1-20-2019

A Comparison of Sampling Approaches for Monitoring Schoolwide Inclusion Program Fidelity

Hank S. Bohanon
Loyola University Chicago, hbohano@luc.edu

Meng-Jia Wu
Loyola University Chicago, mwu2@luc.edu

Recommended Citation

Bohanon, Hank S. and Wu, Meng-Jia. A Comparison of Sampling Approaches for Monitoring Schoolwide Inclusion Program Fidelity. International Journal of Developmental Disabilities, 65, 2: 33, 2019. Retrieved from Loyola eCommons, Education: School of Education Faculty Publications and Other Works, http://dx.doi.org/10.1080/20473869.2018.1546793

A Comparison of Sampling Approaches for Monitoring Schoolwide Inclusion Program Fidelity

Keywords: Program Fidelity, Sampling, Schoolwide, Inclusion, Intervention, Efficiency; Disabilities

Abstract

Including students with disabilities requires schoolwide interventions that are implemented with fidelity (adherence). Collection of fidelity data may become problematic when multiple evidence-based treatments exist in one setting. To address concerns around efficiency of data collection, this study hypothesized that the three sampling approaches (proportional, consensus, stratified random) provided similar levels of agreement with an expert rater, thus warranting the differentiation in their applications based on the needs of the evaluators. Three high schools were randomly assigned to one of three sampling approaches (i.e., proportional, consensus, stratified random) to complete a fidelity measure for the implementation of a schoolwide inclusion approach. Based on the Median Test, each sampling approach did not differ significantly from the scores of an expert rater. The results indicate that schools may have some choice in the methods they use to sample their staff around the measurement of fidelity of implementation.

Introduction

The majority of students with disabilities in the U.S. do not have access to inclusive, general education environments (Kurth et al. 2018). There are 6.7 million children with disabilities in the United States (U.S.) or 13 percent of all students in public schools. Inclusion rates for students with high incidence disabilities, such as learning disabilities, in the U.S. appear to be on the rise. However, only 13 percent of students with multiple disabilities and 16 percent of students with intellectual disabilities spend the majority of the school day in general education settings (National Center for Educational Statistics 2018). Without support, students with disabilities are more likely to be excluded from general education classrooms due to issues such as exhibiting problem behavior (Bowman-Perrot et al. 2011).

Inclusion for students with disabilities in schools, including those with developmental and learning disabilities, requires more than a willing teacher (Choi *et al.* 2017). Successful inclusion involves schoolwide components such as universally provided interventions (Choi *et al.* 2017; Zins *et al.* 1990). Creating supportive schoolwide host environments can enhance the success of students with disabilities (Zins *et al.* 1990). Schoolwide efforts also have a positive impact on learning outcomes for students with disabilities (Choi *et al.* 2017). Students with developmental disabilities benefit from schoolwide supports when settings become more welcoming for social inclusion (Siperstein *et al.* 2017). Schoolwide supports have also provided scaffolding that has led to decreases in problem behavior for students with developmental delay (Loman *et al.* 2017).

The theory for this project is that inclusion requires implementing system-wide change that can be measured using a variety of methods. The inclusion of students with disabilities requires efforts to enhance the overall functioning of the school by improving the allocation of

resources, structures, and the systems that guide the work of the school (Sailor et al. 2017). These systems involve schoolwide efforts to improve the functioning of the school for all students, including students with disabilities. It is essential to measure the level of implementation of schoolwide efforts and compare outcomes for students with disabilities (Algozzine et al. 2016). Collection of schoolwide implementation data may become more difficult if there are multiple initiatives in place (Bohanon and Wu 2014). This study focuses on various methods for obtaining implementation data on schoolwide supports for students with disabilities.

Importance of Fidelity

When interventions are implemented with schools, they can be understood in terms of responsiveness of the participants (how much they engaged in the strategies), the reach of the intervention (how many people participated in the intervention from the target group), the quality of the intervention (how well the strategies were carried out), and the fidelity of the intervention (did implementation follow the manualized procedures) (Lendrum *et al.* 2016). For the purpose of this study, fidelity implementation is defined as the adherence to the factors that are associated with effectively carrying out an intervention based on its design (Glover *et al.* 2010) (Glover & Vaughn, 2010). Fidelity of implementation based on adherence is typically measured using scales, created by researchers, which determine the observance of the procedures and theoretical ideas identified in a model of practice (McHugo 2007).

Measuring the fidelity of implementation, or adherence to design, helps researchers to determine whether the manipulation of critical variables actually occurred, or if status quo was maintained (Herman *et al.* 2012). Tracking the factors that were manipulated during an intervention is often important in larger school settings such as high schools, as they typically

include more staff and students (Molloy *et al.* 2013). Measuring adherence to model implementation also can be useful for enabling district and state staff to determine the needs of their constituents for ongoing training and technical assistance (George *et al.* 2012). For example, fidelity data may indicate that school leadership teams involved with the initiative need more time set aside to meet, or additional training to work more efficiently. Monitoring fidelity data, along with student outcomes, thus becomes an integrative process (Sugai *et al.* 2010) that can lead to the continuous regeneration and sustainability of schoolwide approaches (McIntosh *et al.* 2009).

The ultimate goal for measuring fidelity is determining the relationship between the consistency of the model implemented and student outcomes. Previous research has found that schools with higher fidelity of universal school supports (i.e., teaching behavioral expectations to all students, acknowledging student behavior) are (a) more likely to implement more intensive interventions (e.g., functional behavioral assessments, behavior intervention plans) for students who are most at risk for failure (Eber *et al.* 2009), (b) increase participation by students with disabilities in the general education setting, and (c) reduce the number of students who are placed outside of the school setting (Illinois Positive Behavior Intervention and Support Network 2011).

Collection and integration of fidelity data may become problematic in several cases: when multiple evidence-based treatments exist, when the personal values of the staff impact treatment acceptability, and when treatment efficacy is dependent upon the collaboration of multiple participants (Fiks *et al.* 2010). A significant problem for staff charged with measuring fidelity of tiered approaches is that they are already overburdened with work, including activities that involve assessments (CASEL 2003). Staff that are overwhelmed by demands on their time for data collection may be less likely to complete fidelity assessments. Also, it can be

challenging for school teams to make decisions about implementation when multiple measures are reviewed at the same time (Mercer *et al.* 2017). More research comparing the feasibility and efficiency of different approaches to monitoring program fidelity in schools is needed (Nelson *et al.* 2015).

Sampling Approaches to Fidelity Measurement

To ensure effective collection of fidelity data, schools need reliable processes that enable data collection with minimal burden on staff members. The present study compares different sampling procedures for a schoolwide inclusion fidelity assessment. Each sampling approach required different amounts of time from staff members. Further, it compares the agreement between the multiple approaches that needed different levels of effort for completion. For this research, three sampling approaches were selected. These sampling procedures are: proportional sampling (e.g., surveying the entire staff with the hopes of obtaining an acceptable response rate); consensus (e.g., each school leadership team completes one assessment together); and a random sample of school staff.

The first approach, proportional sampling, involves collecting data from a portion of the entire sample of the setting. This method of sampling data collection involves asking the entire faculty and staff of the school to participate in the survey, with the hope that some reasonable proportion of the sample will respond. This method has been used with the collection of fidelity data related to SWPBS. This sampling approach has demonstrated high levels of agreement with expert driven tools for measuring fidelity (Hagan-Burke 2005; Safran 2006). An advantage of proportional sampling is the collection of perspectives from a larger proportion of stakeholders. By obtaining feedback from a larger body of the participants, this sampling approach could increase the level of buy-in for the intervention if this assessment is part of data collection in

early stages of implementation. Through this sampling approach, additional staff have a "voice" in providing feedback about the intervention (Bohanon and Wu 2014). A limitation of this technique is related to its strength, relies on a larger portion of staff for input.

With regard to self-assessment, some research indicates that staff have lower levels of agreement with expert raters, and that staff typically have a positive bias in their ratings (Hansen *et al.* 2014). Although, there is research that supports that teachers can accurately report their own levels of fidelity of implementation (Sanetti 2009). Self-assessment data can be useful in supporting the implementation of professional development provided to teachers (Desimone 2009; Haslam 2010). These kinds of survey data can help coaches provide performance feedback to school teams. Specifically, these data can be used to celebrate accomplishments and identify areas for improvement for implementation of schoolwide efforts (Castillo *et al.* 2017).

A second sampling approached used with measuring fidelity of implementation involves consensus from a group of people who represent the larger body of staff. This process involves having a team collectively complete a measure of fidelity through self-report. One score for each item is entered on measure based on a discussion of the group. This group can involve 6-10 people in the process. Consensus sampling has been used as an effective process for measuring the fidelity of schoolwide processes (Childs *et al.* 2010; Vincent *et al.* 2010). An advantage for obtaining consensus ratings from your school team is that it limits the demand on the number of people who must complete the tool, and it relies on people who could be potentially most knowledgeable about the approach. A limitation of consensus sampling, like proportional sampling, is that has the same concerns related to the self-report from staff (Hansen *et al.* 2014). Additionally, if the same leadership team of the school is implementing multiple schoolwide initiatives (e.g., social emotional learning, response to intervention), then they could be required

to complete multiple fidelity self-assessments. Asking the same team to complete multiple assessments related to fidelity of implementation could create additional burden on them (CASEL 2003).

A third approach includes stratified random sampling. With this method, the evaluator looks for a small group of individuals that represent the overall target population (Laxton *et al.* 2005; Sprague *et al.* 1996). Stratified random sampling is potentially the most efficient process of the three approaches in that it requires fewer people to complete the instrument. Therefore, it decreases the burden of data collection for the overall staff (CASEL 2003). This approach shares the same limitation as the other two around the self-report (Hansen *et al.* 2014). Additionally, since fewer people are involved in the process, there is less participation from the overall staff. If an evaluator is attempting to obtain buy-in from the school by including as many voices as possible in the data collection process, the stratified random sampling approach would perhaps be the least effective.

Purpose

The primary goal of this project was to determine if differential methods for self-assessment assessing the level of adherence to a schoolwide inclusion intervention would be similar in agreement to that of an expert rater (Hansen *et al.* 2014). If so, then perhaps schoolwide projects could use a variety of sampling techniques depending on the purpose of their assessment. Specifically, the goals of the paper were to describe the differences in agreement between expert rater and responders across three types of sampling approaches, and compare the differences between the expert rater and responders among sampling approaches.

This study hypothesizes that the three sampling approaches (proportional, consensus, stratified random) provide similar levels of agreement with an expert rater, thus warranting the

differentiation in their applications based on the needs (e.g., all-staff input for buy-in, checking for adherence to the intervention across staff roles) of the evaluators and technical assistance providers. This study addresses two specific hypothesizes:

1. H1: There are significant differences in agreement between an expert and staff ratings of fidelity measures for a schoolwide inclusion approach for each of the three data sampling approaches (i.e., proportional sampling, consensus, and random sampling).
2. H2: With regard to the sampling approaches for collecting fidelity data on a schoolwide inclusion approach, = there are differences in the levels of agreement among the expert rater and sample groups.

Methods

Setting

This study took place in a large Midwestern state. The authors were the evaluators for a statewide initiative that helped schools include students with disabilities, including students with intellectual disabilities. The project was supported in part by a State Personnel Development (SPD) grant. This funding was designed to improve and reform the state's approach to personnel preparation and professional development of individuals providing educational services, transition services, and early intervention in order to improve results for children with disabilities. Based on this funding, the state had developed a training, coaching, and technical assistance approach for supporting inclusion of students with disabilities. The study involved three high schools within one urban district served by this project.

The goal of the statewide project was to support evidenced-based instruction for all students, including students with intellectual disabilities. Project staff coached districts and schools to develop a clear vision for inclusion, learning opportunities for all staff, and the

development of a collaborative a culture. The project was based on a schoolwide tiered model of support. Tiered supports provide effective schoolwide interventions to all students (Taylor *et al.* 2017). They also provide additional supports to groups of students who are not successful with schoolwide interventions, as well as additive intensive supports to those who have not responded to previous interventions (Kincaid *et al.* 2016; Goodman and Bohanon 2018). There were also multiple initiatives that were simultaneously being implemented in schools across the state including response to intervention (RtI), schoolwide positive behavior support (SWPBS; Kincaid *et al.* 2016), social and emotional learning (SEL; Taylor *et al.* 2017), and school-based mental health (Bohanon and Wu 2011; Stephan *et al.* 2015). Each of these approaches had their own specific requirements for data collection to be a part of their respective projects.

Participants

Purposive convenience sampling was used to identify the specific schools for this study (Patton 2015). Participating schools needed to be involved in a statewide technical assistance project to support the inclusion of students with disabilities in the general education environment. The schools were also balanced as much as possible for similarity across demographic categories (i.e., comparison-focused sampling). High schools were chosen for this study because of their large staff size and the potential for maximum variation in the perspectives of the staff from each school (Patton 2015). The lead author was the external evaluator for this technical assistance project. The state-level technical assistance team identified the district and schools. The three schools that were selected were part of one large urban district in the Midwest. The average percent of students with disabilities for all three schools was 13.67% (*min* = 12%, *max* = 15%) Table 1 provides a list of other key demographics for each school.

<Insert Table 1 here>

Measurement

The Inclusive Practices Reflection Tool (IPRT) (http://www.hankbohanon.net) was used as the fidelity measure for this study. The format of the IPRT is based on an existing reliable and valid self-assessment for the implementation of SWPBS (Effective Behavior Support Self-Assessment Survey; Safran 2006). The IPRT included 36 items across two types of response prompts. Items in the IPRT address best practices and provided task analysis of effective programming for including students with disabilities. The instrument assesses schoolwide components for inclusion related to resources, structures, practices, use of data, and systems (Sailor et al. 2017). The IPRT is divided into four categories: (a) school-wide settings, (b) classroom settings, (c) non-classroom settings (e.g., hallways, cafeteria), and individual student settings (e.g., individualized supports). Key components of this tool include teachers' use of flexible instructional strategies across content areas (Grossman *et al*. 2014). Previous studies have identified connections between classroom environment factors and value-added scores on standardized assessments. Additional items of the tool included the efficient use of organized classroom routines (e.g., time management, behavior management) which may be responsible for decreasing lost time for instruction by reducing problem behavior (Grossman *et al*. 2014).

Participants could assess their current practices and rate items according to four categories: (a)"*in place*", (b) "*partially in place*", (c) "*not in place*", or (d) "*I don't know*." Table 2 provides the specific items that were associated with each subscale. The responder could also indicate the level of priority for improving each item on the list. Further, each item on the IPRT was aligned with six target areas, based on the state's performance plan for improving outcomes for students with disabilities. These area for improvement included

(1) All students are members of the general education community.

(2) All students have access to and are supported by all of the school's resources.

(3) The school addresses social and emotional skills development as a key part of academic outcomes.

(4) The school develops and uses data systems for decision making and problem solving.

(5) The school makes concentrated and ongoing efforts to involve family members and other community members.

(6) The school/district is committed to system change and developing an effective education system for all.

Analysis of the IPRT's reliability was conducted for this study. Items two and three on the IPRT were related to early childhood settings. These items were not included in the present survey, since the study's sample only included high schools. Thirty-four of the IPRT items were first reviewed using an inter-item correlation matrix. Items 7 and 21 had multiple negative correlations with other items on the IPRT. The corrected item total correlations for items 7 and 21 were .069 and .102 respectively. Field (2013) has recommend removing questions with item correlations of less than $r = .3$. Therefore, items 7 and 21 were removed in this analysis. The reliability indices of Cronbach's Alpha for the 32 items included in the final analysis were .875 (schoolwide), .802 (class-wide), .523 (non-classroom), and .415 (individual support). Non-classroom and individual support factors were not used for analysis due to their lower Alpha levels.

As a part of the evaluation of this state project, an analysis of the validity of the IPRT was conducted regarding the level of inclusion of students with disabilities in the general education classroom for the majority of their school day. During the first year of analysis, schools (n = 8) in the top 25% category of for overall IPRT mean implementation score had a higher percentage of students in included in the general education environment for 80% of the day or more (65.5%) than schools in the bottom 25% category (n = 8, 32.5%). Schools in the bottom 25% group had a greater percentage of students who were removed from the general

education setting for 20% of the day or more than schools in the top 25% implementation group (Project ___, 2008).

On average, 51.3% of the students across the state were included in the general education classroom for the majority of the school day. It was noted that in schools in the top 25% of overall implementation of the project based on the IPRT, 65% of the students were placed in the general education classroom for the majority of the school day. Alternatively, in the lower 25% of schools implementing the project based on the IPRT, forty-one percent of the students were included in general education classroom for a majority of the school day. As a comparison, the state reported that 51.3% of the students to be included in the general education setting for 80% or more of the school day (Henry 2010).

Data Collection Procedures

As previously stated, researchers were the external evaluators for this project. The lead researcher had written consent from the state board of education to use the extant data from the project for analysis and publication. The data from this project were intended for institutional quality assessment, and quality improvement and involved surveys that did not address individual participant behaviors. The data used for this analysis did not include identifiable data for individual participants. The Institutional Review Board for the lead author's institution confirmed that the analysis did not require human subjects approval, as the primary intent of the data was for evaluation purposes.

Three separate data collection procedures were used to collect IPRT data from the schools. Sites were randomly assigned to the following data collection conditions: proportional sampling, consensus, and stratified random sampling. Sampling procedure one was based on a proportional sample of the entire school. The school's leadership team prompted all staff to

complete the IPRT online with the goal of a response rate of at least 40%. A preliminary synthesis of online surveys indicates that the average response rate is 35.1%, 95% CI [34.9, 35.3] in educational settings (Wu *et al.* 2012).

Sample procedure two used consensus ratings of the school's site-based leadership team. This procedure involved the members of the leadership team for the school (6-10 people) completing the IPRT together and submitting one response per item based on the group's consensus. Leadership teams for this state project typically consisted of general and special education teachers, an administrative representative, and curriculum specialists from within the school. The external rater was not a part of the consensus process for the leadership team.

Sampling procedure three included a stratified random sample of individual school staff across four different areas: general education, special education, administration, and support staff (Laxton *et al.* 2005; Sprague *et al.*1996). For the current study, the external consultant (who also served as the expert rater) and the administration of the selected school randomly chose participants to complete the IPRT based on assigned numbers for staff in each category (i.e., general education, special education, support staff, administration). The total sample included 69 participants for the proportional sample, seven participants for the random sample, one consensus response, and one expert rating for each sampling approach (total N = 80).

A separate IPRT was completed by an external (i.e., expert) coach for each school. Schoolwide interventions rely on external coaches to provide feedback to schools to support schools, inclusive of measuring fidelity of implementation (Goodman and Bohanon 2018). This individual was employed by the state technical assistance agency, was familiar with the school, was trained on the use of the IPRT, and followed a written protocol. The IPRT was completed at approximately the same time as the school staff's assessments. The expert's responses were

grounded in attendance in team meetings and in other school-based data. Data were collected in the late spring during a single school year, and at the same time for each school.

Data Analysis

Given the small sample size of the study, non-parametric tests and descriptive statistics were used address the research hypothesizes. Statistical power was not addressed due to the fact that non-parametric tests do not rely on the assumption of normal distribution. The analysis for H1 involved the use of the Median Test (Siegel *et al.*1988). Each sampling group's ratings were compared with the ratings from the expert on both individual items and for the overall score. The purpose of this analysis was to determine if there were significant differences in the median ratings between the expert and the sampling approaches. Non-parametric statistics of this type were appropriate for this study due to the overall small sample size and the concern that the limit of three schools would not represent a normal distribution (Siegel *et al.* 1988). H2 was analyzed using descriptive statistics to compare the level of agreement across sampling approaches for each item for the IPRT.

Results

H1 asked if there were significant differences in agreement between expert raters and staff when using the IPRT for each of three types of data collection approaches. The average scores on the IPRT were: proportional sample, n =69, $\bar{x}$ = 2.85, SD = .56; proportional sample expert n = 1, $\bar{x}$ = 3.09; random sample n =7, $\bar{x}$ = 3.34, SD = .36; random sample expert rater n = 1, $\bar{x}$ = 3.00; consensus sample (consensus score of 6 people) n = 1, $\bar{x}$ = 2.97; and consensus expert n = 1, $\bar{x}$ = 3.00. Overall scores and item-level comparisons were conducted for the expert and the sample at the item level for both the proportional and random sample groups. Due to the low sample size, the Median Test was conducted using only the total rating for the consensus

group. Non-parametric statistics are appropriate to analyze data with small sample sizes (Siegel *et al.* 1988). None of the differences between the expert and the samples were significantly different for any of the sampling approaches (proportional sample $p < .109$, random sample $p < .063$, consensus sample $p < .317$). While the Median Test does not provide the ability to measure the power of the significance, it does allow researchers to determine if there are any underlying differences in the distribution of the data (Siegel *et al.* 1988).

H2 asked if there were differences in the levels of agreement between the expert and the sample among the three different sampling approaches. The proportional and random samples involved a comparison between an expert evaluation and those of multiple raters. Therefore, the percent agreement between the samples and the expert rater was reported as a function of the percentage of items that were not significantly statistically different between the expert rater and the sample. Agreement for the consensus group level was analyzed using one score from the census group and the score of the rater for each item. Level of agreement was analyzed by dividing the number of agreements by the total possible number of agreements. Overall agreement for the IPRT was 50% for proportional, 47% for random, and 63% for consensus sampling approaches. For schoolwide items, agreement for the IPRT was 43% for proportional, 52% for random, and 48% for consensus. For class-wide items, agreement for the IPRT was 71% for proportional, 14% for random, and 100% for consensus. Agreement across all three sampling approaches was evident for items 14, 22, 23, and 34. No agreement between the expert and the sample across all approaches occurred for items 4, 18, and 19. Table 2 provides data about which item ratings were not significantly different between the expert and the sample.

<Insert Table 2 here>

Table 3 provides information about the representation of the participants based on their professional roles. The random sampling approach appeared to provide an advantage for balancing representation of the entire staff over both the consensus and the proportional sampling techniques. For example, the random sampling approach was the only assessment that included data from at least one member of each type of staff (e.g., general education, special education, support personnel, administration). Additionally, the consensus ratings cannot be considered as independent because each responder could have impacted the ratings of the others in the group. This would be particularly of concern if the presence of an administrator influenced the overall response of the group. Further, the proportional sampling approach did not include support personnel or school administration. Finally, proportional sampling appeared to provide an over-representation of general education teachers compared to the other sampling methods.

<Insert Table 3 here>

Discussion

The theory of this paper is that including students with disabilities requires implementing system-wide efforts that can be measured using a variety of sampling methods. Systems-wide efforts need whole-school initiatives that can improve the overall function of the school for all students (Sailor et al. 2017). Including students with disabilities requires schoolwide efforts on the part of teachers, administrators, and parents (Choi et al. 2017). A component of effective schoolwide efforts involves access to data around the fidelity of adherence to the critical factors of the model (Choi et al. 2017; Zins et al. 1990). Measuring fidelity of implementation can allow evaluators and local teams to determine if processes are in place that create more effective host environments for students with intellectual disabilities (Siperstein et al. 2017).

The goals of the paper were to describe the differences in agreement between an expert rater and responders across three types of sampling approaches in the collection of fidelity of

implementation data, and compare the differences between the expert rater and responders among sampling approaches. If differential sampling procedures could provide results with agreement levels that were not significantly different, evaluators might have more sampling options from which to choose. As a result, employing differentiated approaches for gaining needed information could potentially decrease the burden on staff and increase their willingess to participate in the surveys (CASEL 2003).

H1

Each sampling approach in this study did not appear to differ significantly from the ratings of the expert. Although further study is warranted using a larger sample size, it is interesting to note that the lack of significant difference between expert and the samples might indicate that each sampling approach could provide useful data on the treatment adherence of an intervention (Hansen *et al.* 2014). In other words, using any of the three approaches might allow a team to determine if factors related to the design of their interventions were carried out as intended (Glover *et al.* 2010).

In this study, each sampling approach allowed for no more disagreement between the expert rater and sample than would be expected by chance alone. This finding could mean that each sampling approach may provide valuable feedback that school teams could use for decision-making. Having access to critical fidelity information provides district and state staff with the relevant information needed to identify the technical assistance and training needs of their constituents (George *et al.* 2012). While further research and replication is needed to verify these results, the selection of a sampling approach might be guided by the district or staff's intended use of the data (e.g., seeking input from multiple constituents, perceptions of the leadership team).

H2

Each of the sampling approaches used in this study appeared to allow researchers and evaluators to identify the level of treatment adherence for the schoolwide approach within the settings (Molloy *et al.* 2013). Having multiple methods for sampling staff could be useful to districts and school staff who use fidelity data to determine staff needs related to training and support (Molloy et al., 2013). While more research is needed, flexibility in sampling options could provide districts and schools with multiple options for gathering data in order to determine how to adjust their training and technical assistance support. These options could in turn allow for increased access to data related to systematic supports that lead to high and sustained fidelity of implementation over time (Sugai *et al.* 2010).

Completion of the IPRT requires approximately 25 minutes to complete. For the consensus sample, 6 participants completed this tool. While the total amount of time to complete the tool was 25 minutes, this number should be multiplied by the total cost of time for all participants (6 x 25). This means that completion of the IPRT via consensus cost 150 total minutes, or 2.5 hours of combined staff time. The random sample, which included 7 participants, required a total of 175 minutes (7 x 25), or 2.9 hours of combined staff time. In contrast to these shorter amounts of time, the proportional sample, which included 70 participants, required 1,750 minutes (20 x 75), or approximately 29 hours of combined staff time. Based on a total amount of time required to complete the instrument, the consensus and random sampling procedures hold an advantage of requiring less combined time from the overall staff. Again, consideration of the purpose of data collection should be made here. If the evaluator wanted to include the perspectives of as many of the entire staff as possible as a part of the stages for developing buy-in (Bohanon and Wu 2014), the proportional sample might pose a reasonable request of staff

time. However, if the purpose of the data collection procedure is to obtain fidelity data from a representative body, then consensus and random sampling could provide more efficient alternatives.

Having options and not always relying the same group or groups of people for data could be useful for school sites where multiple evidence-based approaches are employed (Fiks *et al.* 2010). For example, a district might select a sampling approach because it is the least disruptive to the staff as a whole (e.g., consensus) or requires the least attention from multiple staff members (e.g., random sample). Likewise, teams could consider using a stratified random sampling approach (Laxton *et al.* 2005; Sprague *et al.* 1996) if they wanted to ensure input from staff with specific content knowledge. Having choices in sampling approaches might provide additional opportunities for schools to collect fidelity of implementation data for problem solving and improving their schoolwide efforts (Castillo *et al.* 2017).

It is notable that some items were rated with complete agreement between the expert rater and the participants in the related sampling group, while other items did not yield agreement. Items that showed agreement across approaches appear to be related to questions of climate for the staff and students (i.e., item 14), as well as those related to observable events or procedures that occurred in classroom and non-classroom settings (i.e., items 22, 23, and 34). It could be that assessors have more flexibility in sampling procedures when it comes to fidelity items that are based on school climate and factors that are observable to the entire staff. While there are concerns with self-report related to positive bias of the responder (Hansen *et al.* 2014), it appeared in this case that the participants could self-report their fidelity of implementation in some instances at similar levels to the expert rater. Therefore, it is possible that schools could

have options for collecting data on schoolwide climate and classroom instruction efforts to improve the implementation of professional development (Desimone 2009; Haslam 2010).

The items that did not prompt agreement across sampling approaches were schoolwide, and dealt with policies for supporting students with disabilities (i.e., items 4, 18, and 19). These items also addressed either transition supports or inclusion. It would appear that none of the sampling approaches led to agreement about polices related to specifically supporting students with special needs. Perhaps other data collection methods that identified permanent products (e.g., lesson plans, professional development calendars, policies) would be more effective for assessing these types of questions.

Limitations and Implications

This study has several limitations. First, it was carried out in a large urban district. There are thus questions of generalizability to rural or suburban settings. Second, the study was carried out in three large high schools. Future studies should address other grade levels in order to identify differences in the agreement for sampling approaches across primary and middle schools. Additionally, no inter-rater reliability data were reported between the expert rater and a second expert rater. However, the same rater was used in each setting with the intention that any observer bias would be consistent among schools. Further, the expert rater was trained by state-based technical assistant providers using a standardized protocol. The expert was provided with training through onsite coaching until his reliability of assessment was equivalent to that of lead trainers for the statewide initiative. Additionally, given the small sample size for at least two of the participant groups, it would be impossible within this study to conduct parametric analysis that would lead to a discussion of effect size. However, the nonparametric Median Test used for this study did not rely on the assumption of heterogeneity of the data, and therefore could be

used with small sample sizes to avoid violations of the assumption of normality of the distribution of the data (Siegel *et al.* 1988). Future studies should involve replications with larger sample sizes to allow for the use of parametric statics and the determination of effect sizes.

It is possible that the lack of significant differences among the expert and the samples across approaches was due to the homogeneity of the schools within the district. However, the assessor was trained and supported by the state-level initiative and was not an employee of the district. The rater's independence likely addressed, at least in part, bias he would have had if he was an employee of the district. Due to these limitations, the results of this study may not generalize to all settings. However, the results might transfer to other settings based on similarities in demographics.

There did not appear to be any significant differences in agreement based on the median scores between the expert and any of the fidelity assessment sampling approaches in this study. Each sampling method performed approximately as well as the expert rater. One possible recommendation might be to rely upon an expert rater with enough depth and training to be able to assess schools' implementation of approaches. However, there are times when the perspectives of those carrying out the intervention are needed. For example, the proportional sample may be useful when attempting to obtain initial buy-in by seeking feedback on the current status of a practice within a school. Perhaps, if a team was particularly interested in the implementation of practices at the classroom level, the proportional sampling might be useful for obtaining initial buy-in and baseline fidelity data.

Conclusion

Although the results of this study are promising, there is still limited information related to which types of sampling approaches are most appropriate when multiple schoolwide

approaches are used. Few studies have compared sampling techniques that assess the fidelity of schoolwide approaches to help include students with disabilities. Future studies also should consider the perspectives of students, families, and community members in the sampling process. Additional research is needed to help districts, staff, and researchers determine the most effective approaches for balancing efficacy and efficiency when it comes to sampling approaches and tracking fidelity of implementation. This paper perhaps provides one small step towards the call for more research comparing the feasibility and efficiency of different approaches to monitoring program fidelity in schools (Nelson *et al.* 2015).

References

Algozzine, B., Morsbach Sweeney, H., Choi, J. H., Horner, R., Sailor, W., McCart, A. B., and Lane, K. L. 2016. Development and preliminary technical adequacy of the schoolwide integrated framework for transformation fidelity of implementation tool. Journal of Psychoeducational Assessment, 35, 302-322.

Bohanon, H. & Wu, M. 2011. Can prevention programs work together? An example of school-based mental health with prevention initiatives. School-Based Mental Health Practice 4 (4), 35-46. [LINK TO ARTICLE:http://ecommons.luc.edu/education_facpubs/1/]

Bohanon, H. & Wu, M. 2014. Developing buy-in for positive behavior support in secondary settings. Preventing School Failure, 58 (4), 1–7. doi: 10.1080/1045988X.2013.798774 http://ecommons.luc.edu/education_facpubs/17/

Bowman-Perrot, L., Benz, M., Hsu, H., Kwok, O., Eisterhold, L. A., & Zhang, D. 2011. Patterns and predictors of disciplinary exclusion over time: An analysis of the SEELS national data set. Journal of Emotional and Behavioral Disorders, 21, 83-96.

Castillo, J. M., Wang, J. H., Daye, J. G., Shum, K. Z., and March, A. L. 2017. A longitudinal analysis of the relations among professional development, educators' beliefs and perceived skills, and response-to-intervention implementation. Journal of Educational and Psychological Consultation, 28, 1-32.

Center for Academic, Social, and Emotional Learning (CASEL). 2003. Safe and sound: An Educational leader's guide to evidence-based social and emotional learning (SEL) programs, 60. Chicago, IL: Collaborative for Academic, Social, and Emotional Learning.

Childs, K. E., Kincaid, D., and Peshak-George, H. 2010. A model for statewide evaluation of a universal positive behavior support initiative. Journal of Positive Behavior Interventions, 12, 198-210.

Choi, J. H., Meisenheimer, J. M., McCart, A. B., and Sailor, W. 2017. Improving learning for all students through equity-based inclusive reform practices: effectiveness of a fully integrated schoolwide model on student reading and math achievement. Remedial and Special Education, 38, 28-41.

Desimone, L. M. 2009. Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. Educational Researcher, 38, 181–199.

Eber, L., Phillips, D., Upreti, G., Hyde, K., Lewandowski, H., and Rose, J. 2009. Illinois positive behavioral interventions & supports (PBIS) network 2008-09 progress report (pp. 245). La Grange Park, IL.

Field, A. 2013. *Discovering statistics using IBM SPSS statistics*: Thousand Oaks, CA: Sage.

Fiks, A. G., and Leslie, L. K. 2010. Partnership in the treatment of childhood mental health problems: A Pediatric perspective. School Mental Health, 2, 93-101.

George, H. P., and Childs, K. E. 2012. Evaluating implementation of schoolwide behavior support: Are we doing it well? *Preventing School Failure: Alternative Education for Children and Youth*, 56, 197-206.

Glover, T. A., and Vaughn, S. 2010. *The Promise of Response to Intervention: Evaluating Current Science and Practice*. New York, NY: Guilford Press.

Goodman, S. & Bohanon, H. 2018. A Framework for Supporting All Students: One-Size-Fits-All No Longer Works in Schools. American School Board Journal, February, 1-4, 2018. https://ecommons.luc.edu/education_facpubs/116/

Grossman, P., Cohen, J., Ronfeldt, M., and Brown, L. 2014. The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, 43, 293-303.

Hagan-Burke, S., Burke, M. D., Martin, E., Boon, R. T., Fore III, C., and Kirkendoll, D. 2005. The internal consistency of the school-wide subscales of the Effective Behavioral Support Survey. *Education & Treatment of Children*, 28, 400-413.

Hansen, W., Pankratz, M., and Bishop, D. 2014. Differences in observers' and teachers' fidelity assessments. *Journal of Primary Prevention*, 35, 297-308.

Haslam, M. B. 2010. *Teacher professional development evaluation guide*. Oxford, OH: NSDC & Maryland State Department of Education.

Henry, P. 2010. Project CHOICES end of the year report. Illinois State Board of Education.

Herman, K. C., Riley-Tillman, T. C., and Reinke, W. M. 2012. The role of assessment in a prevention science framework. *School Psychology Review*, 41, 306–314.

Illinois Positive Behavior Intervention and Support Network. 2011. *FY11 end of the year summary report*. La Grange, IL: Illinois Positive Behavior Intervention Network.

Kincaid, D., Dunlap, G., Kern, L., Lane, K. L., Bambara, L. M., Brown, F., and Knoster, T. P. 2016. Positive behavior support: A proposal for updating and refining the definition. *Journal of Positive Behavior Interventions*, 18, 69.

Kurth, J.A., Morningstar, M.E., Hicks, T.A. and Templin, J., 2018. Exploring the Relationship Between School Transformation and Inclusion: A Bayesian Multilevel Longitudinal Analysis. *Inclusion*, *6*(1), pp.19-32.

Laxton, T. C., and Sprague, J. 2005. Refining the construct of school safety: An exploration of correlates and construct validity of school safety measures. PhD. University of Oregon.

Lendrum, A., Humphrey, N., and Greenberg, M. T. 2016. Implementing for success in school-based mental health promotion: The role of quality in resolving the tension between fidelity and adaptation. In R. Shute & P. Slee (Eds.), Mental health and wellbeing through schools: The way forward (pp. 53–63). London: Taylor and Francis.

Loman, S. L., Strickland-Cohen, M. K., and Walker, V. L. 2017. Promoting the accessibility of SWPBIS for students with severe disabilities. *Journal of Positive Behavior Interventions.*

McHugo, G. J. 2007. Fidelity outcomes in the national implementing evidence-based practices project. *Psychiatric Services*, 58, 1279-1284.

McIntosh, K., Filter, K. J., Bennett, J. L., Ryan, C., and Sugai, G. 2009. Principles of sustainable prevention: Designing scale-up of school-wide positive behavior support to promote durable systems. *Psychology in the schools*, 47. 5-21.

Mercer, S. H., McIntosh, K., & Hoselton, R. 2017. Comparability of fidelity measures for assessing Tier 1 school-wide positive behavioral interventions and supports. The Journal of Positive Behavior Interventions., 19(4), 195-204.

Molloy, L. E., Moore, J. E., Trail, J., Van Epps, J. J., and Hopfer, S. 2013. Understanding real-world implementation quality and "active ingredients" of PBIS. *Prevention Science*, 14, 593–605.

National Center for Educational Statistics. 2018. The condition of education: Children and youth with disabilities. Retrieved from https://nces.ed.gov/programs/coe/indicator_cgg.asp

Nelson, J. R., Oliver, R. M., Hebert, M. A., and Bohaty, J. 2015. Use of self-monitoring to maintain program fidelity of multi-tiered interventions. *Remedial and Special Education*, 36, 14-19.

Project CHOICES. 2008. Annual performance report. Project CHOICES, Illinois State Board of Education.

Patton, M. 2015. *Qualitative evaluation methods (4th ed.).* Beverly Hills, CA: Sage.

Safran, S. P. 2006. Using the effective behavior supports survey to guide development of schoolwide positive behavior support. *Journal of Positive Behavior Interventions*, 8, 3-9.

Sailor, W., Satter, A., Woods, K., McLeskey, J., and Waldron, N. 2017. School improvement through inclusive education. *Oxford Bibliographies in Education*. New York: Oxford University Press.

Sanetti, L. M. H. 2009. Treatment integrity assessment in the schools: An evaluation of the Treatment Integrity Planning Protocol. *School psychology quarterly*, 24, 24-35.

Siegel, S., and Castellan, N. J. 1988. *Nonparametric statistics for the behavioral sciences.* New York, NY: McGraw Hill, Inc.

Siperstein, G. N., Summerill, L. A., Jacobs, H. E., and Stokes, J. E. 2017. Promoting social inclusion in high schools using a schoolwide approach. *Inclusion*, 5, 173-188.

Sprague, J., Colvin, G., and Irvin, L. 1996. *The Oregon School Safety Survey*. Eugene, OR: University of Oregon.

Stephan, S., Sugai, G., Lever, N., and Connors, E. 2015. Strategies for integrating mental health into schools via a multi-tiered system of support. *Child and Adolescent Psychiatric Clinics of North America*, 24, 211-231.

Sugai, G., Horner, R., Fixsen, D., and Blase, K. 2010. Developing systems-level capacity for RtI implementation: Current efforts and future directions. In T. A. Glover & S. Vaughn (Eds.), The promise of response to intervention: Evaluating current science and practice (pp. 286-309). New York, NY: Guilford Press.

Taylor, R. D., Oberle, E., Durlak, J. A., and Weissberg, R. P. 2017. Promoting positive youth Development Through School-Based Social and Emotional Learning Interventions: A Meta Analysis of Follow-Up Effects. *Child Development*, 88, 1156-1171.

Vincent, C., Spaulding, S., and Tobin, T. J. 2010. A Reexamination of the Psychometric Properties of the School-Wide Evaluation Tool (SET). *Journal of positive behavior interventions*, 12, 161-179.

**Wu, M., Young, J. Zhao, L. and Clark, K. 2012. Examining factors that impact the response rate of electronic surveys: An updated meta-analysis. Paper presented at the Annual meeting of American Educational Research Association, Vancouver, BC.

Zins, J. E., and Ponti, C. R.1990. Strategies to Facilitate the Implementation, Organization, and Operation of System-Wide Consultation Programs. *Journal of Educational & Psychological Consultation*, 1, 205.

Table 1

Demographic Information for Participating High Schools

Demographic	% Sample	Random Sample	Consensus
Classroom Teacher FTE	82	101	88
# of Students	1,848	1,832	1,888
School Type	Urban	Urban	Urban
% Caucasian	40	40	43
% African-American	23	28	27
% Hispanic	22	25	23
% Asian	9	3.2	4
% Native American	.1	.3	.2
% Multiracial	4	4	2
% LEP	8	9	1.4
% Low SES	74	69	68
% Attendance Rate	90	80	87
% Mobility Rate	15	10	24
% Chronic Truancy Rate	6	12	20
% Drop Out Rate	10	6	8
% Graduation	57	71	66

Table 2

Significant Differences Between Expert and Sample Groups by Item for IPRT

Item	Sampling		
Schoolwide Settings	P	R	C
1. Academic goals are aligned with State Learning Standards.		X	
4. Partnerships exist to foster transitions.			
5. Decisions regarding graduation are made prior to freshman year.		X	
6. IEPs contain intervention techniques and/or accommodations.			X
8. Natural proportions are maintained.		X	
9. Specialized personnel support all students in the general education.	X		
10. Students with IEPs attend home school.	X	X	
11. Teaming/co-planning meetings use data and are scheduled regularly.			X
12. Professional development occurs for developing an inclusive school.		X	
13. School has a vision/philosophy statement that all students can learn.		X	
14. School encourages interdependence among students and staff.	X	X	X
15. School/District Administrative Leadership support inclusive practices.	X		X
16. School/District Leadership is grounded in data and research.		X	X
17. School-wide problem-solving teams have regularly scheduled meetings.	X		
18. School-wide problem-solving teams use data for ongoing decisions.			
19. Partnerships exist with parents/guardians/family.			
20. There is a venue for informing families/community.			X
21. Parents are invited to, and attend, staff trainings.			
22. There are workshops for parents/guardians/family members.	X	X	X

23. Parents/guardians/family members visit classrooms.	X	X	X
24. Material and equipment support are available to all students.		X	X
25. Faculty/staff evaluate and act on school-wide data as part of meetings.	X		X
Classroom Settings	P	R	C
26. Accommodations are linked to the state Learning Standards.	X		X
27. School personnel use positive behavior strategies.			X
28. Modifications and accommodations are implemented systematically.	X		X
29. Special education teachers support students with IEPs in gen. ed.		X	X
29. General/special education teachers share instructional responsibility.			X
30. All teachers share responsibility to support the learning of all students.	X		X
31. Staff implement research-based strategies for instruction.	X		X
32. Classroom teachers collect and use curriculum-based measures.	X		X
Non-Classroom Settings	P	R	C
33. Students with IEPs participate in typical school activities.		X	X
34. All staff take responsibility for supervision in non-classroom settings.	X	X	X
Individual Student Settings	P	R	C
35. All school staff plan collaboratively to support students.		X	X
36. The IEP contains intervention techniques and/or accommodations.	X		

Notes. The symbol X indicates that there was not a significant difference between the expert rater and the sample on the item. The letter P stands for proportional (40%), R stands for random, and C stands for consensus sampling approaches. The item stems were shortened due to length. A full version of the items can be found at http://www.hankbohanon.net.

Table 3

Representation of Staff

	Total	General Educators		Special Educators		Support/Others		Administrators	
Type of Sampling	N	*n*	%	*n*	%	*n*	%	*n*	%
Consensus	6	2	33	2	33	0	0	2	33
Random	7	2	29	1	14	2	29	2	29
Forty %	70	51	73	14	20	5	7	0	0