Phonological Memory and Broader Language Development: Longitudinal and Etiologic Relations

Robin Leonhardt Peterson

University of Denver

Follow this and additional works at: https://digitalcommons.du.edu/etd

Part of the Clinical Psychology Commons, and the Developmental Psychology Commons

Recommended Citation

Peterson, Robin Leonhardt, "Phonological Memory and Broader Language Development: Longitudinal and Etiologic Relations" (2010). Electronic Theses and Dissertations. 902.
https://digitalcommons.du.edu/etd/902
PHONOLOGICAL MEMORY AND BROADER LANGUAGE DEVELOPMENT:
LONGITUDINAL AND ETIOLOGIC RELATIONS

A Dissertation
Presented to
The Faculty of Social Sciences
University of Denver

In Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

by
Robin L. Peterson
August 2010
Advisor: Bruce F. Pennington
Abstract

The current research investigated the relationship of phonological memory (PM) to vocabulary and syntax learning in school-age children with and without language disorders. Previous research has established that PM and broad oral language skills covary, but disagreement remains about the reason for this association. Opposing theoretical viewpoints emphasize the importance of either bottom-up (PM influences vocabulary and syntax acquisition) or top-down (vocabulary growth influences PM skill) factors. In three longitudinal studies, we tested competing bottom-up and top-down explanations of the PM-broad language link. Study 1 utilized a structural equation modeling approach to understand PM and broad language relations from age 5 to age 8 in population samples from three cultures. Final models varied by culture, with overall results supporting bidirectional relations between PM and vocabulary or syntax. Study 2 used a similar approach to investigate PM and broad language development in children with phonologically-based language disorders: speech sound disorder and reading disability. Results supported a bottom-up account in children with more substantial language delays and a bidirectional account in more mildly affected children. Study 3 used a behavior genetics approach to test for shared etiology of PM and vocabulary deficits in 5- to 8-year-old twins. We found evidence for common influences on PM and vocabulary weaknesses both within and across time. The bottom-up effect appeared to be predominantly influenced by shared genes, while the top-down affect appeared to be predominantly influenced by shared environmental experiences. Across the studies, methodological limitations prevented strong conclusions about the relation between PM and syntax. However, a clear pattern emerged concerning PM and vocabulary knowledge. The relationship of PM to vocabulary in the early school years owes to both bottom-up and top-down factors, with both effects undergoing developmental changes during this age period. We argue that the bottom-up effect gradually wanes with language
development, while the top-down effect emerges as a consequence of learning to read. Future studies to further test these conclusions are proposed.
# Table of Contents

Introduction ............................................................................................................................................. 1

Study 1: Phonological memory and language development in the population .................. 3
Study 2: Phonological memory and language development in children with language disorders ........................................... 5
Study 3: Etiology of phonological memory and broad language: shared or distinct influences .............................................................................................................................. 8

Study 1 ...................................................................................................................................................... 11
  Method .................................................................................................................................................. 11
  Results .................................................................................................................................................. 14
  Discussion .......................................................................................................................................... 24

Study 2 ...................................................................................................................................................... 29
  Method .................................................................................................................................................. 29
  Results .................................................................................................................................................. 33
  Discussion .......................................................................................................................................... 47

Study 3 ...................................................................................................................................................... 50
  Method .................................................................................................................................................. 51
  Results .................................................................................................................................................. 53
  Discussion .......................................................................................................................................... 59

General Discussion ............................................................................................................................... 62

References ............................................................................................................................................. 66
List of Tables

Table 1: Summary of Constructs Assessed at Each Time Point in the ILTS…………………..13
Table 2: Means (and Standard Deviations) by Country within the ILTS…………………….15
Table 3: Correlations among Constructs in the ILTS……………………………………………18
Table 4: Fit Statistics by Country for Model with Bidirectional Effects of PM and Vocabulary……20
Table 5: Fit Statistics by Country for Initial Model with Bidirectional Effects of PM and Syntax…22
Table 6: Fit Statistics by Country for Follow-Up Model with Bidirectional Effects of PM and Syntax…………………………………………………………………..23
Table 7: Effect Sizes for Language and Cognitive Deficits in Each of the Three Disordered Groups, Compared to the Relevant Control Group..............................................37
Table 8: Cross-Lagged Standardized Path Weights in Models Using Different Measures of PM within the DU SSD Sample……………………………………………………………..46
Table 9: Gender and Zygosity by Country within the ILTS……………………………………..52
Table 10: Results of Univariate DeFries-Fulker Analyses of PM and Vocabulary Deficits…………………………………………………………………………………………..54
Table 11: Results of Bivariate DeFries-Fulker Analyses of PM and Vocabulary Deficits at Age 5……………………………………………………………………………………………..56
Table 12: Results of Bivariate DeFries-Fulker Analyses of PM and Vocabulary Deficits Across Time……………………………………………………………………………………………..57
Table 13: Results of Bivariate DeFries-Fulker Analyses of PM and Vocabulary Deficits at Age 5 by Subgroup…………………………………………………………………………………..59
List of Figures

Figure 1: Confirmatory factor analysis of language constructs assessed at age 5 within the ILTS…………………………………………………………………………………………………………………………..18

Figure 2: Initial model of PM-vocabulary relations across time, showing bidirectional effects……19

Figures 3A-C: Final models of PM-vocabulary relation across time in the ILTS, by country……21

Figure 4: Initial model of PM-syntax relations across time, showing bidirectional effects ........22

Figure 5: Follow-up model of PM-syntax relations over time…………………………………………………………23

Figures 6A-C: Final models of PM-syntax relation across time in the ILTS, by country……24

Figure 7: Confirmatory factor analysis of language constructs assessed at age 5 within the DU SSD Sample…………………………………………………………………………………34

Figure 8: Confirmatory factor analysis of language constructs assessed at age 8 within the DU SSD Sample…………………………………………………………………………………………35

Figures 9A-C: Final models of PM-vocabulary relation across time for the three disordered groups……………………………………………………………………………………………….41

Figures 10A-C: Final models of PM-syntax relation across time for the three disordered groups…………………………………………………………………………………………….44
Introduction

A major theoretical issue in the field of language development concerns the separability of different linguistic constructs, such as phonology, semantics, and syntax. The nativist Chomskyan approach has emphasized distinctness, particularly of semantics and syntax (Chomsky, 1965; Pinker, 1991; Rice, Wexler, & Cleave, 1995). In contrast, constructivist approaches, which focus on the development of linguistic forms and assume that general cognitive processes contribute to this development, have emphasized the overlap between various aspects of language (Bates & Goodman, 2001; Elman et al., 1997; Saffran, 2003; Tomasello, Brooks, & Barrett, 1999). While the Chomskyan framework can account for species-universal aspects of language development, research examining individual differences has generally supported a constructivist view by finding high correlations among different subdomains of language skill (Hayiou-Thomas et al., 2006; Smith, McGregor, & Demille, 2006; Thal, Bates, Goodman, & Jahn-Samilo, 1997). The current research uses both longitudinal and etiologic approaches within a constructivist framework to investigate the relationship between phonological memory (PM) and language development. Consistent with previous research, we expect to find that these abilities covary in preschool and school-aged children. However, many questions remain concerning the developmental mechanisms underlying this covariance. Thus, the studies described below further investigate whether the relationship is primarily bottom-up (PM drives broader language), primarily top-down (broader language drives PM), or both.

Numerous studies have demonstrated a link between PM, the component of working memory responsible for keeping phonological information active for brief periods, and vocabulary skill. PM has most often been operationalized with performance on word span tests (e.g., digit span) or with nonword repetition. In developmental samples, the correlation between PM and vocabulary averages approximately 0.3-0.5 (Baddeley, Gathercole, & Papagno, 1998).
Convergent evidence from neuropsychological patients, experimental studies, and foreign language learning further supports a PM-vocabulary link (Baddeley, Papagno, & Vallar, 1988; Masoura & Gathercole, 1999; Papagno, Valentine, & Baddeley, 1991). Based on these data, Baddeley, Gathercole, and colleagues proposed the phonological storage framework, a bottom-up account of the relationship between PM and broader language. On this view, PM serves as a language learning device, and variance in PM skill constrains long-term phonological learning, including the acquisition of vocabulary and morphosyntax (Baddeley et al., 1998; Gathercole, 2006).

A competing theoretical viewpoint holds that vocabulary exerts top-down effects on the quality of phonological representations, which in turn influence performance on PM tasks, particularly nonword repetition (Metsala, 1999; Snowling, Chiat, & Hulme, 1991). The lexical restructuring model of Metsala and Walley (LRM; Metsala & Walley, 1998) is the most developed version of this viewpoint. According to the LRM, early lexical representations are fairly holistic. Only once children’s vocabularies include an increasing number of phonologically similar items (e.g., spat versus slat) is there pressure for lexical representations to become segmented down to the level of the phoneme (see also Fowler, 1991). Thus, vocabulary growth drives the development of increasingly fine-grained phonological representations. Because nonword repetition requires, among other things, the ability to segment phonological information (Snowling et al., 1991), vocabulary growth should lead to improved PM performance.

The LRM is most often cited as a top-down alternative to the phonological storage framework (Gathercole, 2006; Gathercole, Briscoe, Thorn, & Tiffany, 2008). However, other top-down explanations are also possible. For example, there are known to be lexical effects in some kinds of PM tasks, such as memory for relatively word-like nonwords (Gathercole, 1995). It is also conceivable that overall language development supports the emergence of strategies, such as rehearsal, that aid performance on short-term memory tasks. Furthermore, the bottom-up (PM drives vocabulary acquisition) and top-down (vocabulary growth drives performance on PM tasks) accounts of the PM-vocabulary link need not be mutually exclusive (Brown & Hulme, 1996). In fact, empirical results reviewed below have provided evidence for reciprocal relations between
PM and vocabulary, but suggest that the primary direction of effect may vary as a function of developmental stage.

In three studies, we test competing bottom-up and top-down explanations of the relationship between PM and broader language during the early school years. Study 1 investigates the longitudinal PM-broad language relationship in population-based samples drawn from four countries, including speakers of three languages. Study 2 tests the universality of these results by examining the longitudinal relation of PM to vocabulary and syntax in individuals with phonologically-based language disorders. Finally, Study 3 uses a behavior genetics method to determine why PM and broad language deficits co-occur and to compare the etiologic relation of earlier PM and later broad language to that of earlier broad language and later PM. In the sections that follow, we describe each of these studies in more detail.

*Study 1: Phonological memory and language development in the population*

Several longitudinal studies have now tested the relationship between PM and vocabulary development in typically developing children. Gathercole and colleagues (Gathercole & Baddeley, 1989; Gathercole, Willis, Emslie, & Baddeley, 1992) conducted the largest study to date, which also spanned the longest time (age 4 to age 8). Cross-lagged correlations indicated that age 4 PM predicted age 5 vocabulary after accounting for the autoregressive effects of vocabulary at age 4, but the reverse pattern did not hold. However, from ages 5 to 6 and 6 to 8, the direction of effect reversed, with early vocabulary having a larger influence on later PM than vice versa. Thus, during the earliest developmental window tested, results were consistent with the phonological storage framework, but later results were more consistent with the predictions of the LRM. One interpretation of this pattern is that the importance of PM to new word learning wanes with vocabulary size, as individuals become more able to use lexically-supported learning (Gathercole, 2006). Studies with second language learners provide convergent evidence for such a proposal, since the relationship between PM and vocabulary declines with second language proficiency (Cheung, 1996; Masoura & Gathercole, 1999). More recent longitudinal studies of unselected or typically developing samples have also found that PM at age 4 uniquely
predicted vocabulary one year later (Avons, Wragg, Cupples, & Lovegrove, 1998; Bowey, 2001; Gathercole, 1995). However, none of these studies extended beyond age 5.

Further evidence for a causal role of PM in vocabulary acquisition at this age comes from laboratory studies in 4 to 5 year olds showing PM skill predicts the learning of unfamiliar words under controlled conditions (Gathercole & Baddeley, 1990b; Gathercole, Hitch, Service, & Martin, 1997; Michas & Henry, 1994). The relationship of working memory to vocabulary learning appears to be specific to PM, since this research has not supported a link between spatial working memory and later vocabulary (Michas & Henry, 1994).

Baddeley, Gathercole and Papagno (1998) proposed that the role of PM in language learning extends beyond vocabulary development to the acquisition of syntax. Theoretical support for such a position comes from the fact that processing sentence-level syntax places significant demands on phonological working memory (Baddeley, Vallar, Wilson, & Coltheart, 1987; Just & Carpenter, 1992). Furthermore, many morphological rules (rules governing word-level syntax, such as grammatic prefixes and suffixes) include complex phonological components (Joanisse, 2004). For example, the sound of the English past-tense suffix –ed depends on the final phoneme of the root word. A developing child’s ability to discover syntactic rules might thus depend partly on the ability to hold phonological information in working memory (Speidel, 1993).

The relationship between PM and syntax has been investigated less thoroughly than the relationship between PM and vocabulary. One study demonstrated that PM correlated with the diversity of syntactic structures produced by 3 year olds (Adams & Gathercole, 1995). Similarly, a study of artificial grammar learning in 2- to 6-year-old children found that PM predicted performance better than chronological age (Daneman & Case, 1981).

To summarize, there is substantial empirical evidence that PM plays a causal role in vocabulary acquisition from age 4 to age 5. However, it is less clear whether the phonological storage framework accounts for vocabulary acquisition into the early school years. Since children learn up to 3,000 new words during this period (Nagy & Herman, 1987) the current research addresses a basic question about the mechanisms of word learning during an important developmental window. Furthermore, our research will provide the most comprehensive
examination to date of the longitudinal relation between PM and syntax in typically developing individuals.

Study 2: Phonological memory and language development in children with language disorders

If PM constrains language development in typically developing children, then PM deficits might lead to language disorders. In fact, research with several language-disordered populations has pointed to PM as a candidate causal deficit. Children with specific language impairment (SLI) fail to acquire language, including vocabulary and syntax, at the expected rate despite normal-range nonverbal intelligence (Bishop, 1997). SLI is associated with robust PM deficits, and a number of studies have demonstrated that children with SLI have even poorer PM than younger, typically developing children matched for overall language skill, indicating that poor PM is not just a consequence of slow language development (Bishop, North, & Donlan, 1996; Dollaghan & Campbell, 1998; Gathercole & Baddeley, 1990a; Montgomery, 1995). Further, degree of PM impairment correlates with the degree of broader language difficulties in SLI (Bishop et al., 1999). Children with SLI and associated PM deficits also perform poorly on novel word learning tasks under controlled laboratory conditions (Gray, 2004; Rice, Buhr, & Nemeth, 1990).

Difficulties with expressive syntax are a hallmark of SLI. English-speaking children with the disorder characteristically make zero-marking errors in past-tense production (Rice et al., 1995) (e.g., “He walk there” in place of “He walked there.”) Connectionist models have simulated this error pattern with damage to the model’s phonological representations (Hoeffner & McClelland, 1993; Joanisse, 2004). An assumption has been that the phonological damage would also cause PM difficulties (Joanisse, 2004). In one sense, these results are consistent with the hypothesis that PM impairments lead to syntactic deficits in SLI, though the relationship is not specific to PM.

Children with Down Syndrome (DS) are another group with both striking PM impairments and poor vocabulary knowledge (for a review, see Jarrold, Baddeley, & Phillips, 1999). Longitudinal research in individuals with DS has also suggested a link between earlier PM and later syntax (Laws & Gunn, 2004). Because most individuals with DS have moderate mental retardation, however, there are many other possible causes of their slow language development.
The argument for the role of PM is strengthened by comparison to Williams Syndrome (WS). Although most individuals with WS also have mental retardation, they tend to have surprisingly good vocabulary and syntax in comparison to their nonverbal skills, as well as relatively spared PM (Bellugi, Marks, Bihrlle, & Sabo, 1988; Grant et al., 1997). Thus, the contrasting profiles of these two syndromes further support a link between PM and language development.

Most work on the relationship of PM to vocabulary or syntax in language disorders has selected individuals for broad language impairments, and then looked for underlying PM deficits. Less research has examined the later broad language development of children with phonological deficits. One exception is a recent longitudinal study comparing vocabulary development in children selected for poor or age-appropriate PM skill at age 5 (Gathercole, Tiffany, Briscoe, & Thorn, 2005). The children were followed up at age 8, when the poor PM group was subdivided into those with persistent versus resolved PM deficits. The researchers predicted that both subgroups, particularly those with persistent PM deficits, would show slower vocabulary acquisition than controls. This hypothesis was not confirmed. Instead, the persistently poor PM group performed comparably to controls on vocabulary tests at age 8 (as they had at age 5). In contrast, the resolved PM group appeared to have a more general language impairment, evidenced by poorer vocabulary and verbal IQ than controls at both time points. Overall, this study did not provide evidence that PM constrains vocabulary acquisition over the age period studied.

In a follow-up study, the researchers examined the performance of the same sample on a wide range of long-term memory and learning tasks at age 8 (Gathercole et al., 2008). Children with poor age 5 PM performance had deficits in the ability to learn arbitrary verbal material (e.g., to pair words with nonwords or to learn unfamiliar names). However, these children were not impaired when asked to learn meaningful information (e.g., to memorize semantically associated word-word pairs.) The authors argued that PM constrains long-term verbal learning but that with increasing vocabulary size, individuals can better capitalize on existing knowledge structures in their lexicons. Thus, the importance of PM to vocabulary acquisition declines with development. This explanation can account for the normal performance of the poor PM group on some tasks.
This follow-up study did not divide the poor PM group into those with persistent versus resolved deficits, so it is unclear how that dimension would relate to performance on the larger battery of learning tasks.

Another recent study demonstrated a similar pattern among intellectually delayed (ID) individuals (Jarrold, Baddeley, Hewes, Leeke, & Phillips, 2004). The study compared individuals with different degrees of ID to younger, typically developing children matched on vocabulary level. The logic of the study was that if PM constrains the rate of vocabulary acquisition (i.e., a bottom-up account), then older children with ID should perform more poorly on PM tasks than younger vocabulary-matched controls, since they have taken longer to achieve the same vocabulary level. On the other hand, if vocabulary level constrains PM performance (i.e., a top-down account), then the ID and typically developing groups should perform comparably. Among children functioning at a 5-year-old vocabulary level, results were consistent with a bottom-up constraint on vocabulary learning: the ID children had even poorer PM than the younger typically-developing controls. However, among children functioning at an 8-year-old vocabulary level, the groups had similar PM skill, consistent with the top-down account. Thus, studies with atypical populations agree with results from typically developing children and second language learners reviewed above in suggesting that the primary direction of influence between PM and vocabulary depends on developmental level.

In Study 2, we will test the universality of results from the population (Study 1) by studying the PM-broad language relationship in two groups with phonological processing impairments: children with reading disability (RD), and children with a history of speech sound disorder (SSD). These disorders are comorbid, but only a minority of children with one disorder also have the other (Catts, 1993; Pennington & Lefly, 2001; Snowling, Bishop, & Stothard, 2000). RD is characterized by difficulties with accurate and/or fluent printed word recognition. Research has demonstrated that in most cases, literacy difficulties are caused by underlying phonological impairments (for a review, see Vellutino, Fletcher, Snowling, & Scanlon, 2004). However, children with RD also show broader language weaknesses, including in vocabulary and listening comprehension (Keenan, Betjemann, Wadsworth, DeFries, & Olson, 2006; Scarborough, 1990;
Snowling, Gallagher, & Frith, 2003; Stanovich, 1986). Though the defining symptom is different in SSD, the pattern of cognitive impairments is similar. SSD is characterized by difficulty producing developmentally appropriate speech output. Even after this symptom has resolved, individuals show enduring difficulties with a range of phonological processing tasks, as well as broader language weaknesses (Bird & Bishop, 1992; Peterson, Pennington, Shriberg, & Boada, in press; Raitano, Pennington, Tunick, Boada, & Shriberg, 2004; Snowling et al., 2000). Although various hypotheses have been advanced to explain the relationship between phonological and broader language deficits (Scarborough, 2005; Stanovich, 1986), no previous longitudinal study has explicitly modeled the direction of effect.

**Study 3: Etiology of phonological memory and broad language: shared or distinct influences**

To summarize, there is ample evidence for a phenotypic association between PM and broad language skill in both typically developing children and children with language disorders. However, the etiologic basis for this association remains poorly understood. Study 3 will address this question with regard to PM and broader language deficits. Do impairments in these constructs tend to co-occur because of overlapping genes, overlapping environmental experiences, or both? In addition, Study 3 will use an etiologic approach to test bottom-up and top-down explanations for the correlation. We will compare the degree to which shared influences act on earlier PM and later broad language versus earlier broad language and later PM.

Twin designs compare the similarity of monozygotic (MZ) and dizygotic (DZ) twins to estimate the relative contributions of genes and environments to individual differences. Since MZ twins share 100% of their genes and DZ twins share, on average, 50% of segregating genes, the degree to which MZ twins are more similar than DZ twins indicates genetic influence (Plomin, DeFries, McClearn, & McGuffin, 2001). Environmental effects can be broken down into shared environment (shared experiences that make members of MZ and DZ pairs similar, such as home environment) and nonshared environment (unique experiences that make members of a twin pair different). Previous twin studies have found that PM, vocabulary, and syntax are all moderately heritable (Hayiou-Thomas et al., 2006; Kovas et al., 2005; Samuelsson et al., 2005; Samuelsson et al., 2007), indicating that all are subject to genetic effects. However, univariate analyses
cannot address the degree of overlap of such effects—two constructs might have identical heritability estimates due to the action of entirely different genes. In fact, univariate analyses have provided some suggestion for distinct etiologies because phonological processes, including PM, have shown stronger influences of genes than of shared environment, whereas vocabulary, and to a lesser degree, syntax, have often shown the opposite pattern (Dionne, Dale, Boivin, & Plomin, 2003; Hayiou-Thomas et al., 2006; Samuelsson et al., 2005).

Relatively few published studies have used behavior genetic methods to directly estimate the etiologic relationships among various sub-domains of language skill. One twin study investigated the etiology of PM and syntax deficits in a sample overselected for SLI and found no evidence for shared genetic influences (Bishop, Adams, & Norbury, 2006). However, the conclusions that can be drawn from such a null result are limited. In addition, the results may not generalize to the whole range of individual differences or to other language disorders. In fact, a more recent study investigated a wider range of individual differences and reported different results. Overall, both genetic and shared environmental influences operating on a wide range of language measures that included PM, vocabulary, and syntax, were largely shared (Hayiou-Thomas et al., 2006). However, because of its goals and design, that study provided evidence primarily regarding etiologic interrelationships of language skills in general, rather than regarding specific relationships between PM, vocabulary, and syntax.

Longitudinal twin studies can investigate etiologic relationships of different measures across time in order to address causal direction. Dionne and colleagues (Dionne et al., 2003) used this approach to investigate the association between vocabulary and syntax in young children. Etiologic results indicated shared genetic and shared common environment operating on both early vocabulary and later syntax as well as on early syntax and later vocabulary, with magnitude of the effects being similar in both directions. The authors concluded that results supported bidirectional “bootstrapping” of these constructs in toddlers. In Study 3, we will use similar logic to investigate the degree to which shared etiologies contribute to deficits in PM and broad language across time.
Thus, there is some evidence for shared etiology (both genetic and environmental) underlying the correlations between different language skills in development. However, substantial gaps in the literature remain. First, the one study designed to specifically evaluate the relation of PM to broader language (Bishop et al., 2006) examined only the origin of PM and syntax deficits. It did not include measures of vocabulary, which are particularly important for discriminating among the competing theoretical viewpoints outlined above. Second, published studies investigating etiologic relations among sub-domains of oral language skill have all utilized subsets of the same sample (the Twins Early Development Sample from Great Britain). The TEDS studies have reported results across the full range of individual differences, rather than among linguistic deficits. Third, no published research has combined longitudinal and etiologic methods to address PM/broad language relations. Study 3 will proceed in several phases to address 1) the etiology of PM, vocabulary, and syntax deficits; 2) the etiologic relationship of PM to broader language deficits at a given time; and 3) the etiologic relationship of PM to broader language across time. These studies will be conducted in a longitudinal, population-based, international twin sample. By utilizing distinct methodologies (phenotypic and etiologic) across three studies to address the same basic questions about language development, this body of research provides the opportunity for much stronger conclusions than could be drawn from a study utilizing only a single approach.
Study 1

Study 1 uses a structural equation modeling (SEM) approach to examine the relationship between PM and broad language from age 5 to age 8 across the full range of individual differences. Participants are drawn from the International Longitudinal Twin Sample (ILTS), an ongoing behavior genetics study of language and literacy development. The ILTS includes population-based samples of twins from the United States (Colorado), Australia, Sweden, and Norway.

This study will extend previous research in several ways. First, only one study to date has examined the longitudinal relationship between PM and vocabulary in typically developing children beyond age 5 (Gathercole et al., 1992). We will investigate this relationship during early school age in a much larger sample than has been used in any earlier longitudinal study. Second, previous longitudinal studies of the PM-vocabulary relationship have relied on cross-lagged correlations. Our sample size allows use of SEM to compare fit for models with bottom-up, top-down, and bidirectional effects. This more sophisticated analytic approach offers a major advantage over previous studies that have simply compared the magnitude of specific regression weights in order to discriminate among models. Third, previous longitudinal studies of the PM-syntax relationship have focused on special populations and not the full range of individual differences. Finally, our international sample allows us to formally test whether PM-broad language relations are similar across language and cultures.

Method

Participants. This study included participants from all four countries in the ILTS, with Sweden and Norway grouped together as Scandinavia for analyses. Identical and same-sex fraternal twin pairs of preschool age were initially tested prior to entry into kindergarten (mean age = 58.83 months), with follow-up tests at the end of kindergarten, 1st and 2nd grade (mean
ages of 75.53 months, 88.50 months, and 100.49 months, respectively). All children speak the native language of their home country as their first language (i.e., English, Swedish, or Norwegian).

Data from initial testing are available for a total of 951 twin pairs (261 from Australia, 489 from Colorado, and 201 from Scandinavia). Because the study is ongoing, not all twin pairs have complete data, however. Kindergarten data are available for 849 twin pairs, first grade data for 805 twin pairs, and second grade data for 682 twin pairs. Because this study included only phenotypic analyses, one twin was selected at random from each pair to avoid violations of assumptions about independence.

**Procedure.** At preschool age, participants completed a battery of cognitive, language, and preliteracy tests over a period of 5 days. Members of a twin pair were tested separately, either in quiet rooms at their preschool or in their homes. Follow-up testing at the ends of Kindergarten, 1st, and 2nd grades included a battery of cognitive, language, and literacy tests. Testing for each follow-up visit took place in a single session lasting approximately one hour at participants' homes.

**Measures.** The most comprehensive language and cognitive evaluation took place during the initial testing. Two measures are available for each language construct of interest (PM, vocabulary, and syntax) at this time. Due to time constraints, fewer measures are available of the constructs of interest during follow-up testing. Further, each language construct was reassessed at just one later time point. Table 1 provides an overview of which constructs were assessed at which time. Measures changed across time to avoid ceiling effects and to assess newly developing skills.
Table 1  
**Summary of Constructs Assessed at Each Time Point in the ILTS**

<table>
<thead>
<tr>
<th></th>
<th>NIQ</th>
<th>PM</th>
<th>Vocabulary</th>
<th>Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 5 (58.83 months)</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Age 6 (75.53 months)</td>
<td></td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 7 (88.50 months)</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 8 (100.49 months)</td>
<td></td>
<td>●</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Nonverbal IQ.** NIQ was assessed at age 5 using WPPSI Block Design (Wechsler, 1989).

**Phonological Memory.** PM was assessed at age 5 with The Children’s Nonword Repetition Test (Gathercole, Willis, Baddeley, & Emisle, 1994) as well as WPPSI Sentence Memory (Wechsler, 1989). PM was reassessed at age 7 with WPPSI Sentence Memory (Wechsler, 1989). Research has demonstrated that performance on sentence memory tasks is constrained by PM (Willis & Gathercole, 2001), but a reasonable hypothesis is that sentence repetition may be more influenced by broader language skill than are other tests of PM (e.g., digit span). In one of the samples used in Study 2, PM was assessed with sentence repetition in addition to the more traditional digit span and nonword repetition. That study will compare results using each of these measures of PM in order to validate the use of sentence memory in the ILTS.

**Vocabulary.** Vocabulary knowledge was assessed at age 5 with WPPSI Vocabulary (Wechsler, 1989) and a confrontation naming task (Fisher & Glennister, 1992). Vocabulary knowledge was reassessed at age 8 with the Boston Naming test (Kaplan, Goodglass, & Weintraub, 2001).

**Syntax.** Morphosyntactic ability was assessed at age 5 with a modified version of the Wug test (Berko, 1958) as well as a test of productive syntax (McCarthy & Kirk, 1961). This ability was reassessed at age 6 using the Test for the Reception of Grammar (TROG; Bishop, 1989).

**Analyses.** Analyses proceeded in several phases. First, means and standard deviations for demographic variables (age, parent level of education) and individual cognitive variables were compared across the three countries. One-way analyses of variance (ANOVAs) with follow-up
Tukey HSD post-hoc tests were performed to test for significant country differences. Next, a confirmatory factor analysis (CFA) was conducted to test the proposed factor structure of language constructs measures at Time 1. The invariance of this structure across countries was tested. Results from the CFA guided the formation of composite variables that were then used as observed variables in path analyses.

In the final set of analyses, we evaluated competing path models of the longitudinal relationship between PM and vocabulary as well as PM and syntax. Models were initially evaluated separately for the U.S., Australian, and Scandinavian samples, and equivalence was statistically tested. If there was not evidence against equivalence, the samples were combined to increase power. We tested how PM and broader language (vocabulary or syntax) predicted themselves and each other over time. Since not all measures were administered at all time points, cross-lagged paths spanned the shortest possible time frame. Using a nested-model approach that allowed statistical comparison, we evaluated whether either cross-lagged path could be dropped from the model without significant loss of fit, which provided information about whether effects were primarily bottom-up, top-down, or both during the time period studied.

In all SEM analyses (CFAs and path analyses) in this and later studies, we evaluated model fit before interpreting specific parameters. We defined good fit as a non-significant $\chi^2$ ($p > .05$), CFI > 0.95, and RMSEA ≤ .05. We considered adequate fit to be $\chi^2$/df < 5.0, CFI > 0.90, and RMSEA ≤ .08. Finally, we considered a model to have borderline fit if two of the three fit statistics were in the adequate range. In models with good, adequate, or borderline fit, specific path weights were compared to estimate the relative magnitude of the effects of interest.

Results

Descriptive statistics. Table 2 summarizes means and standard deviations by country for demographic variables (age and parent education level) as well as for individual cognitive variables. We conducted one-way ANOVAs with follow-up Tukey post-hoc tests to evaluate cross-country differences. Age-regressed raw scores were used in primary analyses. When available, we also report Scaled Scores for transparency.

14
Table 2
Means (and Standard Deviations) by Country within the ILTS

<table>
<thead>
<tr>
<th></th>
<th>Australia</th>
<th>United States</th>
<th>Scandinavia</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age 5</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (months)</td>
<td>57.25 (3.40)(^c)</td>
<td>58.75 (2.31)(^b)</td>
<td>61.05 (1.75)(^a)</td>
</tr>
<tr>
<td>Mean parent education level (years)</td>
<td>13.51 (1.81)(^b)</td>
<td>14.16 (2.22)(^a)</td>
<td>13.88 (2.90)(^a)(^b)</td>
</tr>
<tr>
<td>WPPSI Block Design</td>
<td>0.47 (1.05)(^a)</td>
<td>-0.17 (0.87)(^b)</td>
<td>-0.20 (1.03)(^b)</td>
</tr>
<tr>
<td>WPPSI Block Design (SS(^1))</td>
<td>12.66 (3.23)(^a)</td>
<td>10.39 (2.55)(^b)</td>
<td>9.14 (3.18)(^c)</td>
</tr>
<tr>
<td>Nonword repetition</td>
<td>0.50 (1.01)(^a)</td>
<td>-0.24 (0.94)(^b)</td>
<td>-0.07 (0.89)(^b)</td>
</tr>
<tr>
<td>WPPSI Sentence Repetition</td>
<td>0.22 (0.97)(^a)</td>
<td>-0.02 (0.95)(^b)</td>
<td>-0.24 (1.09)(^c)</td>
</tr>
<tr>
<td>WPPSI Sentence Repetition (SS)</td>
<td>11.23 (3.00)(^a)</td>
<td>10.30 (2.90)(^b)</td>
<td>9.50 (3.25)(^c)</td>
</tr>
<tr>
<td>WPPSI Vocabulary</td>
<td>0.64 (0.97)(^a)</td>
<td>-0.25 (0.95)(^b)</td>
<td>-0.22 (0.75)(^b)</td>
</tr>
<tr>
<td>WPPSI Vocabulary (SS)</td>
<td>13.59 (3.20)(^a)</td>
<td>10.47 (3.03)(^b)</td>
<td>9.40 (3.01)(^c)</td>
</tr>
<tr>
<td>Confrontation naming</td>
<td>0.46 (0.82)(^a)</td>
<td>-0.21 (1.00)(^b)</td>
<td>-0.08 (1.02)(^b)</td>
</tr>
<tr>
<td>Wug test</td>
<td>0.13 (0.97)(^a)</td>
<td>-0.10 (1.01)(^b)</td>
<td>0.09 (1.00)(^a)(^b)</td>
</tr>
<tr>
<td>Productive syntax</td>
<td>0.07 (1.00)(^a)</td>
<td>0.00 (1.00)(^a)</td>
<td>-0.08 (1.00)(^a)</td>
</tr>
<tr>
<td><strong>Age 6</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (months)</td>
<td>72.87 (4.25)(^c)</td>
<td>75.19 (3.67)(^b)</td>
<td>80.80 (3.53)(^a)</td>
</tr>
<tr>
<td>TROG</td>
<td>0.41 (0.84)(^a)</td>
<td>-0.14 (1.01)(^b)</td>
<td>-0.17 (1.03)(^b)</td>
</tr>
<tr>
<td><strong>Age 7</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (months)</td>
<td>84.14 (4.36)(^c)</td>
<td>89.03 (3.76)(^b)</td>
<td>92.75 (3.81)(^a)</td>
</tr>
<tr>
<td>WPPSI Sentence Repetition</td>
<td>0.15 (0.93)(^a)</td>
<td>0.12 (0.95)(^a)</td>
<td>-0.62 (1.05)(^b)</td>
</tr>
<tr>
<td>WPPSI Sentence Repetition (SS)</td>
<td>10.65 (2.93)(^a)</td>
<td>9.80 (2.91)(^b)</td>
<td>7.34 (3.17)(^c)</td>
</tr>
<tr>
<td><strong>Age 8</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (months)</td>
<td>94.90 (4.40)(^c)</td>
<td>101.24 (3.79)(^b)</td>
<td>104.84 (3.53)(^a)</td>
</tr>
<tr>
<td>Boston Naming Test</td>
<td>-0.02 (0.99)(^a)</td>
<td>0.03 (1.03)(^a)</td>
<td>-0.08 (0.92)(^a)</td>
</tr>
</tbody>
</table>

**Notes.** Age-regressed raw scores are reported unless otherwise noted. \(^{a,b,c}\) Cells with the same superscript within each row do not differ at the p<.05 level (Tukey).

\(^{1}\)Scaled Score
At each time point, there were relatively small but significant differences in age. The Australian sample was the youngest, followed by the U.S. sample, and the Scandinavian sample was the oldest. Scandinavian children start school at a later age than children in the U.S. or Australia. In addition, the U.S. children were tested in summer following the school year while the Australian children were tested in the latter part of the school year (because Australia has shorter summer vacations). There were also some small cross-cultural differences in parent education level, with parents from the U.S. having on average 0.65 more years of education than parents from Australia. Parents from Scandinavia did not differ statistically from either the U.S. or Australian cohort on this variable.

There were also cross-cultural differences on most cognitive variables. In general, the Australian sample obtained higher scores than either of the other two cohorts. When considering age-regressed raw scores, the U.S. and Scandinavian cohort did not differ from each other, except on Sentence Repetition at ages 5 and 7. Poorer performance of the Scandinavian sample on this measure may have arisen because of translation issues. However, when considering Scaled Scores, the Scandinavian cohort tended to perform most poorly of all three groups on both verbal and nonverbal measures. The Australian cohort earned Scaled Scores that were well above average, the U.S. cohort earned Scaled Scores that were very close to national norms, and the Scandinavian cohort earned Scaled Scores that were slightly below average.

There are several possible explanations for these patterns. Superior performance of the Australian cohort may derive in part from a selection bias, since these participants were drawn from a volunteer twin registry. In addition, comparison of standardized cognitive tests that have been normed on both U.S. and Australian populations suggests slightly superior performance for the Australian sample, with effect sizes of around 0.3. The reason for this difference is unknown, but may relate to the more homogenous make-up of Australian society (B. Byrne, personal communication, March 2009). It is initially puzzling that the Scandinavian cohort earned similar age-corrected raw scores but weaker Scaled Scores in comparison to the U.S. sample. It may be that the norming process for the Wechsler instruments has been less rigorous in Sweden and Norway than in the U.S. and Australia, and that these Scaled Scores are therefore less valid (S.
Samuelsson, personal communication, March 2009). Although Scaled Scores were examined in order to determine whether the various cohorts represented true population samples, age-corrected raw scores will be used in all future analyses.

**Data preparation and reduction.** Raw scores were regressed on age and standardized within gender and country. These gender-, country-, and age-corrected scores were examined for outliers, and any scores more than 3 standard deviations (3 SD) from the mean were trimmed to a 3 SD cutoff. This trimming process affected 0.3% of all available data. Variables were then examined for departures from normality, and all had acceptable skew and kurtosis (absolute value < 1).

To reduce the number of variables, a confirmatory factor analysis of the three hypothesized language constructs measured at age 5 was conducted using Amos software (Arbuckle, 1996). Gender-, country-, and age-corrected raw scores were used as the observed variables. We first tested the invariance of the factor structure across countries. Constraining measurement weights to be equal across the three groups did not degrade model fit ($\chi^2(6) = 7.71$, $p > .2$). Thus, the three groups were combined to increase power and the measurement model was evaluated in the full ILTS. This model, which is shown in Figure 1, had borderline fit ($\chi^2(6) = 34.0$, $p < .001$; $\chi^2/df = 5.67$; CFI = 0.99; RMSEA = 0.07). The Vocabulary and Syntax factors were highly correlated (0.90), so we tested a model that collapsed these into a single broad language factor. This change resulted in a significantly poorer fit ($\chi^2$-change(1) = 15.94, $p < .001$) and so this model was rejected. Modification indices were examined, but none suggested a theoretically sensible change. Thus, we accepted the original, theoretically motivated three-factor model. Based on this factor structure, we created composite PM, vocabulary, and syntax variables by averaging the relevant variables. These composites will be used in all future analyses. Table 3 shows the correlations among the key variables of interest, including these three composites in addition to the constructs estimated by a single variable (age 5 NIQ, age 6 syntax, age 7 PM, and age 8 vocabulary).
Figure 1. Confirmatory factor analysis of language constructs assessed at age 5 within the ILTS. Standardized weights are shown.

Table 3
Correlations among Constructs in the ILTS

<table>
<thead>
<tr>
<th>NIQ5</th>
<th>PM5</th>
<th>Voc5</th>
<th>Syn5</th>
<th>Syn6</th>
<th>PM7</th>
<th>Voc8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.34</td>
<td>.36</td>
<td>.32</td>
<td>.33</td>
<td>.22</td>
<td>.27</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.52</td>
<td>.61</td>
<td>.55</td>
<td>.58</td>
<td>.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. All correlations are significant at the p < .001 level.

Model comparison: PM-vocabulary relation. We next tested a series of nested models in which the PM and vocabulary composites predicted themselves and each other over time.
Following previous research, NIQ was included at age 5 as a control variable. We treated the composites as observed variables rather than adopting a latent trait approach for two reasons. First, only two indicators were available for each construct of interest, and latent traits should ideally include at least three indicators. (This limitation may also relate to the relatively poor fit of the three-factor measurement model above). Second, a latent trait approach would lead to some models that were complex for the available sample size, particularly in multiple-group analyses or when the three samples could not be combined.

The initial model, which is shown in Figure 2, included both top-down (age 5 vocabulary \( \rightarrow \) age 7 PM) and bottom-up (age 7 PM \( \rightarrow \) age 8 vocabulary) effects. Note that we did not include a path from age 5 PM \( \rightarrow \) age 8 vocabulary, since this effect would presumably operate through age 7 PM. (In other words, cross-lagged paths always spanned the shortest possible time frame). In a multiple-groups analysis, we tested whether the model was equivalent for the Australian, U.S., and Scandinavian samples. Constraining path weights to be equal across groups resulted in significant loss of fit: \( \chi^2(10) = 24.44, p < .01 \). Thus, we next evaluated this model separately in each sample. Fit statistics by country are shown in Table 4. The model had good fit in the U.S. sample, borderline fit in the Australian sample, and poor fit in the Scandinavian sample.

Figure 2. Initial model of PM-vocabulary relations across time, showing bidirectional effects.
Table 4
Fit Statistics by Country for Model with Bidirectional Effects of PM and Vocabulary (Figure 2)

<table>
<thead>
<tr>
<th>Country</th>
<th>$\chi^2$</th>
<th>$\chi^2$/df</th>
<th>CFI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>$\chi^2(1) = 3.02$, $p &gt; .05$</td>
<td>3.02</td>
<td>.99</td>
<td>.09</td>
</tr>
<tr>
<td>US</td>
<td>$\chi^2(1) = 1.77$, $p &gt; .1$</td>
<td>1.77</td>
<td>1.00</td>
<td>.04</td>
</tr>
<tr>
<td>Scandinavia</td>
<td>$\chi^2(1) = 5.72$, $p &lt; .05$</td>
<td>5.72</td>
<td>.98</td>
<td>.15</td>
</tr>
</tbody>
</table>

Further model testing was carried out separately within each country. We tested a model corresponding to the phonological storage framework by investigating whether the top-down effect (age 5 Vocabulary $\rightarrow$ age 7 PM) could be dropped from the model. In both the Australian and U.S. samples, removing this path resulted in significant loss of fit (Australia: $\chi^2$-change(1) = 18.65, $p < .001$; U.S.: $\chi^2$-change(1) = 42.35, $p < .001$). In the Scandinavian sample, however, this path could be dropped without loss of fit ($\chi^2$-change(1) = 3.22, $p > .05$). We then tested a model corresponding to the LRM. From the initial model, we dropped the bottom-up effect (age 7 PM $\rightarrow$ age 8 vocabulary). This change did not degrade model fit in the Australian sample ($\chi^2$-change(1) = 2.34, $p > .05$) but did significantly degrade fit in both the U.S. and Scandinavian samples (U.S.: $\chi^2$-change(1) = 22.41, $p < .001$; Scandinavia: $\chi^2$-change(1) = 20.43, $p < .001$). Thus, the final model of PM-vocabulary relations varied by country. In the U.S. sample, we rejected the models corresponding to both the phonological storage framework and the LRM, and accepted the initial, bidirectional model. In the Australian sample, we accepted the model corresponding to the LRM. In the Scandinavian sample, none of the models tested had adequate or even borderline fit, so the results must be interpreted cautiously. (In a follow-up analysis, we examined modification indices in this sample, but no changes were suggested.) In sum, although there was not a good-fitting final model of PM-vocabulary relations in Scandinavia, the model that best balanced parsimony with fit corresponded to the phonological storage framework. Final models are shown in Figures 3A-C.
Figures 3A-C. Final models of PM-vocabulary relation across time in the ILTS, by country. Standardized weights are shown.
^p < .1.  *p < .05.  **p < .01.  ***p < .001.

Model comparison: PM-syntax relation. In a related set of analyses, we evaluated the longitudinal relation between PM and syntax. The initial model, which is shown in Figure 4, included both bottom-up (age 5 PM composite → age 6 syntax) and top-down (age 6 syntax → age 7 PM) effects. Multiple-groups analysis indicated that path weights could not be equated...
across countries without significant loss of fit: $\chi^2(10) = 20.34, p < .05$. We therefore tested the model separately within each group. Fit statistics by country are shown in Table 5. The model had poor fit in the Australian and U.S. samples, and borderline fit in the Scandinavian sample.

![Figure 4](image)

*Figure 4.* Initial model of PM-syntax relations across time, showing bidirectional effects.

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Fit Statistics by Country for Initial Model with Bidirectional Effects of PM and Syntax (Figure 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>Australia</td>
<td>$\chi^2(1) = 10.73, p &lt; .01$</td>
</tr>
<tr>
<td>US</td>
<td>$\chi^2(1) = 24.88, p &lt; .001$</td>
</tr>
<tr>
<td>Scandinavia</td>
<td>$\chi^2(1) = 2.50, p &gt; .1$</td>
</tr>
</tbody>
</table>

Before proceeding with further model testing, we examined modification indices to determine if the fit of the initial model could be improved. In both the Australian and U.S. samples, fit of the model could be improved with inclusion of a path from the age 5 syntax composite to age 7 PM. The initial model had included a cross-lagged path from syntax to PM over the shortest possible time frame (i.e., age 6 $\rightarrow$ age 7) for theoretical reasons. We decided to replace this path with the cross-lagged path suggested by the modification index. There are several reasons why age 5 syntax might be a better predictor of later PM than age 6 syntax. First, since two measures of syntax were included at age 5, the earlier measure is likely to be more reliable. Second, the measures used at age 5 tapped expressive morphosyntactic abilities, while the age 6 measure assessed receptive abilities, and the former may be more sensitive to language difficulties.
The second model tested is shown in Figure 5. Multiple-group analysis again indicated that path weights could not be equated across countries without significant loss of fit: $\chi^2(12) = 29.43, p < .01$. Analyses proceeded separately within each sample, with fit statistics reported in Table 6. The model now showed good fit in the U.S. sample. Fit remained poor but was relatively better in the Australian sample. In contrast, in the Scandinavian sample, this change reduced fit from borderline to poor. It is not clear why the specific time-frame of the top-down path should vary by culture; this result may simply reflect error variance. However, for further model testing, we began with the models with the best fit in each country (i.e., Figure 4 for Scandinavia and Figure 5 for Australia and the U.S.).

![Figure 5](image)

Figure 5. Follow-up model of PM-syntax relations over time. The model still includes bidirectional effects, but the path from syntax to PM spans a longer time frame.

<table>
<thead>
<tr>
<th>Country</th>
<th>$\chi^2$</th>
<th>$\chi^2/df$</th>
<th>CFI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>$\chi^2(1) = 5.09, p &lt; .05$</td>
<td>5.09</td>
<td>.98</td>
<td>.13</td>
</tr>
<tr>
<td>US</td>
<td>$\chi^2(1) = 3.71, p &gt; .05$</td>
<td>3.71</td>
<td>1.00</td>
<td>.08</td>
</tr>
<tr>
<td>Scandinavia</td>
<td>$\chi^2(1) = 6.81, p &lt; .01$</td>
<td>6.81</td>
<td>.97</td>
<td>.17</td>
</tr>
</tbody>
</table>

Finally, in parallel analyses to those described in the PM-vocabulary section, above, we investigated whether either the bottom-up path (age 5 PM composite $\rightarrow$ age 6 syntax) or the top-down path (syntax $\rightarrow$ PM; varied by country) could be dropped from the model. In all three
countries, dropping either path significantly degraded fit (all $\chi^2$-change(1) > 4.8; all p-values < .05). Thus, although the final models were not identical across the three samples, in every case the effects of PM and syntax were bidirectional. Final models are shown in Figures 6A-C.

**Discussion**

This study investigated the longitudinal relationship of PM to vocabulary and syntax in population-based samples from three different cultures, including two groups of English speakers.
(from Australia and the United States) and a group of Swedish or Norwegian speakers (from Scandinavia). There were mean differences by culture in performance on most cognitive variables. The Australian cohort demonstrated superior performance to both the U.S. and Scandinavian samples, who were generally similar to each other.

Although we did not expect to find significant cross-cultural differences in the models of PM-broad language relations across time, multiple-group analyses consistently indicated that the three samples could not be combined. Group differences were most striking in models of the relationship between PM and vocabulary. The final model in the U.S. sample, which had very good fit, included bidirectional influence of PM and vocabulary. Further, the cross-lagged paths were similar in magnitude (age 5 vocabulary composite $\rightarrow$ age 7 PM: .27; age 7 PM $\rightarrow$ age 8 vocabulary: .22) and both significant at the $p < .001$ level. In contrast, the final model in the Australian sample, which had borderline fit, was consistent with the predictions of the LRM. It included a top-down influence of earlier vocabulary on later PM but no bottom-up influence of earlier PM on later vocabulary. Unfortunately, none of the models tested in the Scandinavian sample had borderline or better fit, so results may be less meaningful. However, it is notable that the best fitting model in this sample was consistent with the predictions of the phonological storage framework and included only the bottom-up effect. Interestingly, across all three models, NIQ was either not a significant unique predictor of later language skill or showed a negative relationship to later PM or vocabulary after accounting for earlier PM and vocabulary. We conducted a series of follow-up analyses without NIQ in the models, and the pattern of results did not change.

Comparison of results across all three countries indicates that two different developmental processes may be operating. Relative to the Australian cohort, both the U.S. and Scandinavian samples demonstrated more evidence for bottom-up influence of PM on vocabulary from age 7 to 8. This difference may arise from the cross-cultural mean differences discussed above. Previous research has indicated that the role of PM in vocabulary acquisition is particularly important at relatively earlier stages in language development (Cheung, 1996; Gathercole et al., 1992; Jarrold et al., 2004). Thus, there may be a gradual waning of the bottom-
up effect with overall language/cognitive development, and the Australian cohort appears to be furthest along in this process. Such an effect cannot explain all of the cross-cultural differences, however. Relative to the Scandinavian cohort, both the U.S. and Australian samples demonstrated more evidence for a top-down effect of vocabulary on PM from age 5 to age 7. A notable cultural difference between the U.S. and Australia on the one hand and Scandinavia on the other hand relates to early literacy instruction. In Scandinavia, formal education does not begin until age 7, and there is an established tradition that children should not receive any reading instruction before then. In contrast, both the U.S. and Australia favor both informal and formal reading instruction beginning in preschool. As expected, previous results from the ILTS have found that the Scandinavian twins lag well behind their American and Australian peers in terms of early reading ability (Samuelsson et al., 2008). We propose that the process of learning to read relates to the emergence of a top-down effect of vocabulary on PM performance. Of course, an alternate possibility is that that language difference itself (English versus Swedish and Norwegian) relates to the difference in strength of the top-down effect. It would be difficult to discriminate among these possibilities in the current dataset.

Although there were also significant cross-cultural differences in models of the PM-syntax relationship, these differences were less theoretically meaningful. The best-fitting model within each culture included bidirectional influence of PM and syntax over time. In the U.S. sample, the final model had good fit, and the standardized weights of the cross-lagged paths were again similar (age 5 PM composite → age 6 syntax: .20; age 6 syntax composite → age 7 PM: .27) and both significant at the p < .001 level. In the Scandinavian sample, the final model had borderline fit. Although neither cross-lagged path could be dropped from the model without significant loss of fit, the influence of earlier PM on later syntax appeared somewhat stronger than the influence of earlier syntax on later PM (age 5 PM composite → age 6 syntax: .37, p < .001; age 6 syntax → age 7 PM: .24, < .01). The final model in the Australian sample did not achieve borderline fit, so results must be interpreted cautiously. However, it is notable that in this case, the effect of earlier syntax on later PM appeared somewhat stronger than the effect of earlier PM on later syntax. In summary, although the final model in every culture included both bottom-up and top-down effects,
examination of specific path weights revealed a pattern consistent with results from the PM-vocabulary models, above. The strength of top-down and bottom-up effects was relatively balanced in the U.S. In contrast, the bottom-up effect of early PM on later broad language was dominant in Scandinavia, while the top-down effect of early broad language on later PM was dominant in Australia.

Overall, the results from this study indicate that the correlation between PM and broad language development arises from both bottom-up and top-down effects, with each effect undergoing developmental changes during the age period studied. There appears to be a waning of the bottom-up effect with language (or broad cognitive) development, which could owe to at least two different explanations. First, there might be threshold effects for PM. Perhaps a certain level of PM is needed for efficient language learning, but greater PM does not support even faster learning of vocabulary or syntax. Note that on this explanation, PM continues to play a role in language learning at later stages of development, but it may no longer account for individual differences since the large majority of individuals would be over the threshold. We examined scatter plots of PM and broad language by culture to look for obvious threshold effects. All the plots appeared bivariate normal, with no indication of a dramatic change in the importance of PM to language learning.

A second possibility is that the shift owes primarily to increasing general language and/or cognitive ability. For example, once vocabulary reaches a certain size, children may be more able to utilize lexically-based learning strategies, reducing the importance of PM as a language learning device (Gathercole, 2006). One important lexically-based strategy may be reliance on context to infer the meaning of novel words. A future study could experimentally manipulate the extent to which context is available for word learning. For example, children could be asked to learn novel words both in context and out of context. We suggest that when context is not provided, PM should continue to be a strong predictor of word learning for all children. When context is provided, PM should predict efficiency of word learning only for children with relatively smaller vocabularies. Such a result would also be consistent with the recent report by Gathercole and colleagues that 8-
year-old children with poor PM were impaired in their ability to learn arbitrary verbal material, but were normal in the ability to learn meaningful verbal material (Gathercole et al., 2008). Note that the two accounts of the waning influence of PM on language learning are not mutually exclusive.

We also propose a second developmental shift in PM-vocabulary relations, namely that the process of learning to read contributes to the emergence of a top-down effect of vocabulary on PM tasks (perhaps particularly sentence repetition, since that is the later PM measure included in the current study). For example, perhaps more skilled readers can support performance on the sentence memory task using orthographic representations, particularly for words that they know. In Study 2, we will attempt to replicate the primary conclusions from Study 1 and to further test the basis for changing PM-broad language relations with development.
Study 2

The purpose of Study 2 was to extend the findings from Study 1 by examining PM-broad language relations over time in children with either of two language disorders. Both SSD and RD are believed to result from underlying phonological deficits, and the two disorders are known to be associated with broad oral language weaknesses (Keenan et al., 2006; Lewis, Freebairn, & Taylor, 2000; Nathan, Stackhouse, Goulandris, & Snowling, 2004; Peterson et al., in press; Raitano et al., 2004; Scarborough, 1990; Snowling et al., 2003; Stanovich, 1986). However, it is currently unclear whether the phonological deficits (and in particular, PM deficits) cause difficulties acquiring new vocabulary words and syntactic forms as the phonological storage framework would suggest. To test this hypothesis, we used an SEM approach similar to that of Study 1. We evaluated models of the PM-vocabulary and PM-syntax relation in three clinical groups: two groups of children with histories of SSD, and one group of children with RD.

Overall, Study 1 indicated that both bottom-up and top-down effects explain the relation between PM and broad language, with both effects undergoing developmental changes during the early school years. Consistent with previous research, we found strongest evidence for a bottom-up effect among children at a relatively earlier stage in language development. We also found suggestive evidence that the top-down effect emerges over time, perhaps as a consequence of learning to read. In Study 2, we will test for evidence of similar development trajectories in the bottom-up and top-down effects among children with poor phonological development.

Method

Participants. In order to evaluate the relationship of PM to broad language in children with disorders of phonological development, this study included two groups of children with SSD
histories and one group of children with RD. Each of these three groups is described in more detail below.

**U.S. LTS SSD.** Children with histories of SSD were identified within the U.S. portion of the ILTS via a parent questionnaire. This questionnaire was mailed to all families in the Colorado LTS (n = 489 twin pairs), and the response rate was 65.5%. A child was considered to have a positive history of SSD if the parent endorsed that the child had received speech/language therapy and that the child had had difficulties with articulation. A child was considered to have a negative history of SSD if the parent answered “no” to both the speech/language therapy question and the articulation difficulties question. Exclusionary criteria included parent-endorsed poor hearing, cleft palate, risky birth, or more serious developmental problems (i.e., autism or intellectual disability). If the parent answered “yes” to one of the two SSD criteria and “no” to another or if the questionnaire was not returned, SSD history was considered ambiguous and the child was removed from further analyses. Overall, this procedure identified 80/554 children as having a positive history of SSD. Thus, lifetime prevalence of SSD in this sample is estimated at 14.4%, which is similar to rates reported by previous epidemiological studies (Beitchman, Nair, Clegg, & Patel, 1986; Peckham, 1973). Of the 80 children with positive SSD histories, 38 came from twin pairs in which both members of the pair met criteria while 42 were children whose cotwin did not meet SSD criteria. Because of concerns about sample size, all children with SSD histories were used in analyses.

**DU SSD.** The second group of children with histories of SSD was drawn from a longitudinal and genetic linkage study of the relationship between speech and reading conducted at the University of Denver. This sample included 92 children with a history of SSD who completed a battery of language, cognitive, and literacy tests at two time points: age 5 (mean age = 67.91 months) and age 8 (mean age = 99.65 months). All children had a history of speech difficulties by parent report. In addition, SSD participants were required to have a history of speech therapy for speech sound problems or a score at study entry at or below the 30th percentile on the Goldman-Fristoe Test of Articulation (Goldman & Fristoe, 1986). Although only one of these two criteria (intervention history or psychometric performance) was required, the
large majority of SSD participants met both criteria. This sample also included 38 controls with no history of speech or language difficulties. Exclusionary criteria for both the SSD and control groups included a known genetic disorder, intellectual disability, a pervasive developmental disorder, significant birth complications, an acquired brain injury, hearing loss, deficits in the peripheral speech articulators (e.g., cleft palate), or a language other than English spoken in the home.

U.S. LTS RD. Children with RD were identified within the ILTS based on their performance on the TOWRE (Torgesen, Wagner, & Rashotte, 1999), a test of speeded word reading at age 8 testing (end of second grade). The TOWRE includes two subtests: Sight Word Efficiency, which measures speeded reading of real words, and Phonemic Decoding Efficiency, which measures speeded reading of pseudowords. In addition, the test provides two forms (A & B). To maximize reliability, children completed both subtests of both forms, and a composite reading score was derived by averaging the four resulting age-corrected standard scores. RD was defined as performance on this composite reading score in the bottom 10% of participants for each country. This procedure identified a total of 31 children from the Australian sample, 81 children from the U.S. sample, and 25 children from the Scandinavian sample. Results from Study 1 indicated that the three samples should not be combined in model evaluation, and we were concerned about low sample sizes in Australia and Scandinavia. Thus, only children with RD from the U.S. sample were used in further analyses. Of these 81 children, 38 came from twin pairs in which both members of the pair met RD criteria, while 43 were children whose co-twin did not meet criteria. Again because of sample size concerns, all children were used in analyses.

The overlap of the SSD and RD groups within the U.S. LTS was far from complete. A total of 453 participants had relevant data for both diagnoses. Of these, 366 did not fulfill criteria for either disorder, 58 had SSD but not RD, 20 had RD but not SSD, and 9 had both RD and SSD. While these results agree with previous research (Bird, Bishop, & Freeman, 1995; Lewis, Freebairn, Hansen, Ivyengar, & Taylor, 2004; Peterson et al., in press) in demonstrating a significant comorbidity between the disorders ($\chi^2(1) = 6.49, p < .05$), it is clear that the two groups are largely distinct.
Procedure. Procedure for the U.S. LTS SSD and U.S. LTS RD samples are described in Study 1.

For the DU SSD sample, two testing batteries were completed (mean ages 67.91 months and 99.65 months). At each time point, testing was conducted individually at the University of Denver over three testing sessions lasting two hours each.

Measures. The relevant measures for the two U.S. LTS samples are described in Study 1. Measures for the DU SSD sample are described below.

Nonverbal IQ. NIQ was assessed in the DU SSD sample at age 5 with the Pattern Construction and Matrices subtests of the Differential Ability Scales (DAS; Elliott, 1990).

Phonological Memory. PM was assessed in the DU SSD sample at age 5 with a nonword repetition task (Dollaghan & Campbell, 1998), Recall of Digits from the DAS, and the Sentence Imitation subtest from the Test of Language Development-Primary: Third Edition (TOLD-P:3; Newcomer & Hammill, 1997). PM was reassessed at age 8 with the same nonword repetition and Recall of Digits tasks, as well as with the Recalling Sentences subtest from the Clinical Evaluation of Language Fundamentals-Third Edition (CELF-3; Semel, Wiig, & Secord, 1995).

Vocabulary. Vocabulary knowledge was assessed in the DU SSD sample at age 5 with the Picture Vocabulary and Oral Vocabulary subtests of the TOLD-P:3. Vocabulary knowledge was reassessed at age 8 with the Peabody Picture Vocabulary Test-Third Edition (PPVT-3; Dunn & Dunn, 1997) and the Expressive Vocabulary subtest of the Comprehensive Receptive and Expressive Vocabulary Test (CREVT; Wallace & Hammill, 2002).

Syntax. Morphosyntactic ability was assessed in the DU SSD sample at age 5 with the Grammatic Understanding and Grammatic Completion subtests of the TOLD-P:3. This ability was reassessed at age 8 using the Sentence Structure and Word Structure subtests of the CELF-3 as well as an experimental past tense elicitation task (Marchman, Wulfeck, & Ellis Weismer, 1999).

Reading. Reading fluency was assessed in the DU SSD sample at age 8 with the Gray Oral Reading Test—Third Edition (GORT-III; Wiederholdt & Bryant, 1992).
Analyses. There were multiple phases involved in the analyses. First, confirmatory factor analyses were conducted within the DU SSD sample to evaluate the proposed factor structure at both time points, and results guided the formation of composite variables. Second, each of the clinical groups was compared to the relevant control group to confirm that the disorder was associated with deficits in PM and broad language development. T-tests were used to evaluate group differences, and effect sizes (Cohen’s d) were calculated. Third, we evaluated a series of models of the PM-vocabulary and PM-syntax relationships for each of the three disordered samples. Finally, we performed a follow-up analysis within the DU SSD sample designed to validate the use of sentence memory as the primary measure of PM within the ILTS.

Results

Data preparation and reduction. For the LTS samples, please see Study 1. In the DU SSD sample, raw scores for all variables were regressed on age and standardized within gender. These gender- and age-corrected scores were examined for outliers, and any scores more than 3 SD from the mean were trimmed to a 3 SD cutoff. This trimming process affected 0.3% of all available data. Variables were then examined for departures from normality, and all had acceptable skew and kurtosis (absolute value < 3).

To reduce the number of variables, two confirmatory factor analyses were conducted: one with the four hypothesized constructs measured at age 5 (NIQ, PM, vocabulary, and syntax), and one with the three hypothesized constructs measured at age 8 (PM, vocabulary, and syntax). Gender- and age-corrected raw scores were used as the observed variables. Only participants with a positive history of SSD were included. The age 5 model, which is shown in Figure 7, had good fit: \( \chi^2(21) = 26.56, p > .1; \chi^2/df = 1.27; \text{CFI} = 0.98; \text{RMSEA} = 0.05 \). However, the correlation between the Vocabulary and Syntax factors was high (bounded at 1.00), so we tested a model that collapsed these into a single broad language factor. This change did not worsen fit: \( \chi^2\text{-change}(3) = 4.01, p > .05 \), indicating that a more parsimonious three-factor model was also acceptable. In this three-factor model, the correlation between broad language and phonological memory remained high (.83), so we then tested a two-factor model of NIQ and language. This change did significantly worsen fit: \( \chi^2\text{-change}(4) = 29.42, p < .001 \), and so this model was
rejected. For theoretical reasons and to parallel Study 1, primary analyses relied on NIQ, PM, vocabulary, and syntax composites derived by from the initial four-factor model. However, in follow-up analyses we created a single broad language composite based on both the vocabulary and syntax subtests.

Figure 7. Confirmatory factor analysis of language constructs assessed at age 5 within the DU SSD Sample. Standardized weights are shown.

We next tested the initial model of the three hypothesized language constructs measured at age 8. This model, which is shown in Figure 8, had borderline fit: \( \chi^2(17) = 29.73, p < .05; \chi^2/df = 1.75; \) CFI = 0.97; RMSEA = 0.09. The correlation between the Vocabulary and Syntax factors was again very high (.92). Collapsing these into a single broad language factor again did not worsen fit: \( \chi^2\text{-change}(2) = 2.05, p > .05. \) (The CFI and RMSEA for the three- and two-factor models were similar, with the two-factor model also having borderline fit). Thus, we adopted the
same approach as described above. Primary analyses utilized composites created based on the original three-factor structure, but we also created a single broad language composite of both the vocabulary and syntax subtests to be used in follow-up analyses. The final measurement models from both time points were rerun with a slightly larger sample including both participants with an SSD history and controls, and the results were very similar.

![Diagram of factor analysis](image)

**Figure 8.** Confirmatory factor analysis of language constructs assessed at age 8 within the DU SSD Sample. Standardized weights are shown.

**Effect size comparisons.** We next compared the performance of each of the three disordered groups to the relevant control group to confirm that diagnosis was associated with impaired PM and broad language performance. The U.S. LTS SSD group was compared to participants in the U.S. LTS without SSD, the DU SSD group was compared to the control participants recruited into that sample, and the U.S. LTS RD group was compared to participants in the U.S. LTS without RD. Age- and gender-corrected raw scores were used, and when
multiple measures of a construct were available at a given time point, we used composite scores. Because results from Study 1 suggested that the emergence of a top-down effect (broad language→PM) might relate to literacy skill, we also report effect sizes for reading fluency relative to controls at age 8. Reading fluency was measured with the TOWRE (average of Forms A and B) in the LTS samples and with the GORT-III Fluency Score in the DU SSD Sample. For these tasks, age-corrected standardized scores were used.

Results are reported in Table 7. Overall, each disordered group demonstrated poorer performance than controls, but effect sizes varied by group. The U.S. LTS SSD group demonstrated the least impaired language performance, followed by the U.S. LTS RD group, with the DU SSD group demonstrating the most impaired performance. These differences likely relate to variability in the ways the samples were ascertained. Ascertainment of SSD in the LTS was probably the least reliable, since it depended fully on parent report. In contrast, ascertainment of RD in the LTS was reliable, but since it was based wholly on psychometric measurement, this group may have included some children who would not attract clinical concern. It is not surprising that the DU SSD group was the most impaired overall, since members of this group had both a history of clinical concern and poor psychometric performance on the defining symptom (speech production).
Table 7
Effect Sizes for Language and Cognitive Deficits in Each of the Three Disordered Groups, Compared to the Relevant Control Group

<table>
<thead>
<tr>
<th>Age 5</th>
<th>LTS SSD</th>
<th>DU SSD</th>
<th>LTS RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIQ</td>
<td>0.07</td>
<td>0.69***</td>
<td>0.66***</td>
</tr>
<tr>
<td>PM</td>
<td>0.57***</td>
<td>1.64***</td>
<td>0.45***</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.29*</td>
<td>1.27***</td>
<td>0.51***</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.36**</td>
<td>1.10***</td>
<td>0.55***</td>
</tr>
<tr>
<td>Age 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PM</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.24^</td>
<td>--</td>
<td>0.71***</td>
</tr>
<tr>
<td>Age 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PM</td>
<td>0.29*</td>
<td>--</td>
<td>0.58***</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Syntax</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Age 8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PM</td>
<td>--</td>
<td>1.13***</td>
<td>--</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.22^</td>
<td>1.02***</td>
<td>0.52***</td>
</tr>
<tr>
<td>Syntax</td>
<td>--</td>
<td>0.94***</td>
<td>--</td>
</tr>
<tr>
<td>Reading Fluency</td>
<td>0.26^</td>
<td>0.77***</td>
<td>2.55***</td>
</tr>
</tbody>
</table>

^p < .1. *p < .05. **p < .01. ***p < .001.

In addition to overall degree of impairment, the pattern of impaired performance varied by group, with the U.S. LTS SSD group showing the most evidence for a selective phonological deficit. This group was the only disordered group with NIQ scores comparable to controls.

Further, while effect sizes for vocabulary and syntax deficits were in the small range, the PM
deficit was in the moderate range at age 5. In contrast, the U.S. LTS RD group demonstrated comparable (moderate-sized) deficits on PM, broad language and NIQ. The DU SSD group had large deficits on all language constructs and a moderate NIQ deficit. In terms of reading fluency, all three groups demonstrated a significant impairment relative to controls, but the effect was small in the LTS SSD group and large in the other two clinical groups. (The effect was very large in the LTS RD group, which is unsurprising since it was the subgroup-defining measure.)

The differing profiles of the three clinical groups should provide interesting information about the changing PM-broad language relationship with development. First, since all three clinical groups have at least some degree of delayed language development, we expect all to show bottom up effects of PM on vocabulary and syntax that are as large or larger than those found in the full U.S. population (Study 1). These bottom-up effects may be particularly pronounced in the DU SSD group, which has the most delayed language development. In contrast, the LTS SSD group shows only a mild language delay; thus, in this group, the strength of the bottom-up effects can be expected to be quite close to those found in the population.

Reading development is substantially delayed in the DU SSD and LTS RD samples. If the emergence of a top-down effect (broad language→PM) relates to literacy development, then these groups should have a weak or absent top-down effect in comparison to the U.S. population. However, since the LTS SSD group shows only a mild reading delay, we predict at least some top-down influence in this sample. (For this group, Standard Scores for reading fluency were well within the average range.) In sum, the overall cognitive profiles of the DU SSD and LTS RD groups are fairly similar, although the two are drawn from different samples and meet criteria for different diagnoses. We expect both to show a pattern of performance consistent with the phonological storage framework. The LTS SSD is only mildly impaired; this group may well show more bidirectional relations between PM and broad language, consistent with the U.S. population as a whole.

Model comparison: PM-vocabulary relation. U.S. LTS SSD. We began by testing the final PM-vocabulary model showing bidirectional relations from the full U.S. LTS sample (see Figure 3B) in just the U.S. LTS SSD group. The model showed good fit: $\chi^2(1) = 0.81, p > .3; \chi^2/df$
The standardized paths from age 5 vocabulary → age 7 PM and from age 7 PM → age 8 vocabulary were comparable in size (.24 and .27, respectively) and both significant at the p < .05 level. Further, dropping either the bottom-up or top-down path resulted in significant loss of fit ($\chi^2$-change(1) = 6.00, p < .01; $\chi^2$-change(1) = 5.06, p < .05, respectively). Thus, we accepted the bidirectional model and concluded that the longitudinal relationship between PM and vocabulary was similar for this disordered group as for the U.S. population. The final model is shown in Figure 9A.

DU SSD. We began by testing a model in which PM and vocabulary predicted themselves and each other across time. Again, NIQ was included at age 5 as a control variable. The initial model had borderline fit: $\chi^2(1) = 4.74$, p < .05; $\chi^2$/df = 4.74; CFI = 0.99; RMSEA = 0.20. Despite meeting our criteria for borderline fit (two of three indices in the acceptable range), the RMSEA was quite poor. While the path from earlier PM to later vocabulary was significant at the p < .001 level, the path from earlier vocabulary to later PM was not significant (p > .9). We tested a model corresponding to the phonological storage framework by dropping the path from age 5 vocabulary to age 8 PM. This change did not significantly degrade fit ($\chi^2$-change(1) = 0.01, p > .05); further, the RMSEA improved to 0.12, much closer to the acceptable range. Results were very different for the model corresponding to the LRM. From the initial model, we dropped the path from age 5 PM to age 8 vocabulary. This change significantly worsened fit ($\chi^2$-change(1) = 22.92, p < .001) and the RMSEA in this model was even poorer (0.38). We thus accepted the model corresponding to the phonological storage framework, with only a bottom-up effect. As a final step, we added a parameter correlating the error terms for endogenous variables at age 8, to account for time-specific factors (fatigue, etc.) that would not be explained by age 5 variables. This addition significantly improved the model $\chi^2$ ($\chi^2$-change(1) = 4.74, p < .05) and brought overall fit into the good range. The final model describing the PM-vocabulary relationship in the DU SSD group is shown in Figure 9B.

U.S. LTS RD. We next tested the final PM-vocabulary model showing bidirectional relations from the full U.S. LTS sample (see Figure 3B) in just the U.S. LTS RD group. The model had good fit: $\chi^2(1) = 0.72$, p > .3; $\chi^2$/df = 0.72; CFI = 1.00; RMSEA = 0.00. While the path
from earlier PM to later vocabulary was significant at the $p < .01$ level, the path from earlier vocabulary to later PM was not significant ($p > .4$). We next evaluated the model corresponding to the phonological storage framework by dropping the top-down path, and there was not a significant loss of fit ($\chi^2$-change(1) = 0.53, $p > .05$). However, when we dropped the bottom-up path from the initial model, there was a significant loss of fit ($\chi^2$-change(1) = 6.63, $p < .01$). The final model, which is shown in Figure 9C, corresponded to the phonological storage framework and had very good fit.

In summary, results concerning PM-vocabulary relations for the three clinical groups were consistent with our hypotheses. As expected, all three groups showed clear evidence of a bottom-up influence of PM on vocabulary knowledge. Although we cannot formally compare the magnitude of effects across samples, it is notable that the estimate for the bottom-up effect in the DU SSD sample was numerically larger than those in the two LTS samples, despite spanning a substantially longer time frame. This pattern could be consistent with stronger bottom-up effects in the most language-delayed sample. We also predicted that the DU SSD and LTS RD groups would not demonstrate a top-down influence of vocabulary on PM, since both samples showed large impairments in reading fluency. This prediction was confirmed; only the LTS SSD group, which had the mildest deficit in reading skill, demonstrated a top-down effect. The implications of these findings will be addressed in the discussion section.
Figures 9A-C. Final models of PM-vocabulary relation across time for the three disordered groups. Standardized weights are shown.
\(^p < .1. *p < .05. **p < .01. ***p < .001.\)

Model comparison: PM-syntax relation. These analyses proceeded in parallel to the PM-vocabulary analyses described above. In each of the three disordered samples, we began with an initial, bidirectional model of the PM-syntax relation. We then tested whether either cross-lagged path could be dropped from the model without significant loss of fit before choosing a final model depicting the longitudinal relationship between PM and syntax.
**U.S. LTS SSD.** The final PM-syntax model from full U.S. LTS sample (see Figure 6B) had good fit in just the U.S. LTS SSD group: $\chi^2(1) = 0.57, p > .4; \chi^2/df = 0.81; CFI = 1.00; RMSEA = 0.00$. The path from age 5 PM to age 6 syntax was in the large range (.52) and significant at the $p < .001$ level, while the path from age 5 syntax to age 7 PM was smaller (.23) and significant at only the trend level. Dropping the bottom-up path resulted in significant loss of fit ($\chi^2$-change(1) = 14.49, $p < .001$). Dropping the top-down path did not result in a significant loss of fit, but did reduce fit at the trend level ($\chi^2$-change(1) = 3.48, $p < .1$). Furthermore, this change moved the RMSEA into the poor range (.11). Thus, the overall pattern indicated that although the bottom-up effect was somewhat stronger than the top-down effect, a model including both effects was best. The final model for the U.S. LTS SSD group is shown in Figure 10A. Surprisingly, in the final model, earlier PM was a stronger unique predictor of later syntax than was earlier syntax. A similar result was found in the final PM-syntax model for the Scandinavian LTS sample (see Figure 6C). These findings were unexpected and suggest that the tests used to measure syntax at different ages may rely on somewhat different underlying skills. This issue is addressed further in the discussion.

**DU SSD.** We began by testing a model in which PM and syntax predicted themselves and each other across time, controlling for age 5 NIQ. The initial model had poor fit: $\chi^2(1) = 7.54, p < .01; \chi^2/df = 7.54; CFI = 0.98; RMSEA = 0.27$. Results from the PM-vocabulary model in this sample suggested that correlating the error terms at age 8 would improve fit; however, the additional parameter saturated the model, thus preventing an evaluation of model fit. In this saturated model, the path from age 5 PM to age 8 syntax was moderate in size and significant at the $p < .001$ level, while the path from age 5 syntax to age 8 PM was small and nonsignificant. Dropping the top-down path did not significantly degrade fit ($\chi^2$-change(1) = 1.19, $p > .05$) but dropping the bottom-up path did ($\chi^2$-change(1) = 11.17, $p < .001$). The final model in this sample included only the bottom-up effect and had good fit; see Figure 10B.

Because results from the CFAs indicated that the vocabulary and syntax tests could be collapsed into a single broad language factor at both time points in the DU SSD sample, we tested a follow-up model in which PM and broad language predicted each other over time. Not surprisingly,
results closely mirrored findings from both the PM-vocabulary and PM-syntax analyses. The best-fitting, most parsimonious model included a bottom-up effect from PM to broad language but not a top-down effect from broad language to PM.

**U.S. LTS RD.** The final PM-syntax model from full U.S. LTS sample (see Figure 6B) had good fit in just the U.S. LTS RD group: $\chi^2(1) = 0.39, p > .5$; $\chi^2$/df = 0.39; CFI = 1.00; RMSEA = 0.00. The top-down path from age 5 syntax to age 7 PM was in the moderate range (.33) and significant at the $p<.01$ level, while the bottom-up path from age 5 PM to age 6 syntax was smaller (.22) and significant at only the trend level. Dropping the bottom-up path did not significantly reduce fit, but did worsen fit at the trend level ($\chi^2$-change(1) = 2.94, $p < .1$) and also moved the RMSEA out of the good range (to .09). Dropping the top-down path resulted in significant loss of fit ($\chi^2$-change(1) = 9.89, $p < .01$). Thus, in this case, the top-down effect was somewhat stronger than the bottom-up effect, but a model including both effects was again best. The final model for the U.S. LTS RD group is shown in Figure 10C.
In summary, results concerning the PM-syntax relationship in the two groups with SSD largely mirrored the PM-vocabulary results above. The PM-broad language relation was consistently driven by bottom-up effects in the DU SSD sample, which demonstrated large deficits relative to controls on all language and literacy measures. In contrast, in the U.S. LTS SSD sample, which had milder deficits overall, the PM-broad language relation was bidirectional.
However, in the one RD sample with both vocabulary and syntax measures available, results differed somewhat for the two broad language constructs. This inconsistency is addressed in the discussion section.

**Validation of sentence repetition task as a measure of phonological memory.** We conducted a final set of analyses within the DU SSD sample in order to validate the use of a sentence repetition task as a primary measure of PM within the ILTS. Previous studies have more commonly used nonword repetition and/or span tasks to assess PM. Fortunately, all three types of measure were available in the DU SSD sample. We created a composite of “traditional” PM measures (nonword repetition, DAS Recall of Digits) at both time points in this sample. We then reran the analyses evaluating the PM-vocabulary and PM-syntax relationships in this sample using this composite as well as using just the sentence repetition measures, and compared the results.

Results for the PM-vocabulary analysis were very clear-cut. The pattern of findings was nearly identical for the full PM composite (all three measures), the composite of nonword repetition and Recall of Digits only, and the sentence repetition task only. In every case, model testing indicated that the PM-vocabulary relationship owed almost entirely to bottom-up effects within this sample. Further, specific path weights were quite similar for models with different measure of PM. Path weights for the cross-lagged paths in models including bidirectional effects are reported in Table 8.

Results for the PM-syntax analysis were less straightforward. Model testing using the full PM composite or the traditional PM composite (nonword repetition and DAS Recall of Digits) both indicated that the PM-syntax relationship owed primarily to bottom-up effects within this sample. However, when the sentence repetition task was used as the sole measure of PM, model testing indicated that the PM-syntax relationship was bidirectional in nature. It is also notable that overall fit for these models were quite poor.

---

**Table 8**
## Cross-Lagged Standardized Path Weights in Models Using Different Measures of PM within the DU SSD Sample

<table>
<thead>
<tr>
<th>Path</th>
<th>β-weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM5 → Voc8</td>
<td></td>
</tr>
<tr>
<td>Full PM composite</td>
<td>.38***</td>
</tr>
<tr>
<td>Traditional PM composite</td>
<td>.34***</td>
</tr>
<tr>
<td>Sentence repetition only</td>
<td>.33***</td>
</tr>
<tr>
<td>Voc5 → PM8</td>
<td></td>
</tr>
<tr>
<td>Full PM composite</td>
<td>.01</td>
</tr>
<tr>
<td>Traditional PM composite</td>
<td>.05</td>
</tr>
<tr>
<td>Sentence repetition only</td>
<td>.07</td>
</tr>
<tr>
<td>PM5 → Syn6</td>
<td></td>
</tr>
<tr>
<td>Full PM composite</td>
<td>.31***</td>
</tr>
<tr>
<td>Traditional PM composite</td>
<td>.25**</td>
</tr>
<tr>
<td>Sentence repetition only</td>
<td>.56***</td>
</tr>
<tr>
<td>Syn5 → PM7</td>
<td></td>
</tr>
<tr>
<td>Full PM composite</td>
<td>.08</td>
</tr>
<tr>
<td>Traditional PM composite</td>
<td>.12^</td>
</tr>
<tr>
<td>Sentence repetition only</td>
<td>.53***</td>
</tr>
</tbody>
</table>

*Note.* For abbreviations, see Table 3.

^p < .1. *p < .05. **p < .01. ***p < .001.

One explanation for the full pattern of results is that performance on sentence repetition tasks is multiply determined and constrained by both PM and syntactic abilities. Thus, when pitted against measures of vocabulary, a sentence repetition task can provide a reasonable estimate of PM skills. However, these tasks are not distinct enough to provide a “clean” estimate of PM in a model that also includes syntax. In sum, this analysis validated the use of a sentence repetition task to measure PM in some, but not all analyses. This issue will be addressed further in the discussion.

**Discussion**
In this study, we investigated the nature of PM-vocabulary and PM-syntax relations in children with either of two phonologically-based language disorders (SSD or RD). Results from Study 1 indicated that in the population, the correlation between PM and broad language owes to both bottom-up and top-down factors, but that the primary direction of influence varies. We proposed that the bottom-up effect gradually wanes with language development, while the top-down effect may emerge with increasing literacy skill. In Study 2, we tested whether children with impaired language development show strong evidence of bottom-up PM-broad language relations and whether children with impaired literacy development show weaker top-down relations in comparison to their peers with normal development.

We identified children with either SSD or RD in a variety of different ways. Children with SSD histories within the U.S. portion of the ILTS were identified by parent questionnaire. A second sample of children with SSD histories (the DU SSD sample) had both a history of clinical concern and poor performance on an objective test of speech sound production. Children with RD within the U.S. portion of the ILTS were identified on the basis of low test scores only. These differences in ascertainment related to differences in level and pattern of language difficulties by group. The U.S. LTS SSD group had a moderate PM impairment at age 5, mild broad language deficits, a mild deficit in reading fluency, and no evidence for broad cognitive impairments (i.e., no NIQ deficit). The U.S. LTS RD group showed moderate deficits on a broad range of cognitive variables, including PM, broad language, and NIQ, and (by definition) a large deficit in reading fluency. The DU SSD group was the most impaired overall, with large deficits on all language and literacy variables and a moderate NIQ deficit.

Results for the PM-vocabulary analyses were very clear. In the two groups with substantial delays in both language and literacy development (DU SSD, U.S. LTS RD), the final models describing the longitudinal relationship between PM and vocabulary corresponded to the phonological storage framework and included only bottom-up effects. However, in the one group with only mild language and literacy delays (U.S. LTS SSD), the final model included bidirectional relations and was thus very similar to the population model.
These results have several theoretical implications. First, they suggest that in children with phonological impairments, PM skill continues to constrain the acquisition of new vocabulary items well into the school years. In Study 1, we also found evidence for an influence of PM on vocabulary growth in the full U.S. population over this time period. Thus, it is unclear whether the waning of the bottom-up effect is truly delayed in these groups. At least one later time point would be needed to answer this question. Data collection is ongoing in the ILTS, with a measure of receptive vocabulary administered at the end of fourth grade. In addition, a subset of children in the DU SSD sample participated in a follow-up study that included a receptive vocabulary measure (Phinney, 2008). Thus, a future investigation can examine whether PM continues to constrain vocabulary knowledge to age 10 in children with and without language delays. We predict that the bottom-up effect will weaken or disappear in the full U.S. population, but will continue to be evidenced by at least some of the disordered groups. It will be particularly interesting to compare performance of the LTS SSD group with that of the other clinical groups, since that group has the most selective PM impairment.

If there are threshold effects for PM in new word learning, then the waning of the bottom-up effect should be delayed in individuals with a relatively selective PM deficit. If the three disordered groups evidenced a similar delay in the disappearance of the bottom-up effect, that would suggest a threshold effect. If the LTS SSD group performed similarly to the U.S. population through age 10 while the other groups demonstrated a continuing influence of PM on word learning, that would suggest that the waning of the bottom-up effect is more tied to broad language development.

Second, our results are consistent with the proposal that the emerging influence of vocabulary knowledge on PM tasks relates to becoming a fluent reader. However, in our samples, reading ability was somewhat confounded with degree of language impairment. Thus, it is possible that the DU SSD and LTS RD groups failed to show the top-down effect not because of poor literacy but because of weak overall language skill. A future study could clarify the role of reading acquisition in changing PM-vocabulary relations with the “school cutoff design,” which compares children of extremely similar ages who differ by a full year of formal schooling because of just meeting or missing the age-based cutoff for kindergarten entry (e.g., Baltes & Reinert, 1969). Studies using this
design have found larger effects of schooling on literacy than on most oral language abilities (Bowey & Francis, 1991; Christian, Morrison, Frazier, & Massetti, 2000; Crone & Whitehurst, 1999; Ferreira & Morrison, 1994). We predict that an application of the school design to the current research questions would reveal that the two groups have similar developmental trajectories (by age) for the waning of the bottom-up influence, but that the group with more schooling should show an earlier emergence of the top-down effect.

Results for the PM-syntax analyses were more complicated. In the DU SSD sample, the final model of PM-syntax relations mirrored the final model of PM-vocabulary relations (i.e., bottom-up effect only.) This finding was unsurprising given that CFAs had suggested vocabulary and syntax measures could be treated as a single broad language construct in this sample. In the U.S. LTS, though, results of the PM-syntax analyses included some inconsistencies. The current study also generated some methodological questions about the syntax measures available in the ILTS. We conducted a follow-up analysis designed to validate the use of a sentence repetition task to measure PM within the ILTS. In that analysis, we found that nonword repetition, digit span, and sentence repetition tasks behaved similarly as measures of PM in models evaluating PM-vocabulary relations, but not in models evaluating PM-syntax relations.

Across Studies 1 and 2, there have been several other indications of methodological limitations in available syntax measures in the ILTS. In two models, (see Figures 6C and 10A) earlier syntax has been a weaker predictor of later syntax than has earlier PM. This result suggests that the syntax measures at different time points tap different underlying skills. Further, in Study 1, we found that age 5 syntax was a better predictor of age 7 PM than was age 6 syntax in two of three countries. Taken together, there are multiple reasons to question whether theoretically meaningful conclusions can be drawn from the PM-syntax analyses within the ILTS. Thus, in Study 3, we will address only the etiologic relation between PM and vocabulary. Further, the General Discussion will focus primarily on understanding the nature of the PM-vocabulary relationship. Fortunately, evaluation of the predictions of the phonological storage framework versus the LRM depends much more heavily on results including vocabulary measures than those including syntax measures.
Study 3

Study 3 utilizes a behavior genetics approach, the DeFries-Fulker method, to investigate the etiology of PM and vocabulary deficits. This study has several goals. First, previous research in the ILTS has estimated the relative effects of genes and environment on verbal memory and vocabulary across the full range of individual differences, but not for subgroups with language deficits. Current results will address whether PM and vocabulary deficits arise from the same types of influences (genetic or environmental) as do PM and vocabulary scores in the population. Second, this study will further understanding of the basis of the phenotypic correlation between PM and vocabulary for individuals with language deficits. Does the relationship owe primarily to common genes, common environmental experiences, or both? Finally, this study will further explore the direction of effect between PM and vocabulary across time, using a different method than employed in Studies 1 and 2. We will compare the etiologic influences on earlier PM/later vocabulary to the etiologic influences on earlier vocabulary/earlier PM.

The DeFries-Fulker (DF) method (DeFries & Fulker, 1985, 1988) is a regression-based approach suitable for extremes analysis. Probands who fall below a cut-off value in a particular variable are selected. The scores of MZ and DZ co-twins of are then predicted from proband scores in a regression model. The logic of the method is as follows: to the extent that having a poor score is due to genes, the scores of DZ co-twins should regress further back toward the population mean than the scores of MZ co-twins. Effects of shared environment can also be estimated with this method, and are based on the extent to which co-twins resemble probands, regardless of zygosity. Effects of non-shared environments include measurement error and are based on the extent to which MZ probands and their co-twins differ.

Several previous studies have investigated the etiology of verbal memory and vocabulary in the full ILTS. These studies utilized the same measures of age 5 vocabulary, age 7 PM, and
age 8 vocabulary as the current study. The age 5 verbal memory composite used in earlier studies was slightly broader than in the current study; it included a measure of story memory in addition to the nonword repetition and sentence memory measures. (Previous studies have utilized both latent traits and composites, with similar results.)

Samuelsson and colleagues reported on the behavior genetics of the age 5 constructs (Samuelsson et al., 2005). Verbal memory showed large effects of genes ($h^2 = .57$; 95% confidence interval (CI) = .35-.79) and modest effects of shared environment ($c^2 = .29$; 95% CI = .08-.48). At this age, vocabulary showed the opposite pattern, with modest effects of genes ($h^2 = .32$; 95% CI = .06-.56) and a large effect of shared environment ($c^2 = .60$; 95% CI = .38-.81). In a later study (Byrne et al., 2007), age 7 Sentence Memory had moderate genetic effects ($h^2 = .35$, 95% CI = .05-.67), and the effect of shared environment remained modest and was not statistically significant ($c^2 = .24$; 95% CI = .00-.49). Finally, for age 8 vocabulary (Byrne et al., 2009), the effects of genes and shared environment were similar in magnitude and both in the moderate range ($a^2 = .44$; 95% CI = .31-.59; $c^2 = .36$; 95% CI=.22-.49). Overall, these findings indicate that all constructs are heritable. There appear to be stronger effects of shared environment for vocabulary than for verbal memory, particularly at younger ages. There is also an indication of increased genetic influence and waning environmental influence to vocabulary scores with age. The current study will test whether these phenomena also hold true for PM and vocabulary deficits.

**Method**

**Participants.** Twin pairs from all three cultures included in the ILTS participated (total n = 925 pairs). Zygosity was determined with DNA collection from cheek swab samples, or in a minority of cases, by questionnaire (Nichols & Bilbro, 1966). Gender and zygosity by culture are reported in Table 9.

**Measures.** Analyses utilized the age 5 PM and vocabulary composites, the age 7 PM measure, and the age 8 vocabulary measure. Age, gender, and country corrected raw scores were used. See Study 1 for a more detailed description of measures.
Analyses. Three sets of analyses were conducted in order to investigate the etiology of PM and vocabulary deficits both within a time point (at age 5) and across time. For all analyses, we first identified the subset of participants with a deficit in at least one of the constructs. A deficit was defined as a score that fell at least 1.25 standard deviations below the population mean, and probands were selected for having a deficit in either PM or vocabulary (i.e., we selected the bottom 10 percent of the population on each variable.)

Table 9
Gender and Zygosity by Country within the ILTS

<table>
<thead>
<tr>
<th></th>
<th>Australia</th>
<th>United States</th>
<th>Scandinavia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male twin pairs (n)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MZ</td>
<td>80</td>
<td>97</td>
<td>48</td>
</tr>
<tr>
<td>DZ</td>
<td>55</td>
<td>146</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>135</td>
<td>243</td>
<td>88</td>
</tr>
<tr>
<td>Female twin pairs (n)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MZ</td>
<td>71</td>
<td>128</td>
<td>46</td>
</tr>
<tr>
<td>DZ</td>
<td>51</td>
<td>118</td>
<td>49</td>
</tr>
<tr>
<td>Total</td>
<td>122</td>
<td>246</td>
<td>95</td>
</tr>
</tbody>
</table>

In the first set of analyses, we ran univariate DF regressions to examine the relative contributions of genes, shared environment, and nonshared environment ($h^2_g$, $c^2_g$, and $e^2_g$, respectively) to a deficit in PM or vocabulary. We used the basic DF equation:

$$C = B_1P + B_2R + K,$$

where $C$ stands for co-twin’s score on the relevant construct, $P$ stands for the proband’s score, and $R$ represents the coefficient of relationship (1.0 for MZ twins and 0.5 for DZ twins). Data were scaled so that $B_1$ gave an unbiased estimate of heritability ($h^2_g$), and estimates of $c^2_g$ and $e^2_g$ were then derived with an adaptation of Falconer’s method. In these and subsequent DF analyses, twin pairs in which both members met the extreme selection criteria were double entered, and standard errors of the regression coefficients were conservatively
corrected for the number of double-entered pairs (Rodgers & Kohler, 2005; Stevenson, Pennington, Gilger, DeFries, & Gillis, 1993).

In the second set of analyses, we reran the univariate regressions with the inclusion of culture as a covariate in order to test if heritability varied as a function of culture. We used the extended regression equation: \( C = B_1P + B_2R + B_3\text{Culture} + B_4P*\text{Culture} + B_5\text{Culture}*R + K \). The \( B_5 \) coefficient tested whether the heritability of the construct varied as a function of culture. We treated culture as a dichotomous variable, and for each construct, we ran three sets of regressions (Australia vs. U.S., U.S. vs. Scandinavia, and Australia vs. Scandinavia).

Finally, we tested bivariate heritability and shared environmentality of PM and vocabulary both within a time point (at age 5) and across development. The goal of these analyses was to further understand the basis for the phenotypic correlation between PM and vocabulary—does it owe to shared genes, shared environmental experiences or both? The bivariate model is similar to the basic univariate model, except that proband selection is based on one construct, and this value is used to predict the co-twin’s score on a second construct. For example, to determine whether genes that contribute to a deficit in PM at age 5 also influence vocabulary at the same age, we ran the equation: \( C_{\text{Voc}} = B_1P_{\text{PM}} + B_2R + K \). We scaled data so that \( B_1 \) gave an unbiased estimate of \( h^2_{g,xy} \). This value indicates the proportion of a poor score on the co-twin measure (age 5 vocabulary) that is due to genes that also influence the selection measure (age 5 PM). In addition, we computed bivariate shared environmentality (\( c^2_{g,xy} \)) using an adaptation of Falconer’s method.

**Results**

*Univariate DF results.* Univariate results are summarized in Table 10. In general, our findings concerning the etiology of PM and vocabulary deficits agree with previous results from the ILTS across the full range of individual differences. We found that deficits in both constructs were significantly heritable at all time points, indicating that a portion of a poor PM or vocabulary score is due to genetic influence. At ages 5 and 7, heritability of a PM deficit was substantial while the effect of shared environment was small (or null) and nonsignificant. The pattern differed
for age 5 vocabulary. In that case, heritability was moderate, and there was a moderate and statistically significant effect of shared environment. By age 8, heritability of a vocabulary deficit was large; the effect of shared environment remained moderate but was now significant only at the trend level (of note, power was reduced because of the smaller sample size at that age). Estimates of nonshared environment, which includes measurement error, were moderate for most variables.

Table 10  
Results of Univariate DeFries-Fulker Analyses of PM and Vocabulary Deficits

<table>
<thead>
<tr>
<th>Selection measure</th>
<th>PM5</th>
<th>Voc5</th>
<th>PM7</th>
<th>Voc8</th>
</tr>
</thead>
<tbody>
<tr>
<td>n pairs</td>
<td>209</td>
<td>199</td>
<td>189</td>
<td>153</td>
</tr>
<tr>
<td>MZ pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proband mean</td>
<td>-1.75</td>
<td>-1.97</td>
<td>-1.83</td>
<td>-1.75</td>
</tr>
<tr>
<td>Co-twin mean</td>
<td>-1.21</td>
<td>-1.41</td>
<td>-1.17</td>
<td>-1.53</td>
</tr>
<tr>
<td>DZ pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proband mean</td>
<td>-1.69</td>
<td>-1.77</td>
<td>-1.93</td>
<td>-1.74</td>
</tr>
<tr>
<td>Co-twin mean</td>
<td>-0.67</td>
<td>-0.98</td>
<td>-0.62</td>
<td>-1.05</td>
</tr>
<tr>
<td>$h^2_g$ (95% CI)</td>
<td><strong>.59</strong> (.31-.87)</td>
<td><strong>.32</strong> (.02-.62)</td>
<td><strong>.64</strong> (.40-.88)</td>
<td><strong>.53</strong> (.19-.88)</td>
</tr>
<tr>
<td>$c^2_g$ (95% CI)</td>
<td>.10 (0-.38)</td>
<td><strong>.39</strong> (.09-.70)</td>
<td>.00 (0-.25)</td>
<td>.34 (0-.69)</td>
</tr>
<tr>
<td>$e^2_g$</td>
<td>.31</td>
<td>.29</td>
<td>.36</td>
<td>.13</td>
</tr>
</tbody>
</table>

Notes. Heritability and shared environment estimates whose confidence intervals do not include 0 (reported in bold) are statistically significant at the $p<.05$ level. For abbreviations, see Table 3.

Univariate DF results with covariate of culture. Next, we tested whether heritability of the four constructs varied as a function of culture by including country as a covariate in the univariate DF equations. These results can be summarized very simply: There was no evidence for differences in heritability across cultures (U.S. versus Australia: all t-values < 1.6, all $p$-values > .1; U.S. versus Scandinavia: all t-values < 0.6, all $p$-values > 0.6; Australia versus Scandinavia: all t-values < 1.5, all $p$-values > 0.1). Thus, although results from Study 1 had
indicated that the phenotypic influences on PM and vocabulary over time varied across cultures, these results indicate that the relative contributions of genes and environment to individual differences in PM or vocabulary at a given time point are similar across the cultures included in the ILTS. The current finding is very much in keeping with previous etiologic results from the ILTS concerning oral language development (e.g., Byrne et al., 2009; Samuelsson et al., 2005).

**Bivariate DF results.** Finally, we tested bivariate heritability and shared environmentality of PM and vocabulary both within a time point (at age 5) and across time. The cross-time analyses investigated three meaningful variable pairs (age 5 PM/age 8 vocabulary, age 7 PM/age 8 vocabulary, and age 5 vocabulary/age 7 PM). Since all four constructs were significantly heritable, it was sensible to test bivariate heritability of all these variable pairs. PM did not show significant effects of shared environment, but we tested for bivariate shared environmentality nonetheless. Vocabulary deficits did appear to be affected by shared environmental experiences, and proband selection for the different deficits resulted in partially distinct subgroups. Thus, it is possible that particular environmental experiences jointly contribute to poor vocabulary and PM for the subgroup of individuals selected for a vocabulary deficit. We included tests of bivariate shared environmentality with selection on PM for completeness, though given the null univariate results, we did not expect significant bivariate findings in this case.

Results concerning the degree to which shared etiologies contribute to deficits in PM and vocabulary at age 5 are reported in Table 11. Findings varied according to whether the probands was selected for a PM deficit or a vocabulary deficit. There was evidence for a shared genetic etiology only when PM was the selection measure. In that case, bivariate heritability was .38, indicating that 38% of a poor vocabulary score is due to genes that also influence a PM deficit. As expected, $c_{xy}^2$ was small and nonsignificant when age 5 PM was the selection measure. The reverse pattern held with selection on vocabulary. In this case, bivariate heritability was null but bivariate shared environmentality was .44, indicating that 44% of a poor PM score is due to shared environmental experiences that also influence a vocabulary deficit.
Table 11
Results of Bivariate DeFries-Fulker Analyses of PM and Vocabulary Deficits at Age 5

<table>
<thead>
<tr>
<th>Selection measure (X)</th>
<th>PM5</th>
<th>Voc5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-twin measure (Y)</td>
<td>Voc5</td>
<td>PM5</td>
</tr>
<tr>
<td>n pairs</td>
<td>213</td>
<td>195</td>
</tr>
</tbody>
</table>

MZ pairs

| Proband mean (X) | -1.75 | -1.97 |
| Co-twin mean (Y) | -0.99 | -0.78 |

DZ pairs

| Proband mean (X) | -1.69 | -1.77 |
| Co-twin mean (Y) | -0.63 | -0.74 |

\( h^2_{g,xy} \) (95% CI) \( .38 \) (0.07-.70) \( .00 \) (0-.33)

\( c^2_{g,xy} \) (95% CI) \( .18 \) (0-.51) \( .44 \) (.13-.75)

Notes. Estimates whose confidence intervals do not include 0 (reported in bold) are statistically significant at the p<.05 level. For abbreviations, see Table 3.

The cross-time results, which are reported in Table 12, closely mirrored the within-time results. When selection was for a deficit in PM, bivariate heritability was moderate (statistically significant only for the age 7 PM/age 8 vocabulary analysis) and bivariate shared environmentality was small and nonsignificant. Again, when selection was based on a vocabulary deficit, \( h^2_{g,xy} \) was null but \( c^2_{g,xy} \) was moderate and statistically significant.
Table 12
Results of Bivariate DeFries-Fulker Analyses of PM and Vocabulary Deficits Across Time

<table>
<thead>
<tr>
<th>Selection measure (X)</th>
<th>PM5</th>
<th>PM7</th>
<th>Voc5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-twin measure (Y)</td>
<td>Voc8</td>
<td>Voc8</td>
<td>PM7</td>
</tr>
<tr>
<td>n pairs</td>
<td>149</td>
<td>157</td>
<td>168</td>
</tr>
<tr>
<td>MZ pairs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proband mean (X)</td>
<td>-1.75</td>
<td>-1.83</td>
<td>-1.97</td>
</tr>
<tr>
<td>Co-twin mean (Y)</td>
<td>-0.73</td>
<td>-0.87</td>
<td>-0.63</td>
</tr>
<tr>
<td>DZ pairs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proband mean (X)</td>
<td>-1.69</td>
<td>-1.93</td>
<td>-1.77</td>
</tr>
<tr>
<td>Co-twin mean (Y)</td>
<td>-0.48</td>
<td>-0.53</td>
<td>-0.62</td>
</tr>
<tr>
<td>$h_{xy}^2$ (95% CI)</td>
<td>.28 (0-.66)</td>
<td>.41 (.09-.73)</td>
<td>.00 (0-.39)</td>
</tr>
<tr>
<td>$c_{xy}^2$ (95% CI)</td>
<td>.15 (0-.54)</td>
<td>.07 (0-.41)</td>
<td>.38 (.03-.73)</td>
</tr>
</tbody>
</table>

Notes. Estimates whose confidence intervals do not include 0 (reported in bold) are statistically significant at the p<.05 level. For abbreviations, see Table 3.

The degree of asymmetry in our findings is somewhat surprising, though previous studies have also found differing estimates for bivariate heritability depending on selection measure (Gayan & Olson, 2001). Given our relatively small sample size, the asymmetry may be partly due to sampling error. One theoretically meaningful explanation for the difference is that the etiology of PM and vocabulary difficulties vary by subgroup. For example, there may be at least two routes to a poor vocabulary score: one through genes that also influence PM, and one through a relatively impoverished environment. We ran a set of follow-up analyses to test the hypothesis of differential etiologies for poor PM and vocabulary by subgroup. To maximize sample size and because the pattern of bivariate results was so similar at earlier and later ages, these follow-up analyses were based on the age 5 PM and vocabulary composites only.

First, we confirmed that the subgroups of probands selected for PM and vocabulary deficits at age 5 were largely distinct. Of 214 PM probands, 138 (64.5%) were not vocabulary probands. Similarly, of 195 vocabulary probands, 119 (61.0%) were not PM probands. Thus,
although PM and vocabulary deficits co-occurred at higher than chance levels, as expected ($\chi^2(1) = 164.75, p < .001$), nearly two-thirds of the individuals who had one deficit did not have the other. This degree of non-overlap clearly makes possible that distinct etiologic influences are at play in the different proband groups. To evaluate this possibility, we divided individuals with either an age 5 PM deficit or an age 5 vocabulary deficit into three groups: pure PM deficit (i.e., no vocabulary deficit), pure vocabulary deficit (i.e., no PM deficit), and comorbid PM and vocabulary deficits. If there are different etiologies by subgroups, then the pattern of stronger genetic influence with selection on PM and stronger environmental influence with selection on vocabulary should become more pronounced in the pure deficit subgroups. In contrast, the comorbid subgroup would be expected to show effects of shared environment even when PM is the predictor, and to show effects of genes even when vocabulary is the predictor. These analyses must be considered exploratory, because sample sizes are smaller than ideal for bivariate DF regressions (particularly in the comorbid subgroup). The proband subgroups did not differ in terms of gender, culture, zygosity, or parent level of education (all p-values > .1).

Results, which are reported in Table 13, are only partially consistent with our hypothesis of differential etiology by subgroup. With selection on PM, the pattern of greater genetic than environmental effects grew stronger in the pure subgroup, as predicted. Further, in the comorbid subgroup, both heritability and shared environment effects were moderate to large, though neither was statistically significant and the confidence intervals spanned the full range of mathematical possibility. However, with selection on vocabulary, the effect of shared environment became somewhat weaker in the pure subgroup, and there was no indication of a genetic effect in the comorbid group. Overall, results suggest that the degree of asymmetry by selection measure may arise from both meaningful subgroup differences and sampling error. The implications of these results are addressed further in the discussion section.
Table 13
Results of Bivariate DeFries-Fulker Analyses of PM and Vocabulary Deficits at Age 5 by Subgroup

<table>
<thead>
<tr>
<th></th>
<th>Pure PM deficit</th>
<th>Pure vocabulary deficit</th>
<th>Comorbid deficits</th>
<th>Comorbid deficits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection measure (X)</td>
<td>PM5</td>
<td>Voc5</td>
<td>PM5</td>
<td>Voc5</td>
</tr>
<tr>
<td>Co-twin measure (Y)</td>
<td>Voc5</td>
<td>PM5</td>
<td>Voc5</td>
<td>PM5</td>
</tr>
<tr>
<td>n pairs</td>
<td>137</td>
<td>117</td>
<td>76</td>
<td>74</td>
</tr>
<tr>
<td>MZ pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proband mean (X)</td>
<td>-1.67</td>
<td>-1.81</td>
<td>-1.92</td>
<td>-2.23</td>
</tr>
<tr>
<td>Co-twin mean (Y)</td>
<td>-0.69</td>
<td>-0.52</td>
<td>-1.66</td>
<td>-1.33</td>
</tr>
<tr>
<td>DZ pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proband mean (X)</td>
<td>-1.60</td>
<td>-1.74</td>
<td>-1.80</td>
<td>-1.80</td>
</tr>
<tr>
<td>Co-twin mean (Y)</td>
<td>-0.22</td>
<td>-0.45</td>
<td>-1.18</td>
<td>-1.02</td>
</tr>
<tr>
<td>$h^2_{gxy}$ (95% CI)</td>
<td>.55 (.22-.88)</td>
<td>.05 (0-.41)</td>
<td>.42 (0-1.00)</td>
<td>.06 (0-.47)</td>
</tr>
<tr>
<td>$c^2_{gxy}$ (95% CI)</td>
<td>.00 (0-.21)</td>
<td>.23 (0-.61)</td>
<td>.45 (0-1.00)</td>
<td>.54 (.11-.97)</td>
</tr>
</tbody>
</table>

Notes: Estimates whose confidence intervals do not include 0 (reported in bold) are statistically significant at the p<.05 level. For abbreviations, see Table 3.

Discussion

Study 3 addressed three questions: 1) Do deficits in PM and vocabulary arise from the same types of etiologic influences as PM and vocabulary scores in the full population? 2) Does the phenotypic correlation between PM and vocabulary arise primarily from shared genes, shared environments, or both? 3) How does the etiologic relation of earlier PM to later vocabulary compare to that of earlier vocabulary to later PM?

On the first question, it appears that a similar balance of genes and environment cause weaknesses in PM and vocabulary as cause the full range of individual differences. Our findings echo earlier conclusions from two samples that both constructs are heritable and that shared environment contributes more to individual differences in vocabulary than to PM (Hayiou-Thomas
et al., 2006; Samuelsson et al., 2005). Also, consistent with findings from the full ILTS, there was some suggestion that genetic influence on vocabulary grew stronger with age (Byrne et al., 2009), though our confidence intervals for estimates at the two time points were large and overlapping. One notable difference between our results and previous results from the ILTS is that we found slightly higher estimates for genetic contributions to language deficits than have been found across the range of individual differences in language skill. This discrepancy was particularly notable for age 7 PM, though confidence intervals from the current and previous studies were overlapping. Further, the previous study estimated a relatively strong influence of nonshared environment on age 7 PM, suggesting that the lower h² estimate may have owed partly to limited reliability.

The answer to the second question was not straightforward. There was clear evidence for a genetic basis to the PM-vocabulary relationship among a subgroup of individuals. Genes that influenced membership in the PM proband group also impacted co-twins’ vocabulary scores. However, genes that influenced membership in the vocabulary proband group did not carry over to co-twins’ PM scores. Instead, shared environmental experiences explained the phenotypic correlation in this case. A follow-up subgroup analysis strengthened our confidence in the conclusions for the PM proband group, but not the vocabulary proband group. There may be a methodological limitation making it difficult to detect bivariate heritability when vocabulary is the selection measure. If such a limitation is present, it would also impact our ability to answer the third question. However, current results suggest that the influence of PM on vocabulary across time is genetically driven, while the influence of vocabulary on PM across time is due to shared environment.

One interpretation of the full pattern of results is that particular genetic risk factors cause a PM deficit, which over time impairs vocabulary acquisition. Poor vocabulary can also arise from an impoverished environment (or from a combination of genetic and environmental risk factors); this vocabulary deficit may then impair performance on PM tasks. Overall, results from Study 3 agree with Studies 1 and 2 in indicating that the PM-vocabulary relationship is bidirectional, but evidence for the bottom-up effect is stronger in individuals with disordered language development.
It appears that the bottom-up effect owes primarily to genes and the top-down effect primarily to environmental experiences. Future studies with a larger sample size are needed to confirm this interesting result.

An important caveat is that our measures of PM and vocabulary were relatively coarse. With finer-grained measures, we might find that genes predominantly influence particular aspects of PM and vocabulary and that environment predominantly influences other aspects. For example, a recent study found that verbal short-term memory skill predicted ability in one type of word learning task but not another (Jarrold, Thorn, & Stephens, 2009). These researchers reported that learning to distinguish a novel phonological form from close distractors depended heavily on verbal short-term memory, while establishing a mapping between a novel phonological form and a referent did not. It is conceivable that the former process is more subject to genetic influence while the latter relates predominantly to environment (e.g., amount of language input). A future study could use more specific measures within a genetically-sensitive design to investigate this possibility.
General Discussion

The purpose of the current research was to better understand PM-vocabulary and PM-syntax relations in school-age children with and without language disorders. In three studies, we examined the nature of these relationships within the context of two competing theoretical viewpoints: the phonological storage framework, which emphasizes bottom-up influences, and the LRM, which emphasizes top-down factors. Across Studies 1 and 2, several results indicated that methodological issues limited the conclusions that could be drawn from the syntax analyses. Thus, Study 3 included vocabulary as the only broad language construct, and this General Discussion focuses on the meaning of the PM-vocabulary analyses across the three studies.

The current work yielded several novel findings. First, consistent with the predictions of the phonological storage framework, we found clear evidence that PM skill influences vocabulary learning up to age 8. In two of three population samples included in Study 1 and in all three groups with language disorders included in Study 2, the final models of PM-vocabulary relations included a statistically significant path from earlier PM to age 8 vocabulary. The effect size was in the medium range. Previous research has not found a clear effect of PM on vocabulary learning beyond age 5, and there have been some null results for the age range included in the current research (Gathercole et al., 2005; Gathercole et al., 1992). It is likely that a combination of our larger sample size and more sophisticated analytic approach allowed us to detect this moderate effect when earlier studies have not. Results from Study 3 indicated that the bottom-up effect is under partial genetic influence, at least for children with poor phonological development. We found evidence for shared genes operating on earlier PM and later vocabulary among children selected for a PM deficit. Thus, it appears that a weak vocabulary is sometimes caused by a developmental cascade in which risk genes cause an initial PM impairment, which in turn slows
new word learning. Univariate behavior genetics results from the current study and others also indicate that a weak vocabulary can also arise from impoverished environmental input.

Another important set of findings concerned the developmental trajectory of the bottom-up effect. Proponents of the phonological storage framework have argued that the role of PM in new word learning is more important earlier in language development (Gathercole, 2006; Gathercole et al., 2008), with evidence coming from typically developing children, individuals with developmental delay, and second language learners. Our research with cross-cultural and language-disordered populations provided convergent evidence for this claim. In Study 1, the final models of PM-vocabulary relations included a bottom-up effect in the two cultures where mean level of performance on language tasks was relatively weaker (the U.S. and Scandinavia), but not in the culture where mean level of performance was relatively stronger (Australia). Furthermore, results from Study 2 suggested that the magnitude of the bottom-up effect might correlate with degree of language delay. We suggested two (not mutually exclusive) reasons that the bottom-up effect would wane with language development: threshold effects for PM, and a change in word learning strategies (e.g., increased reliance on context). Our results do not directly address which process is primary, in part because the bottom-up effect was still evident in the U.S. population sample at the most recent time point available. However, we will soon have data on PM-vocabulary relations to age 10 in the current samples, which should be informative.

In addition to substantial evidence consistent with the phonological storage framework, we also obtained some evidence for a top-down effect of vocabulary on PM, consistent with the predictions of the LRM. In two of three population samples included in Study 1 and in one of the three language disordered groups included in Study 2, the final model of PM-vocabulary relations included a statistically significant path from age 5 vocabulary to later PM. We proposed that the process of learning to read leads to the emergence of this top-down effect, since it was absent in the least literate groups included in the current research (the Scandinavian early school-age population, the U.S. LTS RD group, and the DU SSD group). However, broad oral language and literacy skill were partly confounded in both Studies 1 and 2, so this proposal awaits confirmation from a study with a different design (e.g., a school cutoff study). Such a study should use
multiple measures of PM to test whether reading skill relates differently to nonword repetition, digit span, and sentence memory tasks. Study 3 suggested that the top-down effect is partly influenced by shared environment, at least for children with vocabulary deficits.

In addition to these theoretical implications, the current research also has some interesting clinical implications. Disorders of broad language development have been associated with poor educational, occupational, and social-emotional outcomes (Hall & Tomblin, 1978; Stothard, Snowling, Bishop, Chipchase, & Kaplan, 1998; Westby & Blalock, 2005), but many children with poor language development do not attract clinical attention until they have already experienced several years of school failure (Bishop & Leonard, 2000; Pennington, 2008). Our results may help support earlier identification and treatment of individuals at risk for such difficulties. For example, since PM skill continues to influence word learning across the early school years, a brief screen of PM skill at school entry could be very informative.

The current research has several limitations that should be addressed by future work. First, it will be important to confirm our key results in a study that has multiple measures of PM and broad language skill at every time point, since the ILTS had only one measure of each construct available beyond age 5, and not all at the same age. Such a follow-up study would also more clearly address the relation of PM to syntax. Second, because the current research drew primarily from population samples, our sample size was more limited in the investigations of language deficits (Studies 2 and 3). Sample size concerns are particularly relevant to Study 3. This study yielded the interesting novel result that the bottom-up influence of a PM deficit on vocabulary is more genetically influenced, while the top-down influence of a vocabulary deficit on PM is more environmentally influenced. However, given the relatively small sizes of our groups, this result may largely reflect error variance; thus, it awaits replication in a larger sample overselected for poor language skill.

In addition to these relatively straightforward methodological limitations, the current work also faces a more general interpretive issue common to most of the research on PM and language development. There is disagreement about whether the relationship of PM to vocabulary and syntax is specific to short-term memory, or whether it reflects the quality of
phonological representations. Proponents of the phonological storage framework emphasize the importance of memory processes (Baddeley et al., 1998; Gathercole, 2006) while top-down accounts assume that vocabulary influences general phonological development, which in turn determines performance on PM tasks (Bowey, 1996; Metsala & Walley, 1998; Snowling et al., 1991). Previous researchers have attempted to differentiate PM from phonological representations with the inclusion of phonological awareness (PA) tasks, which require children to attend to and manipulate sounds in words (Avons et al., 1998; Bowey, 2001; Gathercole et al., 2008; Gathercole, Willis, & Baddeley, 1991; Metsala, 1999). The logic of this approach is that if the PM-broad language association reflects general phonological development, then both kinds of phonological tasks should show similar relations to vocabulary (or syntax). In contrast, if the association is specific to PM, PA should show a different pattern of relations. The problem with this logic is that PM and PA are themselves highly correlated (e.g., Hayiou-Thomas et al., 2006), and most PA tasks place significant demands on working memory. Given this limitation, it is unsurprising that results from these earlier studies have been mixed, and that no clear conclusion has emerged. We propose that a future study compare PM not to PA but to measures designed to tap implicit phonological representations—such as priming and lexical gating tasks (see, for example Boada & Pennington, 2006). Such an approach holds greater promise for understanding the differential relation of quality of phonological representations versus ability to store phonological forms in working memory to broad oral language skill.
References


70


