2018/09/Aspen_research_FINAL_web.pdf

J.E.D.I Collaborative. (2022). *Supporting systemic change in the natural products industry*. Retrieved November 1, 2022 from https://jedicollaborative.com/

Johnson, A. H., Miller, F. G., Chafouleas, S. M., Welsh, M. E., Riley-Tillman, T. C., & 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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kamphaus, R. W., & Reynolds, C. R. (2007). *Behavior Assessment System for Children – Second Edition (BASC-2): Behavioral and Emotional Screening System (BESS)*. Pearson. https://doi.org/10.1037/t29902-000

Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C., Christ, T. J., & Welsh, M. E. (2014). Direct behavior rating as a school-based behavior universal screener: Replication across sites. *Journal of School Psychology, 52*(1), 63–82. https://doi.org/10.1016/j.jsp.2013.11.002

Kim, E. K., Anthony, C. J., & Chafouleas, S. M. (2021). Social, emotional, and behavioral assessment within tiered decision-making frameworks: Advancing research through reflections on the past decade. *School Psychology Review, 51*(1), 1-5. https://doi.org/10.1080/2372966X.2021.1907221

Kozlowski, K. P. (2015). Culture or teacher bias? Racial and ethnic variation in student–teacher effort assessment match/mismatch. *Race and Social Problems*, *7*(1), 43-59. https://doi.org/10.1007/s12552-014-9138-x

Larson, S., Chapman, S., Spetz, J., & Brindis, C. D. (2017). Chronic childhood trauma, mental health, academic achievement, and school-based health center mental health services. *Journal of School Health, 87*(9), 675-686. https://doi.org/10.1111/josh.12541

Leonard, A. M., & Woodland, R. H. (2022). Anti-racism is not an initiative: How professional learning communities may advance equity and social-emotional learning in schools. *Theory Into Practice, 61*(2), 212-223. https://doi.org/10.1080/00405841.2022.2036058

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from Persons' responses and performances as scientific inquiry into score meaning. *The American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Messick, S. (1998). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. *ETS Research Report Series, 2*, i–32. https://doi.org/10.1002/j.2333-8504.1998.tb01797.x

Miles, M.B., Huberman, A. M., & Saldaña, J. (2020). *Qualitative data analysis: A methods sourcebook* (4th ed.). Sage.

Miller, s. (2018). Reframing schooling to liberate gender identity. *Multicultural Perspectives, 20*(2), 70-80. https://doi.org/10.1080/15210960.2018.1447067

National Association of School Psychologists. (2020). *School psychology unified antiracism statement and call to action*. https://www.nasponline.org/resources-and-publications/resources-and-podcasts/diversity-and-social-justice/social-justice/school-psychology-unified-antiracism-statement-and-call-to-action

National Association of School Psychologists. (2021). *Social Justice*. Retrieved November 1, 2022 from https://www.nasponline.org/social-justice

National Center for Education Statistics. (2021). *Digest of Education Statistics, 2019* (NCES 2021-009): *Chapter 2 Elementary and Secondary Education*. U.S. Department of Education, Institute of Education Sciences. Retrieved June 27, 2022 from https://nces.ed.gov/programs/digest/d19/ch_2.asp

National Center for Education Statistics. (2022a). Characteristics of Public School Teachers. *Condition of Education*. U.S. Department of Education, Institute of Education Sciences. Retrieved June 23, 2022 from https://nces.ed.gov/programs/coe/indicator/clr.

National Center for Education Statistics. (2022b). Racial/Ethnic Enrollment in Public Schools. *Condition of Education*. U.S. Department of Education, Institute of Education Sciences. Retrieved June 27, 2022 from https://nces.ed.gov/programs/coe/indicator/cge.

Patrick, S. W., Henkhaus, L. E., Zickafoose, J. S., Lovell, K., Halvorson, A., Loch, S., Letterie, M., & Davis, M. M. (2020). Well-being of parents and children during the COVID-19 pandemic: A national survey. *Pediatrics, 146*(4), e2020016824. https://doi.org/10.1542/peds.2020-016824

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (2017). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child Development*, *88*(5), 1653-1670. https://doi.org/10.1111/cdev.12704

QSR International Pty Ltd. (2020) NVivo (released in March 2020). https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*, 296-320. https://doi.org/10.1207/S15327957PSPR0504_2

Shadish, W., Cook, T., & Campbell, D. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

Smith, R. L., Eklund, K., & Kilgus, S. P. (2018). Concurrent validity and sensitivity to change of Direct Behavior Rating Single-Item Scales (DBR-SIS) within an elementary sample. *School Psychology Quarterly*, *33*(1), 83–93. https://doi.org/10.1037/spq0000209

Starck, J. G., Riddle, T., Sinclair, S., & Warikoo, N. (2020). Teachers are people too: Examining the racial bias of teachers compared to other American adults. *Educational Researcher, 49*(4), 273-284. https://doi.org/10.3102/0013189X20912758

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology, 99*(2), 253–273. https://doi.org/10.1037/0022-0663.99.2.253

U.S. Department of Education. (2020). *OSEP fast facts: Children identified with emotional disturbance*. Retrieved November 1, 2022 from https://sites.ed.gov/idea/osep-fast-facts-children-IDed-Emotional-Disturbance-20

Weathers, E. S. (2019). Bias or empathy in universal screening? The effect of teacher-student racial matching on teacher perceptions of student behavior. *Urban Education*. https://doi-org/10.1177/0042085919873691

Table 1

Participant demographics (n =28)

	n	%
Sex		
Female	23	82.1%
Male	5	17.9%
Race		
Black/African American	3	10.7%
White	24	85.7%
Other	1	3.6%
Degree level		
Bachelor's degree	4	14.3%
Master's degree	19	67.9%
Master's plus	5	17.9%
School level		
Elementary	13	46.4%
Secondary	15	53.6%
Teaching experience		
1-5 years	3	10.7%
6-10 years	9	32.1%
11-15 years	7	25.0%
16-20 years	1	3.6%
21+ years	8	28.6%
School urbanicity [a]		
Rural	12	42.9%
Suburban	9	32.1%
Urban	11	37.9%

Note. Totals to more than 100% because some participants selected multiple descriptors (e.g., some specified working in magnet schools that served multiple communities).

Table 2

Scripted intensity and duration of maladaptive behavior by student race and gender in each ten-minute clip

	White Boy	Black Boy	White Girl	Black Girl
Clip 1	None	None	None	None
Clip 2	Medium duration Medium intensity	Medium duration Medium intensity	None	Short duration High intensity
Clip 3	None	Medium-high duration Low intensity	Short duration High intensity	Medium-high duration Low intensity
Clip 4	Short duration High intensity	None	Medium duration Medium intensity	Medium duration Medium intensity
Clip 5	Medium-high duration Low intensity	Short duration High intensity	Medium-high duration Low intensity	None
Clip 6	Long duration Medium intensity	Short duration Medium intensity	Short duration Medium intensity	Long duration Medium intensity

Note. None indicates that students followed classroom rules for the entirety of the clip. Short duration= 1 minute; medium duration= 5 minutes; medium-high duration = 6 minutes; long duration= 8 minutes. Low intensity= talking out of turn, fidgeting, humming to self while working; medium intensity= whispering with or making faces at others, repeatedly tapping pencils, gently poking others; high intensity= talking back to teacher, yelling at peers, pushing objects or people.

Table 3

Codes for Evidence Beyond Definitions Provided on the Instrument

Code	Definition	Example
Labels	Teacher justifies score using adjectives to describe student (rather than behavior).	"He was a nuisance."
References to classroom experience	Teacher justifies score using teaching experiences outside of the video clips and study setting.	"Some kids need to doodle to pay attention."
Expectations for student behavior	Teacher justifies score by referencing their own expectations for student behavior.	"When someone is actively engaged, they are following along, they are raising their hand, they are participating."
Instructor behavior	Teacher justifies score based on whether the instructor redirected or reprimanded the student or the number of times that happened.	"She had to redirect him 3 times."
Other students' behavior	Teacher justifies score based on how a student's behavior compared to one or more other students.	"She was more focused than any of the others."
Misapplied definitions	Teacher justifies score using behaviors from a different construct's definition.	"Disruptive.. he was off task for that portion."

Table 4

Average score for each focal student on clips with matched intensity and duration of maladaptive behaviors

Elementary				Secondary			
Condition	Academically Engaged	Respectful	Disruptive	Condition	Academically Engaged	Respectful	Disruptive
None				None			
White boy	8.38	9.54	0.77	White boy	8.67	9.87	0.13
Black boy	9.31	9.92	0.23	Black boy	7.47	8.93	1.67
White girl	8.54	9.54	0.54	White girl	9.27	9.73	0.07
Black girl	7.69	9.46	0.85	Black girl	6.93	8.20	0.80
Medium duration – medium intensity				Medium duration – medium intensity			
White boy	4.62	6.77	4.62	White boy	6.27	7.87	3.07
Black boy	5.85	7.31	3.85	Black boy	4.20	5.00	6.40
White girl	5.69	8.23	1.92	White girl	7.67	8.67	2.53
Black girl	5.77	9.23	2.38	Black girl	2.87	4.67	6.87
Short duration – high intensity				Short duration – high intensity			
White boy	7.23	7.54	2.69	White boy	7.73	7.60	2.60
Black boy	6.17	7.42	3.00	Black boy	3.13	2.40	7.00
White girl	6.54	6.38	4.31	White girl	7.60	8.20	2.40
Black girl	7.38	8.00	2.54	Black girl	6.40	6.00	3.93
Medium-high duration – low intensity				Medium-high duration – low intensity			
White boy	5.50	7.08	4.17	White boy	9.00	9.60	0.87
Black boy	6.54	8.92	2.23	Black boy	3.73	4.93	5.67
White girl	5.33	6.50	4.42	White girl	4.67	4.47	5.27
Black girl	4.38	6.92	5.54	Black girl	3.53	5.40	5.67
Average across conditions				Average across conditions			
White boy	6.43	7.73	3.06	White boy	7.92	8.74	1.67
Black boy	6.97	8.39	2.33	Black boy	4.63	5.32	5.19
White girl	6.53	7.66	2.80	White girl	7.30	7.77	2.57
Black girl	6.31	8.40	2.83	Black girl	4.93	6.07	4.33