Measurement Invariance of an Instrument Assessing Sustainability of School-based Universal Behavior Practices

Sterett H. Mercer

University of British Columbia

Kent McIntosh

University of Oregon

M. Kathleen Strickland-Cohen

Texas Tech University

Robert H. Horner

University of Oregon

MANUSCRIPT IN PRINT: Mercer, S. H., McIntosh, K., Strickland-Cohen, M. K., & Horner, R. H. (2014). Measurement invariance of an instrument assessing sustainability of school-based universal behavior practices. *School Psychology Quarterly, 29*, 125-137.

Author Note

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120278 to University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Sterett H. Mercer, Department of Educational and Counselling Psychology and Special Education, University of British Columbia, 2125 Main Mall, Vancouver, BC V6Z 1Z4, Canada. Email: sterett.mercer@ubc.ca

Abstract

Objective: The purpose of the study was to examine the extent to which the School-wide Universal Behavior Sustainability Index: School Teams (SUBSIST; McIntosh, Doolittle, Vincent, Horner, & Ervin, 2009), a measure of school and district contextual factors that promote the sustainability of school practices, demonstrated measurement invariance across groups of schools that differed in length of time implementing School-wide Positive Behavioral Interventions and Supports (PBIS; Sugai & Horner, 2009), student ethnic composition, and student socio-economic status (SES).

Method: School PBIS team members and district coaches representing 860 schools in 14 U.S. states completed the SUBSIST.

Results: Findings supported strong measurement invariance, for all items except one, of a model with two school-level factors (School Priority and Team Use of Data) and two district-level factors (District Priority and Capacity Building) across groups of schools at initial implementation, institutionalization, and sustainability phases of PBIS implementation. Schools in the sustainability phase were rated significantly higher on School Priority and Team Use of Data than schools in initial implementation. Strong measurement invariance held across groups of schools that differed in student ethnicity and SES.

Conclusions: The findings regarding measurement invariance are important for future longitudinal investigations of factors that may promote the sustained implementation of school practices.

Keywords: positive behavior interventions and supports, sustainability, measurement invariance

Measurement Invariance of an Instrument Assessing Sustainability of School-based Universal Behavior Practices

The process of implementing new practices in schools and other agencies has been conceived as occurring in a set of predictable stages of implementation (Fixsen, Blase, Duda, Naoom, & Van Dyke, 2010). Adelman and Taylor (1997) identified these stages as (a) creating readiness, (b) initial implementation, (c) institutionalization, and (d) ongoing evolution, with creating readiness occurring before actual use of the intervention with the intended recipients, and initial implementation, institutionalization, and ongoing evolution occurring as the intervention is used. Across the literature, each of the various conceptualizations of these phases places sustainability as the final phase and ultimate goal of implementation (Adelman & Taylor, 2003; Coburn, 2003; Fixsen, Naoom, Blase, Friedman, & Wallace, 2005). Assessing and predicting sustainability are important goals for research and practice because of its positive effects not only on students (e.g., improved student functioning and long term outcomes; Cook & Odom, 2013; Sanford DeRousie & Bierman, 2012), but also on systems and educators (e.g., improved teacher self-efficacy, organizational health; Baker, Gersten, Dimino, & Griffiths, 2004; Bradshaw, Koth, Bevans, Ialongo, & Leaf, 2008). The purpose of this study is to advance the assessment of school practice sustainability by investigating the measurement invariance of an established sustainability measure across schools at varied stages of PBIS implementation and with varied student ethnic composition and socioeconomic status (SES) levels.

Conceptualization of Sustainability

Sustainability is an elusive concept to measure because the construct is not as straightforward as is often considered (Vaughn, Klingner, & Hughes, 2000). Although sustainability has sometimes been viewed as synonymous with maintenance because it seems on

the surface to be characterized as achieving the same result with the same practice, the actual process of sustainability involves iterative changes over time to make the practice more effective, efficient, and relevant to the context (Castro, Barrera, & Martinez, 2004; McIntosh, Filter, Bennett, Ryan, & Sugai, 2010; McLaughlin & Mitra, 2001). In addition, sustainability is to some extent an outcome in of itself (i.e., continued adherence over time; Han & Weiss, 2005), but it can also be considered to be the potential for a practice to be sustained over time, a constellation of factors that make continued adherence more likely (McIntosh & Turri, in press). This latter conceptualization is more useful for practice because it allows for assessment, prediction, and systems-level interventions for sustainability at the start of implementation, rather than waiting until it has been achieved (Pluye, Potvin, Denis, Pelletier, & Mannoni, 2005).

McIntosh and colleagues (2009) developed a model of sustainability of school-based practices that identifies hypothesized factors and mechanisms by which these practices can be sustained. For example, practice priority (including staff buy-in, administrator support, and integration into daily responsibilities) provides the stimulus to continue implementation, even when considering competing initiatives that are alternatives to the practice. Another factor, collection and use of data for decision making is the mechanism by which school teams engage in ongoing evolution that results in adaptations that improve practices, rather than those that remove their effective components (McLaughlin & Mitra, 2001).

Assessment of Sustainability

To assess the contextual features most closely related to sustainability in the model, McIntosh and colleagues developed a measure, the *School-wide Universal Behavior Sustainability Index: School Teams* (SUBSIST; McIntosh, Doolittle, et al., 2009). The SUBSIST was developed as a self-administered measure of contextual and practice variables predicting

implementation and sustainability of school-wide behavior support practices (e.g., programs, curricula) delivered to all students. Development included review of the items and response process by an expert panel and piloting with school teams (McIntosh et al., 2011), a large-scale study of perceived importance of items by practitioners (McIntosh et al., 2014), a cluster analysis and construct validation (Hume & McIntosh, in press), and large-scale factor analysis with prediction of fidelity of practice implementation (McIntosh et al., 2013). Although developed for use with any universal practice, to date the measure has been validated with School-wide Positive Behavioral Interventions and Supports (PBIS; Sugai & Horner, 2009), a framework for implementing school-based interventions to increase prosocial behavior and decrease problem behavior through environmental redesign, explicit instruction, acknowledgement of prosocial behavior, and team-based use of fidelity of implementation and student outcomes data for continuous improvement. This approach has been implemented in over 18,000 schools in the US and schools in over a dozen countries (Sugai, 2012, October).

Despite this recent research validating the SUBSIST with schools adopting PBIS, additional validation is needed for the measure to contribute to important research in predicting and promoting sustainability. First, examining the measure's factor structure in a larger sample would be helpful in determining the extent to which prior findings (McIntosh et al., 2013) are replicated in an independent, cross-validation sample. Additionally, because the definition of sustainability implies implementation over long periods of time (Lucyshyn et al., 2007), any measure assessing sustainability must have a consistent factor structure at differing periods of time of implementation. Of theoretical and practical interest, sustainability can be assessed for schools at different stages of implementation to identify whether the SUBSIST factor structure holds across initial implementation, institutionalization, and ongoing evolution. Theoretically, it

would be expected that as schools continue with successful implementation and weather various barriers to sustainability, sustainability scores would increase over time, as has been shown with a smaller sample (Hume & McIntosh, in press). Prior to investigating factor score differences across stages of implementation, though, establishing measurement invariance of the SUBSIST is necessary to ensure that observed differences are not artifacts of measurement-related differences (Wu, Li, & Zumbo, 2007).

Measurement invariance of the SUBSIST across stages of implementation may not hold because staff in schools with varying durations of PBIS implementation could interpret or value SUBSIST items differently. Prior research has indicated that school personnel rate perceptions that PBIS is part of systems already in use, integration into new initiatives, family engagement, and staff support as more important for schools in the sustainability phase than initial implementation (McIntosh et al., 2014). Because these particular variables are viewed as more important for schools in the sustainability phase, it is possible that items assessing these areas on the SUBSIST could be differentially related to the School Priority factor, resulting in non-invariance across stages of implementation. Without assessment of invariance based on stage of implementation, any observed differences in scale scores could be due to actual differences or variations in scale psychometrics across implementation stages (Wu et al., 2007). Similarly, in the context of future longitudinal research, it is important to establish measurement equivalence to ensure that differences observed over time are not measurement-related artifacts.

In addition to investigating measurement invariance across groups with varying duration of PBIS implementation, it is important to consider the extent to which SUBSIST measurement invariance holds across schools with varied student populations. The impact of school ethnic composition and culture on the effectiveness of PBIS and the specific practices and interventions

included in local implementations of the PBIS framework has received increased attention in the research literature (Sugai, O'Keeffe, & Fallon, 2012; Vincent & Tobin, 2011), and similar efforts could be observed in the area of sustainability of PBIS. Prior to investigating variability in SUBSIST scores across schools with varied student populations, however, it is important to investigate the extent to which measurement of factors related to sustainability is consistent across these schools.

Purpose of the Study

The purpose of the study was to cross-validate the initial factor structure of the SUBSIST in a larger, independent sample and assess measurement invariance across three periods of time, corresponding theoretically to three stages of implementation: initial implementation, institutionalization, and ongoing evolution. Invariance across time would provide both evidence of the measure's psychometric adequacy for assessing longitudinal sustainability, as well as provide insight regarding how sustainability changes based on the phase of implementation. In addition, measurement invariance across groups of schools that differed in student ethnic composition and SES was assessed. The specific research questions tested were as follows:

1. To what extent is measurement with the SUBSIST invariant across three stages of practice implementation and schools with varied student ethnic composition and SES?
2. Are there mean differences on the subscales of the SUBSIST across stages of implementation and groups of schools with varied student ethnic composition and SES?

Method

Participants and Settings

School PBIS team members and district coaches representing a total of 860 schools implementing PBIS participated in the study. Of the 860 participants, 61% were school PBIS team leaders, 24% were school administrators, 9% were other faculty or staff on the school PBIS team, and 5% were external (e.g., district-level or regional) PBIS coaches. Of the 860 schools, 212 schools (in 149 districts) were in year 0 (planning year) or year 1 (first year of implementation with students), representing the initial implementation stage of the model from Adelman and Taylor (1997). In addition, 410 schools (in 189 districts) were in years 2 to 4 of implementation, representing institutionalization. Finally, 238 schools (in 88 districts) had been implementing PBIS for 5 or more years, representing the ongoing evolution (sustainability) phase. The schools were located in 14 states and represented all 4 U.S. Census Bureau regions. National Center for Education Statistics (NCES) demographic data were available for 98% of schools and are presented by implementation stage group in Table 1.

In addition to implementation stage, schools were divided into groups based on ethnic composition of students and school-level socioeconomic status (SES). For ethnic composition, the NCES Common Core of Data mean school-level percentage of students of color (i.e., non-White) was used to assign schools to groups: 534 schools had fewer than 45% students identified as non-White and 326 schools had 45% or more students identified as non-White. For SES, eligibility for federal Title I funds, an indicator of high numbers or percentages of students from low-SES families, was used to define groups: 335 schools were not eligible and 513 schools were eligible for Title I funds. Title I eligibility was unknown for 12 schools (1% of sample), and these schools were excluded from analyses of measurement invariance across school-level SES.

Measure

The *School-wide Universal Behavior Sustainability Index: School Teams* (SUBSIST; McIntosh, Doolittle, et al., 2009) is a 39-item measure of factors predicting sustained implementation of a school-based practice at a level of fidelity of implementation high enough to continue to meet valued outcomes. Respondents (school team members or external coaches) rate the extent to which each variable is present in their school at the time of response on a 4-point scale from 1 (*not true*) to 4 (*very true*). The measure includes school-level and district-level items.

Evidence of the SUBSIST's psychometric properties come from three studies to date. Results of an expert panel assessment provided evidence of strong content validity (content validity index = .95), and a pilot study showed strong internal consistency ($\alpha = .87$), interrater reliability ($r = .95$), and two-week test-retest reliability ($r = .96$; McIntosh et al., 2011). Results of a larger validation study (McIntosh et al., 2013) included an exploratory factor analysis and concurrent prediction of sustained PBIS implementation. Exploratory analyses indicated a four factor structure, with two school-level factors [School Priority (20 items, $\alpha = .94$) and Team Use of Data (11 items, $\alpha = .94$)] and two district-level factors [District Priority (5 items, $\alpha = .71$) and Capacity Building (3 items, $\alpha = .74$)] representing elements of the practice and its context. SUBSIST items by subscale are presented in McIntosh et al. (2013). Results indicated strong concurrent validity with PBIS implementation, with statistically significant correlations between each factor and PBIS fidelity of implementation scores. A cluster analysis (Hume & McIntosh, in press) identified valid clusters based on use of data and statistically significant correlations with other indicators of sustainability, including number of years implementing, access to district coaching, and school team actions.

Procedure

After obtaining Institutional Review Board approval for the study, the authors worked with several state-level PBIS teams to recruit a large sample of schools at varying years of PBIS implementation to complete the SUBSIST. State PBIS teams recruited any schools implementing or preparing to implement PBIS to participate during existing PBIS training events (either initial or ongoing trainings) and through email contacts. Participation consisted of one member from each school PBIS team actively consenting to complete the SUBSIST through a secure, online survey program.

Data Analyses

Measurement invariance was investigated by fitting a series of multiple-group confirmatory factor analysis (CFA) models. Because items on the SUBSIST have too few response options and exhibited too much negative skew to be considered normally distributed (Lubke & Muthén, 2004), items were specified as ordinal indicators in the CFA models using the theta parameterization and the mean- and variance-corrected weighted least squares (WLSMV) estimator in M*plus* 7 (Muthén & Muthén, 2012). In addition, because schools were nested in districts, standard errors and chi-square tests of model fit were adjusted to account for district-level clustering using the COMPLEX option in M*plus* (Asparouhov, 2005). On several items (ones with only two estimated thresholds in Table 3 or two *df* in Table 4), the full range of response options was not used across all groups. For these items, responses on the lowest two options were combined. Individual SUBSIST items had an average of 6.4% missing data due to participants endorsing items as unknown or not applicable; all available item responses were analyzed using the WLSMV estimator (Asparouhov & Muthén, 2010) and allowed to inform parameter estimates.

To investigate configural invariance, which is the extent to which items load on the same factor across groups, two sets of models were estimated. In the first set, fit of the 4-factor SUBSIST model in each group was investigated separately (e.g., initial implementation, institutionalization, and sustainability for analyses of invariance by implementation stage). In the second set, the fit of multi-group models was investigated, with factor loadings (excluding the first item in each factor that was constrained to equal one) and thresholds freely estimated in each group and all item residuals constrained to equal one and all factor means constrained to equal zero in all groups for model identification. In both sets of analyses, model fit was evaluated based on conventional criteria (Mueller & Hancock, 2010): Comparative Fit Index (CFI) $\geq$ .95 and Root Mean Square Error of Approximation (RMSEA) and its 90% confidence interval < .05.

To test strong measurement invariance, the multi-group models from the tests of configural invariance were compared to models with loadings and thresholds constrained to be equal across groups, with all item residuals constrained to equal one and all factor means constrained to equal zero in one group for model identification. Model fit was compared using the likelihood ratio (LR) chi-square difference test calculated using the DIFFTEST option in M*plus*. Following this global test of strong invariance, possible sources of non-invariance were explored using a backward selection procedure (Kim & Yoon, 2011). Specifically, we fit baseline models with factor loadings and item thresholds constrained to be equal across groups and then freed the factor loadings and thresholds across groups for one item at a time in comparison models. For model identification purposes, the residual variances of the item with freely estimated loadings and thresholds were constrained to equal one in all groups in the comparison models.

In the LR tests comparing the fully-invariant baseline model to each of the comparison models, there is a greater likelihood of Type I errors due to both the number of tests and possible misfit in the baseline model (i.e., if one or more of the items constrained to be invariant are non-invariant; Stark, Chernyshenko, & Drasgow, 2006). To reduce the potential likelihood of Type I errors related to multiple tests, Bonferroni correction of the critical *p* value (.05/39 item tests = .001) for the LR test was used. To account for potential misspecification in the baseline model, the Bonferroni-adjusted critical value was adjusted further using the following equation (Oort, 1998):

$$K_{adjusted} = \left(\frac{\chi_0^2}{K + df_0 - df_{LR}} \right) * K \quad (1)$$

where *K* and df_{LR} are the Bonferroni-adjusted critical chi-square value and degrees of freedom for the LR test, respectively, and χ_0^2 and df_0 are the chi-square value and degrees of freedom in the loading- and threshold-constrained baseline model. Kim and Yoon (2011) found that use of both Bonferroni and Oort (1998) adjustment of critical values of the LR test reduced false positives while maintaining adequate power in simulations of multiple-group ordinal CFA tests of measurement invariance. The combined Bonferroni and Oort (1998) adjustment resulted in chi-square critical values of 28.52 (*df* = 4) and 23.43 (*df* = 6) for the LR invariance tests across PBIS stage of implementation, 20.23 (*df* = 2) and 23.79 (*df* = 3) for tests across school-level ethnicity, and 20.39 (*df* = 2) and 23.99 (*df* = 3) for tests across school-level SES.[1]

Following the tests of measurement invariance across implementation, ethnic composition, and SES groups, differences in latent means on all factors of the SUBSIST were

[1] The following values were used in the calculation of the adjusted critical values: implementation stage, χ_0^2 = 3018.22, df_0 = 2292; ethnicity, χ_0^2 = 2212.05, df_0 = 1499; and SES, χ_0^2 = 2228.54, df_0 = 1498.

investigated in a partial or full measurement invariance model, based on results of the strong measurement invariance analyses, with latent means constrained to equal zero in one group for identification purposes. Standardized mean differences (*d*) were calculated as an effect size measure using the following formula (Hancock, 2001):

$$d = \frac{\left|\bar{Y}_1 - \bar{Y}_2\right|}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}}} \quad (2)$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are latent means in two groups, n_1 and n_2 are sample sizes in the groups, and s_1^2 and s_2^2 are variances of the latent factors.

Results

Results are organized in three main categories: tests of configural invariance, strong measurement invariance, and latent mean differences and factor correlations.

Configural Invariance

Fit indices for the models testing configural invariance are presented in Table 2. In general, model fit of the four-factor SUBSIST model was quite similar across stages of implementation (i.e., initial implementation, institutionalization, or sustainability), school student ethnic composition (i.e., < 45% non-White or ≥ 45% non-White), and school SES (i.e., eligible or not eligible for Title I). RMSEA values (range: .036 - .038) and the 90% C.I. RMSEA for the separate group models were within acceptable limits. CFI values (range: .940 - .950) were approximately equal to values suggesting strong model fit (CFI ≥ .95; Mueller & Hancock, 2010). Similarly, model fit of the multi-group models by stage of implementation, student ethnic distribution, and school SES supported configural invariance, with RMSEA values (range: .035 - .037) indicating acceptable fit and CFI values (range: .945 - .949) approximately equal to conventional criteria for acceptable fit.

Strong Invariance

To investigate strong measurement invariance, the fit of the multi-group configural invariance models was compared to models with factor loadings and item thresholds constrained to be equal across groups. Results are presented separately for analyses by implementation group, school ethnic composition, and school SES.

Stage of PBIS Implementation. Across implementation groups, the configural model fit better than the strong invariance model, $\chi^2_{LR} = 291.71$, $df = 204$, $p < .001$. Despite the statistically significant difference in chi-square between models, changes in CFI (.002) and RMSEA (.002) were negligible and suggested better fit for the strong invariance model. To further investigate potential sources of non-invariance, factor loadings and thresholds were freed one item at a time. Table 3 presents chi-square values for the individual LR tests. The LR chi-square values exceeded the Bonferroni- and Oort-adjusted critical values on only one item (School Priority: Item 17). For this item, threshold values decreased across the implementation groups, indicating that the boundaries between response options were at lower levels of latent School Priority as PBIS implementation time increased. This pattern could be related to the strong and increasing negative skew on the item across groups, initial implementation: -.42, institutionalization: -1.07, sustainability: -1.94. Fit for the final, partially-invariant model with loadings and thresholds free for the non-invariant item continued to be adequate on all indicators other than chi-square, $\chi^2 = 2995.13$, $df = 2286$, $p < .001$, CFI = .952, RMSEA = .033, 90% RMSEA CI [.030, .036]. Standardized factor loadings and item thresholds for this model are presented in Table 3.

School Ethnic Composition. Across schools with $\geq$ 45% and < 45% non-White students, the configural model fit better than the strong invariance model based on the LR test, $\chi^2_{LR} = 170.09$, $df = 107$, $p < .001$; however, both the CFI and RMSEA improved, albeit negligibly (.001

and .002, respectively), in the strong invariance model. Chi-square values for the LR tests of models with loadings and thresholds freely estimated across groups relative to the strong invariance model are presented in Table 4. None of the LR chi-square values exceeded the Bonferroni- and Oort-adjusted critical values. Although the LR test of the configural vs. strong invariance model indicated that the configural model fit better, CFI and RMSEA values suggested that the strong invariance model fit better and strong invariance was supported in tests of individual SUBSIST items. Fit of the strong invariance model was acceptable on most indicators other than chi-square, $\chi^2 = 2212.05$, $df = 1499$, $p < .001$, CFI = .948, RMSEA = .033, 90% RMSEA CI [.030, .036].

School-Level SES. Across schools that differed in Title I eligibility, there was no statistically significant reduction in model fit as factor loadings and item thresholds were constrained to equality in the strong invariance model, $\chi^2_{LR} = 117.22$, $df = 106$, $p = .215$. In addition, none of the item-specific LR chi-square values exceeded the adjusted chi-square critical values, as presented in Table 4. Fit of the strong invariance model was acceptable on indicators other than chi-square, $\chi^2 = 2228.54$, $df = 1498$, $p < .001$, CFI = .950, RMSEA = .034, 90% RMSEA CI [.031, .037].

Latent Variable Correlations and Means

Correlations among the latent variables, factor means, and factor standard deviations in the partially-invariant implementation group model and the strong invariance ethnic composition and SES models are presented in Table 5. In general, all factors were positively and strongly correlated (rs = .53 to .85, $p < .001$) in all groups. Regarding the PBIS implementation model, none of the factor means in the institutionalization group significantly differed from means in the initial implementation group. In contrast, means for School Priority and Team Use of Data were

higher (p = .008 and .001, respectively) in the sustainability group compared to the initial implementation group, and these differences were of small to medium magnitude (d = .36 and .52, respectively). In the school ethnic composition model, the mean of the Team Use of Data factor was lower in schools with 45% or more non-White students (p = .04), but the difference was of small magnitude (d = .24). In the school-level SES model, the mean of the School Priority factor was higher in schools eligible for Title I funds (p = .02), and the difference was also of small magnitude (d = .20). No other latent means differed statistically from reference group means in the models.

Discussion

Overall, results provided additional support for the two school- and two district-level factor structure of the SUBSIST in a sample independent from the one used in McIntosh et al. (2013). In addition, the results supported the four-factor solution across groups of schools that differed in duration of PBIS implementation, student ethnic composition, and student SES. Configural invariance of the SUBSIST was supported by adequate fit of the four-factor model in each group and the multi-group models.

Results also supported strong measurement invariance across implementation groups for all items except one on the SUBSIST and for all items across school-level ethnicity and SES groups. Although change in approximate fit indexes from the configural to strong measurement invariance model was below the commonly-used criterion of change less than .01 on the CFI (Cheung & Rensvold, 2002), use of the backward selection procedure described in Kim and Yoon (2011) with Bonferroni and Oort (1998) adjustment of critical values for the LR test indicated that the item “SW-PBIS is considered to be a typical operating procedure of the school (it has become “what we do here/what we’ve always done”)” had different factor loadings and/or

item thresholds across implementation groups. Logically, one would expect that schools that have been implementing PBIS longer would be more likely to perceive PBIS as a typical practice, and indeed, the item means increased across implementation groups (initial implementation: 2.96, institutionalization: 3.34, sustainability: 3.66) with more pronounced negative skew in groups with greater implementation time. In a related study with the original validation sample, this item was not perceived as important for initial PBIS implementation, but considering PBIS to be a typical operating procedure was rated by school and district PBIS team members as among the most important factors related to sustained implementation (McIntosh et al., 2014). Consequently, it is not surprising that as implementation time increased, it became easier for school teams (i.e., required less School Priority) to rate the item as more true of the school.

Similarly to the results of McIntosh et al. (2013), strong positive correlations were found among the four SUBSIST factors, and the magnitude of correlations appeared to be similar across implementation groups in the final, partially-invariant model that freed factor loadings and thresholds for the "typical operating procedure" item and across groups in the fully-invariant school-level ethnicity and SES models. Inspection of latent factor mean differences indicated no statistically significant differences between the initial implementation and institutionalization groups; however, levels of School Priority and Team Use of Data were greater, with small to medium size differences, in the sustainability compared to the initial implementation group. These findings are consistent with perceptions of school and district PBIS team members that school-level factors, particularly administrator priority, are more important than district-level factors for sustained implementation (McIntosh et al., 2013). In addition, the current finding that the largest mean differences were on Team Use of Data is consistent with prior research

indicating that this factor had the largest independent association with PBIS implementation fidelity while accounting for the other factors on the SUBSIST (McIntosh et al., 2013). Small and statistically significant latent mean differences were also found in the school-level ethnicity and SES models. Schools with 45% or more students identified as non-White had lower scores on Team Use of Data compared to schools with fewer non-White students, and schools eligible for Title I funds had higher scores on School Priority than schools not eligible for funds. Although sometimes poverty is viewed as a strong barrier to sustainability, due to reduced access to resources (Rogers, 2003), staff in schools serving more students in poverty may view preventive practices such as PBIS as more valuable, and thus would rate it as a higher priority.

These findings should be considered in light of several limitations. First, the SUBSIST is a survey measure of perceptions completed by school or district PBIS team members, and it is possible that team member perceptions may not be representative of all team members in the school. This concern is partially tempered by findings of high inter-rater reliability for the SUBSIST in a prior investigation (McIntosh et al., 2011); however, results of the current study would be strengthened by inclusion of reports from multiple school team members and/or the addition of direct observation measures of factors related to sustainability. In addition, although groups that varied in terms of years of PBIS implementation were included in the study, the study was cross sectional. Consequently, the extent to which longitudinal invariance would hold as the same schools continue to implement PBIS is unclear.

The current findings of strong measurement invariance, excluding one item, across implementation groups for the SUBSIST is an important first step toward inclusion of the SUBSIST in a longitudinal investigation examining the dynamic interrelation of factors related to sustainability and fidelity of PBIS. Although Team Use of Data was found to be the most

important factor in relation to implementation fidelity (McIntosh et al., 2013), the predictive power of the SUBSIST factors has not been examined prospectively. In addition to the predictive strength of each SUBSIST factor, future research should also examine the extent to which these associations change across stages of implementation. It is possible that the SUBSIST factors most predictive of sustainability during initial implementation differ from the factors predicting continued implementation after several years of institutionalization. We hope that by setting the stage for future longitudinal research, the current study will contribute to rigorous examination of factors that promote the sustained implementation of evidence-based school initiatives.

References

Adelman, H. S., & Taylor, L. (1997). Toward a scale-up model for replicating new approaches to schooling. *Journal of Educational and Psychological Consultation, 8*, 197-230.

Adelman, H. S., & Taylor, L. (2003). On sustainability of project innovations as systemic change. *Journal of Educational and Psychological Consultation, 14*, 1-25.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12*(3), 411-434. doi: 10.1207/s15328007sem1203_4

Asparouhov, T., & Muthén, B. (2010). Weighted least squares estimation with missing data. Retrieved from http://www.statmodel.com

Baker, S., Gersten, R., Dimino, J. A., & Griffiths, R. (2004). The sustained use of research-based instructional practice: A case study of peer-assisted learning strategies in mathematics. *Remedial and Special Education, 25*, 5-24.

Bradshaw, C. P., Koth, K., Bevans, K. B., Ialongo, N., & Leaf, P. J. (2008). The impact of school-wide positive behavioral interventions and supports on the organizational health of elementary schools. *School Psychology Quarterly, 23*, 462-473.

Castro, F. G., Barrera, M., & Martinez, C. R. (2004). The cultural adaptation of prevention interventions: Resolving tensions between fidelity and fit. *Prevention Science, 5*, 41-45.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. doi: 10.1207/s15328007sem0902_5

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher, 32*(6), 3-12.

Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children, 79*, 135-144.

Fixsen, D. L., Blase, K. A., Duda, M. A., Naoom, S. F., & Van Dyke, M. (2010). Implementation of evidence-based treatments for children and adolescents. In J. R. Weisz & A. E. Kazdin (Eds.), *Evidence-based psychotherapies for children and adolescents* (pp. 435-450). New York: Guilford.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: Synthesis of the literature*. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).

Han, S. S., & Weiss, B. (2005). Sustainability of teacher implementation of school-based mental health programs. *Journal of Abnormal Child Psychology, 33*, 665-679.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*(3), 373-388. doi: 10.1007/bf02294440

Hume, A. E., & McIntosh, K. (in press). Construct validation of a measure to assess sustainability of school-wide behavior interventions. *Psychology in the Schools*.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*(2), 212-228. doi: 10.1080/10705511.2011.557337

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*(4), 514-534. doi: 10.1207/s15328007sem1104_2

Lucyshyn, J. M., Albin, R. A., Horner, R. H., Mann, J. C., Mann, J. A., & Wadsworth, G. (2007). Family implementation of positive behavior support with a child with Autism: A longitudinal, single case experimental and descriptive replication and extension. *Journal of Positive Behavior Interventions, 9*, 131-150.

McIntosh, K., Doolittle, J., Vincent, C. G., Horner, R. H., & Ervin, R. A. (2009). *School-wide universal behavior sustainability index: School teams*. Vancouver, BC: University of British Columbia.

McIntosh, K., Filter, K. J., Bennett, J. L., Ryan, C., & Sugai, G. (2010). Principles of sustainable prevention: Designing scale-up of school-wide positive behavior support to promote durable systems. *Psychology in the Schools, 47*, 5-21. doi: 10.1002/pits.20448

McIntosh, K., Horner, R. H., & Sugai, G. (2009). Sustainability of systems-level evidence-based practices in schools: Current knowledge and future directions. In W. Sailor, G. Dunlap, G. Sugai & R. H. Horner (Eds.), *Handbook of positive behavior support* (pp. 327-352). New York: Springer.

McIntosh, K., MacKay, L. D., Hume, A. E., Doolittle, J., Vincent, C. G., Horner, R. H., & Ervin, R. A. (2011). Development and initial validation of a measure to assess factors related to sustainability of school-wide positive behavior support. *Journal of Positive Behavior Interventions, 13*, 208-218. doi: 10.1177/1098300710385348

McIntosh, K., Mercer, S. H., Hume, A. E., Frank, J. L., Turri, M. G., & Mathews, S. (2013). Factors related to sustained implementation of schoolwide positive behavior support. *Exceptional Children, 79*, 293-311.

McIntosh, K., Predy, L. K., Upreti, G., Hume, A. E., Turri, M. G., & Mathews, S. (2014). Perceptions of contextual features related to implementation and sustainability of school-

wide positive behavior support. *Journal of Positive Behavior Interventions, 16*, 29-41. doi: 10.1177/1098300712470723

McIntosh, K., & Turri, M. G. (in press). Positive behavior support: Sustainability and continuous regeneration. In C. R. Reynolds, K. J. Vannest & E. Fletcher-Janzen (Eds.), *Encyclopedia of special education: A reference for the education of children, adolescents, and adults with disabilities and other exceptional individuals* (4th ed.). Hoboken, NJ: Wiley.

McLaughlin, M. W., & Mitra, D. (2001). Theory-based change and change-based theory: Going deeper, going broader. *Journal of Educational Change, 2*, 301-323.

Mueller, R. O., & Hancock, G. R. (2010). Structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 371-384). New York: Routledge.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles: Author.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 5*(2), 107-124. doi: 10.1080/10705519809540095

Pluye, P., Potvin, L., Denis, J.-L., Pelletier, J., & Mannoni, C. (2005). Program sustainability begins with the first events. *Evaluation and Program Planning, 28*, 123-137.

Rogers, E. (2003). *Diffusion of innovations* (5th ed.). New York: The Free Press.

Sanford DeRousie, R. M., & Bierman, K. L. (2012). Examining the sustainability of an evidence-based preschool curriculum: The REDI program. *Early Childhood Research Quarterly, 27*, 55-65.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306. doi: 10.1037/0021-9010.91.6.1292

Sugai, G. (2012, October). *School-wide PBIS: An effective framework for teaching and learning*. Paper presented at the National Forum on School-wide Positive Behavior Support, Chicago, IL.

Sugai, G., & Horner, R. H. (2009). Defining and describing schoolwide positive behavior support. In W. Sailor, G. Dunlap, G. Sugai & R. H. Horner (Eds.), *Handbook of positive behavior support* (pp. 307-326). New York: Springer.

Sugai, G., O'Keeffe, B. V., & Fallon, L. M. (2012). A contextual consideration of culture and school-wide positive behavior support. *Journal of Positive Behavior Interventions, 14*(4), 197-208. doi: 10.1177/1098300711426334

Vaughn, S., Klingner, J., & Hughes, M. (2000). Sustainability of research-based practices. *Exceptional Children, 66*, 163-171.

Vincent, C. G., & Tobin, T. J. (2011). The relationship between implementation of school-wide positive behavior support (SWPBS) and disciplinary exclusion of students from various ethnic backgrounds with and without disabilities. *Journal of Emotional and Behavioral Disorders, 19*(4), 217-232. doi: 10.1177/1063426610377329

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*(3), 1-25.

Table 1

School Demographics by Stage of Implementation

Stage	*M* or % (*SD*) Initial Implementation (*n* = 209)	Institutionalization (*n* = 408)	Sustainability (*n* = 233)
Enrollment	517.62 (375.84)	533.28 (344.24)	565.62 (299.66)
Title 1 Eligible	63% (48)	64% (48)	52% (50)
% Non-White	35% (33)	43% (32)	39% (25)
Grade Level			
Elementary	61%	72%	68%
Middle	25%	16%	22%
High	14%	12%	10%

Note. Data obtained from National Center for Education Statistics for 98% of schools.

Table 2

Model Fit in Tests of Configural Invariance

Model	Group	χ^2	*df*	CFI	RMSEA	90% RMSEA
Stage	Initial Implementation	894.69	696	.940	.037	.029 - .044
	Institutionalization	1075.90	696	.949	.036	.032 - .041
	Sustainability	913.91	696	.942	.036	.029 - .043
	Combined	2840.79	2088	.949	.035	.032 - .039
Ethnicity	< 45% non-White	1303.57	696	.941	.040	.037 - .044
	$\geq$ 45% non-White	953.58	696	.950	.034	.028 - .039
	Combined	2128.33	1392	.947	.035	.032 - .038
Title I	Eligible	1208.01	696	.946	.038	.034 - .041
	Not Eligible	1020.33	696	.942	.037	.032 - .042
	Combined	2189.09	1392	.945	.037	.034 - .040

Note. n = 860 for analyses by implementation stage, n = 860 for analyses by ethnicity, and n = 848 for analyses by school Title 1 eligibility.

Table 3

Factor Loadings and Thresholds for the Partially Invariant Model and Likelihood Ratio Tests of Measurement Invariance Across Stage of Implementation

Level	Factor	Item	λ	γ_1	γ_2	γ_3	χ^2_{LR}	df_{LR}	p
School	School Priority	1	.47	-1.67	-0.77	--	1.92	4	.751
		2	.58	-1.77	-0.38	--	5.82	4	.213
		3	.64	-2.51	-1.59	-0.45	9.41	6	.152
		4	.62	-2.74	-1.58	-0.14	3.51	6	.743
		5	.51	-0.50	0.86	1.80	3.58	6	.733
		6	.67	-2.01	-0.77	--	16.35	4	.003
		7	.64	-2.78	-1.86	-0.65	11.34	6	.079
		8	.37	-2.19	-1.55	-0.76	8.08	6	.232
		9	.59	-2.31	-0.93	--	7.75	4	.101
		10	.70	-3.00	-1.49	0.51	15.11	6	.019
		11	.69	-2.17	-1.34	-0.17	10.85	6	.093
		12	.45	-1.76	-0.81	0.35	14.67	6	.023
		13	.70	-3.61	-2.44	-0.60	11.58	6	.072
		14	.63	-2.21	-1.01	0.05	14.26	6	.027
		15	.76	-2.86	-1.56	0.38	3.44	6	.752
		16	.62	-2.68	-0.97	--	4.73	4	.316
		17[a]	.80	-2.60	-0.86	0.77	64.95*	6	.000
			.73	-2.73	-1.51	-0.06			
			.80	-3.26	-2.27	-0.52			
		18	.57	-2.87	-1.50	0.24	13.48	6	.036
		19	.55	-2.44	-1.30	0.20	4.36	6	.628
		20	.59	-2.30	-0.59	--	1.70	4	.791
	Team Use of Data	1	.64	-2.97	-1.59	0.04	9.84	6	.131
		2	.72	-1.79	0.09	--	23.29	4	.000
		3	.51	-2.32	-1.63	-0.86	19.21	6	.004
		4	.52	-2.13	-1.12	-0.22	7.70	6	.261
		5	.63	-2.24	-1.31	-0.24	5.86	6	.439
		6	.67	-2.37	-1.29	0.07	7.38	6	.287
		7	.84	-3.06	-1.49	0.21	15.06	6	.020
		8	.80	-2.18	-0.80	0.41	8.80	6	.185
		9	.84	-1.79	-0.68	0.73	23.12	6	.001
		10	.88	-4.31	-2.16	-0.03	13.51	6	.036
		11	.59	-2.19	-1.34	0.04	8.52	6	.202
District	District Priority	1	.55	-1.38	-0.38	0.93	9.06	6	.170
		2	.69	-2.93	-1.39	-0.06	6.86	6	.334
		3	.55	-1.83	-0.87	0.15	9.75	6	.135
		4	.79	-2.24	-0.53	1.30	10.59	6	.102
		5	.71	-2.46	-1.14	0.02	18.84	6	.004
	Capacity Building	1	.60	-2.19	-1.17	-0.27	2.35	6	.885
		2	.78	-3.04	-1.75	-0.40	8.45	6	.207

3	.62	-1.79	-0.80	0.13	9.23	6	.161

Note. n = 860. χ^2_{LR} = likelihood ratio chi-square, df_{LR} = degrees of freedom for likelihood ratio test. Presented factor loadings are standardized.

[a]Group-specific loadings and thresholds are presented for this item.

[*]Chi-square value exceeds Bonferroni- and Oort-adjusted critical value.

Table 4

Likelihood Ratio Tests of Measurement Invariance by School-Level Ethnicity and SES

			Ethnicity			Title 1		
Level	Factor	Item	χ^2_{LR}	df_{LR}	p	χ^2_{LR}	df_{LR}	p
School	School Priority	1	10.67	2	.005	7.60	2	.022
		2	2.38	2	.305	.64	2	.727
		3	2.79	3	.425	11.75	3	.008
		4	1.39	3	.707	3.60	3	.308
		5	4.66	3	.199	1.95	3	.583
		6	3.62	3	.306	2.01	3	.570
		7	1.77	3	.623	1.47	3	.689
		8	11.90	3	.008	4.19	3	.242
		9	3.34	3	.343	3.31	3	.346
		10	7.47	3	.058	2.07	3	.559
		11	2.18	3	.535	4.89	3	.180
		12	7.28	3	.063	1.02	3	.797
		13	4.52	3	.210	2.55	3	.467
		14	11.84	3	.008	3.02	3	.388
		15	5.80	3	.122	8.22	3	.042
		16	1.98	3	.578	1.40	2	.496
		17	3.78	3	.286	3.76	3	.288
		18	3.08	3	.380	1.05	3	.790
		19	19.10	3	.000	6.22	3	.101
		20	12.82	3	.005	2.58	3	.461
	Team Use of Data	1	2.39	3	.495	4.21	3	.239
		2	1.25	3	.740	8.59	3	.035
		3	2.31	3	.512	6.25	3	.100
		4	2.12	3	.548	3.18	3	.365
		5	1.54	3	.674	2.33	3	.507
		6	8.37	3	.039	2.11	3	.549
		7	1.84	3	.606	1.65	3	.649
		8	2.68	3	.444	2.66	3	.447
		9	5.85	3	.119	2.05	3	.562
		10	3.71	3	.294	3.43	3	.331
		11	2.81	3	.422	1.37	3	.713
District	District Priority	1	4.30	3	.231	3.06	3	.382
		2	.70	3	.873	1.84	3	.606
		3	15.91	3	.001	4.25	3	.236
		4	5.41	3	.144	4.66	3	.198
		5	2.63	3	.452	5.56	3	.135
	Capacity Building	1	13.60	3	.004	3.50	3	.321
		2	10.94	3	.012	9.01	3	.029
		3	.94	3	.816	2.99	3	.393

Note. n = 860 for analyses by ethnicity, and n = 848 for analyses by school Title 1 eligibility.

χ^2_{LR} = likelihood ratio chi-square, df_{LR} = degrees of freedom for likelihood ratio test.

Table 5

Factor Correlations, Standard Deviations, and Means by Group

Group	Factor	SP	TUD	DP	*M*	*SD*
Initial Implementation	School Priority	--			.00[a]	.53
	Team Use of Data	.78***	--		.00[a]	.84
	District Priority	.77***	.64***	--	.00[a]	.65
	Capacity Building	.61***	.66***	.72***	.00[a]	.75
Institutionalization	School Priority	--			.01	.56
	Team Use of Data	.80***	--		.13	1.04
	District Priority	.70***	.62***	--	.02	.69
	Capacity Building	.69***	.75***	.74***	-.04	.81
Sustainability	School Priority	--			.20**	.58
	Team Use of Data	.85***	--		.47**	.97
	District Priority	.62***	.53***	--	.07	.65
	Capacity Building	.58***	.64***	.62***	.01	.76
< 45% non-White	School Priority	--			.00[a]	.67
	Team Use of Data	.81***	--		.00[a]	1.15
	District Priority	.72***	.64***	--	.00[a]	.70
	Capacity Building	.62***	.75***	.69***	.00[a]	.90
≥ 45% non-White	School Priority	--			-.04	.75
	Team Use of Data	.84***	--		-.28*	1.23
	District Priority	.67***	.61***	--	.06	.68
	Capacity Building	.68***	.66***	.77***	-.11	.88
Title I Not Eligible	School Priority	--			.00[a]	.58
	Team Use of Data	.79***	--		.00[a]	1.01
	District Priority	.69***	.57***	--	.00[a]	.68
	Capacity Building	.71***	.76***	.77***	.00[a]	.84
Title I Eligible	School Priority	--			.11*	.56
	Team Use of Data	.83***	--		.12	.98
	District Priority	.68***	.62***	--	.10	.66
	Capacity Building	.60***	.67***	.68***	.10	.94

Note. SP = School Priority, TUD = Team Use of Data, DP = District Priority.

[a]Parameter constrained for model identification.

*$p < .05$, **$p < .01$, ***$p < .001$