Effects of a Year Long Supplemental Reading Intervention for Students with Reading Difficulties in Fourth Grade

Jeanne Wanzek

Vanderbilt University

Yaacov Petscher

Florida State University

Stephanie Al Otaiba

Brenna Rivas

Francesca Jones

Southern Methodist University

Shawn Kent

University of Houston

Christopher Schatschneider

Florida State University

Paras Mehta

University of Houston

Wanzek, J., Petscher, Y., Al Otaiba, S., Rivas, B. K., Jones, F. G., Kent, S. C., Schatschneider, C., & Mehta, P. (2017). Effects of a year long supplemental reading intervention for students with reading difficulties in fourth grade. *Journal of Educational Psychology*, 109, 1103-1119.

Author note

Jeanne Wanzek, Department of Special Education, Vanderbilt University; Yaacov Petscher, Florida Center for Reading Research, Florida State University; Stephanie Al Otaiba,

Simmons School of Education , Southern Methodist University; Brenna K. Rivas, Simmons School of Education, Southern Methodist University, Francesca G. Jones, Simmons School of Education Southern Methodist University; Shawn C. Kent, Department of Educational Leadership and Policy Studies, University of Houston; Christopher Schatschneider, Florida Center for Reading Research and Department of Psychology, Florida State University; Paras Mehta, Department of Psychology, University of Houston.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A150269 to Vanderbilt University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Correspondence should be addressed to Jeanne Wanzek, Department of Special Education, Vanderbilt University, 110 Magnolia Circle, Nashville, TN 37203.

Email: jeanne.wanzek@vanderbilt.edu

Abstract

Research examining effective reading interventions for students with reading difficulties in the upper elementary grades is limited relative to the information available for the early elementary grades. In the current study, we examined the effects of a multicomponent reading intervention for students with reading comprehension difficulties. We employed a partially nested analysis with latent variables to adequately match the design of the study and provide the necessary precision of intervention effects. We examined the effects of the intervention on students' latent word reading, latent vocabulary, and latent reading comprehension. In addition, we examined whether these effects differed for students of varying levels of reading or English language proficiency. Findings indicated the treatment significantly outperformed the comparison on reading comprehension (ES = 0.38), but no overall group differences were noted on word reading or vocabulary. Students' initial word reading scores moderated this effect. Reading comprehension effects were similar for English learner and non-English learner students.

Educational Impact and Implications Statement

This study examined the effects of a multi-component reading intervention for students with reading difficulties in fourth grade. Findings indicated students receiving the intervention made greater gains in reading comprehension than students who did not receive the intervention. This finding was similar for students who were English learners or non-English learners. However, students with higher initial word reading scores benefitted more from the intervention. These findings suggest students receiving the intervention made progress in closing the gap between their current level of performance and expected levels of performance in reading comprehension.

Effects of a Year Long Supplemental Reading Intervention for Students with Reading Difficulties in Fourth Grade

Students with reading difficulties can benefit from supplemental reading instruction provided in small groups; reading interventions at the elementary level have demonstrated power for preventing and remediating many reading difficulties (Blachman et al., 2004; Mathes et al., 2005; O'Connor, Fulmer, Harty, & Bell, 2005; Torgesen et al., 1999; Vellutino et al., 1996). However, research examining effective reading interventions for students with reading difficulties in the upper elementary grades is limited relative to the information available for the early elementary grades (Wanzek, Wexler, Vaughn, & Ciullo, 2010). The need for effective reading interventions for students with reading difficulties in the upper elementary grades is essential given the large numbers of students who continue to struggle with reading at these grade levels (National Center for Educational Statistics, 2016).

Reading Interventions for Upper Elementary Students

The research available on reading interventions related to upper elementary students with reading difficulties demonstrates positive effects for interventions providing instruction in comprehension or word recognition (Wanzek et al., 2010). Higher effects were noted for interventions related specifically to comprehension instruction. For example, large mean effects across comprehension measures were noted in two experimental studies of comprehension strategy instruction for students with reading difficulties (Mason, 2004; Miranda et al., 1997). However, the upper elementary research, including these comprehension interventions, has also largely examined intervention effects on proximal, researcher-developed measures. In fact, 15 of the 24 studies synthesized by Wanzek et al. (2010) employed only researcher-developed measures. Researcher-developed measures often result in higher effects than standardized

measures of the same constructs (Scammacca et al., 2007; Swanson, Hoskyn, & Lee, 1999). Thus, the lack of information on the effects of providing comprehension interventions on standardized measures represents a gap in the knowledge base on upper elementary reading interventions.

Additionally, Wanzek et al. (2010) reported that most of research thus far on upper elementary reading interventions for students with reading difficulties has been conducted with relatively brief interventions (e.g., 15 min sessions; less than 6 weeks) that examined single instructional strategies (e.g., main idea strategy). These studies provide important information regarding effective practices that could be incorporated in reading interventions to accelerate student learning. Knowledge of student outcomes when effective practices for various reading components are put together to form more comprehensive interventions for struggling readers is also needed.

In fact, some of the highest effects in the upper elementary reading intervention literature have come from multicomponent interventions (Wanzek et al., 2010). Though there are only a few of these studies in the literature (e.g., O'Connor et al., 2002; Ritchey, Silverman, Montanaro, Speece, & Schatschneider, 2012; Therrien, Wickstrom, & Jones, 2006; Vadasy & Sanders, 2008; Wanzek & Roberts, 2012), the findings suggest the possible importance of addressing multiple reading components in reading intervention for these older students. Three of these studies demonstrated moderate to large, significant effects on norm-referenced measures of comprehension or broad reading achievement (O'Connor et al., 2002; Therrien et al., 2006; Vadasy & Sanders, 2008). The effect sizes ranged from 0.37 to 1.87. The interventions in these studies included instruction in reading comprehension along with additional instruction in word reading (O'Connor et al., 2002), fluency (O'Connor et al., 2002; Therrien et al., 2006; Vadasy &

Sanders, 2008), and/or vocabulary (Vadasy & Sanders, 2008). The findings suggest students with reading difficulties at the upper elementary level may benefit most when interventions focus on multiple elements of reading, providing opportunities for students to integrate reading practices to read and understand text. In an earlier synthesis of interventions for students with learning disabilities, Swanson et al. (1999) reported the highest effects for interventions that combine direct instruction of content with strategy instruction. Most of the multiple component reading interventions conducted at the upper elementary level have incorporated both types of instruction. Several other syntheses for older students confirm the value of multi-component interventions (Kamil et al., 2008; Scammacca et al., 2007; Torgesen et al., 2007).

The previous research also suggests some differential effects for English learners (ELs) with reading difficulties relative to their non-EL peers (Kieffer, 2008). In particular, ELs are at a markedly greater risk of late-emerging (after Grade 3) reading difficulties (Kieffer, 2010; 2014), suggesting reading foundation skills such as word reading may be mastered more easily. But, many ELs may struggle later with understanding texts that have more complex syntax, vocabulary, or background knowledge needs. Previous fourth grade interventions have noted higher effects for ELs in reading intervention on word reading measures but not on comprehension or vocabulary measures (Wanzek & Roberts, 2012). Thus, examining the differential effects of ELs with a multi-component, comprehension focused reading intervention program could provide additional evidence regarding for whom a reading intervention is most valuable.

Passport to Literacy

One multi-component reading intervention that is widely used in schools across the United States is Passport to Literacy. Passport to Literacy is a packaged program that applies

principles of behavioral learning theory and cognitive psychology (Flavell, 1992; Palincsar & Brown, 1984), providing explicit instruction and strategies for reasoning in the foundational skills of reading (e.g., decoding, word reading) as well as reading comprehension and vocabulary. Semi-scripted lessons are built sequentially to help students acquire missing foundational reading skills, increase background knowledge, and build strategies for comprehending text.

Although Passport to Literacy is widely used, there is a lack of independent research on the program's effectiveness. We conducted one initial study of the Passport to Literacy intervention with fourth grade students (Wanzek et al., in press). This study was the first causal study conducted on Passport to Literacy and also the first to examine outcomes on standardized measures of reading achievement. Fourth grade students scoring below the 30th percentile in reading comprehension (n = 221) were randomly assigned to receive the standard implementation of the Passport the Literacy intervention or typical school services. The intervention was provided in small groups of four to seven students for 30 min, 4 days a week throughout the school year (M = 90.45 lessons). There were no effects for Passport to Literacy on standardized measures of word reading or fluency, but small effects were noted on standardized measures of reading comprehension (ES = 0.14 to 0.28). Exploratory analyses indicated the intervention effects differed by students' comprehension abilities. Students' exhibiting low levels of comprehension demonstrated no increased benefit of the Passport to Literacy standard intervention. In other words, the multi-component Passport to Literacy intervention demonstrated average increased outcomes on reading comprehension, but was least effective for students with the lowest comprehension levels.

In the current study, we build upon this previous study to examine the effects of Passport to Literacy with a larger sample. This larger sample allows for a more sophisticated analysis that matches the design of the study taking into account the differing clustering structures of the treatment and comparison groups. In addition, the larger sample allows us to be more precise in measuring student reading achievement through the use latent variables. By using latent variables, the impact and exploratory analyses reflect a stronger test of theory as effects are less due to assessment-specific outcomes and more to the theoretical overlap among them. Finally, the larger sample included a large enough sample of ELs to examine other possible associations that may explain the differential effects noted in the first study.

Study Purpose

The purpose of this study was to examine the effects of the standard implementation of the Passport to Literacy intervention for students with reading comprehension difficulties. We sought to examine the effects of this multi-component intervention on students' word reading, vocabulary, and reading comprehension. In addition, we examined whether these effects differed for students with varying levels of reading or English language proficiency. Specifically, we examined:

1) What are the effects of Passport to Literacy on students' word reading, vocabulary, and reading comprehension?
2) Do these effects differ by initial reading achievement or English language level?

Based on the previous study of the intervention, we hypothesized that students with reading difficulties receiving the Passport to Literacy intervention would outperform students receiving typical school services in reading comprehension and not in word reading or vocabulary. We also hypothesized that students with higher initial levels of reading achievement

on word reading, fluency, or comprehension would benefit more from the intervention. Based on previous reading intervention work for ELs we hypothesized more benefits of the multi-component intervention for ELs on word reading outcomes than for their non-EL peers.

Method

Participants

Four hundred fifty-one fourth-grade students who scored at or below the 30th percentile on the reading comprehension subtest of the Gates-MacGinitie Reading Tests (GMRT; MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2006) were selected for the study. The students came from 16 public elementary schools located across six school districts in three states. One school district was located in a large, urban metropolitan area; one district was located in a mid-size city; and four districts were located in rural areas. Male students made up 49% of the sample. With regards to ethnicity, 46% of the students were identified as Hispanic. Of those who reported language status, 13.2% of the total sample was flagged as having a primary language other than English or as currently receiving EL services. All schools provided only instruction in English. The racial composition of the sample was 35% Black, 44% White, 17% American Indian, 1% Asian, and 2% multiracial. Eighty-five percent of the students qualified for low income or free or reduced lunch programs. Fifteen percent were identified as having a disability. The majority of students with a disability were identified with a learning disability or a speech/language disability. There were no differences in any of the demographics between the two study groups.

A total of 40 students (9% of total sample) withdrew from their respective schools after the screening test. Attrition was 12% (n = 27) in the treatment group and 6% (n = 13) in the comparison group. By applying guidelines set forth by What Works Clearinghouse (2014), it

was observed that the overall attrition of 9% and differential attrition of 6% falls into a category of *low attrition*, which is operationalized as a condition where the balance between overall and differential attrition, "…is expected to result in an acceptable level of bias even under the conservative assumptions" (pg. 12).

Procedures

Screening and assignment. Research staff screened all consented fourth grade students at the 16 schools during the fourth or fifth week of school using the reading comprehension subtest of the GMRT. All students scoring at or below the 30th percentile on this measure were identified for the study and randomly assigned within school to treatment (Passport; n = 226) or comparison (n = 225) using stratification on the screening measure.

Students assigned to the treatment group were subsequently assigned within school to small groups of four to seven students (a total of 43 groups across schools). Each treatment group received the Passport to Literacy intervention daily for 30 min sessions for 25 weeks. Students assigned to the comparison group received the typical services provided by the school.

Data collection. Following screening, pre-test measures were administered at the end of September and beginning of October to all participants. Post-test assessments were administered in early May, within 2 weeks of the intervention completion. Assessments were counterbalanced by measure and were administered by trained research assistants blind to condition and assignment. Prior to pretesting and post-testing, assessment staff were required to demonstrate 100% accuracy in administration and scoring on all measures. Further, all measures were double-scored and double-entered by two, independent research staff.

We observed students' school provided reading instruction. First, we collected data on students' core, classroom reading instruction (tier 1) in the fall and in the spring in order to

understand the type and amount of reading instruction students received in their classrooms. Observers were trained to use the Instructional Content Emphasis Instrument-Revised (ICE-R; Edmonds & Briggs, 2003) to record what was taught, how long it was taught, and the instructional grouping used for teaching. Following the guidelines of the ICE-R, specific instructional activities were coded if they lasted for at least 1 min. Content categories included phonemic awareness, phonics/word recognition, fluency, vocabulary/oral language development, comprehension, spelling, text reading separate from other instruction, and non-literacy activities (e.g., other academic instruction, non-instructional time). Observers also coded instructional groupings as whole class, small-group, pairs, independent activity/assignment, or individualized instruction. Student engagement for the overall observation was coded using a three point rubric (3 = high engagement, 1 = low engagement). Finally, observers assigned a global quality of instruction rating for the overall observation based on a 4-point Likert scale ranging from weak (rating of 1) to excellent (rating of 4). This global instructional quality variable considered a teacher's use of direct and explicit language, modeling, students' opportunities for practice, specific feedback, monitoring and encouragement of engagement, scaffolding of tasks, and pacing throughout the lesson.

We used a multiple-step training process to establish interrater reliability for the tier 1, classroom reading instruction observations in fall and again in the spring before each round of observations began. Initially, each observer was instructed on the meaning of each code/indicator and provided specific examples. Next the coding process was modeled by the principal investigator of the project using a short video segment of reading instruction from another project. Finally, each observer practiced coding using several novel video segments that were subsequently discussed with the principal investigator. Each observer established 90% or

higher coding accuracy with the principal investigator (i.e., gold standard approach) on a separate video segment of reading instruction. Observers reestablished reliability prior to spring observations with new video segments. All coders were required to be above 90% reliability at each time point. Exact interrater reliability across coders and time periods was 95.1%.

To identify any supplemental reading instruction/intervention, research staff completed brief interviews with classroom teachers regarding additional reading support beyond core reading instruction for each participating student. Each semester teachers indicated the session time, frequency, grouping, implementer, and implementer's credentials. All supplemental intervention sessions in both study conditions were audio recorded at three time points during the school year (fall, winter, and spring); recordings of instruction were then coded using the ICE-R measure to describe any interventions students received.

In addition, the fidelity of implementation of the Passport to Literacy intervention was monitored monthly via direct observations of lessons with a measure specific to the required components of the Passport to Literacy intervention. Interventionists were observed and scored on implementation of each activity, student academic engagement, and quality of instruction for each lesson component. The scale for implementation ranged from 0 (teacher did not complete elements of component) to 3 (all or nearly all required elements completed), while engagement and instructional quality were also rated from 1 (*weak engagement or quality*) to 3 (*excellent engagement or quality*). Instructional quality indicators included ongoing monitoring, redirection of off-task behavior, positive and corrective feedback, organization of materials, and appropriate selection of additional items for practice when needed. Each observer obtained a minimum reliability of 90% in comparison to a gold standard rating by the project coordinator prior to formal data collection; across three observers, reliability was 95.3%.

Description of Instruction

Tier 1, classroom reading instruction. Data from observations of core reading instruction received by all participating students indicated that the length of reading classes was, on average, 75.40 min (*SD* = 26.34). Within this instruction, activities devoted to reading comprehension and vocabulary development were most prevalent, accounting for nearly 35 min (46%) of total time. Instruction devoted to word analysis/decoding was minimal (< 1 min [< 1% of time]), while time spent in reading of connected text and/or reading fluency practice was approximately 9 min (12% of time) daily. Of note, approximately 15 min (20% of time) was spent in differentiated instructional activities where students in the class were engaged in different activities simultaneously. The additional 14 min (19%) of time was spent in other types of activities (e.g., transitions). Core reading instruction primarily occurred as whole-class instruction (approximately 45 min or 60% of time on average). Just less than 10 min (13%) of instructional time consisted of students working independently on the same activity, while approximately 8 min (11%) was spent in either small-group or paired instructional activities. Generally, the global ratings of instruction for the core classroom instruction were suggestive of high average instructional quality (*M* = 3.17, *SD* = .59). Similarly, academic engagement by students during core reading instruction was rated as high (*M* = 2.78, *SD* = .55).

School-provided supplemental instruction. A total of 130 students (*n* = 62 treatment [27%]; *n* = 68 comparison [30%]) also received supplemental intervention provided by their respective schools for all or part of the year. Teacher reports indicated that this supplemental reading intervention was most often delivered by classroom teachers (20%) or other certified teachers (43% of students) with eight interventions (18%) delivered by a paraprofessional or a volunteer, and 6 interventions (14%) delivered by speech-language pathologists. Interventions

most often held sessions between 31-50 min (70%) with 16% of the interventions meeting between 21-30 min and 10% between 10-20 min. Seventy percent of the interventions were held in group sizes of one to five students. Nine students received two supplemental interventions during the school day.

Across the two years, based on recordings of this instruction, intervention sessions averaged 28.34 min (SD = 13.78). The most frequent instructional activities involved those related to comprehension of text (*M* = 8.27 min, *SD* = 7.60) with about 29% of intervention time, as well as vocabulary and oral language development (*M* = 4.45 min, *SD* = 5.90) for about 16% of intervention time. Text reading without other instruction occurred for approximately 6 min (*M* = 6.43 min, *SD* = 5.1) or 23% of intervention time, and students received phonics/decoding instruction for an average of 3.84 min (*SD* = 7.86) or 14% of intervention time. Minimal instruction (0-4% of intervention time) was focused on oral reading fluency practice (*M* = .53 min, *SD* = 1.71), spelling (*M* = 1.22 min, *SD* = 3.27), or phonemic awareness (*M* = .04 min, *SD* = .23). During the additional reading intervention, an average of 1.86 min (*SD*=3.74) or 7% of instructional time was spent in other academic instruction. About 4% of the intervention time was spent in non-instructional activities (*M* = 1.04 min, *SD* = 3.68). The mean rating of instructional quality for students who received supplemental reading instruction was 2.83 (*SD* = .47) and student engagement was also high (*M* = 2.65, *SD* = .36). Table 1 provides information on this typical school instruction in comparison to the treatment intervention sessions.

Passport to Literacy intervention. We provided the standard implementation of the Passport to Literacy intervention program at the fourth-grade level to students in the treatment condition. Passport to Literacy is designed to be used as a supplemental reading intervention provided in small groups daily for 30 min sessions for 1 school year (up to 120 lessons). We

scheduled the intervention sessions with the school/teachers outside of their core, classroom reading instruction block, typically during the time that schools had already designated for intervention/enrichment.

The Passport to Literacy intervention is broken into 12, 10-day adventures, with each lesson targeting phonics and word recognition, fluency, vocabulary, and comprehension. To monitor students' mastery of content and progress on oral reading fluency, checkpoints are designed at the fifth and 10th lesson of each adventure. The sequence of instruction began with an *Adventure Starter* activity (approximately 3 -5 min) to build background knowledge by linking the lessons and readings to the adventure. Then, lessons included two major components; the first, *Word Works*, or word study, taught students to read and understand unknown multisyllabic words using strategies to break words down into smaller parts, including affixes, roots, and syllabication. For the first six weeks, the Word Works instruction was 20 min and also included more basic word reading skills such as letter/sound identification, decoding, sight word reading, word families, and spelling instruction. In subsequent lessons, Word Works was reduced to 5 min, but also included a brief 2 min *Warm-Up* where students received additional word study practice through review and application of previously learned letter combinations, sight words, spelling rules, and word endings.

Then, during the second component, *Read to Understand*, students were taught the meaning of vocabulary words introduced during Word Works, as well as comprehension skills and strategies to apply while reading fiction and non-fiction. For example, lessons offered explicit instruction in previewing, setting purpose, text structure and evaluation, making inferences and taking perspectives, drawing conclusions, author's purpose, sequencing, main idea, summarizing, independent reading fix-up strategies, teacher and reader questioning, and

making connections within and across texts. In the first six weeks, instruction in the Read to Understand component lasted 10 min and in subsequent lessons, was increased to 25 min. Lessons also included a brief focus on fluency (reading with appropriate accuracy, rate, and expression) during the text reading.

Intervention teachers and training. A total of seventeen teachers, hired by the research team, were responsible for teaching the Passport to Literacy lessons. All the teachers had a bachelor's degree, four (33.3%) had obtained a master's degree in education, and one had a Ph.D. Twelve of the interventionists were certified teachers and one was a counselor. The other four had degrees in non-education areas. All intervention teachers were female. Three teachers identified themselves as Hispanic ethnicity. In terms of race, 11 (65.7%) teachers were White and five teachers (29.4%) were Black and one chose not to fill in the information.

Prior to the start of instruction, intervention teachers participated in approximately 8 hr of training over the course of two days. Training provided by the project coordinators at each site, allowed interventionists to become oriented to the project, familiarize themselves with the Passport to Literacy intervention program and instructional routine, practice implementation of lessons, and discuss positive behavior supports. Once intervention sessions with students were initiated, twice monthly coaching visits were conducted by the project coordinators. These visits allowed teachers to receive feedback on implementation as well as discuss any questions or concerns. Finally, monthly meetings with all intervention teachers were held at each site to provide continued support and ensure fidelity of implementation.

Intervention implementation and fidelity. The total number of Passport to Literacy lessons covered for each of the intervention groups ranged from 83 to 106 sessions. For those individual students who remained in the school for the duration of the intervention, the number

of lessons attended ranged from a low of 58 sessions to a high of 106 sessions (*M* = 93.79, *SD* = 7.82).

As noted earlier, each intervention teacher recorded three intervention lessons during the year, and these recordings were coded for instructional content and quality using the ICE-R to directly compare the instructional elements in Passport and the school-provided interventions. On average, the treatment session instruction was 28.56 min (*SD* = 4.07) in length. Instruction focused on developing students' reading comprehension (*M* = 11.80 [41% of intervention time], *SD* = 5.65) and vocabulary/oral language ability (*M* = 6.05 [21% of intervention time], *SD* = 4.81). During treatment lessons, students engaged in text reading for 4.72 min (*SD* = 2.43) or 17% of intervention time, decoding and word reading activities for 3.29 min (*SD* = 3.11) or 12% of intervention time, and practiced spelling for just over 1 min (M = 1.32, *SD* = 2.34) or 5% of intervention time. Explicit instruction in oral reading fluency was observed for 0.26 min (*SD* = 0.92) or 1% of intervention time, on average. During treatment lessons, less than 1 min (1%) of time was considered either non-instructional in nature (*M* = 0.18, *SD* = 0.64) or focused on instruction in another academic area such as writing or grammar (*M* = .27, *SD* = 0.83). Ratings of instructional quality indicated high-average quality (*M* = 3.37, *SD* = .62) and on average, intervention students were engaged during instruction (*M* = 2.85, *SD* = .43).

In terms of direct fidelity of implementation to the Passport to Literacy lessons, mean implementation ratings for each tutor implementation were high, ranging from 2.71 to 3.00, across the lesson components. Similarly, mean ratings of student academic engagement (2.85 to 3.00) and quality of tutor instruction (2.76 to 3.00) for each component were high.

Dependent Measures

Project staff blind to condition assessed students' word reading, decoding, vocabulary, reading fluency, and reading comprehension in the fall and spring. Due to the high correlation between students' word reading and oral reading fluency (see Table 2), we included only the word reading measures in the dependent variables, but examined possible moderation of students' fluency on outcomes.

Woodcock-Johnson III Tests of Achievement (WJIII; Woodcock, McGrew, & Mather, 2001). To assess word reading and comprehension, we selected four individually administered subtests from the nationally standardized WJIII. The letter-word identification subtest measures recognition of real words, and begins with individual letters. The word attack subtest measures decoding skill and includes items that are pseudowords, which begin with a few single letter sounds and progress to decoding of complex pseudowords. The picture vocabulary test asks students to name pictured objects increasing in difficulty. The passage comprehension subtest measures how well students can read text with missing words, presented as a cloze procedure in which students read the sentences silently and are asked to supply the missing word. Test authors report that test-retest reliability for these four subtests at fourth grade are .81, 85, .77, and .86 respectively.

Dynamic Indicators of Basic Early Literacy Skills -6th Edition (DIBELS; Good & Kaminski, 2002). To assess student's ability to read connected text with speed and accuracy, we administered the oral reading fluency (ORF) subtest from DIBELS. Students read three separate passages aloud for 1 min and the total number of correct words read per minute from the passage is considered the oral reading fluency rate. Test-retest reliabilities for ORF with elementary age students range from .92 to .97; alternate-form reliability across passages from the same level is reported as .89 to .94 (Good et al., 2004).

GMRT (**MacGinitie et al., 2006**). The GMRT is a group-administered, norm-referenced test. We administered the vocabulary and comprehension subtests. The fall reading comprehension scores were used to screen students for inclusion in the study. Vocabulary presents words in context. The student chooses the correct meaning of the target word. Comprehension provides students with reading passages and multiple choice questions. Questions address facts, inferencing, and drawing conclusions. Test-retest reliabilities are above .85. Construct validity estimates range from .79-.81.

Analytic Approach

For both research questions, a longitudinal, multilevel structural equation modeling (ML-SEM) framework was used to estimate primary and conditional impacts. A structural equation model approach is useful as it minimizes the limitation of measurement error inherent to individual observed measures by leveraging the common variance across multiple assessments of a construct. Common specifications of the ML-SEM for randomized controlled trials include latent factors of pretest and posttest measures at both a lower-level unit, such as students, and at an upper-level unit (e.g., classrooms). Similar to multilevel models of observed outcomes, the ML-SEM includes the regression of posttest on pretest but in this case with latent variables. Estimation of the treatment effect may occur through one of two common approaches. One methodology includes the simple regression of the posttest on k-1 dummy codes for a grouping variable, where k is the number of treatment arms, to reflect whether an individual received the intervention or not. An alternative approach does not include a variable for treatment status, but rather tests for group differences through a multiple group invariance approach. In this instance the test of impact is estimated by inspecting the posttest means for invariance between groups when constraining other parameters of the model to be equal (e.g., loadings, residual variances,

regression of posttest on pretest). The difference in standardized posttest means between groups then represents the standardized effect size difference. ML-SEMs have received fair attention in the literature as of late (e.g., Goddard, Goddard, Kim, & Miller, 2015; Heck & Thomas, 2015) as a method to not only overcome measurement issues but also in increasing power to detect effects due to latent variables increasing reliability of the measured construct. A known property of effect sizes is that they are negatively related to unreliability of measurement. Subsequently, with greater precision in measurement through the latent variable, it is possible to detect larger effects that may not be possible with observed variable error.

Despite the increasing prevalence of ML-SEM in the literature for testing treatment effects, a limitation in application has been to randomized designs where not all units are nested. In partially nested randomized controlled trials (PN-RCT; Baldwin, Bauer, Stice, & Rohde, 2011; Lohr, Schochet, & Sanders, 2014), only some individuals are nested within a group. For the present study, the partial nesting is observed where students receiving the intervention were all nested within small groups but the comparison students were not. Baldwin et al. noted that in their review of studies with PN-RCT designs, researchers frequently ignored this structure to the detriment of standard error estimation. Although robust methods have been proposed that model observed measures for PN-RCT designs, less attention has been given to the treatment of PN-RCT data in the ML-SEM context. Sterba et al. (2014) presented an approach within Mplus that allows an individual to match the ML-SEM methodology to the PN-RCT design. However, a limitation of reported approaches for observed and latent variable approaches for PN-RCT data is that they involve the introduction of ancillary variables into the data, as well as additional model specifications (e.g., adjusting estimation of the denominator degrees of freedom for observed variables) that are not possible to implement across commonly used software.

A more naturalistic approach to treating PN-RCT data is to view the nesting structure through *n*-level SEM (*n*SEM; Mehta & Neale, 2005) which easily accommodates complex nesting. Within *n*SEM, observed and latent variables may be used across multiple levels. The concept of *level* in *n*SEM takes on unique meaning differing from multilevel modeling. That is, a level typically refers to a unit of clustering for one set of observations within another unit such as students nested within classrooms. A level in *n*SEM refers to this type of nesting but further describes any meaningful, nominal grouping of individuals such as male or female, students eligible for free/reduced lunch or not, or those who received an intervention or not. This more flexible use of level allows us to more naturalistically situate the PN-RCT design in the *n*SEM framework. Consider a sample *n*SEM model in Figure 1 that is relevant to the current study. Note that there are four boxes that are each representative of participant groupings. Pertaining to students, there are two levels of groupings one for the Passport students (level 1) and one for comparison students (level 2). Small group represents a nesting structure for only the Passport students (level 3) and Classrooms represent the nesting of students from both student groups in classrooms (level 4). Figure 1 then represents a 4-level partially nested, cross-classified SEM where the comparison students are nested within classrooms and the Passport students are cross-classified by small groups and classrooms.

At this point, it may useful to provide an introduction to more specific components of the model. For both the Passport and comparison levels, the SEM specifies that there is a posttest (η_1^1 for Passport and $\eta_1^2$for comparison), where the superscript notation denotes the level for the parameter and the subscript denotes the parameter number. Thus, η_1^1 is the first level-1 latent variable, (i.e., the Passport posttest latent variable) and $\eta_1^2$is the first level-2 latent variable for the comparison group at the posttest. η_2^1 then is the second latent variable for the Passport group

(i.e., the pretest) and η_2^2 is the pretest latent variable for the comparison group. The latent variables in Passport are indicated by the four measures Y_1^1 to Y_4^1, two at pretest and the same two at posttest, as are the latent variables for comparison group indicated by the same measures Y_1^2 to Y_4^2. Each of the observed measures has a residual (θ) and loading (λ). Note that the loading subscripts are the same from posttest to pretest and between the Passport and comparison groups. This specification denotes that the model constrains the estimated values to be equal across groups, as it does also for the residual variances and the regression of the posttest latent construct on the pretest (β). Across all four levels, there are latent means (α) and variances (ψ). As a multilevel model, only the latent means at the student levels (i.e., Passport and comparison) are estimated; they are fixed at 0 at the small group and classroom levels. Similar to a longitudinal SEM, the pretest means (not reflected in the diagram) are set at 0 and the variances are fixed at 1. This specification is so that the means at the posttest are standardized such that the difference between α_1^1 and α_1^2 is the standardized treatment effect.

The model building process for the PN-RCT *n*SEM occurred in two phases with four models each. Phase 1 was focused on testing longitudinal invariance of the loadings and intercepts and phase 2 tested between-level posttest invariances. Within phase 1, three models were tested: 1) Freed loadings and intercepts across pretest and posttest latent variables in treatment and comparison groups (Model 1); 2) Invariant loadings and freed intercepts across pretest and posttest latent variables in treatment and comparison groups (Model 2); 3) Invariant loadings and intercepts across pretest and posttest latent variables in treatment and comparison groups (Model 3). These steps were necessary to evaluate whether a fully invariant model for intercepts and loadings was plausible such that the latent means are reflective of actual latent mean differences and not loading/intercept structure differences. For phase 2, five models were

tested to test for posttest invariance across combinations of the treatment, comparison, and small group levels: 1) Freed loadings and intercepts across treatment, comparison, and small group levels (Model 4); 2) Invariant loadings and freed intercepts between treatment and comparison levels (Model 5); 3) Invariant loadings and intercepts between treatment and comparison levels (Model 6); 4) Invariant loadings and freed variances between treatment and small group levels (Model 7); and 5) Invariant loadings, intercepts, pretest means, and variances across treatment, comparison, and small group levels (Model 8). Each set of eight models were applied to reading comprehension, word reading, and vocabulary outcomes. Exploratory analyses in the study tested whether EL status, pretest, letter-word identification, or oral reading fluency moderated the relation between treatment status and posttest performance. Model comparisons were made using the deviance statistic as well as the AIC and BIC indices. A log-likelihood difference test was used for hypothesis testing of model differences.

Results

Descriptive Statistics and Correlations

A preliminary review of the data for missingness (Table 2) showed that complete data were available for the fall GMRT-RC measure (n = 412), but missing data rates varied from .7% to 20.4% for other measures. The reason for the high level of missing data on the Fall GMRT Vocabulary measure was it was not administered in one site in Year 1. Little's missing completely at random (MCAR) test suggested that all missing data met reasonable assumptions for MCAR [$\chi^2(81) = 77.99$, $p > .500$]; thus, using full information maximum likelihood for model estimation was appropriate and would not negatively bias results.

Table 2 presents the full sample student performance results on the individual measures of reading comprehension, word reading, and vocabulary at fall and spring and Table 3 reports

means and standard deviations by treatment condition. Students' scores on the measures were consistently higher at the spring compared to fall. Correlations among the measures in the fall ranged from r = .12 between WJIII picture vocabulary and word attack to r = .77 between WJII word attack and letter-word identification. Spring correlations ranged from r = .26 between WJIII picture vocabulary and GMRT reading comprehension to r = .79 between WJII word attack and letter-word identification. Stability coefficients from fall to spring ranged from r = .32 for GMRT reading comprehension to r = .82 for WJII letter-word identification, suggesting moderate to high stability in relative rank orders of individuals over time.

Tests of Invariance

Results from the tests of invariance are presented in Table 4. For the first phase of invariance testing, which was related to longitudinal invariance between pretest and posttest between the treatment and comparison groups, results consistently demonstrate that imposing incremental equality constraints on the intercepts and loadings did not significantly denigrate fit. This step is important as it suggests that the means and loadings didn't differ by forcing them to be equal across groups. For reading comprehension, the difference in deviance between Models 2 and 3 was negligible (Δ-2LL = 0.65) and not statistically significant (p = .723). Similarly, no significant differences were observed between Models 2 and 3 for word reading (Δ-2LL= 0.30, p = .861) or vocabulary (Δ-2LL = 0.87, p = .647). Phase 2 invariance testing in the posttest invariance among the treatment, comparison, and small groups (Models 4-8) show that no substantive difference was observed in the deviance statistic. In fact, the largest difference in deviance between Model 4 (the least restrictive model) and Model 8 (the most restrictive model) was for reading comprehension where the deviance difference was < 4 points with 6 degrees of freedom, a non-significant finding. When comparing the final two models, no significant

differences were observed for reading comprehension (Δ-2LL = 4.18, p = .652), word reading (Δ-2LL = 1.72, p = .886), or vocabulary (Δ-2LL = 0.78, p = .978).

*n*SEM Primary Impact Model Results

Primary impact model results for the three latent outcomes of reading comprehension, word reading, and vocabulary related to the first research question are presented in Figures 2 and 3. Using a similar methodology for comparing the factor analytic models, the impact analyses tested constrained and freed estimate versions of the *n*SEM in Figure 1. In the constrained version of the model, the latent post-test means for the Passport and comparison groups (i.e., α_1^1 and α_1^2; Figure 1) were constrained to be equal. This constraint was relaxed for a second model test of mean difference. A log-likelihood difference test was used for hypothesis testing of model differences. The model comparison for reading comprehension (Table 4) showed that the model with freed posttest means fit better than the model with constrained means (Δ-2LL = 9.47, Δdf = 1, p < .001). Figure 2 shows that controlling for the pretest relation to posttest (β = 1.08), the standardized mean posttest value was α = 1.26 for the Passport group and α = 0.88 for the comparison group, a statistically significant difference. The effect size of Passport for latent reading comprehension outcomes is calculated as the difference between these two scores, or 0.38. No significant differences were observed between the constrained and freed posttest means models for latent word reading (p = .280) or latent vocabulary (p = .480). Further, no substantive primary impacts for Passport were observed for word reading (Δα = 0.06; Figure 3 top), nor was there an impact on vocabulary (Δα = 0.08; Figure 3 bottom).

*n*SEM Exploratory Modeling Results

To address the second research question, exploratory analyses evaluated the moderation of treatment effects based on EL status and selected baseline measures (i.e., pretest, letter-word

identification, and oral reading fluency). As previously noted, two methods are frequently employed to test for treatment effects in SEM studies including the inclusion of *k*-1 dummy codes or multiple groups. In a similar manner, moderation of treatment effects can be tested by including interaction terms in a regression model, or by using the multiple group method. The moderators for our exploratory analyses were a combination of continuous (i.e., baseline/pretest, letter-word identification, and oral reading fluency) and categorical (i.e., EL). As such, two different approaches were used for tests of moderation.

Three baseline moderation models were tested. The first moderation model, which we call baseline moderation model, tested the impact of the autoregressive, latent pretest construct and whether the relation between latent pretest and posttest varied by group. By releasing the β in Figure 1 to be freely estimated for the Passport and comparison groups, and comparing the fit of this model to the primary impact model where the β in Figure 1 is constrained to be the same between the two groups, a test is provided as to whether baseline performance moderates the treatment effect. The second and third moderation models, which each used single-item indicators of letter word identification and ORF, was done by first creating a single-item indicator latent construct for the moderator of interest (i.e., where the loading was fixed at 1.0 and the residual variance was set at a reliability adjusted estimate of the sample variance). This factor was set as a predictor of the latent posttest, identical to the β parameter in Figure 1, as well as set to covary with the latent pretest for both Passport and comparison groups. Estimation for this type of model required two steps; first, the path from the baseline measure was constrained to be equal between Passport and comparison groups. Fit from this model was compared to a model where the β constraint was freed for estimation. Improved fit for a freed model provided evidence for moderation.

Results for the three tests of moderation for each outcome are reported in Table 5. For latent reading comprehension, no moderation was observed for baseline latent reading comprehension (Δ-2LL = 0.00, p = 1.00) or baseline oral reading fluency (Δ-2LL = 1.00, p = .321), but statistically significant moderation was estimated for baseline letter-word identification (Δ-2LL = 14.87, p < .001) such that students with higher initial word reading scores performed better on reading comprehension in the treatment. No significant moderation was observed for any of the selected moderators for either latent word reading or vocabulary outcomes (Table 5).

For the EL indicator, moderation was tested by fitting the factor models from Figure 1 separately for EL and non-EL students and evaluating Passport and comparison group post-test mean differences using constrained and freed post-test means similar to the primary impact model. Relevant results for the EL student model (Table 5 and Figure 4) showed no statistically significant difference in posttest means were observed for reading comprehension (p = .068), word reading (p = .108), or vocabulary (p = .841); however, the mean effect size difference in Figure 4 shows small effects in favor of Passport for latent word reading (Δα = 0.54 - 0.35 = 0.19) and latent reading comprehension (Δα = 1.42 – 1.04 = 0.38). No effect of Passport was observed for EL students on latent vocabulary (Δα = 0.01). A statistically significant effect of Passport was estimated for non-EL students on reading comprehension (p = .009; Table 4) with an effect size of Δα = 0.39 (Figure 5). No significant effects were estimated for latent word reading (p = .729) or vocabulary (p = .362); however, different from the other analyses, a small effect on vocabulary was estimated (Δα = 0.13; Figure 5).

Discussion

In this study, our aim was to contribute to the relatively limited body of research on effective comprehensive reading interventions to improve reading comprehension for upper elementary students by extending our prior work examining the effects of a widely used, multicomponent, upper elementary reading intervention (Wanzek et al., in press). The present study adds uniquely to the existing literature by employing a large sample, using latent variables based on standardized reading measures, and by using a relatively more sophisticated data analytic method (*n*SEM) to address differences in nesting within the treatment and comparison groups. In addition, the larger sample also allowed us to examine additional moderators such as initial baseline reading performance and EL status to learn more about for whom the intervention was most effective. The treatment was implemented with a high degree of fidelity that included approximately 94 sessions. Thus, the study is not only rigorous in design, but is one of the most extensive to date for this grade level; providing a fairly optimal test of the possible effects of implementing this multicomponent intervention at the fourth grade level.

Our first research question addressed main effects of the multicomponent intervention on reading comprehension, word reading, and vocabulary. Consistent with our hypothesis that students with reading difficulties receiving the intervention would outperform students receiving only typical school services in reading comprehension, we did find a significant effect of the intervention on reading comprehension with an effect size of 0.38. However, we found no significant effects on word reading (ES = 0.05) or on vocabulary (ES = 0.08). The magnitude of the effects on comprehension are slightly larger than in our previous study of the Passport to Literacy intervention, which found effect sizes on the individual measures that comprised our latent variable in the present study (i.e., WJIII passage comprehension [ES = 0.14] and the GMRT [ES = 0.28]). It is noteworthy that 0.38 exceeds the effect size criteria of 0.25 for

substantively important impact from the What Works Clearinghouse (2014). Based on the mean standard scores, students in the comparison group appeared to make expected progress (one year's worth of progress) in reading comprehension, while students in the treatment group accelerated their learning. In other words, students in the comparison group didn't fall any further behind while students in the treatment group made some progress towards closing the gap between their current level of performance and expected levels of performance in reading comprehension. Importantly, neither group of students demonstrated on grade level performance at the end of the intervention, although the accelerated learning in reading comprehension for students in the treatment group is promising. We found no significant differences between study groups on word reading or vocabulary. Thus, our findings suggest participation in Passport to Literacy can improve student reading comprehension; a finding consistent with our initial work (Wanzek et al., in press).

That we found no main effects for word reading or vocabulary is important, particularly as it is consistent with our prior study (Wanzek et al., in press) and suggests that for students with weak comprehension, participating in Passport to Literacy would likely move the dial on only on reading comprehension. This is likely because, although the program is multi-component, it focuses primarily on reading comprehension, with relatively limited word work or in-depth vocabulary instruction. Our observations indicated that, as designed, on average more than 40% of the treatment intervention was devoted to explicit instruction in reading comprehension. In contrast, the percentages of implemented intervention devoted to vocabulary, text reading, decoding, and spelling were 21%, 17%, 12%, and 5% respectively. The quality for this instruction was fairly high as well, indicating students received explicit, systematic instruction in reading comprehension. This high quality, comprehension emphasis in the

intervention may explain the reading comprehension outcomes students realized. In other words, the fact that Passport to Literacy has its benefits largely in the area of reading comprehension may be related to the focus of the intervention. The effect sizes for reading comprehension in the present study are larger than those in our prior study (effect sizes ranged from 0.14 to 0.28 in the prior study), but are smaller than effect sizes reported in two other multicomponent interventions. Specifically, for reading comprehension measures, Vadasy and Sanders (2008) reported an effect size of 0.50 and O'Connor et al.'s (2002) effect sizes ranged from 1.39 to 1.46. By contrast, Ritchey et al. (2012) found no significant differences on a standardized measure of reading comprehension, but did report an effect size of 0.56 on a researcher-developed measure of comprehension strategy use.

In our previous study of the effects of Passport to Literacy (Wanzek et al., in press) with a smaller sample, we suggested that our pattern of effects (significant effects for reading comprehension, but not for word reading or vocabulary) might be related to the amount of time attributed to narrative and expository comprehension and word reading during the lessons; with an average of 12 min of reading comprehension instruction and 6 min of vocabulary instruction in a typical half hour lesson, compared to 3 min of decoding or word reading instruction. In contrast, the interventions in the O' Connor et al. (2002) and Vadasy and Sanders (2008) studies included relatively more fluency practice than in the current study, perhaps allowing students to access greater amounts of text for improving their overall reading comprehension. The samples in the studies by O'Connor et al. as well as Vadasy and Sanders presented with lower overall word recognition and fluency abilities initially as well. Ritchey et al. (2012) emphasized fluency and expository comprehension, but for a briefer period of time (24 sessions) than O'Connor et al., Vadasy and Sanders, or the current study. The brief time period makes it difficult to directly

compare the relationship between the instruction in the intervention and findings to these other more lengthy studies. However, the current findings seem to align with the differences in intervention focus, length of intervention, and results of the previous studies.

Our second research question addressed moderation, to help inform for whom the intervention was effective. We hypothesized, based on exploratory findings from our previous study, that students with low levels of initial comprehension might demonstrate less growth than students with better initial comprehension. However, with our larger sample and using latent variables, we found no moderation effects for initial status on comprehension, suggesting the intervention was equally beneficial for students at all levels of initial comprehension. This is encouraging for practice as the intervention, with its relative emphasis on comprehension, can assist all levels of struggling, upper elementary students in improving their reading comprehension. There was also no moderation of the intervention effects for reading comprehension based on students' initial reading fluency, a finding that aligns with O'Connor et al (2002), though O'Connor et al. categorized students into lower or higher fluency students based on a break point. We examined moderation of oral reading fluency differences as a continuous variable. The intervention was equally beneficial in improving reading comprehension for students at all levels of initial reading fluency. However, we did find that initial individual differences in word reading ability significantly moderated the effect of the treatment, with students who entered the intervention at lower levels of word recognition making less progress in reading comprehension than students who entered the intervention with higher levels of word reading. An implication for schools is that these students with low word reading may require a reading intervention that incorporates more word study before they can fully benefit from an intervention that emphasizes reading comprehension. The relatively brief

intensive word study provided at the beginning of the Passport to Literacy intervention may not be enough for students with low word recognition to make the same gains as those entering with higher levels of word recognition. Torgesen et al. (2001) implemented an intensive reading intervention largely focused on word recognition for students with very low initial word reading skills and reported significant gains in standard scores across word reading and reading comprehension. The lack of control group in the Torgesen et al. study makes it difficult to compare effect sizes to other studies, but an intensive intervention with a heavier emphasis on word recognition is likely needed for students with the lowest word recognition abilities at the upper elementary level. To summarize, the Passport to Literacy intervention provided improvements in students' reading comprehension beyond the typical school services for students at varying levels of initial reading comprehension or reading fluency but who had relatively higher levels of word reading ability.

Encouragingly, the effects of the intervention on reading comprehension were similar for EL and non-EL students (ES = 0.38 and 0.39, respectively), suggesting the intervention is equally beneficial and appropriate for ELs to improve their reading and understanding of English text. Practical benefits of the intervention were noted in relation to word reading for the EL students, but this was not a significant moderation. Previous work reviewed by Baker et al. (2014) demonstrated that both younger ELs (K-1) and older ELs (Grades 6-8) benefit from small group multicomponent reading interventions in terms of word reading and comprehension. Wanzek and Roberts (2012) also noted EL status moderated effects on word attack and word identification with the EL students performing better than non-EL students following intervention. These higher effects occurred regardless of the emphasis of the intervention (e.g., comprehension emphasis, word recognition emphasis).

Limitations

Although our study was rigorous, there are always limitations involved with school-based research. To ensure a strong test of the efficacy of the Passport to Literacy intervention, we trained research staff to implement the intervention with a high degree of fidelity and dosage consistent with the publisher's recommendations. Thus, similar effects may or may not be achieved by school personnel depending on implementation. We also recruited schools that were diverse and served students from low socioeconomic backgrounds, so our findings might not generalize to schools serving students from higher socioeconomic backgrounds. The majority of our ELs in our study were Hispanic and our findings may not generalize to students from other language backgrounds, particularly those with orthographies that are very different than English. Further, effect sizes are interpretable relative to the comparison condition in the participating schools where very few struggling readers received supplemental interventions as a part of their typical practice.

Implications and Directions for Future Research

Teachers and school leaders face challenges in identifying effective reading interventions for students in the upper elementary grades, particularly given the high numbers of students who continue to struggle with reading after third grade (National Center for Educational Statistics, 2016). The increased demands placed on students beginning in fourth grade may cause a slowing of actual versus expected growth for some students (Chall & Jacobs, 1983). Therefore, fourth grade teachers are often faced with the challenge of providing intervention not only for students with previously identified reading difficulties that have not been adequately remediated, but also students with late-emerging reading difficulties (Compton, Fuchs, Fuchs, Elleman, & Gilbert, 2008).

The current study suggests that a multicomponent intervention emphasizing comprehension instruction can allow students to accelerate their reading comprehension outcomes. Without such interventions, particularly given the limited emphasis within core classroom instruction to support learning to read in fourth grade, students who do not read proficiently could face serious and ongoing consequences, not only in reading language arts, but across content areas.

On the one hand, the positive effects for reading comprehension found in our study extend the limited evidence base on effective multicomponent reading interventions for upper elementary students. On the other hand, the lack of effects for word reading or vocabulary underscores the need for more research on intensive interventions for fourth grade students with the most severe reading difficulties. For example, there is an even more intensive level of the Passport to Literacy intervention, which the publisher recommends for students in need of more intensive levels of instruction. It is more intensive in that students are served in smaller groups and for a longer session and includes additional instruction, including instruction in reading fluency that has been more emphasized in previous work (O'Connor et al., 2002; Vadasy & Sanders, 2008). It is possible this extended intervention will be more potent than the standard implementation of the Passport to Literacy intervention, providing the additional emphasis without decreasing the time spent on comprehension. To guide schools' intervention implementation for the upper elementary grades, additional research is needed to identify appropriate interventions, describe for whom they are effective, and also to examine the relative benefits of interventions with increasing intensity to meet adequately meet the varying needs of students.

References

Baker, S., Lesaux, N., Jayanthi, M., Dimino, J., Proctor, C. P., Morris, J., . . . Newman-Gonchar, R. (2014). *Teaching academic content and literacy to English learners in elementary and middle school* (NCEE 2014-4012). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education.

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods, 16,* 149-165. doi: 10.1037/a0023464

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., & Shaywitz, S. E. (2004). Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *Journal of Educational Psychology, 96*(3), 444-461.

Chall, J.S., & Jacobs, V.A. (1983). Writing and reading in the elementary grades: Developmental trends among low-SES children. *Language Arts, 60*, 617–626.

Compton, D. L., Fuchs, D., Fuchs, L. S., Elleman, A. M., & Gilbert, J. K. (2008). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences, 18,* 329-337. doi:10.1016/j.lindif.2008.04.003

Edmonds, M., & Briggs, K. (2003). The instructional content emphasis instrument: Observations of reading instruction. In S. Vaughn & K. L. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 31-52). Baltimore, MD: Brookes Publishing Co.

Flavell, J. H. (1992). Cognitive development: Past, present, and future. *Developmental psychology, 28,* 998-1005. doi:10.1037/0012-1649.28.6.998

Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology, 91*(4), 579-593.

Goddard, R., Goddard, Y., Kim, E. S., & Miller, R. (2015). A theoretical and empirical analysis of the roles of instructional leadership, teacher collaboration, and collective efficacy beliefs in support of student learning. *American Journal of Education, 121*, 501-530.

Good, R. H., & Kaminski, R. (2002). Dynamic Indicators of Basic Early Literacy Skills 6th Edition (DIBELS). Eugene, OR: Institute for the Development of Educational Achievement. Retrieved from http://dibels.uoregon.edu/

Good, R. H., Kaminski, R. A., Shinn, M., Bratten, J., Shinn, M., Laimon, D.,…..& Flindt, N. (2004). *Technical adequacy of DIBELS: Results of the early childhood research institute on measuring growth and development* (Technical Report, No. 7). Eugene, OR: University of Oregon.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. Routledge.Kamil, M. L., Borman, G. D., Dole, J., Kral, C. C., Salinger, T., & Torgesen, J. (2008). *Improving adolescent literacy: Effective classroom and intervention practices. IES practice guide (NCEE 2008-4027).* National Center for Education Evaluation and Regional Assistance.

Kieffer, M. J. (2008). Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. *Journal of Educational Psychology, 100*(4), 851-868. doi: 10.1037/0022-0663.100.4.851.

Kieffer, M. J. (2010). Socioeconomic status, English proficiency, and late-emerging reading difficulties. *Educational Researcher, 39*(6), 484-486. doi: 10.3102/0013189X10378400.

Kieffer, M. J. (2014). Morphological awareness and reading difficulties in adolescent spanish-speaking language minority learners and their classmates. *Journal of Learning Disabilities (Austin), 47*(1), 44-53.

Lohr, S., Schochet, P.Z., and Sanders, E. (2014). *Partially Nested Randomized Controlled Trials in Education Research: A Guide to Design and Analysis (NCER 2014-2000).* Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncer/pubs/20142000/pdf/20142000.pdf /

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2006). *Gates-MacGinitie Reading Tests* (4th ed.). Rolling Meadows, IL: Riverside Publishing.

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly, 40*(2), 148-182.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*, 259-284. doi: 10.1037/1082-989X.10.3.259

National Center for Educational Statistics (2016). *National assessment of educational progress: The nation's report card.* Washington, DC: U.S. Department of Education.

O'Connor, R. E., Bell, K. M., Harty, K. R., Larkin, L. K., Sackor, S. M., & Zigmond, N. (2002). Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty. *Journal of Educational Psychology, 94*, 474-485. doi:10.1037/0022-0663.94.3.474

O'Connor, R.,E., Fulmer, D., Harty, K. R., & Bell, K. M. (2005). Layers of reading intervention in kindergarten through third grade: Changes in teaching and student outcomes. *Journal of Learning Disabilities, 38*(5), 440-455.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1,* 117-175. doi: 10.1207/s1532690xci0102_1

Ritchey, K. D., Silverman, R. D., Montanaro, E. A., Speece, D. L., & Schatschneider, C. (2012). Effects of a tier 2 supplemental reading intervention for at-risk fourth-grade students. *Exceptional Children, 78,* 318-334. doi: 10.1177/001440291207800304

Scammacca, N., Roberts, G., Vaughn. S., Edmonds, M., Wexler, J., Reutebuch, C. K., & Torgesen, J. K. (2007), Interventions for adolescent struggling readers: A meta-analysis with implications for practice. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Sterba, S. K., Preacher, K. J., Forehand, R., Hardcastle, E. J., Cole, D. A., & Compas, B. E. (2014). Structural equation modeling approaches for analyzing partially nested data. *Multivariate behavioral research*, *49*(2), 93-118.

Swanson, H.L., Hoskyn, M., & Lee, C. (1999). *Interventions for students with learning disabilities: A meta-analysis of treatment outcomes.* New York: Guilford.

Therrien, W. J., Wickstrom, K., & Jones, K. (2006). Effect of a combined repeated reading and question generation intervention on reading achievement. *Learning Disabilities Research & Practice, 21,* 89-97. doi:10.1111/j.1540-5826.2006.00209.x

Torgesen, J. K., Houston, D. D., Rissman, L. M., Decker, S. M., Roberts, G., Vaughn, S., … Lesaux, N. (2007). *Academic literacy instruction for adolescents: A guidance document*

from the Center on Instruction. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Lindamood, P., Rose, E., Conway, T., & Vadasy, P. F., & Sanders, E. A. (2008). Repeated reading intervention: Outcomes and interactions with readers' skills and classroom instruction. *Journal of Educational Psychology, 100,* 272-290. doi: 10.1037/0022-0663.100.2.272

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*(4), 601-638.

Wanzek, J., & Roberts, G. (2012). Reading interventions with varying instructional emphases for fourth graders with reading difficulties. *Learning Disability Quarterly, 35*(2), 90-101.

Wanzek, J., Petscher, Y., Al Otaiba, S., Kent, S. C., Schatschneider, C., Haynes, M…., & Jones, F. G. (in press). Examining the effects of a standardized treatment for fourth graders with reading difficulties. *Journal of Research on Educational Effectiveness*.

Wanzek, J., Wexler, J., Vaughn, S., & Ciullo, S. (2010). Reading interventions for struggling readers in the upper elementary grades: A synthesis of 20 years of research. *Reading and writing, 23*(8), 889-912. doi: 10.1007/s11145-009-9179-5

What Works Clearinghouse. (2014). *Procedures and standards handbook* (version 3.0). Washington, D. C.: U. S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement*. Itasca, IL: Riverside.

Table 1

Average intervention instructional time in minutes and percent of time by study condition

Instructional Component	Passport Intervention		School-Provided Intervention	
	# of min	% of total time	# of min	% of total time
Phonics and Word Recognition	3.29	12	3.84	14
Spelling	1.32	5	1.22	4
Reading Fluency	.26	1	.53	2
Vocabulary/Oral Language	6.05	21	4.45	16
Comprehension	11.80	41	8.27	29
Non-instructional Text Reading	4.72	17	6.43	23
Other Academic Instruction	.27	1	1.86	7
Non-instruction	.18	1	1.04	4

Table 2

Descriptive statistics and correlations for study measures

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Fall GMRT RC	1.00												
2. Fall WJ PC	.32	1.00											
3. Fall WJ LWID	.30	.60	1.00										
4. Fall WJ WA	.26	.52	.77	1.00									
5. Fall GMRT Voc	.39	.49	.52	.41	1.00								
6. Fall WJ PV	.14	.51	.25	.12	.33	1.00							
7. DIBELS ORF	.29	.51	.70	.62	.46	.13	1.00						
8. Spring GMRT RC	.32	.46	.38	.32	.43	.23	.44	1.00					
9. Spring WJ PC	.35	.64	.54	.43	.50	.43	.47	.47	1.00				
10. Spring WJ LWID	.29	.60	.82	.72	.49	.21	.69	.39	.61	1.00			
11. Spring WJ WA	.24	.49	.76	.76	.44	.19	.60	.30	.50	.79	1.00		
12. Spring GMRT Voc	.31	.55	.51	.41	.64	.34	.49	.64	.53	.54	.46	1.00	
13. Spring WJ PV	.17	.52	.33	.16	.39	.74	.23	.26	.54	.36	.26	.43	1.00
Mean	440.61	481.92	484.78	490.32	445.93	486.44	80.35	456.69	487.54	493.01	495.90	462.06	491.11
SD	19.37	12.16	18.97	16.55	27.51	12.41	26.87	24.13	9.66	17.85	14.40	30.67	11.91
N	412	409	409	409	328	409	410	405	404	404	404	406	404
% Missing Data	0%	0.7%	0.7%	0.7%	20.4%	0.7%	0.5%	1.9%	1.9%	1.9%	1.9%	1.5%	1.9%

Note. GMRT RC = Gates-McGinitie Reading Comprehension. WJ PC = WJ-III Passage Comprehension. WJ LWID = WJ-III Letter Word Identification. WJ WA = WJ-III Word Attack. GMRT Voc = Gates-MacGinitie Vocabulary. WJ PV = WJ-III Picture Vocabulary. All correlations statistically significant at least $p < .05$.

Table 3

Descriptive statistics of measures by condition

	Passport			Comparison		
Measure	*N*	M	SD	*N*	M	SD
Fall GMRT RC	199	439.96	19.96	213	441.23	18.82
Fall WJ PC	199	481.52	11.67	210	482.30	12.61
Fall WJ LWID	199	484.43	18.82	210	485.12	19.14
Fall WJ WA	199	488.91	16.99	210	491.65	16.03
Fall GMRT Voc	159	444.87	27.09	169	446.93	27.93
Fall WJ PV	199	486.85	12.98	210	486.05	11.84
Fall DIBELS ORF	198	78.11	25.58	212	82.44	27.91
Spring GMRT RC	198	459.25	23.93	207	454.23	24.11
Spring WJ PC	198	488.12	9.35	206	486.98	9.93
Spring WJ LWID	198	492.79	17.14	206	493.23	18.54
Spring WJ WA	198	495.47	14.67	206	496.31	14.21
Spring GMRT Voc	198	462.08	31.87	208	462.04	29.55
Spring WJ PV	198	491.70	11.97	206	490.54	11.85

Note. GMRT RC = Gates-MacGinitie Reading Comprehension. WJ PC = WJ-III Passage Comprehension. WJ LWID = WJ-III Letter Word Identification. WJ WA = WJ-III Word Attack. GMRT Voc = Gates-MacGinitie Vocabulary. WJ PV = WJ-III Picture Vocabulary. ORF = Oral Reading Fluency.

Table 4

Confirmatory factor analysis model fit comparison for latent reading comprehension, word reading, and vocabulary

Outcome	Model	-2LL	*df*	AIC	BIC	Δ-2LL	Δ*df*	*p*
Reading Comp	Model 1	6766.20	12	6790	6847			
	Model 2	6766.21	11	6788	6840			
	Model 3	6766.85	9	6784	6827	0.65	2	.723[a]
	Model 4	6578.35	18	6614	6698			
	Model 5	6578.35	17	6612	6692			
	Model 6	6578.36	16	6610	6686			
	Model 7	6578.39	18	6612	6692			
	Model 8	6582.57	12	6606	6662	4.18	6	.652[b]
Word Reading	Model 1	6650.71	12	6675	6731			
	Model 2	6650.22	11	6672	6724			
	Model 3	6650.56	9	6673	6716	0.3	2	.861[a]
	Model 4	6364.37	18	6400	6485			
	Model 5	6364.36	17	6398	6478			
	Model 6	6364.62	16	6397	6472			
	Model 7	6364.37	17	6398	6478			
	Model 8	6366.09	12	6390	6446	1.72	5	.886[b]
Vocabulary	Model 1	6283.66	12	6308	6363			
	Model 2	6283.65	11	6305	6356			
	Model 3	6284.52	9	6303	6344	0.87	2	.647[a]
	Model 4	6933.29	18	6969	7054			
	Model 5	6933.29	17	6967	7047			
	Model 6	6933.47	16	6965	7041			
	Model 7	6933.29	17	6967	7047			
	Model 8	6934.07	12	6958	7014	0.78	5	.978[b]

Note. -2LL = -2*log likelihood. AIC =Akaike Information Critera. BIC = Bayes Information Criteria. Comp = Comprehension. [a] Model is compared to Model 2, [b] Model is compared to Model 7.

Model 1= Treatment-comparison, pretest-posttest freed loadings and intercepts

Model 2= Treatment-comparison, pretest-posttest, invariant loadings, freed intercepts

Model 3= Treatment-comparison, pretest-posttest, invariant loadings and intercepts

Model 4= Treatment-comparison -small group freed loadings and intercepts

Model 5= Treatment-comparison invariant loadings, freed intercepts

Model 6= Treatment-comparison invariant loadings and intercepts

Model 7= Treatment-small group invariant loadings, freed variances

Model 8= Treatment-small group-comparison invariance loadings, intercepts, means, and variances

Table 5

Fit comparison for primary impact models and moderation with EL, baseline, letter-word identification, and oral reading fluency

Outcome	Type	Model	-2LL	*df*	AIC	BIC	Δ-2LL	Δ*df*	*p*
Reading Comprehension	Impact	Constrained	13167.32	16	13199	13285			
		Freed	13157.85	17	13192	13284	9.47	1	.002
	EL Moderation	Constrained	3466.47	16	3498	3563			
		Freed	3463.15	17	3497	3566	3.32	1	.068
	Non-EL Moderation	Constrained	9654.43	16	9686	9768			
		Freed	9647.69	17	9682	9768	6.74	1	.009
	Baseline Moderation	Constrained	13164.84	16	13197	13283			
		Freed	13164.83	17	13198	13290	0.01	1	.920
	LWID Moderation	Constrained	16545.13	24	16593	16728			
		Freed	16560.00	25	18610	18750	14.87	1	.000
	ORF Moderation	Constrained	16875.00	24	16923	17058			
		Freed	16874.00	25	16923	17064	1.00	1	.320
Word Reading	Impact	Constrained	12485.20	16	12517	12603			
		Freed	12484.05	17	12518	12609	1.15	1	.284
	EL Moderation	Constrained	3323.66	16	3356	3421			
		Freed	3321.07	17	3355	3424	2.59	1	.108
	Non-EL Moderation	Constrained	9124.90	16	9157	9239			
		Freed	9124.78	17	9159	9245	0.12	1	.729
	Baseline Moderation	Constrained	12486.67	16	12519	12605			

		Freed	12486.65	17	12521	12612	0.02	1	.888
	LWID Moderation	Constrained	-	-	-	-	-	-	-
		Freed	-	-	-	-	-	-	-
	ORF Moderation	Constrained	16032.00	24	16080	16215			
		Freed	16031.00	25	16081	16222	1.00	1	.320
Vocabulary	Impact	Constrained	12826.17	16	12858	12943			
		Freed	12825.67	17	12859	12950	0.50	1	.480
	EL Moderation	Constrained	3025.1	16	3057	3119			
		Freed	3025.06	17	3059	3125	0.04	1	.841
	Non-EL Moderation	Constrained	9679.59	16	9712	9793			
		Freed	9678.76	17	9713	9799	0.83	1	.362
	Baseline Moderation	Constrained	12825.66	16	12858	12943			
		Freed	12825.15	17	12859	12950	0.51	1	.480
	LWID Moderation	Constrained	16237	24	16285	16418			
		Freed	16235	25	16285	16424	2.00	1	.157
	ORF Moderation	Constrained	16550	24	16598	16732			
		Freed	16549	25	16600	16739	1.00	1	.320

Note. -2LL = -2*log likelihood. AIC =Akaike Information Critera. BIC = Bayes Information Criteria. EL = English learner. LWID = Letter word identification. ORF = Oral reading fluency. LWID moderation was not tested for the latent word reading outcome as it was part of the latent variable itself and included in the pretest construct.

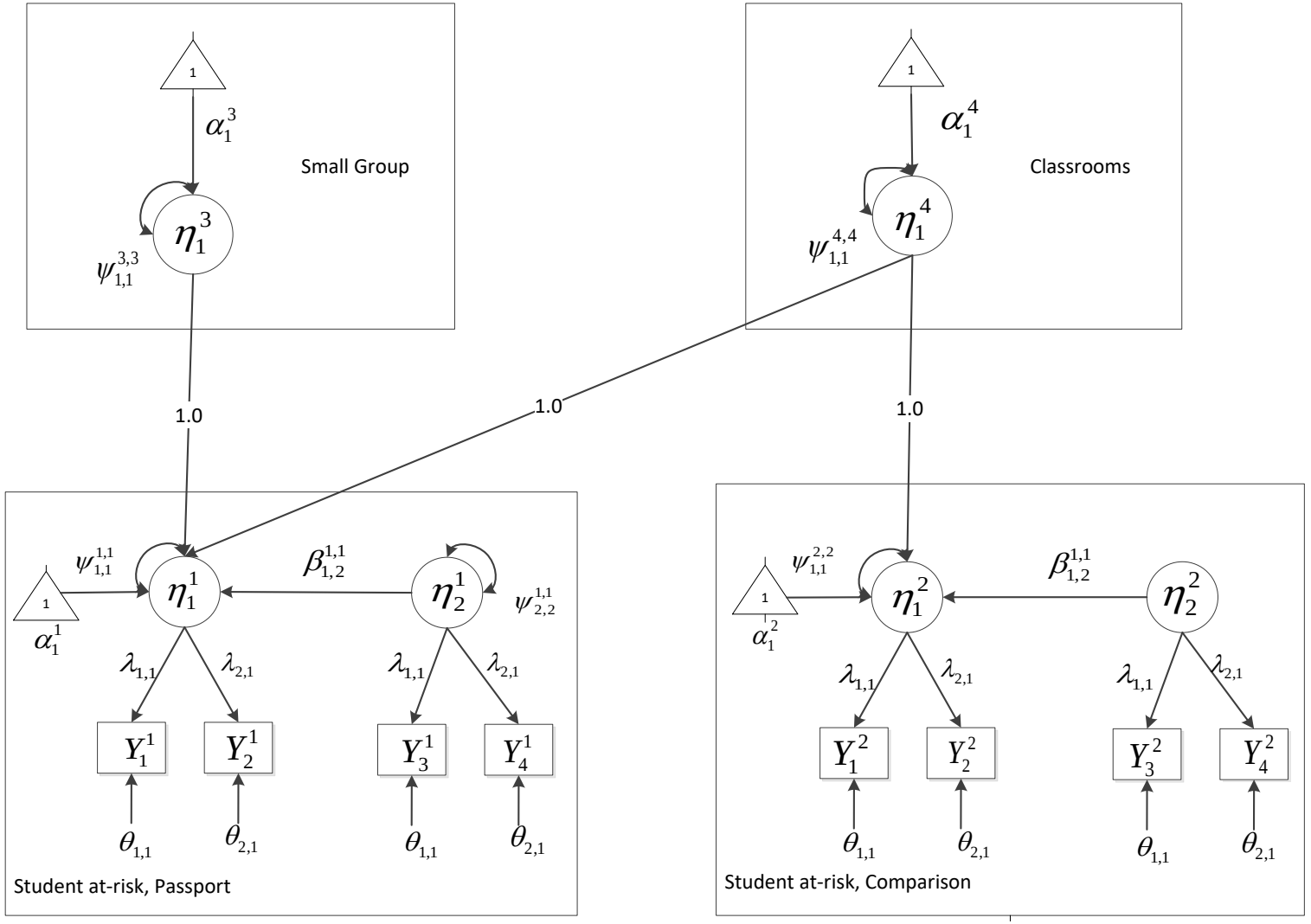

Figure 1. Sample *n*-level structural equation measurement model for partially nested designs.

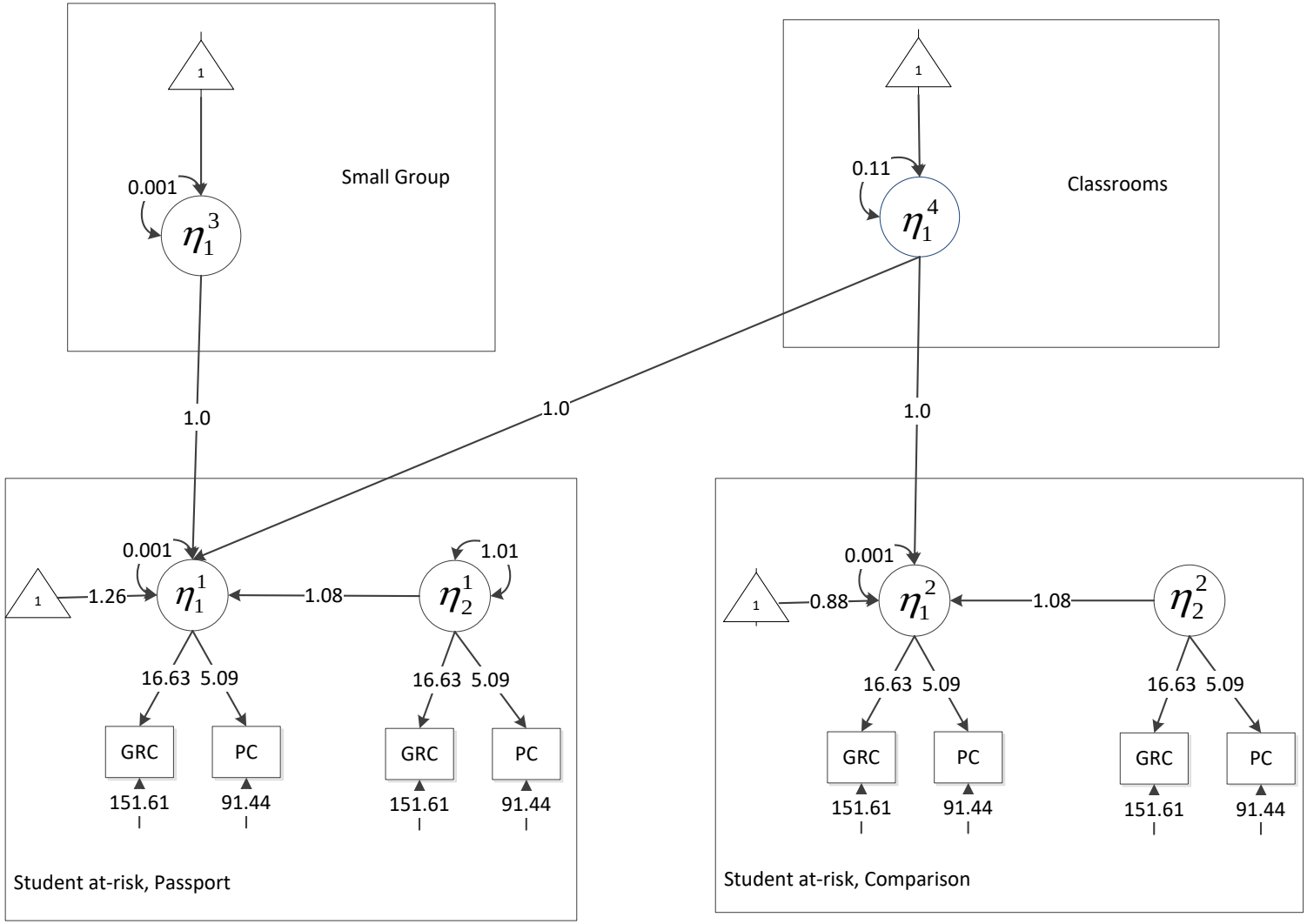

Figure 2. Primary impact *n*-level structural equation models for partial nested randomized controlled trial for reading comprehension.

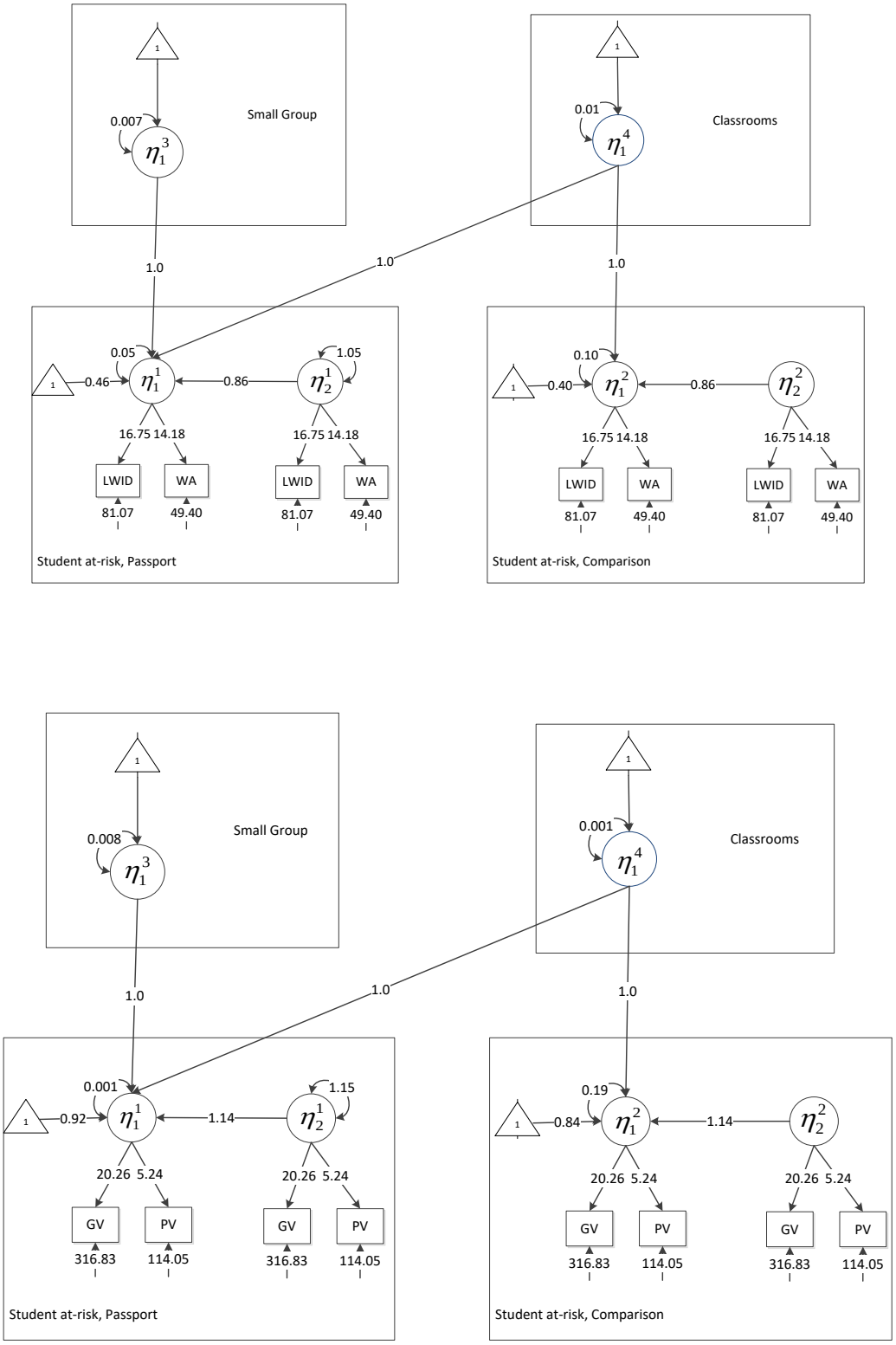

Figure 3. Primary impact *n*-level structural equation models for partial nested randomized controlled trial for word reading (top) and vocabulary (bottom).

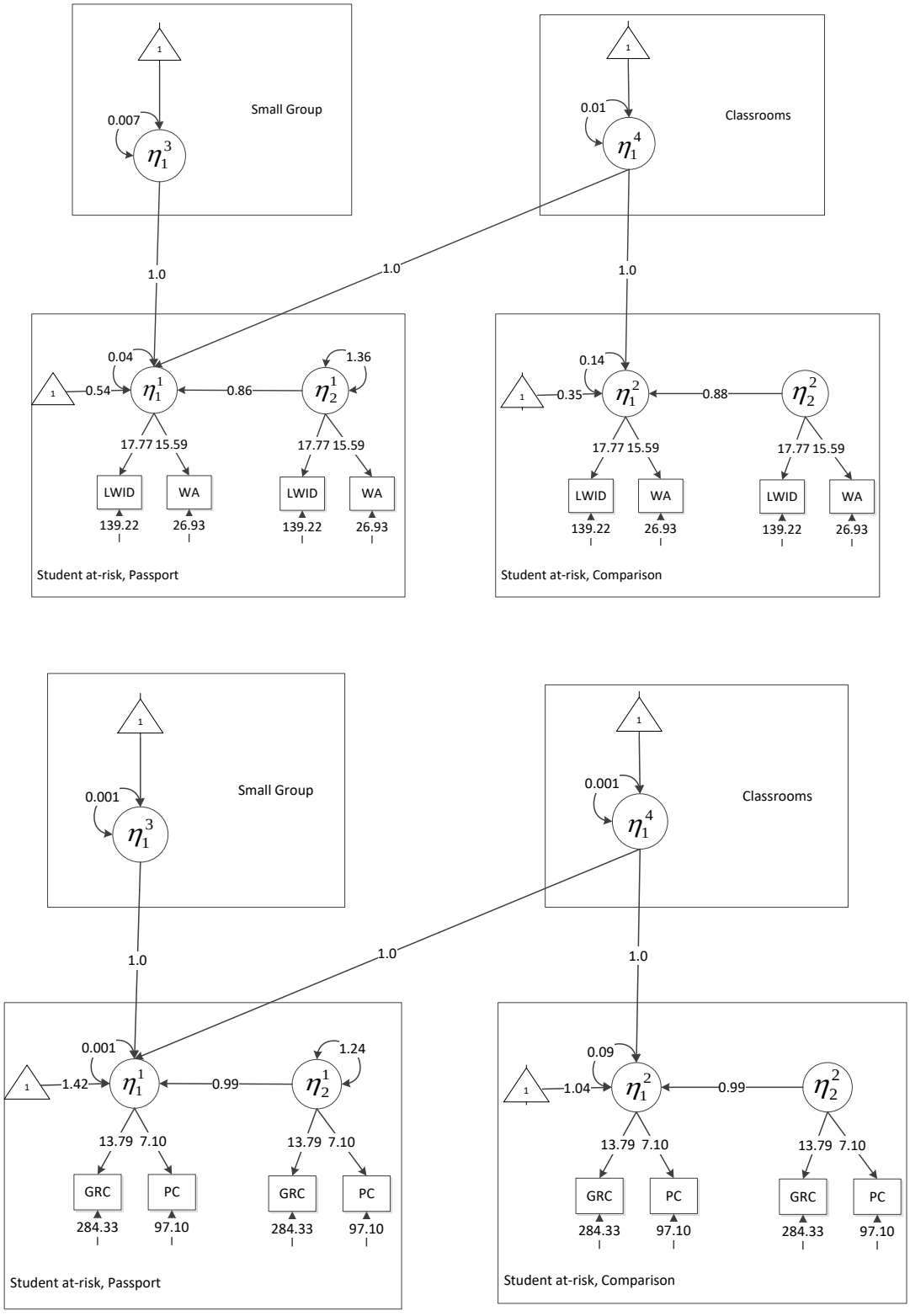

Figure 4. Exploratory *n*-level SEM for English Learners on word reading (top) and reading comprehension (bottom)

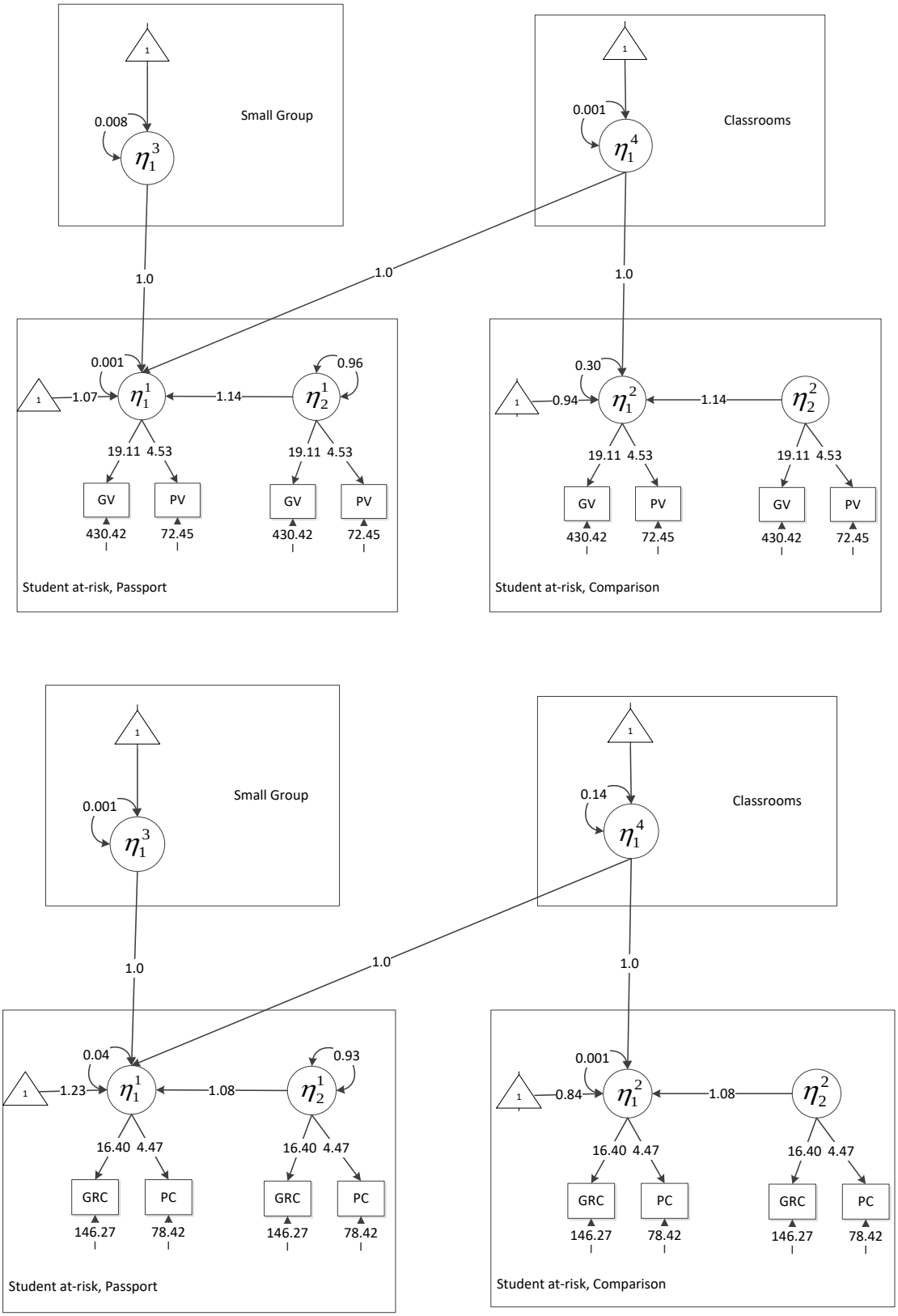

Figure 5. Exploratory *n*-level SEM for non-English Learners on vocabulary (top) and reading comprehension (bottom).