Hammill Institute on Disabilities

Assessment for Effective Intervention
2021, Vol. 46(4) 316–321

DOI: 10.1177/1534508420944231
aei.sagepub.com
SAGE

A Comparison of Teacher and Student Ratings in a Self-Monitoring Intervention

Allison Bruhn, PhD[1], Sheila Barron, PhD[1], Bailey Copeland, MA[2], Sara Estrapala, MA[1], Ashley Rila, MA[1], and Joseph Wehby, PhD[2]

Abstract

Self-monitoring interventions for students with challenging behavior are often teacher-managed rather than self-managed. Teachers direct these interventions by completing parallel monitoring procedures, providing feedback, and delivering contingent reinforcement to students when they monitor accurately. However, within self-monitoring interventions, the degree to which teachers and students agree in their assessment of students' behavior is unknown. In this study, a self-monitoring intervention in which both teachers and students rated the students' behavior, we analyzed 249 fixed interval ratings of behavior from 19 student/teacher pairs to determine the relationship between ratings within and across teacher/student pairs. We found a strong correlation overall (r =.91), although variability existed within individual pairs and student ratings tended to be higher than teacher ratings. We discuss implications for practice, limitations, and future directions.

Keywords

interventions, technology, progress monitoring

Self-monitoring (SM) is an antecedent-based strategy in which students are taught to recognize the occurrence of a specific behavior and record the extent to which that behavior occurs at predetermined times. Theoretically, SM is effective because it prompts students to be intentional about exercising control over their behavior (i.e., self-regulation; Bandura, 1991). Research indicates SM has been successful in improving students' academic and behavioral outcomes (Bruhn et al., 2015). One argument for SM interventions is that if they are truly student-managed, students will become more self-reliant and independent while also reducing the cost and burden associated with teacher-managed interventions (Briesch & Chafouleas, 2009). A meta-analysis of SM interventions for students with autism found greater student involvement resulted in stronger effects (Davis et al., 2016). In contrast, reviews of SM studies have revealed SM interventions are often reliant on the teacher to manage external contingencies such as delivering feedback and reinforcement (Briesch & Chafouleas, 2009). On one hand, this may be viewed as a limitation of the extent to which SM is really self-managed by the student. Conversely, teacher involvement has been used as one way to improve the accuracy of students' SM, while also promoting generalization across settings (Peterson et al., 2006).

In a review of 41 peer-reviewed articles on SM for students with problem behavior, Bruhn and colleagues (2015) reported 13 studies including contingent reinforcement for students' SM accuracy (i.e., student ratings [SR] matched teacher ratings [TR] either exactly or within a range during the same time period). SM with reinforcement for matching accuracy has resulted in decreases in off-task and disruptive behaviors (e.g., Freeman & Dexter-Mazza, 2004). Relatedly, researchers have found that once students were deemed accurate with SM and accuracy checking and reinforcement were removed, on-task behavior continued to improve (Peterson et al., 2006). Conversely, Ardoin and Martens (2004) found accuracy matching decreased disruptive behavior, but when matching was removed, behavior worsened.

Regardless of the effects of accuracy matching and its unique contribution to SM interventions, across these studies, the teacher's rating is the presumed standard for accuracy (e.g., Chafouleas et al., 2012). In some cases, these data may be used to make decisions about student responsiveness to SM. In a recent study, 13 elementary teachers and one of their students completed ratings of students' behavior during

[1]The University of Iowa, Iowa City, IA, USA
[2]Vanderbilt University, Nashville, TN, USA

Corresponding Author:
Allison Bruhn, The University of Iowa, N252 Lindquist Center, Iowa City, IA 52242, USA.
Email: allison-bruhn@uiowa.edu

one instructional classroom activity (Bruhn et al., 2019). The length of the time between ratings (e.g., every 5 min) and total session length varied by teacher (e.g., 45 min). Teachers used their ratings of behavior to (a) determine whether students were responding to the SM intervention and (b) make intervention adaptations (e.g., increasing SM interval length). According to multilevel modeling of teachers' rating data, students improved their positive behaviors significantly ($p < .001$) from baseline to intervention.

As students progress through SM interventions, which include teachers completing parallel procedures to check for accuracy and make data-based decisions, teacher support may be faded to promote maintenance and generalization. To continue tracking student progress without parallel data, teachers may have to rely on students' SM data to evaluate on-going response to intervention. For teachers and researchers who view teachers' data as the standard for accuracy, they may be hesitant to rely on students' SM data for fear it may be unreliable. To this end, the purpose of this *brief report* is to examine the degree to which teacher and student ratings completed as part of an SM intervention are related. Research questions (RQ) include the following:

Research Question 1 (RQ1): Across all sessions and teacher/student pairs, is there a correlation between average teacher and average student ratings?
Research Question 2 (RQ2): Are average teacher and average student ratings significantly different?
Research Question 3 (RQ3): To what extent is there agreement between teacher and student within sessions for individual student–teacher pairs?

Method

Participants and Setting

The Institutional Review Boards at two universities and three school districts approved this study. Participants included teachers and students from two school districts (A and B) in a Midwest state that is noncategorical for special education services (i.e., students are not labeled under the 13 disability categories) and one urban district (C) from a Southern state. One middle school from District A (rural), three elementary schools from District B (small city), and one middle school from District C (urban) participated. Teachers of Grade 3–8 consented to participate and identified students who might benefit from behavioral SM (e.g., frequent off-task behavior, high rates of office discipline referrals, behavior goal on individualized education program). Then, we obtained parental consent and student assent, and teachers completed the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) on the consented student. Students who scored in the borderline or abnormal range for hyperactivity/inattention, conduct problems, or total difficulties screened into the study. In total, 17 teachers and 18 students participated in the study. One student participated with the same teacher in two different settings, and thus, we analyzed each setting separately. One teacher completed procedures with two different students (each at different times and in separate settings). Thus, the analysis indicates 19 student/teacher combinations (see Table 1).

Measures and Procedures

SDQ. The SDQ is a behavioral rating scale consisting of 25 items rated on a 0–2 scale (i.e., never, sometimes, always) that are used to assess student risk across five domains: *hyperactivity/inattention, emotional symptoms, conduct problems, peer problems*, and *prosocial behavior*. The first four domains constitute an aggregate score for *total difficulties*. The SDQ was originally validated for ages 4–17 years. It has demonstrated high correlations with the Rutter Questionnaire (Rutter, 1967) and the Child Behavior Checklist (Achenbach, 1991); while also evidencing adequate internal consistency ($\alpha = .64–.89$; Hill & Hughes, 2007).

Percentage of positive behavior: Teacher and student ratings. Teachers and students used a noncommercially available, author-developed mobile application (MoBeGo) on an iPad to rate student behavior. We used MoBeGo rather than traditional paper forms because this app was being tested as part of an externally funded research and development project funded by the Institute of Education Sciences (510-14-2540-00000-13607400-6200-000-00000-20-0000). As part of the iterative development process of MoBeGo, we aimed to determine the extent to which teacher and student ratings were similar to each other.

Prior to completing ratings, teachers first met with research assistants (RAs) to complete a 1-hr training sequence during which they determined (a) students' problem behaviors, (b) positive replacement behaviors to monitor, (c) the class or activity for monitoring, and (d) SM interval lengths. Teachers programmed the app to these specifications during the training. Teachers had the option to select from positive behaviors from the default settings in the app or input their own behaviors. The behaviors had accompanying operational definitions in the form of a question (e.g., Be Responsible = Did the student work carefully on the assigned task and ask for help if needed?). Teachers could select as few as one or as many as five behaviors, although generally, they selected three. Teachers and RAs discussed various classroom scenarios and how behaviors might look during these scenarios.

After programming behaviors into the app, teachers selected a target class period (e.g., seventh period math) or instructional activity (e.g., reading rotations) for the student to self-monitor. Teachers selected the class or activity during which the student most often displayed the problem

Table 1. Participant Demographics.

Variables	District A (One Middle School)		District B (Three Elementary Schools)		District C (One Middle School)	
	Teachers $n = 6$	Students $n = 6$	Teachers $n = 7$	Students $n = 7$	Teachers $n = 4$	Students $n = 5$
Gender						
Male	0	6	2	2	1	1
Female	6	0	5	5	3	4
Ethnicity						
White	6	5	6	3	3	2
Black	0	0	1	2	1	2
Hispanic	0	1	0	2	0	1
Disability Status						
General Ed	1	0	6	4	4	4
Special Ed	5	6	1	3	0	1
SDQ total difficulties risk						
Very high		3		3		2
High		1		0		2
Slightly raised		2		3		1
Close to average		0		1		0

Note. SDQ = Strengths and Difficulties Questionnaire.

behavior. Each behavior was rated on a fixed interval, selected by the teacher, for the duration of the class period or instructional activity. For instance, if math instruction occurred for 45 min and the teacher selected a 5-min interval length, then they had up to nine opportunities for ratings. Teachers customized interval length to suit individual student need (e.g., severity of problem behavior, student age) and instructional context. An audio prompt from the app signaled the interval was over and it was time to rate. Ratings followed a 5-point numerical scale with accompanying anchors (*0 = never, 1 = a little, 2 = sometimes, 3 = a lot, 4 = always*; see Figure S1 in Supplemental Appendix). The app automatically calculated and graphed an aggregate percentage of positive behavior (PPB) by summing the total number of points earned, dividing by the total points possible, and multiplying by 100. Using the previous example, if the teacher rated two behaviors, there was a possibility of 72 points (two behaviors × four points × nine ratings). Previous research has indicated moderate to high correlations between teachers' ratings of students' positive behavior and systematic direct observation of academic engagement (r = .61–.91; Bruhn et al., 2018), as well as high interrater reliability between teachers and RAs, using the same 5-point scale (r = .82–.91; Bruhn et al., 2018).

Once teachers completed the training, they began rating their student's behavior during the same instructional period for 3 consecutive days (i.e., baseline). Following baseline, after class was over, the teacher and RA trained the student in the classroom. This included teaching the student about the programmed behaviors by reviewing operational definitions (e.g., examples and nonexamples), discussing why these behaviors are important to classroom success, and asking the student how they would rate given different classroom scenarios. Students learned how to use the various features and functions of the app (e.g., where to touch the iPad to rate behaviors, how to view behavior definitions). Next, students practiced rating behaviors with the app based on hypothetical scenarios. Scenarios included examples or nonexamples of the programmed behaviors, and then students practiced rating behaviors using the app until they demonstrated 100% accuracy using the app's functions and indicated they were comfortable with procedures and definitions.

The next day, both the teacher and student rated the student's behavior during the same instructional period using the same interval length and procedures the teacher used during baseline. During the intervention condition, students rated first. Immediately after rating, students passed the device to their teacher, and the teacher rated the student's behavior for that same interval. After both students and teachers completed ratings independently, they viewed both ratings before starting the next interval (see Figure S1 in Supplemental Appendix). Although they viewed these ratings together to see how ratings aligned, teachers did not deliver planned reinforcement for accuracy matching. Teachers had the option to provide specific feedback on the ratings (e.g., "You did a great job with . . . " and "I see we

both rated you a 3 . . . "). This continued for each interval until the end of the session. We used the total PPB from TR and SR from each completed session for data analysis. Across the 19 teacher/student pairs, the number of completed sessions for each pair ranged from 6 to 27 (median = 12) resulting in 249 sessions with a PPB from both TR and SR.

Data Analysis

RQ1: Correlation between average teacher and student ratings. To determine whether there was a correlation between the average TR and the corresponding average SR for teacher/student pair, we first calculated the average rating across sessions. We then plotted the averages and calculated Pearson's correlation. Given research suggesting students can be trained to accuracy (e.g., Ardoin & Martens, 2004), we hypothesized a moderate to high correlation between average ratings.

RQ2: Difference between teacher and student ratings. To examine the overall degree of agreement between TR and SR, we conducted a paired samples *t* test. We used this to determine whether there was a significant difference between TR and SR. Based on previous research (e.g., Ardoin & Martens, 2004), we hypothesized students would rate themselves higher, but the difference would not be significant.

RQ3: Within teacher–student pair agreement. First, we used linear regression separately for each teacher/student pair to examine the degree of the relationship within TR and SR individual pairs. Second, we used mixed model analysis to measure the degree of relationship across students. We hypothesized correlations would vary by individual pair, but the relationship would be significant.

Results

RQ1: Correlation Between Average Teacher and Student Ratings

We found a strong positive relationship ($r = .91$) between the average TR and average SR obtained for each student (see Figure S2 in Supplemental Appendix). This finding indicates that both teachers and students scored behaviors similarly, on average, across observations. One student, who had very low TR and SR, appears in the scatterplot as an outlier. Repeating the analysis without the outlier showed little change in the correlation coefficient ($r = .86$).

RQ2: Difference Between Teacher and Student Ratings

In 14 of 19 cases, on average, students rated themselves higher (see Figure S3 in Supplemental Appendix). Specifically, students rated about 4.6 points higher than the corresponding TRs, which was a statistically significant difference ($t[18] = 2.85$, $p = .01$).

RQ3: Within Teacher–Student Pair Agreement

In 14 of 19 cases, TR and SR demonstrated a moderate to high correlation ($r = .52$–$.96$). We did not observe this trend in five cases (see Figure S4 in Supplemental Appendix). Ratings for Student 1 showed no association ($r = .05$) due to consistently high SR. For Student 7 (not shown), we did not calculate a correlation because the student consistently gave perfect ratings. The moderate correlations for Student 16 ($r = .39$) and Student 19 ($r = .45$) were also hampered by low variability in the SR. Thus, in these four cases, the lack of a strong correlation appears to be a result of low variability in SR, which were consistently higher than TR. The fifth case, Student 18, may be the most interesting ($r = .47$). Student 18 showed high agreement when the teacher provided a high rating but showed greater variability in self-ratings when the teacher provided a low rating.

Mixed model analysis confirmed the earlier findings. That is, we found a significant positive relationship between teacher and student ratings ($p = .0001$). Second, we found considerable variability in the strength of the relationship across individual students ($p = .0001$).

Discussion

The role of the teacher in SM interventions primarily has been to check the accuracy of students' SM. These teacher-completed monitoring procedures can yield data to monitor students' response to intervention. As students demonstrate a positive response to intervention and teacher support is faded, then teachers may use student SM data to determine whether the students' behavior change is maintaining. Thus, the purpose of this study was to determine the degree to which TR and SR are similar, as this information is pertinent to teachers for a number of reasons. If teachers are resistant to relying on students' SM data, these findings may shed light on the accuracy with which students can self-monitor. Following this logic, if teacher and student data are similar, then this supports fading teacher involvement so that the intervention can be truly student-managed (e.g., Peterson et al., 2006).

In general, the results of this brief report indicate average TR and SR are strongly correlated. Despite this strong correlation, students tend to rate themselves 4.6% points higher. This is not surprising given previous research indicating students tend to rate themselves more positively (e.g., Ardoin & Martens, 2004). Although the difference is statistically significant, this relatively small difference begs the question as to whether this is a practically significant difference. It is unclear whether a TR of 75% and an SR of 79.6% is meaningfully different within the context of the classroom.

When examining individual cases, nearly 75% of cases indicated moderate to high correlations. In terms of practical implications, if the SR and TR indicate the student was demonstrating high PPB over time and there was strong agreement between SR and TR, then fading teacher support may be warranted. This transfer of intervention management can be done by gradually reducing the number of intervals the teacher rates each day, or by reducing the number of days the teacher completes ratings (e.g., every other day). For SM interventions that include reinforcement for accuracy matching, the matching discussion and subsequent reinforcement can be faded and removed (e.g., Peterson et al., 2006).

However, periodic TR allow the teacher to continue monitoring for student accuracy while the student develops independence over time. For some students, however, continual teacher involvement is necessary. Specifically, in the cases demonstrating weak correlations between TR and SR, further training and supports may be needed to develop accuracy. This could involve the teacher and student reviewing the definitions of the behaviors to ensure understanding. In addition, the teacher could provide contingent reinforcement to the student for accurate ratings (Ardoin & Martens, 2004; Chafouleas et al., 2012; Peterson et al., 2006).

Limitations and Future Directions

Although the current study is unique in that we compared TR and SR of behavior, this study is not without limitations. First, though the sample size is relatively small, it is similar to, and in some cases larger than, samples of similar studies comparing TR to outside observer ratings on comparable behavioral rating scales (Bruhn et al., 2018; Riley-Tillman et al., 2008). While the sample includes nearly equal numbers of students with and without disabilities and equal numbers of boys and girls, future research should include larger samples that are more diverse in terms of race/ethnicity. A second limitation is the use of a 5-point scale to generate a total PPB. Whereas some research suggests the scale gradient does not affect rating reliability (Briesch et al., 2013), it is possible the extent to which teachers and students agree will vary based on the instrument being used and the behavior being assessed. In the future, researchers should consider examining how different scale gradients or a dichotomous monitoring system (e.g., yes/no) impact agreement between teachers and students. Finally, we purposefully did not mandate teachers provide reinforcement for accuracy matching, though they did have the option to provide feedback to students. It is plausible that reinforcement for accuracy would have resulted in different results (i.e., higher agreement). Despite this, findings suggest TR and SR are comparable without accuracy-based reinforcement.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for this paper was provided by the *Institute of Education Sciences* (510-14-2540-00000-13607400-6200-000-00000-20-0000) and the *Iowa Measurement Research Foundation* (05202017).

Supplemental Material

Supplemental material for this article is available online at https://journals.sagepub.com/doi/suppl/10.1177/1534508420944231.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 profile*. Department of Psychiatry, The University of Vermont.

Ardoin, S., & Martens, B. (2004). Training children to make accurate self-evaluations: Effects on behavior and the quality of self-ratings. *Journal of Behavioral Education*, *13*, 1–23.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, *50*, 248–287.

Briesch, A. M., & Chafouleas, S. M. (2009). Review and analysis of literature on self-management interventions to promote appropriate classroom behaviors (1988–2008). *School Psychology Quarterly*, *24*(2), 106–118.

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.

Bruhn, A. L., Barron, S., Fernando, J. A., & Balint-Langel, K. (2018). Extending the direct behavior rating: An examination of schoolwide behavior ratings and academic engagement. *Journal of Positive Behavior Interventions*, *201*(1), 31–42.

Bruhn, A. L., McDaniel, S., & Kreigh, C. (2015). Self-monitoring interventions for students with behavior problems: A review of current research. *Behavioral Disorders*, *40*(2),102–121.

Bruhn, A. L., Rila, A., Mahatmya, D., Estrapala, S., & Hendrix, N. (2019). The effects of data-based, individualized interventions for behavior. *Journal of Emotional and Behavioral Disorders*, *28*(1), 3–16.

Chafouleas, S. M., Sanetti, L. M. H., Jaffery, R., & Fallon, L. M. (2012). An evaluation of a classwide intervention package involving self-management and a group contingency on classroom behavior of middle school students. *Journal of Behavioral Education*, *21*, 34–57.

Davis, J. L., Mason, B. A., Davis, H. S., Mason, R. A., & Crutchfield, S. A. (2016). Self-monitoring interventions for students with ASD: A meta-analysis of school-based research. *Review Journal of Autism and Developmental Disabilities*, *3*, 196–208.

Freeman, K. A., & Dexter-Mazza, E. T. (2004). Using self-monitoring with an adolescent with disruptive classroom behavior: Preliminary analysis of the role of adult feedback. *Behavior Modification*, *28*(3), 402–419.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *38*, 581–586.

Hill, C. R., & Hughes, J. N. (2007). An examination of the convergent and discriminant validity of the Strengths and Difficulties Questionnaire. *School Psychology Quarterly*, *22*, 380–406.

Peterson, L. D., Young, K. R., Salzberg, C. L., West, R. P., & Hill, M. (2006). Using self-management procedures to improve classroom social skills in multiple general education settings. *Education and Treatment of Children*, *29*(1), 1–21.

Riley-Tillman, T. C., Chafouleas, S. M., Sassu, K. A., Chanese, J. A., & Glazer, A. D. (2008). Examining the agreement of direct behavior ratings and systematic direct observation data for on-task and disruptive behavior. *Journal of Positive Behavior Interventions*, *10*(2), 136–143.

Rutter, M. (1967). A children's behavior questionnaire for completion by teachers: Preliminary findings. *Journal of Child Psychology and Psychiatry*, *8*, 1–11.