Article

Hammill Institute on Disabilities

Assessment for Effective Intervention
2016, Vol. 41(3) 155–171

DOI: 10.1177/1534508415619732
aei.sagepub.com
SAGE

Exploring an Ecological Model of Perceived Usability Within a Multi-Tiered Vocabulary Intervention

Sabina R. Neugebauer, EdD[1], Sandra M. Chafouleas, PhD[2], Michael D. Coyne, PhD[2], D. Betsy McCoach, PhD[2], and Amy M. Briesch, PhD[3]

Abstract

The present study examines an ecological model for intervention use to explain student vocabulary performance in a multi-tiered intervention setting. A teacher self-report measure composed of factors hypothesized to influence intervention use at multiple levels (i.e., individual, intervention, and system level) was administered to 54 teachers and 48 interventionists conducting vocabulary interventions with different levels of instructional intensity with 553 kindergarten students. The reliability and validity of the measure in the context of a specific multi-tiered intervention was explored. Of particular interest was the potential explanatory power of system-level factors, over and above intervention-specific measures of fidelity, to explain student performance in a multi-tiered context. Results indicate that the climate of the school system predicted student performance in the Tier 1 context, and intervention feasibility predicted student performance in Tier 2. However, the intervention-specific fidelity measure was not a significant predictor. This research provides supporting evidence for the use of ecological models of intervention implementation to capture factors that influence intervention use and performance in multi-tiered settings.

Keywords

early literacy, rating scales, interventions

Since the reauthorization of the Individuals With Disabilities Education Act (IDEA) in 2004, multi-tiered systems of support (MtSS) have increasingly been adopted in schools, with more than 90% reporting use of some elements of MtSS. MtSS frameworks aim to prevent students from falling behind in the school curriculum by providing effective core (Tier 1) classroom instruction for all students and more intensive supplementary (Tier 2) intervention for students who do not respond to classroom instruction (National Center on Response to Intervention, n.d.). One recommendation for assuring the success of MtSS is to adopt systematic, evidence-based curricula at Tier 1 and Tier 2 (Fuchs, Mock, Morgan, & Young, 2003), on the basis that these programs will improve equity and reduce variability in implementation. Moreover, a critical feature of MtSS is that students have participated in high-quality instruction at each Tier. Put differently, if teachers do not use the Tier 1 program, it is difficult to know whether students who do not respond need Tier 2 or just better Tier 1 (Fuchs, Fuchs, & Vaughn, 2008). Research supports the idea that it matters how much of an intervention teachers implement: Thorough implementation of interventions has a strong relation with students' academic performance (Durlak & DuPre, 2008).

Thus, an important issue is whether teachers implement programs as intended. Generally, researchers are particularly interested in reducing variability in implementation because their goal is to observe effects specific to their intervention. Therefore, they examine implementation with a checklist or observation tool that measures adherence to specific features of the intervention (Briesch, Chafouleas, Neugebauer, & Riley-Tillman, 2013). This provides important information about implementation quality, and should help in the detection of intervention effects.

However, within the authentic school—or ecological—setting, fidelity can only capture some factors that link the intervention and student achievement. Teachers implement interventions in ecological contexts in which many forces operate on teachers' ability and desire to use interventions, each of which may affect student achievement in the

[1]Loyola University, Chicago, USA
[2]University of Connecticut, Storrs, USA
[3]Northeastern University, Boston, MA, USA

Corresponding Author:
Sabina R. Neugebauer, Loyola University Chicago, 820 N. Michigan Avenue, Lewis Towers, 11th Floor, Chicago, IL 60611, USA.
Email: sneugebauer@luc.edu

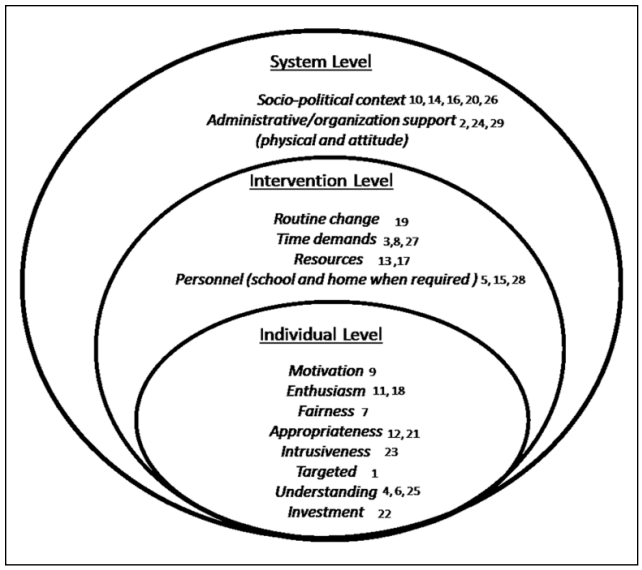

Figure 1. A schematic representation of the three ecological levels targeted in the present study and drawn from the literature base on school-based factors at the individual, intervention, and system level.
Note. Each listed area is targeted in URP-IR items. The corresponding item(s) is listed next to the target area. URP-IR = Usage Rating Profile–Intervention Revised.

intervention separate from teachers' fidelity (Riemer, Rosof-Williams, & Bickman, 2005). This article examines a tool designed to capture teachers' perceptions of multiple factors thought to influence whether they will use a program as intended to (a) establish the reliability of the measure and (b) determine whether the measure relates to student achievement after accounting for fidelity of implementation in both Tier 1 and Tier 2 settings.

Factors That Influence Intervention Use

Our intent in the present investigation was to take an ecological approach to intervention use (Sanetti & Kratochwill, 2009). Factors that are potentially relevant for inclusion in an ecological model of intervention implementation at the individual, intervention, and environmental level have been researched extensively in school-based settings (see Figure 1).

To date, the majority of commonly used measures of intervention use focus primarily on intra-individual factors, namely, treatment acceptability. The acceptability literature focuses on whether participants feel the treatment is fair, reasonable, appropriate for the target problem, intrusive, and consistent with their general understanding of what the treatment should be (Kazdin, 1980). Extensive research has been conducted on treatment acceptability, with conflicting and in many cases weak findings regarding the association between acceptability and use (Sterling-Turner & Watson, 2002). Measures focused on intra-individual factors do not consider the multiple systems in which interventions are situated, systems that are even more interactive in a multitiered system of support. Furthermore, the mixed results of studies on the relationship between acceptability and use suggest that other factors should be considered.

Over the past several decades, researchers have identified hundreds of variables that are believed to influence treatment integrity. In synthesizing the literature, Sanetti and Kratochwill (2009) organized these variables into four different levels of influence: interventionist, intervention, organization, and external environment. At the interventionist level (i.e., at the level of the individual), researchers have most commonly focused on whether an individual views an intervention as fair, reasonable, or appropriate (Kazdin, 1980) and educational stakeholders' motivation and enthusiasm for implementing the intervention (Perepletchikova & Kazdin, 2005). Several studies have also shown that an individual's understanding of the intervention affects acceptability (Reimers, Wacker, & Koeppl, 1987).

At the intervention level, there has been considerable focus on the feasibility of implementing the intervention. Feasibility includes whether there is enough time required to execute the intervention and whether the resources or materials needed for the intervention are available. The degree to which interventions interrupt business-as-usual routines also affects perceptions of feasibility (Perepletchikova & Kazdin, 2005).

At the organization level, factors associated with intervention use include systems variables, such as available administrative support for an intervention, the beliefs and attitudes about interventions in the school and the broader community (Broughton & Hester, 1993), the school climate or the level of support from parents, peers, and other educational stakeholders (Sanetti & Kratochwill, 2009). Finally, external environment level factors include considerations such as whether there is support from community stakeholders or bureaucratic obstacles to implementation (Sanetti & Kratochwill, 2009). Factors at the organization and external environment level are collectively considered systems-level factors (Briesch et al., 2013).

Although researchers tend to agree that singular explanations for intervention use are not sufficient (Orne & Binik, 1989) and acknowledge several conceptual models that incorporate multiple factors across ecological levels (Sanetti & Kratochwill, 2009), there are few comprehensive measures that simultaneously assess these factors. One exception is the Usage Rating Profile–Intervention Revised (URP-IR), which was designed to be a teacher self-report measure (Briesch et al., 2013). The URP-IR includes six factors at multiple ecological levels (i.e., Acceptability, Understanding, Feasibility, Family–School Collaboration, System Climate, and System Support). The URP-IR addresses factors at the individual, intervention, and system level (see Figure 1, for a schematic representation of how

URP-IR items map on to the literature base on factors at different ecological levels; and Table 2, for a description of each factor). The individual-level *Acceptability* factor measures whether someone views an intervention as fair and appropriate (Kazdin, 1980) and something they would be enthusiastic about implementing (Perepletchikova & Kazdin, 2005). The individual-level *Understanding* factor captures the extent to which participants know and comprehend how to implement the intervention (Reimers et al., 1987). The intervention-level *Feasibility* factor assesses whether participants feel that implementing the intervention as described is manageable, given existing demands (Perepletchikova & Kazdin, 2005). The intervention-level *Family–School Collaboration* factor focuses on the importance of shared-decision making, support, and collaboration from parents in the success of an intervention (Durlak & DuPre, 2008). This factor addresses the extent to which participants believe that family–school collaborations are necessary for the intervention to be successfully utilized. The URP-IR also includes system-level factors that capture support and attitudes within the school community (Broughton & Hester, 1993). The *System Climate* factor assesses whether participants feel the intervention is compatible with the school environment. The *System Support* factor captures whether participants feel they need external support to use the intervention. By addressing factors at these multiple levels, the URP-IR allows us to explore intervention use in multi-tiered settings. The URP-IR is more sensitive to system-level factors that are part of MtSS, than measures that consider fewer factors related to intervention use.

The URP-IR captures multiple factors that not only compose an ecological model but also consolidate overlapping constructs in the literature to create a manageable instrument. To elaborate, Sanetti and Kratochwill's (2009) broad review of the literature identified up to 12 factors believed to influence treatment integrity at each of four levels of influences. However, incorporating all of these potential factors into one comprehensive measure would significantly limit its usability given the number of items required. To address this measurement issue, the URP-IR aggregates across overlapping factors. For example, the factors in Sanetti and Kratochwill's model of complexity, time, material resources, and ease of implementation are conceptualized as separate factors despite capturing related and interacting aspects of an intervention. By contrast, the URP-IR subsumes these factors under one factor, feasibility. In addition, the URP-IR does not account for external environment factors (e.g., educational mandates, and bureaucratic barriers), instead focusing more explicitly on the school system.

Existing research indicates that the URP-IR relates to intervention use, but exploration to date has involved only behavioral interventions (Briesch et al., 2013). The present study represents a critical first step in using this recently validated measure to investigate an ecological model of intervention use in the context of a specific multi-tiered academic intervention.

Intervention Use Factors and Student Outcomes

Generally, it is assumed that effective interventions, implemented as intended, increase student outcomes (Noell, 2008). However, in specific contexts, this relationship has been found to be more complicated (Durlak & DuPre, 2008). Some deviations may in fact make an intervention more socially valid in a given intervention context. For example, in some settings, extended time spent on certain program components may be advantageous for populations with greater or less need in a specific domain (Sanetti & Kratochwill, 2009). Multi-tiered systems may encourage deviations or adaptations as a function of additional logistical demands such as coordinating intervention content, resources, and scheduling across tiers.

Furthermore, tiered systems of support, by definition, encourage instruction that is responsive. At the core of this model is pedagogy that is adaptive with regard to pace and intensity to the learners' needs and that facilitates communication across those administering support. Research indicates that tiered systems rely enormously on a positive school climate and educational stakeholder relationships to provide responsive instruction (Bean & Lillenstein, 2012). In the present study, we hypothesized that in a context that demands flexibility across tiers, system-level factors would also be associated with student performance outcomes for interventions conducted at Tier 1 and Tier 2. In a model focused on altering instruction to the learners' needs, it is reasonable to think that teachers who have more positive reports of system-level support may be teachers who are better able to provide responsive instruction, and are able to flexibly adapt to the time, resource, and content demands of a tiered model.

The present investigation was specifically focused on the application of a tool to measure factors associated with intervention use across ecological levels in the ecological setting of a vocabulary intervention. Our focus on vocabulary learning reflects research that indicates that students' early language and literacy skills are highly predictive of future academic outcomes (Dickenson & Caswell, 2007), and early vocabulary development is a critical area for preventing a cycle of low literacy skills where students fall progressively behind (Stanovich, 1986). Furthermore, vocabulary is quickly becoming an area of increasing interest for MtSS educators, as it has been a traditionally neglected area for intervention (Biemiller & Slonim, 2001; Gersten et al., 2009). Vocabulary as a less-studied literacy component represents more complexity for intensifying instruction given that vocabulary knowledge is a vast knowledge set compared with other literacy related skills

that are more finite, for example, there are only, approximately, 44 phonemes, and 26 letter names (Stahl, 2011). Furthermore, vocabulary interventions are well suited for examination with the URP-IR in that the developers of the URP-IR use the construct of usage to capture both "initial use as well as use that is habitual or customary" (Briesch et al., 2013, p. 82), which is compatible with the incremental nature of vocabulary knowledge (Nagy & Scott, 2000). For vocabulary interventions, teacher intervention usage (shorter and longer term use) is particularly germane as a potential indicator of student outcomes given that word knowledge depends on sustained instruction and word exposure (Beck, McKeown, & Kucan, 2002).

An open question in the present investigation was whether the URP-IR would be valid and reliable in the context of this specific vocabulary intervention and if so, whether it would support the potential association of factors that influence use and student vocabulary performance in settings of different instructional intensity. Thus, the present study explored the utility of the URP-IR measure in a kindergarten multi-tiered vocabulary intervention context and examined whether teachers' self-perceptions of the usability of an intervention (i.e., responses to the URP-IR) explained unique variance in student performance on a curriculum-specific vocabulary outcome with teachers and interventionists implementing different degrees of instruction intensity (i.e., Tier 1 or Tier 2 instruction). Of particular interest was whether system-level factors that capture the level of functioning in a system of coordinated intervention would explain additional variance in student performance beyond traditional intervention-specific measures of implementation fidelity. We conducted secondary analyses using data from a quasi-experimental study in which one group of kindergarten students received whole-class vocabulary instruction (Tier 1) and another group of students received intensive small-group vocabulary intervention that supplemented the classroom instruction (Tier 2). The following two research questions guided our investigation:

Research Question 1: Is the URP-IR valid and reliable in the context of a specific academic multi-tiered intervention?

Research Question 2: Do URP-IR factors, specifically those that represent system-level influences, explain additional variance in students' curriculum-specific vocabulary performance in (a) Tier 1 and (b) Tier 2 settings beyond an intervention-specific measure of implementation fidelity?

Method

Participants

The study was conducted across schools in three states, two in the Northeast and one in the Western United States.

Table 1. Sample Numbers by Site, School, and Teacher/Interventionist.

Population	Site 1	Site 2	Site 3
Schools per site (Total *N* = 23)	10	11	2
Teachers across schools per site (Total *N* = 54)	29	23	2
Average Tier 1 student per class (Total *N* = 360)	7	6	5
Interventionists across schools per site (Total *N* = 48)	26	20	2
Average Tier 2 student per class (Total *N* = 193)	4	4	3

Note. The number of students listed under each cell for this secondary data analysis reflects the larger quasi-experimental study design. To elaborate, Tier 2 instruction, by design, includes a small group of three to four students. The number of students listed under the Tier 1 sample includes students who were at risk as well as those who were typically developing, all of whom were receiving Tier 1 instruction only. As part of the larger quasi-experimental study, on average, only three to four students from each condition, that is, control (typically achieving), comparison (at risk but not receiving Tier 2), and intervention (receiving Tier 2), were included and randomly selected as part of the investigation.

Twenty-three schools participated with a total of 54 Tier 1 kindergarten classroom teachers and 48 Tier 2 interventionists instructing a total of 553 kindergarten students. Based on an initial *Peabody Picture Vocabulary Test* (PPVT) assessment administered before the commencement of the study, 380 students were determined to be at risk of language and literacy difficulties (i.e., performing below a standard score of 92; Simmons et al., 2011) with 193 of those students randomly assigned to receive Tier 1 classroom instruction plus supplemental Tier 2 intervention. The remaining 187 at-risk students, plus 173 students scoring within the normal range of vocabulary scores (i.e., 95–105), received Tier 1 instruction only. This range of scores in students receiving Tier 1 instruction is consistent with the spread of scores seen in most Tier 1 classrooms (Gersten et al., 2009). Table 1 shows the number of students per site, school, and classroom. Approximately 23% of the sample students were White, 16% were Black, 36% were Latino/a, and less than 1% were Asian or Native American. More than 20% of the students were multi-racial. Student ages ranged from 4 years and 8 months to 6 years and 8 months for students receiving Tier 2 instruction and 4 years and 8 months to 6 years and 9 months for students receiving Tier 1 instruction only. There were no statistically significant differences in age between students receiving different degrees of instructional intensity. Teachers in this sample were approximately 70% White, 22% Black, and 8% Latino/a, and the interventionist sample was 62% White, 28% Black, 5% Latino/a, and 5% reported as "Other." In the Tier 1 teacher and Tier 2 interventionist sample, 88% and 33%, respectively, had an advanced degree. Tier 1 teachers were those already employed by their school, and Tier 2

interventionists were school designated interventionists including special education teachers, reading teachers/specialists, and paraprofessionals, depending on the site.

Procedures

Classroom teachers received 1 day of training with the research team before the start of the school year on the Elements of Reading Vocabulary Program (Beck & McKeown, 2004), which uses storybook reading to teach five weekly vocabulary words. Interventionists received 2 days of professional development on the program and booster trainings when needed.

All students in the core classroom participated in interactive lessons using words that are intended for more mature language users but capture concepts that students already understand. The program was implemented for 15 to 20 min per day, 5 days a week, for 20 weeks. Students participating in the Tier 2 supplementary program received a researcher-developed vocabulary intervention in groups of three to four students from the same class, 30 min a day for 4 days a week, in conjunction with the 20 weeks of the core classroom program (Coyne et al., 2015). During the small-group intervention, students engaged in more in-depth instruction on three of the five words learned that week in the core classroom. The more intensive instruction in small groups included explicit interventionist modeling, increased opportunities for corrective feedback, and additional scaffolding activities. These scaffolded activities included multiple opportunities to hear, say, and use the target words, as well as student-friendly definitions of target words accompanied by activities with examples and non-examples of the words used with pictures and manipulatives. Additional information about teacher and interventionist training, and the intensive supplementary vocabulary intervention, can be found in Cuticelli, Coyne, Ware, Oldham, and Loftus Rattan (2015) where the efficacy of these Tier 1 and Tier 2 programs on student curriculum vocabulary is supported.

Implementation fidelity was documented with both teachers and interventionists through two sets of observation occasions over the course of the 20-week intervention, at approximately the first and third quarter of the interventions. Mid-way through the intervention, teachers and interventionists completed the URP-IR. At the conclusion of the 20 weeks, students were administered both standardized language measures (measures also assessed at pre-test) and the curriculum-specific, researcher-developed vocabulary measure to assess student progress.

Measures

URP-IR. The URP-IR is an empirically validated self-report questionnaire to assess six factors (i.e., Acceptability, Understanding, Family–School Collaboration, Feasibility, System Climate, and System Support) that capture individual, intervention, and system influences that affect the quality of use and maintenance over time. This measure uses a 6-point Likert-type scale ranging from *strongly disagree* to *strongly agree* and was determined to be a reliable and valid measure of behavioral intervention usage with $\alpha = .95$ for Acceptability, $\alpha = .80$ for Understanding, $\alpha = .79$ for Family–School Collaboration, $\alpha = .84$ for Feasibility, $\alpha = .91$ for System Climate, and $\alpha = .72$ for System Support (Briesch et al., 2013). Two substantive adjustments were made to subscale items to make them specific to the study intervention. The first adjustment is that items were changed from the hypothetical language (e.g., "I would") used in the original study to reflect actual predicted intention (e.g., "I will"). The second adjustment is that item wording was changed to focus on vocabulary (e.g., "The intervention is an effective choice for addressing a variety of vocabulary difficulties" rather than "This intervention is an effective choice for addressing a variety of problems").

Intervention specific vocabulary knowledge. The primary outcome of interest, curriculum vocabulary, was measured by a researcher-developed expressive vocabulary assessment. At pre-test and post-test, the students in both groups were individually assessed on their knowledge of 20 words introduced in the *Elements of Reading* program. Each student was asked, "What does the word ___ mean?" Responses were given two points for a complete response, one point for a partial or related response, and zero point for an unrelated response or no response. Inter-rater reliability was set at 90% during the training process. This coding scheme has historically demonstrated high inter-reliability. In a previous study conducted by the authors using the same procedures, 20% of the assessment protocols that were scored by research assistants were randomly selected and independently scored by the project director with an agreement of 100% (Coyne et al., 2010). We chose to focus on curriculum-specific target measures because of our interest in vocabulary growth as a function of teachers' instruction. Therefore, we were specifically interested in growth in taught words. In the present study, the target vocabulary measure demonstrated acceptable reliability with the current sample ($\alpha = .84$) and demonstrated convergent validity with the Understanding Spoken Paragraphs ($r = .42$, $p < .01$) subscale from the Clinical Evaluation of Language Fundamental IV, a core subscale for evaluating students' mastery of word meanings (Semel, Wiig, & Secord, 2003).

The *Peabody Picture Vocabulary Test–4* (PPVT-4; Dunn & Dunn, 2007) is a commonly used standardized norm-referenced, individually administered test of receptive language and vocabulary. Each test item contains four illustrations. The student is asked to point to the picture that best represents the meaning of the word presented by the examiner. Test–retest reliability for the PPVT-4 is .77, and

alternate form reliability is .82. The PPVT-4 was used to determine students' eligibility for the supplementary intensive intervention.

The *Expressive Vocabulary Test–2* (EVT-2; K. T. Williams, 2007) is a standardized, norm-referenced, individually administered test of expressive language and vocabulary. Students view a picture and are asked to respond with a one-word answer to a question prompt (e.g., to provide a label or synonym for a picture). Test–retest reliability is .95, and the alternate form reliability is .87. The split-half reliability is .94. The EVT-2 was 100% co-normed with the PPVT-4; their correlation is $r = .82$. The pre-test EVT-2 was used as a control variable to increase our confidence in our findings that we have accounted for individual differences that might influence the target vocabulary outcomes.

The correlations among the different vocabulary measures in the present study were positively correlated. The EVT-2 and PPVT-4 were strongly positively correlated at pre-test ($r = .66$) and post-test ($r = .66$) and were moderately positively correlated with the expressive curriculum-specific measure at pre-test ($r = .33$ and $r = .43$ for the EVT-2 and PPVT-4, respectively) and post-test ($r = .38$ and $r = .35$ for the EVT-2 and PPVT-4, respectively).

Procedural Fidelity Checklist. To assess teachers' and interventionists' fidelity of implementing the vocabulary intervention in Tier 1 and Tier 2 settings, we used an observation checklist used in previous research that was determined to be valid and reliable (Coyne et al., 2010). Inter-observer agreement was found to be a coefficient kappa-*n* of .93 for the vocabulary activities. Trained research assistants used this fidelity checklist to mark whether curriculum-specific activities were completed, materials used, and pedagogy implemented. There were four or five items per activity for a total of 16 or 20 items for the Tier 1 and Tier 2 settings, respectively. The protocol was tailored to reflect these aspects of the intervention and was scored based on the percentage of observed fidelity. An average implementation fidelity score across items, activities, and observations was calculated and used as a covariate. The teacher and interventionist versions demonstrated acceptable reliability ($\alpha = .87$ and $\alpha = .81$, respectively).

Data Analysis

Research Question 1: Validity and reliability of the URP-IR in an academic context. To explore the first research question regarding the reliability and validity of the URP-IR in an academic intervention context, data were analyzed to examine the hypothesized factor structure using confirmatory factor analysis procedures. Factor analyses were employed using weighted least squares means and variance adjusted (WLSMV) estimation techniques with MPLUS 6.11. This estimation technique, which treats items as categorical as opposed to continuous, was used for two reasons. First, WLSMV was used to ensure consistency with the procedures used in the original validation study of the URP-IR (Briesch et al., 2013). Second, an exploration of the item-level data indicated that item descriptive statistics were not skewed but approximately a handful of items were kurtotic (i.e., ratio > 3). WLSMV adjusts for non-normality and is a robust estimation that yields less biased chi-square statistics and standard errors (Finney & DiStefano, 2006). Thus, this estimator was used as an alternative estimation technique with an adjustment method for the non-normal data. Curriculum-specific expressive vocabulary knowledge was also examined for multivariate normality for students receiving Tier 1 and Tier 2 instruction. An exploration of the plots of normality and measures of skewness and kurtosis indicated that the data were approximately normally distributed, with the Shapiro-Wilk test indicating that the *p* value was greater than an alpha level of .05, and therefore the null hypothesis that the data came from a normally distributed population could not be rejected.

To address missing data issues, the pattern of missing variables was examined using descriptive and correlational statistics to identify whether a consistent classroom or student-level characteristic was predictive of missing data. After exploring these data, the "missingness at random" (MR) assumption did not appear to be violated. Thus, we used the full information maximum likelihood (FIML) method (Dempster, Laird, & Rubin, 1977). Less than 10% of data were missing for URP-IR scores and student outcome data, with only 2.7% and 7.8% missing, respectively. Reliability estimates were calculated using Cronbach's alpha to determine the potential reliability of the URP-IR's different subscales.

Research Question 2: Associations among URP-IR factors and Tier 1 and Tier 2 student performance. The second research question regarding the potential explanatory power of the URP-IR in predicting student vocabulary performance was analyzed using SAS version 9.3 with a three-level, multilevel model using the PROC MIXED statement nesting students within classrooms within schools to explore student performance outcomes (Level 1) as a function of both Level 1 predictors (student previous performance based on the EVT-2 and curriculum-based word knowledge at pretest) and Level 2 predictors (teacher and interventionist URP-IR scores and implementation fidelity) all nested within schools (Level 3). All variables at the student level and classroom level were centered at their (grand) means. Grand mean centering was chosen over group mean centering so that all fitted models would be mathematically equivalent to the raw score model (Raudenbush & Bryk, 2002). We wanted to appropriately adjust for student-level effects across the sample for the interpretation of classroom-level URP-IR effects on student achievement. Before estimating

models to explain student target word performance, we explored the intraclass correlations (ICCs) of the outcome and question predictors (i.e., the URP-IR subscales) to determine whether a three-level model was necessary.

Model building. After determining that there was significantly large between-school variability, our three-level model building commenced with an unconditional model followed by stepwise integration of models with our controls (previous performance on curriculum and standardized vocabulary measures) first, followed by our covariate (implementation fidelity), and finally our question predictor (URP-IR scores). All teacher-level interaction terms were explored; that is, models including individual interaction terms for implementation fidelity and each URP-IR subscale were fit. Interaction terms were not significant and thus were not included in subsequent models.

We conducted model building for our two samples. First, models were estimated for students receiving Tier 1 instruction alone with their core classroom teachers (i.e., normally achieving students and students who met the criteria for at-risk status but who were not randomly assigned to Tier 2 instruction). Second, models were estimated for those students who were determined to be at risk of language and literacy difficulties and were randomly assigned to receive Tier 2 with an interventionist. In the following, we present our results from the measurement portion of this investigation, descriptive statistics for our measures, and our final models.

Results

The URP-IR in Multi-Tiered Academic Settings

The six-factor model was explored using confirmatory factor analyses. Goodness-of-fit indices were examined to determine whether this six-factor model exhibited acceptable fit. We present multiple statistics to quantify the degree of model fit to further support the acceptable or unacceptable fit of the model. Specifically, we explored absolute fit statistics (χ^2), as well as a statistic that accounts for model parsimony (root mean square error of approximation [RMSEA]; Steiger & Lind, 1980), and finally a preferred statistic of comparable fit, the comparative fit index (CFI; Brown, 2006). In conjunction, these various fit indices provide additional confidence regarding the fit of the hypothesized model. We used the following criteria to evaluate the fit indices: a CFI > .90 (Bentler, 1992), and a RMSEA < .10 (Byrne, 2009). The six-factor model demonstrated acceptable fit, $\chi^2(105) = 677.66$, $\chi^2 / df = 1.87$, CFI = .94, RMSEA (90% confidence interval [CI]) = .09 [.08, .10] with our sample of 102 teachers and interventionists. We examined our modification indices for potential changes to the model for model improvement, but given the acceptable fit of the model and our aim to use the originally validated model, we decided to keep the model with all of its original items loading on the hypothesized factors (see Note 1). Figure 2 displays the conceptual measurement model for the URP-IR, showing the six factors and their respective component items. To further determine whether our model adequately fit the data, we explored a one factor model, $\chi^2(102) = 1237.25$, $\chi^2 / df = 3.28$, CFI = .84, RMSEA (90% CI) = .15 [.14, .16] that demonstrated worse fit than the six-factor model based on fit statistics and a corrected chi-square difference test (see Note 2), $\chi^2(15, N = 102) = 408.96$, $p < .0001$. A second-order latent factor model was also explored (the second-order factors included latent variables for individual, intervention, and systems). However, due to sample size constraints, the model would not converge. Given that the six-factor model exhibited a better fit based on our fit statistics, we proceeded using this model.

Each of the factor loadings was statistically significant (estimate to standard error ratio > 1.96). Inter-factor correlations were positively and significantly correlated with the exception of Family–School Collaboration and System Support, which were positively correlated with each other and negatively correlated with the other factors. The positive correlation between these two factors also emerged in the confirmatory factor analysis conducted in the original URP-IR validation study, that is, $r = .15$ $p < .01$ (Briesch et al., 2013). Briesch and colleagues' confirmatory factor analysis also found a negative correlation between System Support and all the other URP-IR factors. Accordingly, the negative correlation between the System Support subscale and the other subscales in the present investigation was expected, as higher scores on this subscale would indicate potentially negative attitudes toward the support provided for the intervention. It stands to reason that teachers who strongly agreed with the following System Support items, "I need additional professional development beyond that provided to implement this intervention" and "I need additional resources beyond those provided to carry out this intervention," are teachers who also found the intervention to be less feasible, acceptable, harder to understand, and felt the school climate was not hospitable to such an intervention.

The finding that the Family–School Collaboration subscale was negatively correlated with the other subscale was not expected because neither intervention included a family component. This subscale was the least correlated with other subscales in the original study, with vignettes that included a family component. However, in authentic school contexts, this factor may also capture school stakeholders' feelings about support broadly, with items such as "A positive home–school relationship is needed to implement this intervention" and "Regular home–school communication is needed to implement the intervention procedures" tapping a general sense that the intervention requires more than educational stakeholders were provided with, which they may

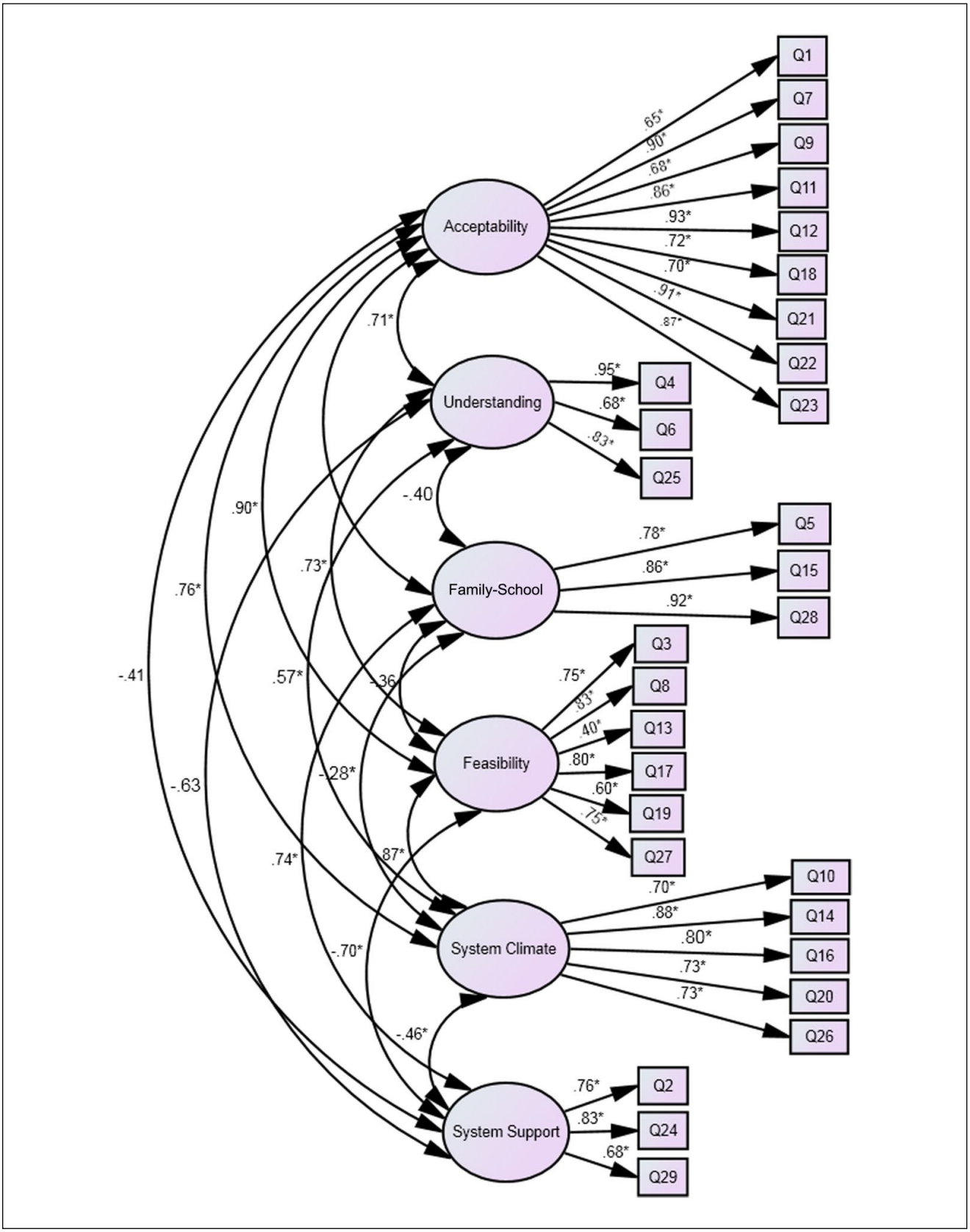

Figure 2. Confirmatory factor analysis conceptual model for the final measurement model.

attribute to an absence of family–school collaborations (i.e., it was needed but was not a built in part of the intervention). Teachers and interventionists who feel the intervention is usable as is will not feel an additional component, be it a Family–School support or another kind, is needed. Thus, while this finding was unexpected, the language in the Family–School Collaboration items regarding the absence of what is "needed to implement the intervention" supports the negative relationship between this subscale and the other subscales.

The variance components for each of the six factors were also statistically significant, indicating that the amount of variance accounted for by each factor was significantly different from zero. Due to the small sample size, we were unable to explore measurement invariance across teachers and interventionists; all estimated models did not converge. However, given that both groups were administered the same items and a model including both groups converged with acceptable fit, the present measurement model was used for further analyses (Little, 2013).

Reliability estimates for confirmatory factor analysis sample. Reliability coefficients for the six subscales are presented in Table 2. The majority of the subscales demonstrated acceptable levels of internal consistency reliability (range $\alpha = .71$–$.90$), with Understanding and Systems Support approximating acceptable levels (>.65). When disaggregating by core teachers and interventionists, most subscales were moderately correlated with one another, indicating that the subscales are distinct (see Table 3). Notably, the correlations between the URP-IR subscales and the fidelity measure were neither significant nor highly associated in either the Tier 1 instruction only or Tier 2 intervention context.

Descriptive Statistics for the URP-IR in Multi-Tiered Settings

Before examining the multi-level models exploring students' vocabulary performance, we first examined the descriptive statistics for student vocabulary performance (see Table 4). By design, students in the Tier 2 setting (participants determined to be at risk for language and literacy difficulties) scored, on average, below students in the Tier 1 only setting, across initial vocabulary measures. With regard to the URP-IR subscales, teachers and interventionists on average reported that they "agreed" with statements regarding intervention Acceptability, Understanding, Feasibility, and System Climate across instructional settings, with interventionists in the Tier 2 setting "strongly agreeing" with their understanding of the intervention features. As might be expected given the design of the intervention, teachers and interventionists, on average, reported lower scores on the Family–School Collaboration and Systems Support subscales. The procedural fidelity in the Tier 1 and Tier 2 settings was on average very high, .83 and .95, respectively.

Explanatory Power of the URP-IR for Student Performance in Tier 1

To explore the potential association between the URP-IR and student performance on the curriculum vocabulary measure, several multi-level models were fit to the classroom level only (Tier 1) data which included 54 teachers and 380 students. As a preliminary step to understanding potential differences between schools and within schools (i.e., between classrooms) on the question predictors (URP-IR subscales), ICCs were estimated for each variable before beginning our model building.

Table 5 displays the within- and between-school residuals as well as the ICCs for the variables of interest. Of particular importance for the present investigation is that the majority of the variance (i.e., > 50%) in each of the system-level variables (System Support and System Climate) was predominantly attributable to between-school differences. This finding supports the ecological validity and use of the URP-IR to capture school-level factors. The majority of the variance in the system-level subscales was attributable to between-school differences. However, the other subscales varied in whether their variance was predominantly attributable to between-school (Acceptability and Feasibility subscales) or within-school variance (Understanding and Family–School subscales).

Given the large amounts of between-school variance on the question predictor, we proceeded with model building using a three-level model. The first model to be fit, the unconditional model, revealed an ICC of .28; that is, 28% of the variance in target vocabulary knowledge post-intervention was attributable to between-school differences. Furthermore, 72% of the variance in target vocabulary knowledge post-intervention was attributable to within-school differences. This high degree of within-school variability increases the importance of exploring student and classroom-level variables. Standardized vocabulary knowledge and pretest target word knowledge were included in the model because, when within- and between-teacher ICCs were explored, 12% of the variance in the target outcome was attributable to between-teacher differences and 88% was attributable to student-level variables.

Model building began with control variables, followed by the central covariate, and finally the question predictors. At each stage that additional variables were added, potential interaction terms were explored. For a priori hypothesis testing, statistical significance was set at $p < .05$. None of the tested interaction terms were found to be statistically significant. A few of the URP-IR subscales were highly correlated ($r \geq .60$; Tabachnick & Fidell, 2013). These

Table 2. Summary of Reliability Statistics for the URP-IR.

Subscale	Items	Description	α
Fidelity	1–20 Interventionist 1–16 Teacher	Assesses fidelity of implementing the vocabulary intervention using an observation checklist	.81 .86
Individual			
Acceptability	1. This intervention is an effective choice for addressing a variety of vocabulary difficulties. 7. The intervention is a fair way to address the child's vocabulary difficulties 9. I am not interested in implementing this intervention. 11. I have positive attitudes about implementing this intervention. 12. This intervention is a good way to address the child's vocabulary difficulties 18. I will implement this intervention with a good deal of enthusiasm. 21. This intervention will not be disruptive to other students 22. I am committed to carrying out this intervention. 23. The intervention procedures easily fit in with my current practices.	Assesses how acceptable the participant finds the intervention to be and how enthusiastic the participant is about implementing it.	.90
Understanding	4. I understand how to use this intervention. 6. I am knowledgeable about the intervention procedures. 25. I understand the procedures of this intervention.	Assesses the extent to which participants understand the intervention procedures and how to implement the intervention.	.68
Intervention			
Family–School	5. A positive home–school relationship is needed to implement this intervention. 15. Parental collaboration is required to use this intervention. 28. Regular home–school communication is needed to implement the intervention procedures.	Assesses the extent to which participants believe that family–school collaborations are necessary for the intervention to be successfully utilized.	.84
Feasibility	3. I will be able to allocate my time to implement this intervention. 8. The total time required to implement the intervention procedures is manageable. 13. Preparation of materials needed for this intervention will be minimal. 17. Material resources needed for this intervention are reasonable. 19. This intervention is too complex to carry out accurately. 27. The amount of time required for record keeping will be reasonable in this intervention.	Assesses whether participants felt that implementing the intervention as described was feasible, given existing demands.	.71
System			
System Climate	10. My administrator will be supportive of my use of this intervention. 14. Use of this intervention will be consistent with the mission of my school. 16. Implementation of this intervention is well matched to what is expected in my job. 20. These intervention procedures are consistent with the way things are done in my system. 26. My work environment is conducive to implementation of this intervention.	Assesses whether participants felt the intervention is compatible with the school environment.	.78
System Support	2. I will need additional resources beyond those provided by the EVI team to carry out this intervention. 24. I will need consultative support beyond that provided by the EVI team to implement this intervention. 29. I will require additional professional development beyond that provided by the EVI team to implement this intervention	Assesses whether participants felt they needed external support to use the intervention.	.65

Note. The Procedural Fidelity Checklist has two Cronbach's alpha coefficients: One for the interventionist version and one for the teacher version. URP-IR = Usage Rating Profile–Intervention Revised; EVI = Early Vocabulary Intervention.

Table 3. Correlations Among Tier 1 Teachers' and Tier 2 Interventionists' Measures.

Measure	Fidelity	Acceptability	Understanding	Family–School	Feasibility	System Climate	System Support
Fidelity	1.00	−.07	−.06	−.04	.36	.15	−.18
Acceptability	−.10	1.00	.53**	−.11	.71**	.62	−.18
Understanding	−.03	.46**	1.00	−.16	.37*	.20	−.27
Family−School	.07	−.23	−.27	1.00	−.20	−.07	.51**
Feasibility	.26	.60**	.41**	.26	1.00	.60**	−.43**
System Climate	−.07	.52**	.33*	−.30*	.60**	1.00	−.33**
System Support	−.17	−.51**	−.31*	.55**	−.50**	−.37*	1.00

Note. Correlations among teacher measures are displayed above the diagonal and interventionists correlations among measures are portrayed below the diagonal.
***$p < .0001$.

Table 4. Descriptive Statistics For All Kindergarten Classrooms.

	Students receiving Tier 1 Only (n = 360)		Students receiving Tier 1 + Tier 2 (n = 193)		
Variable	*M*	*SD*	*M*	*SD*	*t* test
Student-level variables					
Pre-general vocabulary knowledge					
Pre-PPVT-4	90.88	10.13	82.12	5.7	12.98***
Pre-EVT-2	93.53	11.17	87.33	9.54	6.74***
Pre-target word knowledge	1.44	1.91	0.93	1.47	3.30**
Female	0.50	0.50	0.45	0.50	1.11
Teacher-level variables					
Fidelity	0.83	0.21	0.95	0.07	−3.47***
Acceptability	4.71	0.76	5.21	0.52	−7.74***
Understanding	5.11	0.59	5.58	0.40	−6.93***
Family−School collaboration	3.42	1.20	2.95	1.13	3.21**
Feasibility	4.87	0.62	5.15	0.53	−3.53**
System Climate	4.78	0.63	4.97	0.73	−3.76***
System Support	2.12	0.84	1.90	0.70	0.359*
Female	0.94	0.23	0.95	0.23	−0.081

Note. URP-IR subscale scores ranged from 1 to 6. All students were in kindergarten classrooms. Students in Tier 1 classrooms were composed of those at risk not participating in Tier 2 instruction and normally achieving students. Students in the Tier 1 + Tier 2 classrooms are those that are at risk and receiving Tier 2 instruction. Students receiving Tier 2 intervention who are at risk are explored as a separate sample in subsequent analyses and students only receiving Tier 1 (at risk and normally achieving) are considered as a separate sample in subsequent analyses. PPVT = *Peabody Picture Vocabulary Test*; EVT = *Expressive Vocabulary Test*.
***$p < .0001$.

Table 5. Intraclass Correlation Coefficients Across URP Subscales for Teachers in the Tier 1 Context and Interventionists in the Tier 2 Context.

	Tier 1			Tier 2		
Variable name	Between school	Within school	ICC	Between school	Within school	ICC
Post-target word	21.02	53.85	.28	49.34	62.29	.44
Acceptability	0.47	0.14	.77	0.22	0.11	.67
Understanding	0.18	0.21	.46	0.10	0.06	.63
Family–School	0.61	0.99	.38	0.55	0.66	.45
Feasibility	0.34	0.14	.70	0.15	0.16	.48
System Climate	0.36	0.13	.73	0.34	0.26	.57
System Support	0.44	0.36	.55	0.27	0.25	.53

Note. URP = Usage Rating Profile; ICC = intraclass correlation.

Table 6. The Final Multi-Level Models for Post-Intervention Expressive Target Word Measure for Tier 1 and Tier 2 Groups.

	Tier 1	Tier 2
Variable	Final model	Final model
Fixed effects		
Intercept	9.98 (0.77)***	14.42 (1.54)***
EVT	0.24 (0.03)***	0.26 (0.07)***
Pre-target vocabulary	0.98 (0.20)***	0.30 (0.44)
Implementation fidelity		26.04 (16.91)
Acceptability		
Understanding		
Feasibility		5.00 (2.37) *
System Climate	3.13 (1.04)**	0.54 (2.1)
System Support		0.72 (2.21)
Random effects		
Level 1	30.18 (2.79)**	41.28 (6.29)***
(Class)	7.91 (3.60)**	20.93 (13.09)*
(School)	5.82 (4.33)~	34.69 (20.42)*
Goodness-of-fit		
−2LL	1,793.3	908.4

Note. EVT = *Expressive Vocabulary Test* ; −2LL = −2 log likelihood.

strong correlations are consistent with both theory (e.g., feasibility is believed to be highly predictive of acceptability) and previous empirical work involving the URP-IR; however, for model building, these correlations have the potential to inflate standard errors and CIs. Therefore, to avoid unstable and potentially untrustworthy estimates of the parameter coefficients for individual parameters (M. N. Williams, Grajales, & Kurkiewicz, 2013), we only retained URP-IR factors that were statistically significantly correlated with the outcome in models that included controls (Chatterjee & Hadi, 2012). The implementation fidelity measure was not statistically significant in a model that solely contained the vocabulary control variables. In addition, when estimated in a model with all the URP-IR subscales, a chi-square difference test indicated that the removal of the fidelity measure did not make for a worse fitting model ($\Delta\chi^2 = .010$, $df = 1$, $p < .97$). As such, it was removed from the final model. The final multi-level model included both control predictors of students' pre-intervention EVT scores ($b = .24$, $SE = .77$, $p < .001$) and pre-target vocabulary scores ($b = .98$, $SE = .20$, $p < .001$). In Table 6, we present our final model for Tier 1.

The System Climate factor, which measures the degree to which teachers felt the intervention was compatible with their school environment, was the only factor that explained significant variability in performance ($b = 3.13$, $SE = 1.04$, $p < .001$). The final model's goodness-of-fit statistic (−2 log likelihood), 1793.3, provided a parsimonious model. These findings indicate that, for every 1 point higher on the System Climate subscale of the URP-IR, students scored an average of 3.13 points higher at post-test when controlling for students' pre-test vocabulary. That is, in schools with teachers who tended to overall "agree" with items on the System Climate subscale, items such as "my work environment is conducive to implementation of this intervention" had students who on average performed 3.13 points (out of 20 total points) higher than their peers in schools with teachers who tended to "slightly agree" with these same items.

Explanatory Power of the URP-IR for Student Performance in Tier 2

Similar to the preliminary analyses conducted in the Tier 1 setting, we conducted initial analyses to explore the ICCs of the URP-IR subscales to further examine the appropriate multi-level model to fit to the data. Table 5 shows the within and between-school residuals as well as the ICCs for the variables of interest in the Tier 2 setting. Compared with the Tier 1 setting, the system-level variables (System Support and System Climate) exhibited large between-school differences, with more than 50% of the variance in the URP-IR system-level variables attributable to school-level differences. This is in contrast to the individual- and intervention-level subscales that varied in whether their variance was predominantly attributable to between school (Acceptability and Understanding subscales) or within-school variances (Feasibility and Family–School subscales). To explore the potential association between the URP-IR and student performance on the target vocabulary measures in the Tier 2 intervention context, the same basic multi-level models were fit to the Tier 2 classroom-level data, which included 48 interventionists and 193 students. Similar to model building in the Tier 1 context, the first model to be fit was the unconditional model. This model revealed an ICC of .44; that is, 44% of the variance in target vocabulary knowledge post-intervention was attributable to between-school differences. In other words, in these more intensive instructional environments, there was less within-school variability in student performance at post-test than in the less intensive instructional environments which had more within-school variability. Similar to the Tier 1 model, student-level controls including the standardized vocabulary knowledge measure as well as pretest target word knowledge were included in all models because when within- and between-teacher ICCs were explored, 19% of the variance in the target outcome was attributable to between-teacher differences and 81% was attributable to student differences.

The fixed effects in the hypothesized model for the Tier 2 contexts contained the same predictors and were tested using the same set of control variables, covariates, and question predictors. Consistent with the Tier 1 setting, none of the tested interaction terms were found to be statistically significant. To avoid issues of multi-collinearity and present a parsimonious model (as described in the Tier 1 model building),

we only retained URP-IR factors that were statistically significantly correlated with the outcome in models that included controls (i.e., System Climate and System Support). In Table 6, we present our final model. The final multi-level model included both control predictors of students' pre-intervention EVT scores ($b = .26$, $SE = .07$, $p < .0001$) and pre-target vocabulary scores ($b = .30$, $SE = .44$, *ns*). Although the pre-target vocabulary score was not found to be statistically significant, we elected to keep this variable in the model, based on a theoretical rationale. The fidelity measure was not statistically significant ($b = 26.04$, $SE = 16.90$, $p <$ *ns*) but was correlated with the outcome and a chi-square difference test indicated that the fidelity parameter should be retained ($\Delta\chi^2 = 3.8$, $df = 1$, $p < .05$); thus, it was also included in the final model. Of the URP-IR factors, Feasibility was statistically significant ($b = 5.00$, $SE = 2.37$, $p < .05$) in predicting student performance. The final model's goodness-of-fit statistic (−2 log likelihood), 908.4, provided a parsimonious model. These findings indicate that for every 1 point higher on the Feasibility subscale of the URP-IR, students scored an average of 5 points higher on curriculum-specific vocabulary at post-test, when controlling for initial vocabulary knowledge pre-intervention. In other words, interventionists who tended to overall "agree" with items on the Feasibility subscale items such as "I would be able to allocate my time to implement this intervention" had students who on average performed 5 points (out of 20 total points) higher than their peers in classrooms with interventionist who tended to "slightly agree" with the same feasibility items.

Discussion

The present study explored the relationship between teachers' self-reported facilitators and barriers to intervention use that exist at the individual, intervention, and system level and student vocabulary outcomes in a multi-tiered instructional context. First, this study examined the reliability and validity of an ecological instrument of social validity, the URP-IR, in an academic multi-tiered setting. Second, this study explored whether there were different associations among factors at these multiple ecological levels (i.e., individual, intervention, system level), with particular focus on the potential association between system-level factors, and student vocabulary performance for Tier 1 and Tier 2 teachers and interventionists.

Reliability and Validity of the URP-IR in Academic Settings

Results from the present study provide modest, yet promising, support that the six-factor URP-IR is a valid assessment in early vocabulary intervention contexts. Confirmatory factor analysis indicated that the factor structure identified previously in studies on behavioral interventions also held in the present study (Briesch et al., 2013). A model that included Acceptability and Understanding (i.e., individual-level factors), Feasibility and Family–School Collaborations (intervention-level factors), as well as System Climate and System Support (school system-level factors) demonstrated acceptable model fit. The Systems Support and Understanding subscales exhibited lower reliability, $\alpha = .65$, and $\alpha = .68$, respectively, than the other four subscales (i.e., $\alpha > .70$). One explanation for the lower levels of reliability for these subscales is that they were composed of only three items compared with the five- to nine-item compositions of other subscales. Of note is that the valid application of these subscales is a function of the nature of the intervention, with some subscales being less relevant for certain interventions (e.g., the Family–School Collaborations subscale in the present analysis).

The URP-IR was designed to be an ecologically valid measure of factors that contribute to the quality of use and maintenance over time. The potential sensitivity and applicability of this ecological measure also emerged in our exploration of the amount of subscale variance accounted for by between-school and within-school differences. Findings revealed that for those subscales conceptualized as assessing individual- or intervention-level variables, a large percentage of variance was attributable to within-school variability (although the proportion varied across Tier 1 and Tier 2 settings). By contrast, the greater part of the variance in the systems subscales was attributable to between-school differences, as would be expected.

Associations Among the URP-IR and Student Outcomes

Subscales from the URP-IR were associated with curriculum-specific vocabulary performance for students receiving Tier 1 instruction only and students receiving Tier 1 instruction and Tier 2 intervention. The URP-IR subscales that significantly explained additional variance in student performance differed as a function of the instructional setting.

An important finding in the present study is that factors hypothesized to influence use at the individual, intervention, and system levels may be more or less predictive of student outcomes, as a function of the degree of instructional intensity. In the less intensive Tier 1 setting, the only significant predictor of student performance was a system-level factor when controlling for students' previous vocabulary performance. Specifically, teachers' perceptions of the compatibility of the intervention with their school context, a more philosophical dimension of support, helped to explain student performance. System-level factors were not found to predict student performance in the Tier 2 context. A potential explanation for the different contributions of system-level factors across tiers may be that the Tier 2

interventionists at some schools were less connected to the school system or had more peripheral roles in the community than the Tier 1 teachers and thus were less affected by the system's values, beliefs, and philosophies. Therefore, it is possible that these findings reflect the different roles and responsibilities of these educational stakeholders beyond just reflecting the demands of the intervention contexts.

Implementation fidelity was not found to be a significant predictor of student performance. However, we do not conclude from this that implementation fidelity is not a central ingredient for increasing performance, but instead that interventionists and teachers received extensive training as part of this program and thus were able to meet a more basic assessment of program fidelity. It is possible that a more sensitive measure assessing implementation beyond basic program features might have predicted additional variance in student performance at post-test. A possible alternative explanation for the absence of an association between implementation fidelity and performance despite a statistically significant relationship between the Systems Climate subscale and performance may be that in the case of core vocabulary curriculum, teachers who feel vocabulary instruction is valued by their school (high Systems Climate scores) may integrate vocabulary instruction throughout the day. These teachers may not only be "faithful" to the intervention during the scheduled intervention time, but as core kindergarten teachers who see their students throughout the school day and whose "dose" of instruction is relatively mild may also choose to emphasize vocabulary across contexts beyond the allotted program vocabulary time. The practices in the Tier 1 intervention involved less explicit instruction and word play than Tier 2, and thus these Tier 1 practices enacted during the program period of the day could easily have been adopted outside of the planned period. Indeed, studies of lower elementary education contexts indicate that language differences between students are a function of the amount and diversity of children's language experiences by teachers or caregivers (see review in Hoff, 2006).

In the Tier 2 contexts, only interventionists' perceptions of the feasibility of the intervention explained variance in student gains, when controlling for students' previous level of standardized and curriculum-specific vocabulary performance. One explanation for the lack of a significant relationship for the other subscales could be the consistently high levels of Acceptability and Understanding, as well as the high perceptions of System Climate and correspondingly low levels (negatively scored) of need for System Support. However, counterevidence for this explanation is that the Feasibility subscale exhibited a high mean score across the sample and a comparably small standard deviation (SD = .53).

The absence of a relationship between the two system-level factors and performance was counter to our original hypothesis regarding the importance of system-level factors in multi-tiered systems. However, the Feasibility subscale, while not considered by the developers of the URP-IR as a system-level subscale, does include items that under other taxonomies might be incorporated as system-level indicators. Sanetti and Kratochwill (2009), in their summary of the variables identified by scholars as influencing intervention implementation, include both "access to needed materials and supplies" and "daily time allocated for planning and implementation" as factors at the "organization level" (p. 449). Items in the Feasibility subscale such as "Preparation of materials needed for this intervention is minimal" or "I am able to allocate my time to implement this intervention" may more aptly serve as indicators conceptualized by Sanetti and Kratochwill as associated with the larger school organization, as opposed to simply reflecting intervention-level factors. As such, the significant association between interventionists' reported feasibility scores and student performance should also be viewed through a system-level framework because feasibility in such contexts may be inextricability linked to the general feasibility of educational practice in the confines of the school system. The finding that feasibility explained student performance is consistent with literature demonstrating that even effective treatments may be unsuccessful if they require substantial changes to teachers' normal routines within their system (Perepletchikova & Kazdin, 2005). Sensitivity to the magnitude of change has been described as a primary factor in successful implementation of an intervention and outcome gains (Riley-Tillman & Chafouleas, 2003). This relationship may be corroborated in our findings, which support the potential importance of this intervention-level/organization-level factor to performance, specifically in more intensive tiered contexts.

A potential pathway to explain the explanatory power of the Systems Climate and Feasibility subscales for student performance for teachers and interventionists, respectively, may be via teacher quality of delivery, and not strictly dosage or adherence to the intervention (as demonstrated by the absences of a relationship between the fidelity measure and student performance). It is reasonable that Tier 1 teachers who felt the intervention was more compatible with the school environment, and Tier 2 interventionists who felt the intervention was easy to implement within their given demands, had more buy-in, and perceived flexibility, which in the case of this multi-tiered context, may have led to increased quality of instructional delivery and by extension higher student performance. Dane and Schneider (1998) suggested that quality of delivery may be an important component of treatment integrity. To explore the veracity of this potential mechanism, future work should collect observational and teacher self-report data on interactive features of teachers' method of instructional delivery in conjunction with the URP-IR.

Limitations

Due to our analytic choice to focus on multi-tiered settings, we cannot completely disaggregate intervention intensity from student characteristics in the Tier 2 setting. All students who received Tier 2 intervention were identified as at risk and thus exhibited lower standardized vocabulary scores than their peers who were not at risk. Therefore, our conclusions regarding intervention usage factors that explain additional variance in students' Tier 2 performance reflect the intensity of the intervention, implementation, and the student population. It is unknown whether the explanatory power of URP-IR factors might differ as a function of student ability in Tier 2 contexts. However, given that schools are increasingly adopting multi-tiered frameworks of instruction, the present design is more compatible with the kinds of practice-based questions teachers and administrators will be asking when trying to establish what interventions are best suited for their local school context. The Tier 1 analysis does include students at risk and students not at risk, and thus allows this student characteristic to be separable from intervention setting.

We are cautious in our interpretations about our findings that only feasibility of the intervention explained significant variance in students' gains on the curriculum-specific vocabulary measure in the Tier 2 instructional context. To implement these vocabulary programs, we provided substantial intervention training/support to interventionists and developed strong relationships with our collaborators, which likely increased system-level resources and attitudes about the intervention. Although these behaviors bolstered the internal validity of the study, they also pose limitations for the generalization of our findings to other settings. These contextual variables may explain the lack of a significant relationship between several URP-IR factors and student performance. However, we believe that in the absence of significant relationships across several subscales, future work is warranted to explore ways of improving the factor structure of the URP-IR through the addition of supplementary items and item revisions.

Implications for Practice

Administrators, school psychologists, and others responsible for directing the implementation of interventions in schools should be sensitive to additional factors such as school's philosophical climate in addition to more traditional aspects of intervention implementation such as feasibility and integrity. Our findings support using the URP-IR for face-to-face consultations to examine concerns with interventions. For example, in the present investigation, teachers providing Tier 1 instruction, on average, reported lower levels of Acceptability than other subscales and lower levels than interventionists. This information is useful from a program evaluation perspective for potential revisions to the Tier 1 intervention to make it more acceptable to teachers. Moreover, this knowledge can be constructive at the classroom level as these data can facilitate conversations with teachers around adjustments that can be made in their classroom to suit their instructional needs. In addition, the present study contributes to the existing literature that has largely hypothesized a relationship between measures of social validity and performance by offering testing of the presumed relationship (Briesch et al., 2013). Furthermore, this study adds to the small literature base on intervention use in early intervention academic settings. The use of the URP-IR measure for academic settings offers teachers and researchers a potential tool to capture and disaggregate by level the factors that may prevent successful intervention use. Ecological assessments allow for more valid evaluations of system-level reform efforts that include multi-tiered approaches to instruction. Future research should explore how the URP-IR may be beneficial in planning and evaluating intervention efforts across different instructional tiers.

Authors' Note

The content and positions in this article do not necessarily represent those of the funding agency.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by Grant R324A110135, National Center for Special Education Research, from the Institute of Education Sciences (IES), U.S. Department of Education.

Notes

1. Of the modification indices explored, Item 23 exhibited a high modification index. While this item was maintained in the present analysis to examine the originally validated model, we would recommend that future work explore an improved model where Item 23 is modeled as an indicator of system-level constructs.
2. A corrected chi-square difference test, with the DIFFTEST option in MPLUS, was used. A traditional chi-square difference test cannot be used with models that use the WLSMV estimation technique, as indicated by the MPLUS User Manual and work conducted by Asparouhov and Muthen (2006).

References

Asparouhov, T., & Muthen, B. (2006). Robust chi-square difference testing with mean and variance adjusted test statistics. *Mplus Web Notes*, *10*, 1–5.

Bean, R., & Lillenstein, J. (2012). Response to intervention and the changing roles of schoolwide personnel. *Reading Teacher*, *65*, 491–501.

Beck, I. L., & McKeown, M. G. (2004). Direct and rich vocabulary instruction. In J. F. Baumann & E. J. Kame'enui (Eds.), *Vocabulary instruction* (pp. 13–27). New York, NY: Guilford Press.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York, NY: Guilford Press.

Bentler, P. M. (1992). *EQS structural equations program manual*. Los Angeles, CA: BMDP.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations. *Journal of Educational Psychology*, *93*, 498–520.

Briesch, A. M., Chafouleas, S. M., Neugebauer, S. R., & Riley-Tillman, T. C. (2013). Assessing influences on intervention use: Revision of the Usage Rating Profile. *Journal of School Psychology*, *51*, 81–96.

Broughton, S. F., & Hester, J. R. (1993). Effects of administrative and community support on teacher acceptance of classroom interventions. *Journal of Educational and Psychological Consultation*, *4*, 169–177.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Byrne, B. M. (2009). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York, NY: Taylor & Francis.

Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example* (5th ed.). Hoboken, NJ: John Wiley.

Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli, R., Ruby, M., Crevecoeur, Y., & Kapp, S. (2010). Direct & extended vocabulary instruction in kindergarten: Investigating transfer effects. *Journal of Research on Educational Effectiveness*, *3*, 93–120.

Coyne, M. D., McCoach, D. B., Ware, S., & Rattan, S. (2015, March). *Evaluating the efficacy of a supplemental kindergarten vocabulary intervention implemented within a tiered instructional framework*. Paper presented at the Society for Research on Educational Effectiveness, Washington, DC.

Cuticelli, M., Coyne, M. D., Ware, S. M., Oldham, A., & Loftus Rattan, S. (2015). Improving vocabulary skills of kindergarten students through a multi-tier instructional approach. *Intervention in School and Clinic*, *50*, 150–156.

Dane, A., & Schneider, B. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23–45.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*, 1–38.

Dickenson, D., & Caswell, L. (2007). Building support for language and early literacy in preschool classrooms through in-service professional development: Effects of the Literacy Environment Enrichment Program (LEEP). *Early Childhood Research Quarterly*, *22*, 243–260.

Dunn, D. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test: Manual*. Bloomington, MN: NCS Pearson.

Durlak, J. A., & Dupre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–350.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: Information Age Publishing.

Fuchs, D., Fuchs, L. S., & Vaughn, S. (Eds.). (2008). *Response to intervention: A framework for reading educators*. Newark, DE: International Reading Association.

Fuchs, D., Mock, D., Morgan, P., & Young, C. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for learning disabilities construct. *Learning Disabilities Research & Practice*, *3*, 157–171.

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. D. (2009). *Assisting students struggling with reading: Response to intervention and multi-tier intervention in primary grades*. Washington, DC: U.S. Department of Education Institute of Educational Sciences.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*, 55–88.

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

Kazdin, A. E. (1980). Acceptability of alternative treatments for deviant child behavior. *Journal of Applied Behavior Analysis*, *13*, 259–273.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.

Nagy, W., & Scott, J. (2000).Vocabulary processes. In M. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 269–284). Hillsdale, NJ: Lawrence Erlbaum.

National Center on Response to Intervention. (n.d.). Available from http://www.rti4success.org/

Noell, G. H. (2008). Research examining the relationships among consultation process, treatment integrity, and outcomes. In W. P. Erchul & S. M. Sheridan (Eds.), *Handbook of research in school consultation: Empirical foundations for the field* (pp. 315–334). Mahwah, NJ: Erlbaum.

Orne, C. M., & Binik, Y. M. (1989). Consistency in adherence among regiment demands. *Health Psychology*, *8*, 27–43.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*, 365–383.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Chicago, IL: Scientific Software International.

Reimers, T. M., Wacker, D. P., & Koeppl, G. (1987). Acceptability of behavioral treatments: A review of the literature. *School Psychology Review*, *16*, 212–227.

Riemer, M., Rosof-Williams, J., & Bickman, L. (2005). Theories related to changing clinician practice. *Child and Adolescent Psychiatric Clinics of North America*, *14*, 241–254.

Riley-Tillman, T. C., & Chafouleas, S. M. (2003). Using interventions that exist in the natural environment to increase treatment integrity and social influence in consultation. *Journal of Educational and Psychological Consultation*, *14*, 139–156.

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review*, *38*, 445–459.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4)*. Toronto, Ontario, Canada: Harcourt Assessment.

Simmons, D. C., Coyne, M., Hagan-Burke, S., Kwok, O., Simmons, L., Johnson, C., . . . Crevecoeur, Y. (2011). Effects of supplemental reading interventions in authentic contexts: A comparison of kindergarteners' response. *Exceptional Children*, *77*, 207–228.

Stahl, K. A. D. (2011). Applying new visions of reading development in today's classrooms. *Reading Teacher*, *65*, 52–56.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 360–407.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Sterling-Turner, H. E., & Watson, T. S. (2002). An analog investigation of the relationship between treatment acceptability and treatment integrity. *Journal of Behavioral Education*, *11*, 39–50.

Tabachnick, B. G., & Fidell, L. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Williams, K. T. (2007). *Expressive vocabulary test (2nd ed.): Manual*. Minneapolis, MN: Pearson Assessments.

Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*, *18*, 1–14.