Longitudinal Stability of Reading-Related Skills and their Prediction of Reading Development

Jacqueline Hulslander, **Richard K. Olson**, **Erik G. Willcutt**, and **Sally J. Wadsworth**
University of Colorado at Boulder

Abstract

Individual differences in word recognition, spelling, and reading comprehension for 324 children at a mean age of 16 were predicted from their reading-related skills (phoneme awareness, phonological decoding, rapid naming and IQ) at a mean age of 10 years, after controlling the predictors for the autoregressive effects of the correlated reading skills. There were significant and longitudinally stable individual differences for all four reading-related skills that were independent from each of the reading and spelling skills. Yet the only significant longitudinal prediction of reading skills was from IQ at mean age 10 for reading comprehension at mean age 16. The extremely high longitudinal latent-trait stability correlations for individual differences in word recognition (.98) and spelling (.95) left little independent outcome variance that could be predicted by the reading-related skills. We discuss the practical and theoretical importance of these results and why they differ from studies of younger children.

The present study explores the predictors of individual differences in word recognition, reading comprehension, and spelling from mean age 10 years to mean age 16 years. Heterogeneity in skills correlated with literacy outcomes, such as phonological awareness and phonological decoding, has long been recognized among children who have difficulty learning to read (Boder, 1973), as well as among children across the normal range of reading ability (Bryant & Impey, 1986), raising their potential as possible developmental predictors. Previous studies aiming to predict individual differences in reading development have typically focused on development across the early grades, with initial assessments administered during the first years of schooling. These studies often find that reading-related skills such as phonological awareness, phonological decoding, or rapid naming improve the prediction of early reading development even after controls for initial literacy skills. In addition to the practical advantage of offering better prediction of reading development beyond that predicted from initial reading levels, the independent prediction from reading-related skills is often taken as evidence to support early interventions that are focused on improving performance on those skills.

Studies of older children typically show significant correlations between reading and reading-related skills that are similar to those found for younger children. Thus, we might expect that such reading-related skills would also improve prediction of reading across the later grades, with implications for intervention similar to those noted for younger children. Importantly, this predictive power needs to be independent of initial reading level. When children begin to read, reciprocal interactions between their development of reading and their development of related skills such as phonological awareness may confound our understanding of the causal pathways (Perfetti, 1985; Perfetti, Beck, Bell, & Hughes, 1987). One way to avoid this potential confound

is to see if a given skill at initial test predicts reading at a follow-up assessment *after* controlling for the "autoregressive" effect of reading at initial assessment (Gollob & Reichardt, 1987).

There has been little previous exploration of the unique prediction from reading-related skills for reading development in older children, beyond what can be predicted from their initial reading levels. It is possible that by the time children are in the middle grades, they have established reading development trajectories that are no longer independently influenced by reading-related skills. A related possibility is that reading and reading-related skills become so intertwined in the later grades that their independent variance is largely due to measurement error. Thus, at any given level of reading ability there will be at least some individual differences in performance on reading-related skills, but if these differences are due only to measurement error, there is no reason to expect that they would independently predict development in reading across the later grades. Therefore, in the present study, we begin by establishing the reliability and longitudinal stability of individual differences in the reading-independent variance for our reading-related predictors across an average testing interval of five and a half years. Then, after establishing their reliability, we explore their prediction of development in word recognition, spelling, and reading comprehension across the later grades. Since the present study included the reading-related skills of phoneme awareness, phonological decoding, rapid naming, and IQ, we will review the results of previous longitudinal studies that have included one or more of these predictors while controlling for autoregressive effects.

Several studies of unselected population samples have explored the unique prediction of development in young children's word recognition from initial individual differences in phonological awareness after controlling for the autoregressive effects of word recognition at initial assessment. Wagner et al. (1997) reported a unique contribution of phonological awareness to individual differences in word recognition over successive two-year intervals from kindergarten through fourth grade (see also Parrila, Kirby, & McQuarrie, 2004, for similar results for first to third grade). However, Torgesen, Wagner, Rashotte, Burgess, and Hecht (1997) did not find a unique contribution from third-grade phonological awareness to fifth-grade word decoding efficiency. de Jong and van der Leij (2002) did not find a unique contribution of phonological awareness to word decoding speed for first- to third-grade Dutch children, and the authors suggested that their null result may have been due to the language's relatively high orthographic regularity. To our knowledge, the only study of reading outcomes in the later grades that has investigated the unique longitudinal prediction of phonological awareness has been Scarborough (1998). For fifty-five participants tested in the second and eighth grades, a single phonological awareness measure did not add any significant predictive power to initial reading level, even when combined with verbal memory, rapid naming, and IQ. Analyses of a reading-disabled sub-sample also failed to find a unique predictive role for phonological awareness. Despite some negative evidence, Bowey (2008) recently called for additional longitudinal work, hypothesizing that some effect of relative phonological weakness must be present to explain the shift from a higher proportion of "Chinese" (exception word reading superior to nonword reading) to "Phoenician" (the reverse pattern) readers at fourth grade to an equal ratio of reading styles at eighth grade.

Studies of remedial interventions for reading disabilities have also explored the unique contribution of individual differences in phonological awareness to subsequent reading performance. Wise, Ring, & Olson (1999, 2000) found that phonological awareness predicted unique variance in word recognition gains from the beginning to the end of a school year for second- to fifth-grade poor readers who participated in several different computer-based training programs. Phonological awareness has also been associated with development in word recognition during remedial reading instruction delivered by trained teachers in the first grade (Hatcher & Hulme, 1999) and during the two years following training for 8 to 10 year-old very poor readers (Torgesen, et al., 2001).

Phonological awareness is commonly assumed to have its beneficial effect for development in word recognition though its influence on phonological decoding skills, typically assessed through the oral reading of nonwords. The ability to phonologically decode printed nonwords is thought to promote development in word recognition by providing a "self teaching" capability when the reader encounters unfamiliar printed words (Share, 1995). Phonological awareness and phonological decoding are usually very highly correlated, and both skills are significantly lower on average in groups of children with English reading and spelling disabilities when compared with younger normally progressing groups matched on reading or spelling raw scores (Friend & Olson, 2008; Johnston & Morrison, 2007; Rack, Snowling & Olson, 1992). Therefore, individual differences in phonological decoding may also predict unique variance in the development of word recognition or spelling after controlling for autoregressive effects. Surprisingly, we found only one published study that tested this hypothesis. Torgesen et al. (2001) reported a significant unique prediction of individual differences in word recognition from phonological decoding across their intensive two-month training period for 8 to 10 year-old poor readers, and from the end of training to a two year follow-up assessment. A recent reanalysis of the Wise et al. (1999; 2000) studies also found that phonological decoding significantly and uniquely predicted development in word recognition for children with reading disabilities in the second through fifth grades (J. Ring, personal communication, January 25, 2007).

Rapid naming of letters, numbers, pictures, and colors is another important reading-related skill that predicts subsequent reading development (Wolf & Bowers, 1999). In an unselected sample, Wagner et al. (1997) reported a significant unique prediction from rapid naming of letters and numbers in kindergarten to development in word recognition in second grade and from first to third grade (see also Parrila et al., 2004, for rapid color naming from first to third grade). However, Wagner et al. did not find a unique prediction from rapid naming in second grade to word recognition in fourth grade and Torgesen et al. (1997) found no unique prediction from third-grade rapid naming to fifth-grade word decoding accuracy and fluency. Furthermore, Meyer, Wood, Hart, and Felton (1998) found that third-grade rapid naming predicted fifth- and eighth-grade word reading only in a subset of the sample scoring below the 10th percentile on third-grade word reading, and Scarborough (1998) reported that second-grade rapid naming predicted eighth-grade literacy outcomes only in a reading-disabled subgroup. Taken together, these studies suggest that the unique prediction of development in word recognition from rapid naming may be limited to children at the lowest levels of reading development, and therefore we might not expect to see any unique prediction across the later grades in the present study.

IQ or general cognitive ability has played a controversial role in the definition of and research on reading disabilities, including its unique ability to predict individual differences in response to instruction. Fuchs and Young (2006) discussed the views of Siegel (1989) and others that "IQ is irrelevant to the definition of learning disabilities," and they reviewed the evidence for Siegel's specific claim that IQ is unrelated to response to instruction. In fact, the Fuchs and Young review of studies that tested IQ as a unique predictor of response to instruction, including the study by Wise et al. (1999), found that the majority of studies did find significant, unique prediction from IQ, particularly for gains in reading comprehension. However, it is important to note that all of the studies reviewed included children in the early grades who participated in remedial reading programs, with the fifth grade being the highest grade level in any study. Looking at outcomes in the later grades, Scarborough (1998) found that second-grade full-scale IQ contributed unique variance to eighth-grade spelling, but not word identification or passage comprehension, only in an RD sub-sample, and, unexpectedly the direction of the effect was inverse such that lower IQ tended to indicate higher spelling. Because of these inconsistent results, the current study seeks to clarify the relationship between IQ, its

subscales, initial reading performance, and literacy achievement in an even older group of readers.

The possibility for reading-related skills to uniquely predict development in reading skills across any interval will be influenced in part by the amount of variance in reading at the final assessment that remains unexplained by variance in reading at the initial assessment. Clearly, if the initial and final reading assessments are perfectly correlated, there is no reading outcome variance that remains to be explained by a reading-related skill after controlling for the autoregressor. Of course the longitudinal correlations for measures of reading skills are never perfect, but they do tend to be quite high across the early grades if the measures are reliable. Consistency in reading skills across the early grades has been estimated from consistency in categorical classifications of poor and good readers, such as Juel's (1988) finding that a child classified as a poor reader in grade 1 had a .88 probability of being similarly classified in grade 4. Since reading ability is normally distributed in the population (Rodgers, 1983), it is appropriate to assess the stability of individual differences in reading ability across the normal range through longitudinal correlations. For example, de Jong and van der Leij (2002) reported a longitudinal correlation of .69 for word recognition from the first to the third grade for children learning to read Dutch. Parrila et al. (2004) reported more modest correlations from the first semester of first grade to assessments in the second through fifth grades, but their longitudinal correlations for second grade assessments ranged from .93 with third grade to .81 with fifth grade assessments (see also Wagner et al., 1997). Thus, by the second grade, individual differences in word recognition are highly though not perfectly stable through the fifth grade, constraining the amount of additional variance in reading outcomes that can be explained by individual differences in reading related skills after controlling for the autoregressor.

The present study utilized data collected as part of the Colorado Longitudinal Twin Study of Reading Disability (LTSRD) (Wadsworth, DeFries, Olson & Willcutt, 2007), a follow-up of participants in the Colorado Learning Disabilities Research Center (CLDRC) (DeFries et al., 1997). Data from a subset of the LTSRD sample were used to explore the longitudinal prediction of individual differences in the development of word recognition, spelling, and reading comprehension across the later grades. The reading-related skills of phoneme awareness, phonological decoding, rapid naming, and IQ were included as predictors. All of the reading and related skills, except for reading comprehension, were assessed with at least two measures that were used as indicator variables for latent trait models that reduce the influence of individual test measurement error. To our knowledge, this is the first study to assess the longitudinal stability of latent traits for reading and spelling or to use latent traits for reading-related skills to predict development in latent traits for reading after controlling for initial reading level.

The literature described above leads us to formulate several predictions. Because essentially the same test battery was used at initial and follow-up assessments, we were able to explore the longitudinal stability of the factor structure among the individual measures. Based on previous longitudinal studies of reading comprehension across the early grades (cf., Catts, Hogan, & Adlof, 2005), we predicted that the loading of reading comprehension on a word reading/spelling factor would decline across the initial and follow-up assessments, but that reading comprehension would maintain or increase its relation with IQ. In longitudinal prediction analyses, such a result would leave room for independent effects of IQ on later comprehension. Conversely, we expect that there may be a decline over the later grades in the unique predictive power of phonological awareness and phonological decoding for reading and spelling when compared to the results of studies on younger children, given the results of Wagner et al. (1997), Scarborough (1998), and de Jong and van der Leij (2002). While Scarborough (1998) most closely addressed the questions raised here, the current study offers insight into these processes even further into schooling. In addition, the power provided by the

larger sample is important when some null results are hypothesized. Where significant results are hypothesized, as for IQ and comprehension, subscale information is available for specifying the nature of effects. Finally, this study improves on previous research by employing multiple measures modeled as latent traits, removing the possibility that correlated errors may explain relations among measures.

Methods

Participants

Participants were 324 children who were tested in the LTSRD (Wadsworth et al., 2007). Participants were originally tested between the ages of 8 and 13.5 years of age (mean = 10.2, SD = 1.3) in the CLDRC (DeFries, et al., 1997), and were retested an average of 5 years and 6 months later (SD = 8.6 months) when they were between 13.5 and 19 years old (mean = 15.8, SD = 1.4). For participation in the CLDRC, twin pairs and their non-twin siblings were ascertained from school records from 27 Colorado school districts. Participants who were initially tested at the CLDRC between September 1996 and December 2001 were invited for follow-up testing. Fifty-five percent of families contacted agreed to participate in follow-up testing.

One-hundred and fourteen of the individuals included in the present study had a broadly defined history of reading difficulties based on school records and/or parent report, though many of the twins with this school history had only mild reading deficits when tested in the laboratory and many of their co-twins and non-twin siblings had no reading deficits. Furthermore, individuals that participated in follow-up testing tended to have higher average IQ and reading scores at their initial assessment than the full initial sample (Wadsworth et al., 2007). As a result, when the school-history and no-school-history samples were combined, individual differences in reading standard scores were normally distributed with means and standard deviations that approximated those for the tests' norming samples. Therefore, this combined sample of individuals with positive and negative histories of reading difficulties is appropriate for our main analyses of individual differences across the normal range.

Measures

All participants were administered an extensive test battery for reading and related cognitive skills during two testing sessions at the University of Colorado. The subset of the entire battery of measures that was used in the current analyses is described here. For those tests for which revisions have been published (see below for the PIAT-R, WRAT-3, and WISC-III/WAIS-III), the newer versions were used at follow-up, and all data transformations and analyses are based on standard scores. Unadjusted raw and standard scores were used for descriptive analyses only. Correlational, factor, and predictive analyses are based on age-adjusted z-scores standardized across the full sample of the current study. These z-scores were trimmed at +/− 3 SD to minimize the undue influence of outliers, and follow-up data for nonpublished tests were power-transformed (Judd & McClelland, 1989, chap. 16) to limit ceiling effects and provide approximately normal distributions (details below).

Word Recognition

Timed Word Recognition Test (TWRT) (Olson, Forsberg, Wise, & Rack, 1994)—
The TWRT consists of words presented on a computer screen in order of increasing difficulty, as assessed in an independent sample. Responses are considered correct only when the correct pronunciation of the word is initiated within two seconds of stimulus onset. Testing continues through a list of 182 items until the participant fails to answer 10 of the last 20 items correctly within the time limit or the end of the list is reached. Raw scores are based on the last word

read. Test-retest reliability is .93. At retest, data were squared before age-adjustment and standardization to reduce the effects of a skew of −0.78 (SE = .14).

Peabody Individual Achievement Test Word Recognition (PIAT Rec) (Dunn & Markwardt, 1970; Markwardt, 1989)—The PIAT Rec task presents words of increasing difficulty in rows across a page. The participant reads the words aloud in sequence until 5 of the last 7 items are missed or the end of the list is reached. Standard scores are provided. The published correlation between the PIAT and the PIAT-R for this task is .88. Published test-retest reliabilities are .89 for the PIAT and .96 for the PIAT-R.

Phoneme Awareness

Phoneme Deletion (PhoDel) (Olson et al., 1994)—The Phoneme Deletion task consisted of 6 practice and 40 test trials presented via CD player in which the participant repeated a nonword and was then asked to say it again, deleting a specified phoneme to form a real word ("say prot – now say prot without the/r/"). Participants were given two seconds for repetitions and four seconds for deletions, as signaled by a warning tone on the CD. Raw data for this task consist of percent correct scores. Chronbach's alpha is .93. At retest, data were squared before age-adjustment and standardization to reduce the effects of a skew of −1.11 (SE = .14).

Phoneme Segmentation and Transposition (Phoneme S&T) (Olson, Wise, Connors, Rack, & Fulker, 1989)—The Phoneme S & T task required participants to play a word game similar to "Pig Latin" where they take the first phoneme off the front of a word, move it to the end of the word, and add a long "a" sound. A percent correct score was calculated for the 45 test items. Reliability estimated from the correlation with a composite Phoneme Deletion measure is .78. At retest, data were squared before age-adjustment and standardization to reduce the effects of a skew of −.91 (SE = .14).

Phonological Decoding

Phonological Choice (Pho Choice) (Olson et al., 1994)—The Pho Choice task consisted of 60 items requiring participants to select which of three nonwords would sound like a real word (beal *bair* rabe). Raw data consist of percent correct scores. Reliability estimated from the correlation with Oral Nonword Reading is .80. At retest, data were squared before age-adjustment and standardization to reduce the effects of a skew of −1.08 (SE = .15).

Oral Nonword Reading (Nonwords) (Olson et al., 1994)—The Nonwords task consisted of reading 45 one-syllable (ter, strale) and 40 two-syllable (vogger, strempick) nonwords aloud. Percent correct scores were calculated for each task. Test-retest reliability is .86. At retest, data were squared before age-adjustment and standardization to reduce the effects of a skew of −1.32 (SE = .15).

Spelling

Word-Pseudohomophone Choice (WPh Choice) (Olson et al., 1994)—The WPh Choice task required participants to distinguish a real word from a nonword with the same pronunciation (rane *rain*). There were 80 items, and raw scores were based on percent correct. Split-half reliability is .93. At retest, data were squared before age-adjustment and standardization to reduce the effects of a skew of −1.36 (SE = .14).

Peabody Individual Achievement Test Spelling Recognition (PIAT Spelling) (Dunn & Markwardt, 1970; Markwardt, 1989)—PIAT Spelling requires participants to choose the correct spelling of an orally presented target word from four orthographically and

often phonologically similar printed alternatives. Target words are administered in increasing order of difficulty, and the task is discontinued if the participant makes five errors in any seven consecutive responses. Standard scores are provided. The published correlation between the PIAT and the PIAT-R for this task is .76. The published test-retest reliabilities are .65 for the PIAT and .88 for the PIAT-R.

Wide Range Achievement Test Spelling Production (WRAT Spelling) (Jastak & Wilkinson, 1984; Wilkinson, 1993)—The WRAT-R Spelling test consists of 45 items, and the WRAT-3 consists of 40 items. Stimuli are administered in increasing order of difficulty. The test is discontinued after 10 consecutive spelling errors and standard scores are provided. The published alternate form reliability is .90.

Comprehension

Peabody Individual Achievement Test Reading Comprehension (PIAT Comp) (Dunn & Markwardt, 1970; Markwardt, 1989)—PIAT Comp requires participants to silently read short one- or two-sentence passages and choose one of four pictures that express the meaning of the sentence(s). Standard scores are provided. The PIAT-R includes some more complicated paragraphs, and its published correlation with the original PIAT is .79. The published test-retest reliabilities are .64 for the PIAT and .88 for the PIAT-R.

General Intelligence

Wechsler Intelligence Scale (WISC-R, WISC-3, WAIS-3) (Wechsler, 1974, 1991, 1993)—At initial testing, all participants were administered the WISC-R. At follow-up, participants were administered the WISC-3 or WAIS-3 as appropriate for their age. All subtests were administered, allowing the calculation of performance, verbal, and full-scale IQ. The published correlation for the WISC-R and WISC-3 is .89. Published test-retest reliabilities for all versions of full-scale IQ are .96–.97.

Rapid Naming

Rapid Naming (RAN) (Decker, 1989; after Denckla & Rudel, 1976)—Participants were presented four separate cards, each displaying randomly arranged letters, numbers, colors, or pictures. Participants were presented with the cards one at a time, and asked to name the items from left to right as quickly as possible. Raw scores were equivalent to the number of items named on each card in 15 seconds. A conservative estimate of reliability for our RAN measure is its correlation with the Denkla and Rudel test at .85 (Compton, Olson, DeFries, & Pennington, 2002).

Analyses

The primary analyses, including means comparisons, longitudinal stability correlations, exploratory and confirmatory factor analyses, and the final longitudinal predictions for individual differences in reading skills were all based on the full sample to maximize power. However, since the sample consists of twin pairs and their school-age siblings, there is a partial lack of independence within families. Therefore, we also randomly selected one member of each family and repeated the longitudinal prediction analyses with this independent sub-sample (N = 153) to more conservatively assess their statistical significance.

Where multiple measures of a construct were available, latent traits were created using Amos software for structural equation modeling (Arbuckle, 2005). This method models the common variance among the individual measures, and the results differ from composite results in that individual measures may be weighted differently and measure-specific error is removed

from the correlation. The age-adjusted, standardized, and trimmed versions of each variable were utilized when modeling latent traits.

Results

Descriptive Statistics

Table 1 reports the means and standard deviations for the individual measures at each time point. Analyses were limited to individuals who had data for a given measure at both time points so as to be most comparable to the stability analyses that follow. For measures without published norming data, raw scores are reported to indicate absolute gains in performance over time. On average, scores for each of these measures increased over the testing interval ($p < .05$), indicating that this group did experience absolute growth in these areas. Scores for standardized tests are reported as standard scores, and as such are adjusted for age at time of testing. Mean standard scores for PIAT Comp, PIAT Spelling, Verbal IQ, and Performance IQ were slightly but significantly lower at the second testing point, whereas scores for PIAT Rec and WRAT Spelling were slightly but significantly higher ($p < .05$). These differences may be due in part to the more recent norms applied to the tests administered at follow-up compared to those at initial testing.

Correlations

Correlations among measures are listed in Table 2. All correlations are based on age-adjusted *z*-scores or, when multiple measures were available, the latent traits thereof. Correlations among measures at initial testing are shown above the diagonal, and correlations at follow-up are shown under the diagonal. The same-measure correlations across the testing interval, which averaged five and a half years, are presented on the diagonal. All correlations are significant at the $p < .01$ level.

Longitudinal correlations ranged from .65 for the single measure of comprehension to 1.0 for the latent traits of Phoneme Awareness and Full Scale IQ. Although all measures within each time-point were significantly related, correlations varied in magnitude between .25 and .95. Furthermore, both similarities and differences can be seen between the pattern of correlations at initial testing and that at follow-up. These relations among measures across constructs and time are explored further in factor analyses.

Factor Analyses

Age-adjusted, standardized data were subjected separately by time point to exploratory factor analysis with promax rotation, an oblique method that allows the factors to correlate, as expected based on the correlation results. Table 3 reports the structure matrix for the four meaningful factors extracted in each analysis. Loadings with substantial unique contributions, as indicated by the pattern matrix, are displayed in bold. At each time-point, only the first three factors have Eigen values greater than one, but the fourth factor, characterized by high IQ loadings, is also reported because of its interpretability and marginal Eigen value greater than 0.8. However, interpretation of the results should be tempered with the fact that this final factor does not capture a large proportion of the overall variance.

The first factor at initial testing can be characterized as a literacy factor, with high loadings from word and nonword reading, comprehension, and spelling measures, and moderate loadings from verbal IQ and phoneme awareness measures. The pattern matrix for this factor confirms that this factor is characterized by unique contributions from word reading, spelling, comprehension, and verbal IQ. The second factor can be considered a phonological factor, with unique contributions from phoneme awareness and phonological decoding measures and high overall loadings for these measures along with word reading and moderate loadings for spelling

and comprehension. The third factor is characterized by high loadings, both unique and overall, on the rapid naming measures. The highest loading for the fourth factor comes from performance IQ, with moderate loadings for verbal IQ and comprehension. As expected, these factors are correlated, and r values range from .70 for the literacy and phonological factors to .24 for the RAN and IQ factors.

When these initial results are compared to those from follow-up testing, generally similar variables and loadings characterize the factors, indicating that relations among these skills are largely stable over time. Interestingly, differences in unique factor loadings across time did appear for the comprehension measure. At younger ages, presumably when decoding limits understanding, the comprehension measure loads most highly with the word-level literacy measures. At older ages, comprehension becomes most closely associated with IQ. In order to quantify the stability of the factor structure, factor scores were created. The literacy factor at initial testing correlated .88 with the first factor at follow-up, and the phonological factor correlated .89 with the second factor at follow-up. The RAN and IQ factors captured the smallest and least stable variance, with longitudinal correlations of .29 and .13, respectively.

Confirmatory factor analysis provides a direct significance test of the hypothesis that the factor structure was similar over time. The results of the exploratory analysis of initial testing data were used to fit a confirmatory model to data at each time point, as shown in Figure 1. In the interest of simplicity, the RAN variables were not included because they did not load heavily with any of the reading or language measures and they did not appear to show significant change over time. A path was included in the confirmatory model if its loading in pattern matrix of the exploratory analysis was greater than .5 at either time-point. The standardized solution is displayed, with single-headed arrows representing factor loadings and double-headed arrows representing covariance. Observed variables are shown as rectangles and latent variables, which correspond to the factors, are displayed as ovals. Although not explicitly shown in the figure in the interest of simplicity, each observed variable was assumed to have unique error variance. Latent variables at initial testing were assumed to be correlated.

Model fitting results are listed in Table 4. Model 1 is a test of structural invariance; it assumes the same factors and paths, but not necessarily loadings, exist at each time point. The model fit was good to acceptable based on the Comparative Fit Index (CFI), for which values greater than .9 are conventionally taken to indicate significant improvement over a null model and values over .95 are considered "good" (Hu & Bentler, 1999). Model 2 tests metric invariance by assuming corresponding factor loadings are equivalent at initial and follow-up testing. This model fits the observed data significantly less well than Model 1. To determine which loadings contributed to this loss of fit, we tested models constraining only one pair of loadings at a time. When compared against Model 1, only one submodel, listed as Model 3, resulted in significant loss of fit (nonsignificant submodels are not shown). The factor loading that was significantly variable over time was that between PIAT Comp and the Literacy latent trait, which is primarily based on word-level reading and spelling. Inspection of Table 3 indicates that although PIAT Comp loads most highly with Literacy at initial testing, this measure shares more variance with IQ at follow-up. In Model 4, therefore, all factor loadings were constrained to be equal except for the one from Literacy to PIAT Comp which contributes significantly to loss of fit. The fit of Model 4 to the data did not significantly differ from Model 1, but was more parsimonious in its assumption that most factor loadings did not differ over time.

Stability of Residuals

In order for reading-related skills to provide meaningful predictions of literacy outcomes beyond the autoregressive effect, they must be stable and reliable. If only noise remains after controlling for variance in the initial reading measures, null results in the developmental analyses would be both unsurprising and uninteresting. Therefore, we explored the longitudinal

stability of the residual variance in potentially predictive skills after controlling for the initial literacy measures. Residual variables were created by regressing out the effects of the various initial literacy measures (Word Recognition latent trait, Spelling latent trait, and PIAT Comprehension) from variance in latent traits of the concurrently measured related skills (Phoneme Awareness, Phonological Decoding, IQ, and RAN). Table 5 displays the correlations for these residuals from initial to follow-up testing. All of these longitudinal correlations are highly significant ($p < .001$), indicating substantial stability. It is important to note that the error variance of the individual measures is not included in these regression residuals based on the latent traits.

Developmental Cholesky Models

The reliability of the residuals for the reading-related skills from initial to follow-up tests is a limiting factor for their prediction of reading development. Because the residual variance in reading-related skills relative to literacy outcomes is substantially reliable, we investigated the predictive value of initial reading-related skills above and beyond the autoregressive effect of the literacy variables at initial testing. Even though the longitudinal correlations in Table 2 suggest that initial literacy will be a very strong predictor of later literacy, the variance around mean difference scores reported in Table 1 indicates additional individual variation in development over time that could potentially be predicted by other language or cognitive measures.

Because of the ability to model our multiple measures as latent traits, thereby separating measure-specific error variance, we chose to use a specific type of structural equation model, the Cholesky decomposition, implemented using maximum-likelihood estimation in the Amos program (Arbuckle, 2005). A template Cholesky model is displayed in Figure 2. Cholesky analyses can be thought of as hierarchical regressions with latent traits. The latent trait F1 captures the variance common among the literacy variable at initial testing, the predictive skill of interest, and the literacy variable at follow-up. The factor loading on the F1-to-follow-up path (3-1) represents the autoregressive effect. The F2 latent trait captures the variance common to the predictive skill of interest and the follow-up outcome, *independent* of the initial literacy variance, as if it were the second step in a hierarchical regression. Thus, a significant factor loading from F2 to follow-up (3-2) represents a significant unique contribution of the reading-related skill to variance in the follow-up literacy measure after controlling for initial scores on that same literacy measure. The factor loading for the F3 latent trait (3-3) represents variance in the follow-up literacy measure that is unrelated to either initial literacy measure or the predictive measure.

Twelve models were tested: three literacy measures by four reading-related skill measures. The three literacy measures were a latent trait for Word Recognition (consisting of loadings from PIAT Rec and the TWRT), a latent trait for Spelling (WPh Choice, PIAT Spelling, and WRAT Spelling), and the single PIAT Comprehension measure. The four reading-related skill measures were latent traits for Phoneme Awareness (Phoneme S & T and Pho Del), Phonological Decoding (Pho Choice and Nonwords), IQ (Verbal and Performance IQ), and RAN (Letters, Numbers, Pictures and Colors). Models using PIAT Comprehension as the predicted variable differ from the template in Figure 2 in that there is only a single measure. These Comprehension models were run twice, once with no error removed and once with error preset based on the published reliability of the test. Models using Spelling and RAN measures differ from the template in Figure 2 because the reading-related skill latent trait consists of not two, but three and four observed measures, respectively.

Table 6 shows the results for the Cholesky models tested. The models fit the data well, as indicated by CFI values ranging from .956 to .999. When the factor loadings displayed are squared, the resulting value can be interpreted as the percentage variance explained in each

latent trait. Significant F1 factor loadings for all models indicate that a substantial proportion of variance is shared among all variables. Loadings on the autoregressive paths indicate high stability for the literacy measures, especially those modeled as latent traits: initial outcome scores explained 96% of the variance in follow-up Word Recognition and 90% of the variance in follow-up Spelling. For Word Recognition and Spelling analyses, the only component skill to contribute significant variance to follow-up outcome independent of these large autoregressive effects was IQ for Word Recognition. However, the magnitude of the unique variance explained is quite small, less than 1%.

Table 6 displays the results for the single PIAT Comprehension measure when no error is removed. The autoregressive effect was substantial but less strong, with initial scores explaining 41% of the variance in follow-up scores. Three reading-related skills significantly predicted follow-up Comprehension above and beyond the autoregressive effect. Initial Phoneme Awareness and initial Phonological Decoding each explained 2–2.6% unique variance in follow-up PIAT Comprehension. IQ predicted 13% of the variance in follow-up Comprehension independent of what is predicted by initial Comprehension. Because Phoneme Awareness, Phonological Decoding, and IQ are correlated, we explored whether these unique predictive variances were independent of one another. When IQ was entered in a four-factor Cholesky model before Phoneme Awareness or Phonological Decoding, those measures no longer contributed significant unique variance to follow-up Comprehension. Conversely, when Phoneme Awareness or Phonological Decoding was entered before IQ, IQ still explained more than 11.5% of the variance in follow-up comprehension. Thus, the contributions of Phonological Decoding and Phoneme Awareness to Comprehension appear to be explained by their relation with IQ.

When the Comprehension models were run with error variance preset to .36 for Initial PIAT, based on a published reliability of .64, and .22 for Follow-up PIAT-R, based on a published reliability of .88, patterns for the factor loadings were similar. The unique predictions of Phonological Decoding and Phonological awareness fell below significance. The unique prediction of IQ was smaller but still significant, explaining 4% of the variance after controlling for the autoregressive effect.

Initial RAN measures did not explain significant unique variance in any of the follow-up literacy measures; all of their relations to follow-up outcome were shared with initial literacy.

Parallel Cholesky analyses were conducted for a sub-group of participants consisting of one randomly selected member from each family (N = 153). The pattern of significant and nonsignificant effects was identical to that for the full sample with two exceptions. The unique prediction of Comprehension by Phoneme Awareness and Phonological Decoding only reached marginal levels of significance in the independent sample (*p* = .06 and *p* = .10 respectively). Importantly, the significant prediction of Comprehension by IQ did replicate in the non-family sample at the *p* < .05 level, despite the substantial reduction in sample size.

Discussion

The primary goal of the present study was to see if individual differences in the reading-related skills of phoneme awareness, phonological decoding, rapid naming, and IQ could uniquely predict individual differences in the development of reading, spelling, and reading comprehension across the later grades. To accomplish this goal, we analyzed data from twins and their siblings who participated in the Colorado Learning Disabilities Research Center when they were initially tested on a broad range of reading and related skills at an average age of 10.2 years, and were subsequently tested on the same measures in the Colorado Longitudinal Twin Study of Reading Disability at the average age of 15.8 years. Before we discuss the

prediction results from the reading-related skills, we will first discuss the results of several preliminary analyses that focused on longitudinal stability of the measures and their factor structure.

The vast majority of children showed significant improvement in raw scores on all measures across the average five-and-a-half year testing interval. In contrast, the children's standard scores on the standardized reading, spelling, and IQ measures were generally quite consistent across the testing interval, reflecting consistent rates of reading development on average, in comparison to same-age peers in the tests' norming samples. The individual-difference stability correlations for our latent traits of word recognition (.98) and spelling (.95) were very high. In contrast, the stability correlation for our single PIAT reading comprehension measure was significantly lower (.65). At least part of this lower stability correlation may be due to a lower reliability for this single measure at each test occasion (.64 and .88, respectively). The stability correlations for the latent traits for the reading-related skills of phoneme awareness (1.0), phonological decoding (.93), full scale IQ (1.0), and rapid naming (.82) were also quite high. Although these correlations generally indicate a high level of longitudinal stability for most measures, they do not necessarily indicate that the relations among the different measures are stable across the testing interval, so we turn to that question next.

Our exploratory factor analyses suggested a remarkably high level of consistency in factor structure across the testing interval, with the notable exception of reading comprehension. In a confirmatory factor analysis based on the results of the exploratory analysis, there was a significant loss of model fit when we equated all the factor loadings from initial to final test, but not when we allowed the loading of reading comprehension on the first literacy factor, defined primarily by word reading and spelling, to vary across the testing occasions. Thus, it was clear that reading comprehension was significantly more linked to variance in word recognition and spelling when the children were younger than when they were older. This effect can also be seen in the correlation analyses; comprehension correlated more highly with word recognition and spelling at initial testing than at follow-up. Our results show that the developmental shift in the dependence of comprehension on word recognition that has previously been reported in a number of studies with younger children is still in effect across the later grades.

Now we turn to the main goal of our study, to assess the prediction of reading development across the later grades from reading-related skills after controlling for the autoregressor effect of reading at initial test. A critical first step toward this goal was to determine the longitudinal stability of individual differences in reading-related skills that were independent from the reading and spelling measures at each test occasion. The longitudinal stability of the residual variance in reading related skills is important for two reasons. The first is that residual variance in a predictor variable may simply be measurement error after controlling for highly correlated reading skills. The second is that if the residual predictor variance at initial testing is expected to influence individual differences in reading development across the later grades, it should be at least moderately consistent across that interval. This is what we found when we assessed the longitudinal stability correlations for the residuals of each predictor with each reading skill: all of the stability correlations in Table 5 were greater .7 and all were highly significant.

In spite of the significant and longitudinally reliable residual variance in our reading-related predictor latent traits of phoneme awareness, phonological decoding, IQ, and rapid naming, there was only one statistically significant prediction from initial test to follow-up for the word recognition and spelling latent traits. That was from IQ (primarily the verbal sub-scales) to follow-up word recognition, but even that one significant prediction accounted for less than one percent of the variance. Thus, on average, children's profiles of reading-related skills that were independent from their word-reading and spelling skills had practically no influence on

the development of those skills across the later grades. We will consider the implications of these results for theory and practice after discussing the very different outcome for reading comprehension.

In contrast to the essentially null results for the longitudinal prediction of word recognition and spelling from reading-related skills after autoregressor control, there were significant longitudinal predictions for individual differences in reading comprehension from phoneme awareness, phonological decoding, and IQ. However, since IQ was significantly correlated with both phoneme awareness and phonological decoding, we also assessed the independent contributions of each predictor after controlling for the other two predictors. Both phoneme awareness and phonological decoding lost their significance after controlling for IQ. In contrast, the remaining variance in reading comprehension that was predicted by IQ after controlling for phoneme awareness (11.6%) or phonological decoding (11.8%) was highly significant, and only slightly below the 13% of variance predicted by IQ without these controls. Even in a conservative test assuming high error variance in the comprehension measures as suggested by their published reliabilities, IQ remained a significant unique predictor. A related result reported by Torgesen et al. (1997) was that 10% of the unique variance in Woodcock (1987) reading comprehension test performance in the fifth grade was predicted by the children's vocabulary score from the Stanford-Binet IQ test (Thorndike et al., 1986) in the third grade. The present results show that verbal skills including vocabulary that are independent from reading comprehension at an initial test occasion also predict the development of reading comprehension across the later grades.

The lack of longitudinal prediction from children's reading-related skills for the development of word recognition and spelling across the later grades stands in sharp contrast to their significant prediction of word reading and spelling across the early grades that we reviewed in the Introduction. Should we conclude that phonological awareness and phonological decoding have no causal relation to the development of word recognition or spelling in older children, when in fact they are significantly correlated with word recognition and spelling at both initial and final tests? Researchers who have previously faced this conundrum have argued that once children move beyond the very early grades, the causal influence their reading-related skills have on the development of word recognition has already been fixed in their word-reading and spelling ability. Thus, when we control for initial word-reading level beyond the very early grades, we are also controlling for the causal effects of the reading-related skills that have already defined the trajectory for future development. Once this trajectory is set in early reading development, only *new* causal influences from reading-related skills that emerge later in development can be detected from their independent prediction after controlling for the autoregressive effects of prior reading (de Jong & van der Leij, 2002; Torgesen et al., 1997).

New causal influences might include an intervention during the later grades to remediate poor readers' deficits in phonological awareness and decoding, since these skills tend to be lower than expected from poor readers reading and spelling skills (Friend & Olson, 2008; Rack, Snowling, & Olson, 1992). Interventions to improve poor readers' phonological skills in the early grades do show significant benefits for subsequent reading development (c.f., Torgesen et al., 1999, Wise et al., 2000). However, the present study found that the residual variance in phonological skills did not predict individual differences in reading development across the later grades. Perhaps consistent with this lack of prediction across the later grades, both Wise et al. (2000) and Torgesen et al. (2001) found no unique advantage from phonological training compared to intervention conditions with little or no phonological training for poor readers past the third grade. Therefore, phonological interventions for poor readers in the later grades may not serve well as *new* influences that would significantly change their development of word reading and spelling or reading comprehension. It is important to note that there may well be other reading-related skills, such as morphological knowledge, that could be significant

predictors of reading and spelling development and useful targets for intervention in the later grades. Furthermore, results may differ for reading outcomes not measured by this study. For example, growth in reading fluency was predicted by the non-alphanumeric RAN performance of Norwegian pre-readers (Lervåg & Hulme, in press).

The remarkably high levels of stability for individual differences in word reading, spelling, and reading-related skills across the later grades in the present study are consistent with strong genetic influences on individual differences in these reading and related skills that are evident as early as the end of first grade (Byrne et al., 2007; Harlaar, Spinath, Dale, & Plomin, 2005), and continue through the later grades (Gayán & Olson, 2003). Cultural influences on reading development (i.e., schools and peer groups) may also be fairly consistent across the later grades in our sample. Nevertheless, we are only studying the current state of stability for individual differences in reading development. The high levels of stability we have found do not constrain what could be. Regardless of the consistency of genetic and cultural influences on learning rates for reading and related skills across development in the present sample, broad cultural changes in the quantity and quality of reading practice could significantly improve reading development across the later grades for most children, even if the longitudinal stability of their performance relative to their peers remained very high.

Acknowledgments

References

Arbuckle, JL. Amos 6.0 [computer software]. Spring House, PA: Amos Development Corp; 2005.

Boder E. Developmental dyslexia: A diagnostic approach based on three atypical reading and spelling patterns. Developmental Medicine and Child Neurology 1973;15:663–687. [PubMed: 4765237]

Bowey JA. Is a "Phoenician" reading style superior to a "Chinese" reading style? Evidence from fourth graders. Journal of Experimental Child Psychology 2008;100:186–214. [PubMed: 18054365]

Bryant PE, Impey L. The similarity between normal readers and developmental and acquired dyslexics. Cognition 1986;24:121–137. [PubMed: 3791918]

Byrne B, Samuelsson S, Wadsworth S, Hulslander J, Corley R, DeFries JC, Quain P, Willcutt E, Olson RK. Longitudinal twin study of early literacy development: Preschool through Grade 1. Reading and Writing: An Interdisciplinary Journal 2007;20:77–102.

Catts, HW.; Hogan, TP.; Adlof, SM. Developmental changes in reading and reading disabilities. In: Catts, HW.; Kamhi, AG., editors. The connections between language and reading disabilities. Mahwah, NJ: Erlbaum; 2005.

Compton DL, Olson RK, DeFries JC, Pennington B. Are all RAN created equal? Comparing the relationships among two different formats of alphanumeric RAN and various word reading skills in normally achieving and reading disabled individuals. Scientific Studies of Reading, 6 2002;343:368.

Decker SN. Cognitive processing rates among disabled and normal young adults: A nine year follow-up study. Reading and Writing: An Interdisciplinary Journal 1989;2:123–134.

DeFries JC, Filipek PA, Fulker DW, Olson RK, Pennington BF, Smith SD, et al. Colorado Learning Disabilities Research Center. Learning Disability Quarterly 1997;8:7–19.

Denckla MB, Rudel RG. Rapid "automatized" naming (R.A.N.): Dyslexia differentiated from other learning disabilities. Neuropsychologia 1976;14:471–479. [PubMed: 995240]

de Jong P, van der Leij A. Effects of phonological abilities and linguistic comprehension on the development of reading. Scientific Studies of Reading 2002;6(1):51–77.

Dunn, LM.; Markwardt, FC. Peabody Individual Achievement Test. Circle Pines, MN: American Guidance Service; 1970.

Friend A, Olson RK. Phonological spelling and reading deficits in children with spelling disabilities. Scientific Studies of Reading 2008;12(1):90–105.

Fuchs D, Young CL. On the irrelevance of intelligence in predicting responsiveness to reading instruction. Exceptional Children 2006;73(1):8–30.

Gayán J, Olson RK. Genetic and environmental influences on individual differences in printed word recognition. Journal of Experimental Child Psychology 2003;84:97–123. [PubMed: 12609495]

Gollob HF, Reichardt CS. Taking account of time lags in causal models. Child Development 1987;58:80–92. [PubMed: 3816351]

Harlaar N, Spinath FM, Dale PS, Plomin R. Genetic and environmental influences on word recognition abilities and disabilities: A study of 7 year old twins. Journal of Child Psychology and Psychiatry 2005;46:373–384. [PubMed: 15819646]

Hatcher P, Hulme C. Phonemes, rhymes, and intelligence as predictors of children's responsiveness to remedial reading instruction. Journal of Experimental Child Psychology 1999;72:130–153. [PubMed: 9927526]

Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling 1999;6(1):1–55.

Jastak, S.; Wilkinson, GS. The Wide Range Achievement Test-Revised: Administration manual. Wilmington, DE: Jastak Associates, Inc.; 1984.

Johnston RS, Morrison M. Toward a resolution of inconsistencies in the phonological deficit theory of reading disorders: Phonological reading difficulties are more severe in high-IQ poor readers. Journal of Learning Disabilities 2007;40(1):66–79. [PubMed: 17274548]

Judd, CM.; McClelland, GH. Data Analysis: A model-comparison approach. New York: Harcourt Brace Jovanovich; 1989.

Juel C. Learning to read and write: A longitudinal study of 54 children from first through fourth grades. Journal of Educational Psychology 1988;80:437–447.

Lervåg A, Hulme C. Rapid Naming (RAN) taps a basic neuron-developmental constraint on the development of reading fluency. Psychological Science. in press.

Markwardt, FC, Jr. Peabody Individual Achievement Test-Revised. Circle Pines, MN: American Guidance Service; 1989.

Meyer MS, Wood FB, Hart LA, Felton RH. Selective predictive value of rapid automatized naming in poor readers. Journal of Reading Disabilities 1998;31:106–117.

Olson, R.; Forsberg, H.; Wise, B.; Rack, J. Measurement of word recognition, orthographic, and phonological skills. In: Lyon, GR., editor. Frames of Reference for the Assessment of Learning Disabilities: New views on measurement issues. Baltimore: Paul H. Brookes Publishing Co; 1994. p. 243-277.

Olson RK, Wise B, Connors F, Rack J, Fulker D. Specific deficits in component reading and language skills: Genetic and environmental influences. Journal of Learning Disabilities 1989;22:339–348. [PubMed: 2738467]

Parrila R, Kirby JR, McQuarrie L. Articulation Rate, Naming Speed, Verbal Short-Term Memory, and Phonological Awareness: Longitudinal Predictors of Early Reading Development? Scientific Studies of Reading 2004;8(1):3–26.

Perfetti, CA. Reading Ability. New York: Oxford University Press; 1985.

Perfetti CA, Beck I, Bell LC, Hughes C. Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. Merrill-Palmer Quarterly 1987;33:283–319.

Rack JP, Snowling MJ, Olson RK. The nonword reading deficit in developmental dyslexia: a review. Reading Research Quarterly 1992;27(1):28–53.

Rodgers B. The identification and prevalence of specific reading retardation. British Journal of Educational Psychology 1983;53:369–373. [PubMed: 6652037]

Scarborough H. Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid naming, and IQ. Annals of Dyslexia 1998;48:115–136.

Share DL. Phonological recoding and self teaching: Sine qua non of reading acquisition. Cognition 1995;55:151–218. [PubMed: 7789090]

Siegel LS. IQ is irrelevant to the definition of learning disabilities. Journal of Learning Disabilities 1989;21(5):469–478. [PubMed: 2794763]

Thorndike, RL.; Hagen, EP.; Sattler, JM. Guide for administering and scoring the Stanford-Binet Intelligence Scale. 4. Chicago: Riverside; 1986.

Torgesen JK, Alexander AW, Wagner RK, Rashotte CA, Voeller KS, Conway T. Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. Journal of Learning Disabilities 2001;34(1):33–58. [PubMed: 15497271]

Torgesen JK, Wagner RK, Rashotte CA, Burgess S, Hecht S. Contributions of phonological awareness and rapid automatic naming ability to the growth of word-reading skills in second- to fifth-grade children. Scientific Studies of Reading 1997;1:161–185.

Torgesen JK, Wagner RK, Rashotte CA, Rose E, Lindamood P, Conway T, Garavan C. Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. Journal of Educational Psychology 1999;91(4):579–593.

Wadsworth SJ, DeFries JC, Olson RK, Willcutt EG. Colorado Longitudinal Twin Study of Reading Disability. Annals of Dyslexia 2007;57(2):139–160. [PubMed: 18060583]

Wagner RK, Togesen JK, Rashotte CA, Hecht SA, Barker TA, Burgess SP, et al. Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. Developmental Psychology 1997;33:468–479. [PubMed: 9149925]

Wechsler, D. Examiner's Manual: Wechsler Intelligence Scale for Children – Revised. New York: The Psychological Corporation; 1974.

Wechsler, D. Examiner's Manual: Wechsler Intelligence Scale for Children. 3. San Antonio, TX: The Psychological Corporation; 1991.

Wechsler, D. Examiner's Manual: Wechsler Intelligence Scale. 3. San Antonio, TX: The Psychological Corporation.; 1993.

Wilkinson, GS. Examiner's Manual: Wide Range Achievement Test. 3. Wilmington, DE: Jastak Associates – Wide Range Inc; 1993.

Wise BW, Ring J, Olson RK. Training phonological awareness with and without attention to articulation. Journal of Experimental Child Psychology 1999;72:271–304. [PubMed: 10074381]

Wise BW, Ring J, Olson RK. Individual differences in gains from computer-assisted remedial reading with more emphasis on phonological analysis or accurate reading in context. Journal of Experimental Child Psychology 2000;77:197–235. [PubMed: 11023657]

Wolf M, Bowers PG. The double-deficit hypothesis for the developmental dyslexias. Journal of Educational Psychology 1999;91:415–438.

Woodcock, RW. Woodcock Reading Mastery Test – Revised. Circle Pines, MN: American Guidance Services; 1987.

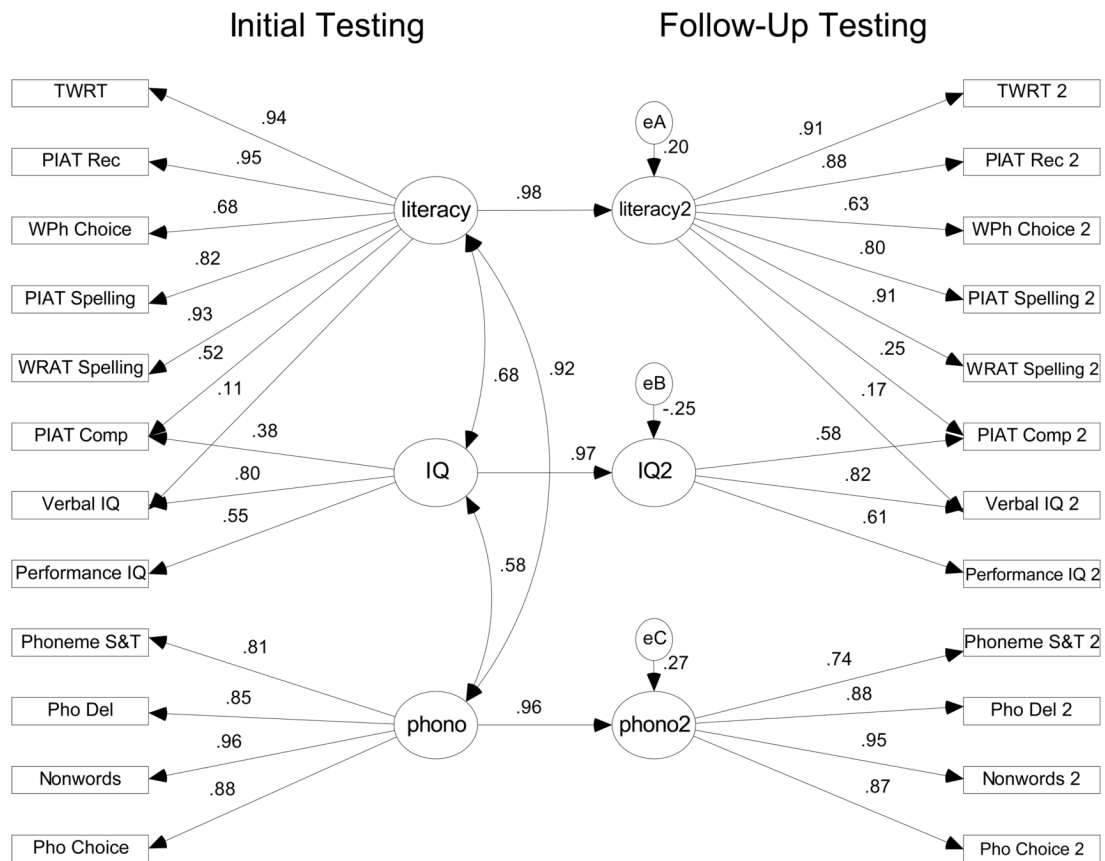**Figure 1.**
Longitudinal confirmatory factor analysis for Model 1: structural invariance and metric invariance between initial and follow-up testing.
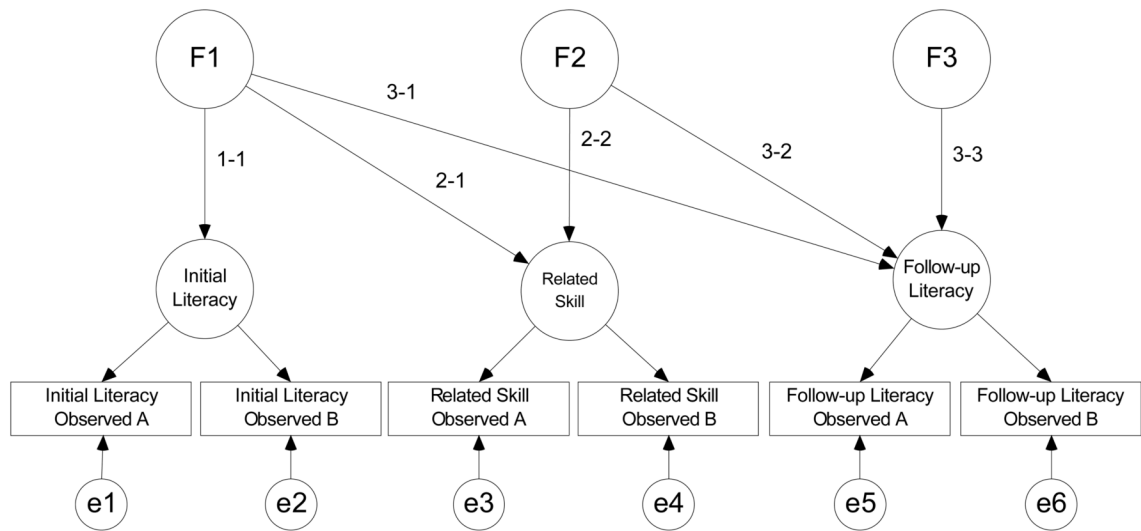
Figure 2.
Schematic drawing of Cholesky decomposition model.

Table 1

Descriptive Statistics

Measure	N	Initial Mean (SD)	Follow-up Mean (SD)	Difference (SD, range)
TWRT[1]	302	97.16 (39.89)	149.87 (36.49)	53.44 (23.34, −3:130)
PIAT Rec[3]	323	102.54 (13.04)	104.14 (15.30)	1.60 (8.66, −17:33)
Phoneme S & T[2]	303	61.28 (26.68)	76.24 (18.82)	13.96 (19.07, −39:73)
Pho Del[2]	313	63.11 (23.41)	79.03 (18.88)	15.73 (13.73, −20:60)
Nonwords[2]	270	62.94 (25.47)	81.62 (15.90)	18.68 (15.13, −6:68)
Pho Choice[2]	262	64.18 (19.16)	79.35 (14.78)	15.17 (13.32, −17:62)
WPh Choice[2]	318	79.99 (11.20)	88.53 (7.82)	8.54 (9.63, −14:40)
PIAT Spelling[3]	320	101.51 (13.31)	99.20 (12.53)	−2.31 (10.57, −29:40)
WRAT Spelling[3]	322	98.51 (16.69)	100.34 (13.27)	1.83 (9.22, −19:30)
PIAT Comp[3]	320	105.78 (13.64)	103.24 (14.45)	−2.54 (11.70, −40:38)
Verbal IQ[3]	322	108.72 (14.08)	101.94 (13.64)	−6.77 (8.83, −32:17)
Performance IQ[3]	323	105.43 (12.69)	103.10 (13.33)	−2.33 (10.46, −29:26)
RAN Colors[1]	324	21.43 (4.17)	29.57 (5.10)	8.15 (4.27, −4:29)
RAN Pictures[1]	324	18.79 (3.38)	23.76 (3.78)	4.96 (3.43, −5:15)
RAN Letters[1]	321	28.74 (6.92)	39.16 (6.81)	10.43 (6.25, −7:30)
RAN Numbers[1]	322	31.05 (6.32)	41.69 (7.63)	10.63 (6.24, −8:32)

[1] Raw Score,

[2] Percent Correct,

[3] Standard Score.

Table 2

Correlations Within and Across Testing Sessions

Measure	1.	2.	3.	4.	5.	6.	7.
1. Word Recognition Latent Trait	**.98**	.85	.93	.95	.81	.73	.57
2. Phoneme Awareness Latent Trait	.82	**1.0**	.95	.81	.70	.70	.44
3. Phono. Decoding Latent Trait	.90	.91	**.93**	.89	.74	.67	.51
4. Spelling Latent Trait	.93	.79	.84	**.95**	.74	.72	.59
5. PIAT Comprehension	.62	.53	.53	.56	**.65**	.65	.44
6. Full Scale IQ	.73	.73	.64	.63	.65	**1.0**	.46
7. RAN Latent Trait	.42	.36	.44	.44	.31	.45	**.82**

Note: Correlations among initial testing measures above the diagonal. Correlations among follow-up testing measures below the diagonal. Longitudinal correlations along the diagonal.

Table 3

Comparison of Factor Loadings across Time for Full Sample

Measure	Initial Testing				Follow-Up			
	F1	F2	F3	F4	F1	F2	F3	F4
TWRT	**.89**	.78	.47	.35	**.88**	.71	.43	.62
PIAT Rec	**.89**	.83	.46	.35	**.85**	.74	.32	.61
Phoneme S & T	.57	**.91**	.28	.36	.54	**.86**	.23	.52
Pho Del	.65	**.89**	.37	.31	.69	**.91**	.31	.50
Nonwords	.81	**.92**	.40	.28	.79	**.89**	.37	.48
Pho Choice	.79	**.84**	.39	.29	.75	**.83**	.40	.47
WPh Choice	**.82**	.48	.30	.19	**.78**	.43	.21	.24
PIAT Spelling	**.86**	.64	.47	.32	**.84**	.58	.29	.51
WRAT Spelling	**.90**	.79	.46	.29	**.90**	.74	.40	.53
PIAT Comp	**.81**	.64	.37	.53	.61	.43	.28	**.82**
Verbal IQ	**.71**	.49	.33	**.68**	.61	.54	.36	**.90**
Performance IQ	.34	.38	.21	**.84**	.24	.53	.31	**.77**
RAN Letters	.56	.31	**.79**	−.01	.37	.37	**.86**	.25
RAN Numbers	.34	.35	**.85**	.06	.34	.35	**.86**	.26
RAN Colors	.29	.37	**.74**	.39	.28	.30	**.81**	.33
RAN Pictures	.29	.09	**.69**	.49	.15	.09	**.78**	.29

Note. Structure matrix for promax rotation shown. Loadings with substantial unique contributions as indicated by the pattern matrix are displayed in bold.

Table 4

Confirmatory Factor Analysis Model Fit

	χ^2	df	RMSEA	CFI	$\Delta\chi^2$(df)
1. Structural invariance, metric variance	918.8	242	.093	.92	
2. Structural invariance, metric invariance	939.2	253	.092	.92	20.4(11) [*]
3. Structural invariance, metric variance except for Literacy→ PIAT Comp	931.3	243	.092	.92	12.5(1) [*]
4. Structural invariance, metric invariance except for Literacy→PIAT Comp	929.2	252	.091	.92	10.4(10)

Note. RMSEA = root-mean-square error of approximation. CFI = comparative fit index.

[*]
$p < .05$ when compared to Model 1

Table 5

Longitudinal Correlations of Latent Trait Residuals

	Controlling for Concurrent:		
Residuals of:	**Word Recognition Latent Trait**	**Spelling Latent Trait**	**PIAT Comprehension**
Phoneme Awareness	1.0	1.0	1.0
Phonological Decoding	.78	.93	.90
IQ	.95	1.0	.71
RAN	.79	.79	.79

Table 6

Cholesky Standardized Factor Loadings Predicting Follow-up Outcomes

	Initial Related Skill (F2)											
	Phoneme Awareness			Phonological Decoding			IQ			RAN		
	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
Initial Word Recognition	1.0*			1.0*			1.0*			1.0*		
Initial Related Skill	.85*	.53*		.93*	.35*		.72*	.69*		.57*	.82*	
Follow-up Word Recognition	.98*	.01	.19*	.98*	.02	.20*	.98*	.08*	.18*	.98*	−.02	.19*
Initial Spelling	1.0*			1.0*			1.0*			1.0*		
Initial Related Skill	.81*	.59*		.89*	.46*		.71*	.70*		.59*	.81*	
Follow-up Spelling	.95*	−.07	.31*	.94*	.07	.32*	.95*	.06	.29*	.95*	−.01	.30*
Initial Comprehension	1.0*			1.0*			1.0*			1.0*		
Initial Related Skill	.67*	.74*		.73*	.69*		.77*	.64*		.44*	.90*	
Follow-up Comprehension	.65*	.16*	.74*	.65*	.14*	.75*	.64*	.36*	.67*	.65*	.02	.76*

*
$p < .05$