Published in final edited form as:

Sci Stud Read. 2014 ; 18(5): 347–362. doi:10.1080/10888438.2014.904870.

Longitudinal Stability of Phonological and Surface Subtypes of Developmental Dyslexia

Robin L. Peterson¹, Bruce F. Pennington², Richard K. Olson^{3,4}, and Sally Wadsworth³

¹Children's Hospital Colorado

²University of Denver

³University of Colorado Boulder

⁴Linköping University

Abstract

Limited evidence supports the external validity of the distinction between developmental phonological and surface dyslexia. We previously identified children age 8 to 13 meeting criteria for these subtypes (Peterson, Pennington, & Olson, 2013), and now report on their reading and related skills approximately 5 years later. Longitudinal stability of subtype membership was fair and appeared stronger for phonological than surface dyslexia. Phonological dyslexia was associated with a pronounced phonological awareness deficit, but subgroups otherwise had similar cognitive profiles. Subtype did not inform prognosis. Results provide modest evidence for the validity of the distinction, although not for its clinical utility.

Introduction

Children with developmental dyslexia can be subtyped according to various criteria (Castles & Coltheart, 1993; Morris, Stuebing, et al., 1998; Bosse, Tainturier, & Valdois, 2007; Bednarek, Seldana, & Garcia, 2009; Wolf & Bowers, 1999). Of any subtyping scheme, the distinction between developmental phonological and developmental surface dyslexia has arguably garnered the most empirical support. Children with phonological dyslexia are very poor at reading pseudowords, but have relatively spared exception word reading, while children with surface dyslexia show the opposite pattern. Numerous studies across languages have documented that substantial percentages of children with dyslexia meet criteria for either phonological or surface dyslexia (Castles & Coltheart, 1993; Jimenez, Rodriguez, & Ramirez, 2009; Manis, Seidenberg, Doi, & McBride-Chang, 1996; Olson, Kliegl, Davidson, & Foltz, 1985; Sprenger-Charolles, Cole, Lacert, & Serniclaes, 2000; Sprenger-Charolles, Siegel, Jimenez, & Ziegler, 2011; Stanovich, Siegel, & Gottardo, 1997; Ziegler et al., 2008). However, very little research has investigated whether children identified as belonging to one subtype or another continue to exhibit that pattern over time. The purpose of the present investigation was to test the longitudinal stability of developmental phonological and surface dyslexia over a 5-year period.

The phonological and surface subtypes have been interpreted in light of two competing models of single word reading: the dual-route model (Ellis & Young, 1988) and the triangle model (Seidenberg & McClelland, 1989). The dual-route cascaded model (DRC; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) is a fully specified computational instantiation of the dual-route framework, which was originally developed largely to account for the behavioral double dissociation of *acquired* phonological versus acquired surface dyslexia observed among previously skilled adult readers who had sustained brain damage. The DRC includes two procedures for reading words aloud. In the direct (lexical) route, orthographic input selects an entry in the orthographic lexicon via interactive activation, which in turn activates the appropriate phonological output. The indirect, or nonlexical, route, takes orthographic input, parses it into graphemes, converts the graphemes into their corresponding phonemes via a set of explicit rules, and then assembles these phonemes into a word for output. Both routes are invoked in parallel in response to a word stimulus, and both routes contribute to successful reading of regular words. However, only the lexical route can successfully read exception words, since these words break rules of grapheme-phoneme correspondences. Only the nonlexical route can read pseudowords, since these have not been encountered before and are not in the orthographic lexicon. According to the DRC, both acquired and developmental phonological dyslexia arise from differential damage to the nonlexical route, while both acquired and developmental surface dyslexia arise from differential damage to the lexical route.

Harm and Seidenberg (1999) developed an implementation of the triangle model that offered an alternate account of developmental dyslexia subtypes. While the DRC is a static model originally designed to account for skilled adult reading, Harm and Seidenberg's model (henceforth, the HS model) learned to establish mappings between orthographic inputs and a phonological output network through training. The HS model read regular words, exception words, and pseudowords via a single procedure. No explicit rules were specified, but the model demonstrated the ability to extract regularities in grapheme-phoneme mappings by reading pseudowords reasonably successfully. Harm and Seidenberg simulated developmental phonological dyslexia with various types of damage to the phonological network, consistent with the prevailing view that most cases of developmental dyslexia are caused by impairments in phonological representations. Based on empirical findings that children with surface dyslexia perform similarly to younger, typically developing readers on a variety of reading-related tasks (Manis et al., 1996; = Sprenger-Charolles et al., 2000; Sprenger-Charolles et al., 2011; Stanovich et al., 1997), surface dyslexia was conceptualized as "reading delay dyslexia." The HS model successfully simulated surface dyslexia by slowing overall reading acquisition in four different ways—providing less training, reducing the learning rate, degrading the orthographic input, and removing some of the hidden units. Each of the manipulations causing either dyslexic profile (phonological damage and reading delays) tended to impact *both* pseudoword and exception word reading, but one was impacted relatively more than the other. This result agrees with the well-documented finding that most children with dyslexia demonstrate a mixed profile, though they may be differentially impaired with one type of stimulus.

A later version of the HS model implemented a semantic network (Harm & Seidenberg, 2004) and provided two routes from print to naming: one from orthography to phonology, and one from orthography through semantics to phonology. The model differed from the DRC in that the division of labor between the two routes was not “clean”; instead, both routes contributed to successful reading of regular words, exception words, and pseudowords. This model was not used to simulate developmental dyslexia subtypes.

To summarize, the dual-route and triangle frameworks offer competing explanations for the phonological and surface subtypes of developmental dyslexia. These accounts differ in how discrete they assume the underlying deficits to be, as well as in whether they posit that the proximal causes of poor reading are fully encapsulated within the reading system. Despite their disagreements, the models agree that the subtype distinction is valid. Surprisingly little evidence supports this claim. Perhaps the strongest support comes from a comparison of the subgroups on reading-related tasks: children with phonological dyslexia are differentially impaired at phoneme awareness, while children with surface dyslexia are differentially impaired at tasks emphasizing orthographic knowledge (Jimenez et al., 2009; Manis et al., 1996; Manis et al., 1999; Stanovich et al., 1997). However, support for distinct etiologies, brain bases, prognoses, or treatment responses is quite limited (e.g., Castles, Datta, Gayan, & Olson, 1999; Gustafson, Ferreira, & Ronnberg, 2007), and on balance, the evidence indicates that the subtypes represent two ends of a continuum rather than discrete categories (Castles et al., 1999; Griffiths & Snowling, 2002; Olson et al., 1985). Clearly, if the distinction is valid, the subtypes should show reasonable longitudinal stability. Both models could account for some proportion of children switching categories over time, due to changes in more distal factors influencing reading acquisition (e.g., increased exposure to print, specialized remediation, etc.). Indeed, there is some evidence that intervention and type of instruction would influence subtype membership (Olson, 2011). A finding of very low longitudinal stability, however, would suggest that the distinction is not particularly meaningful.

Manis and colleagues have conducted the only longitudinal studies of phonological and surface dyslexia of which we are aware (Manis & Bailey, 2008; Manis et al., 1999). These researchers identified children with phonological or surface dyslexia using a z-score subtraction method in 3rd grade and followed them for two years. Phonological dyslexia was very stable over time, and also became increasingly common with development. In contrast, surface dyslexia showed little stability from one year to the next. At each age, subtype membership was associated with a characteristic pattern on reading-related tasks. The authors thus argued that although surface dyslexia was short-lived, it was nonetheless valid, and suggested that environmental factors (e.g., exposure to print, tutoring) that varied from year to year likely explained the instability in this group. We hope to extend the findings of Manis and colleagues by examining a much broader age range over a longer time frame, which puts us in a better position to detect developmental effects. In addition, as in our initial subtyping study, we identify both “pure” and “relative” cases (sometimes also referred to as “classical” and “regression” or “hard” and “soft” cases) using the regression outlier method, which is the most widely used approach in the literature.

Overview of the Current Investigation

Our current focus is on the longitudinal stability of imbalanced reading profiles in individual children. A complementary question concerns the stability of reading-related constructs across the full range of individual differences. Previous research in the current sample used a structural equation modeling approach to demonstrate that individual differences in single word reading, phonological decoding, and orthographic skills were extremely stable over five years, as were their relations to one another (Hulstlander, Olson, Willcutt, & Wadsworth, 2010). However, it does not necessarily follow that patterns within individuals failing at one end of the distribution remain constant with development (Pennington et al., 2012). Understanding what happens to these individuals over time is important both because it will constrain theoretical models of reading development and difficulties, and also because this information is clinically relevant and may help inform accurate identification, prognosis, and appropriate treatment.

In our earlier subtyping study (Peterson et al., 2013) in the Colorado Learning Disabilities Research Center (CLDRC; DeFries et al., 1997) we identified children aged 8 to 13 meeting criteria for phonological or surface dyslexia. Phonological coding (PC; equivalent to nonlexical route function) and orthographic coding (OC; equivalent to lexical route function) were each assessed with several measures. Previous studies have most often assessed PC with nonword reading and OC with exception word reading. We assessed PC with a nonword reading test and a phonological choice test (described further below). An exception word reading measure was not available for the majority of the sample. Instead, we assessed OC with orthographic and homonym choice measures (again, described further below). A subsample of participants had completed an exception word measure, which allowed us to validate the use of our less traditional subtype defining measures for OC. There are undoubtedly strengths and weaknesses to our use of orthographic/homonym choice in place of exception word reading. On the one hand, neither of these tasks require explicit articulation of a phonological code and so might be less likely to tap into the phonological processes impaired in phonological dyslexia, thus leading to a stronger chance of discovering dissociated profiles. On the other hand, since both are forced-choice tasks, they may more heavily emphasize skills such as inhibition and sustained attention that are not the focus of the present study.

In our initial subtyping study, the regression outlier procedure was used to identify children with “pure” phonological dyslexia (PC below age expectations and OC within normal limits), “pure” surface dyslexia (OC below age expectations and PC within normal limits), relative phonological dyslexia (deficits in both processes, and PC worse than expected based on OC), and relative surface dyslexia (deficits in both processes, and OC worse than expected based on PC). A subset of these children returned for further testing an average of 5 years later, as part of the Colorado Longitudinal Twin Study of Reading Disability (LTSRD; Wadsworth, DeFries, Olson, & Willcutt, 2007). In the present investigation, we applied the same subtyping criteria to these individuals to evaluate the stability of word reading profiles over this time period and thus inform the validity of the distinction. We addressed the following specific questions:

1. What is the longitudinal stability of the phonological and surface subtypes of developmental dyslexia from middle childhood into adolescence?
2. Does subtype membership at initial assessment affect literacy prognosis?
3. Does the pattern of performance on neuropsychological tasks both within and across time support the validity of the subtype distinction?

Method

Participants

Participants included 72 participants with dyslexia and 65 controls from the CLDRC who were included in our earlier subtyping study (Peterson et al., 2013) and who returned for follow-up as part of the LTSRD (Wadsworth et al., 2007). Participants were originally tested between the ages of 8 and 13.5 ($M = 10.01$, $SD = 1.50$) and were retested an average of 5.32 years later ($SD = 0.82$ years) when they were between 12 and 22 years of age ($M = 15.33$, $SD = 1.78$). For participation in the CLDRC, twin pairs were ascertained from school records from 27 Colorado school districts. Participants who were initially tested in the CLDRC between 1996 and 2003 were invited for follow-up testing. Fifty-five percent of families contacted agreed to participate in follow-up testing. Criteria for dyslexia in the original study included a school history of dyslexia plus psychometric evidence of dyslexia based on a discriminant function analysis (DeFries, 1985). Controls had no history of reading problems and did not meet psychometric criteria for dyslexia. When both members of a twin pair met criteria for either the dyslexic or control group, one was chosen at random for the initial study. Sample characteristics are provided in Table 1.

Data at follow-up were available for 16.5% of participants with dyslexia included in the original subtyping study as well as for 20.4% of control participants. Control participants who were retained were similar to control participants who were not retained in terms of age, gender, race (defined as Caucasian or other), mother years of education, reading skill, and Full Scale IQ at initial assessment. Participants with dyslexia who were retained were similar to those who were not retained in terms of gender, race, and Full Scale IQ at initial assessment. However, retained participants with dyslexia were slightly younger at initial assessment (retained: $M = 9.85$ years, $SD = 1.58$; excluded: $M = 10.32$ years, $SD = 1.58$; $t(435) = 2.30$, $p = .02$; had mothers with slightly more education (retained: $M = 14.51$, $SD = 2.50$; excluded: $M = 13.87$, $SD = 2.21$; $t(430) = 2.30$, $p = .02$), and had slightly better reading (as measured by a word recognition composite variable) at initial assessment (retained: $M = -2.16$, $SD = .68$; excluded $M = -2.41$, $SD = .83$; $t(435) = 2.29$, $p = .02$) in comparison to participants with dyslexia who were not retained. Because the equations for subtype membership are all based on the performance of the control group, we do not expect these small differences in the dyslexic group to make a substantial impact on the pattern of results.

Measures

At both time points, all participants were administered an extensive battery of reading and reading-related tasks at the University of Colorado, and the subset of measures described

below was used in the current analyses. All variables were examined for departures for normality, and were transformed if necessary as detailed below. Age-adjusted z-scores standardized against control participants were calculated, and composites were created for each construct by averaging relevant z-scores.

Literacy—Single-word reading was assessed with the Peabody Individual Achievement Test Reading Recognition subtest at initial assessment (PIAT Rec; Dunn & Markwardt, 1970) and with the updated version of this subtest (PIAT-R Rec; Markwardt, 1989) at follow-up. The published correlation between the two tasks is .88. Published test-retest reliabilities are .89 for the PIAT and .96 for the PIAT-R. In addition, the time limited word recognition test (TLWRT; Olson, Forsberg, Wise, & Rack, 1994) was administered as a second measure of single word reading at both time points. Test-retest reliability is .93.

The Wide Range Achievement Test-Revised Spelling subtest (WRAT-R Spelling; Jastak & Wilkinson, 1984) provided a measure of spelling production at initial assessment, and the updated version of this subtest provided a measure of spelling production at follow-up (WRAT-III Spelling; Wilkinson, 1993). The published alternate form reliability is .90.

Reading comprehension was assessed with PIAT Reading Comprehension (PIAT Comp) at initial assessment and the updated version (PIAT-R) at follow-up. The published correlation between the tasks is .79. The published test-retest reliabilities are .64 for the PIAT and .88 for the PIAT-R.

Subtyping measures—At both time points, phonological coding (PC) was assessed with two tasks. First, a phonological choice task (Olson et al., 1994) included 60 items requiring participants to select which of three pseudowords would sound like a real word (e.g., *beal/bair/rabe*). At retest, data were squared to reduce the effects of skew of -1.04. The second task was a nonword reading task including 85 items of varying difficulty in two blocks, with percentage correct scores calculated for each block. At retest, data were squared to reduce the effects of negative skew (block 1: -1.43; block 2: -1.19). A conservative estimate of reliability for a composite of the PC tasks was estimated by examining monozygotic (MZ) twin correlations. MZ correlations were .83 at initial assessment and .82 at follow-up. The correlation between the PC composite at initial assessment and follow-up was .85.

At initial assessment, orthographic coding (OC) was assessed with two tasks. First, an orthographic choice task (Olson et al., 1994) included 80 real word/pseudohomophone pairs (e.g., *easy-eazy*, *salmon-sammon*) presented in two blocks, with percentage correct scores calculated for each block. Second, a homophone choice task (Olson, Forsberg, & Wise, 1994) required participants to select which of two homophones presented on the computer screen answered a question asked orally by the computer (“Which is a flower?” *rose/rows*). There were 65 items. At follow-up, OC was assessed with the orthographic choice task only, as the homophone choice task was not included in the LTSRD battery. At retest, scores for block 1 were squared to reduce the effects of a negative skew of -1.94. The MZ twin correlation for the OC composite at initial assessment was .71. The MZ twin correlation for the OC composite at follow-up (based on both blocks of orthographic choice) was .66, which falls in the “questionable” range. Reliability appeared to be reduced by ceiling effects

on the first block of orthographic choice. The MZ correlation for block 2 only was .70, which falls in the “acceptable” range. Furthermore, the correlation between the OC composite at initial assessment with the OC composite at follow-up was .56, but this improved to .63 if only block 2 was included at follow-up. Thus, only block 2 data were used in further analyses.

General Intelligence—Different versions of the Wechsler Intelligence Scale were used to assess general intelligence. At initial testing, participants completed the WISC-R (Wechsler, 1974). At follow-up, participants completed the WISC-3 or WAIS-3 (Wechsler, 1991, 1993) as appropriate for their age. The published correlation for the WISC-R and WISC-3 is .89. Published test-retest reliabilities for all versions of full-scale IQ are .96 to .97.

Reading-related skills—At both time points, participants completed two measures of phonological awareness (PA): pig Latin (Olson, Wise, Connors, Rack, & Fulker, 1989) and phoneme deletion (Olson et al., 1994). The pig Latin task required participants to take the first sound from the front of a word, put it at the end, and add the sound /ay/. For example, “boat” would become “oat-bay”. The child received 5 demonstrated examples, 9 practice trials with feedback, and 45 experimental trials with no feedback. Cronbach's alpha is .93. At retest, data were squared to reduce the effects of a skew of -1.52. The phoneme deletion task consisted of 6 practice and 40 test trials, and required subjects to repeat a nonword, then remove a specific phoneme (when done correctly, a real word resulted—e.g., “Say ‘prot.’ Now say ‘prot’ without the ‘/r/’.”) A conservative estimate of reliability estimated from the correlation with a composite phoneme deletion measure is .78. At retest, data were squared to reduce the effects of a skew of -1.01.

At both time points, participants completed four measures of Rapid Automatized Naming (RAN; Decker, 1989, after Denckla & Rudel, 1976). Participants were presented with a series of stimuli (letters, numbers, colored circles, or pictures of objects) and then asked to identify the items orally as quickly as possible. Raw scores were equivalent to the number of items named on each card in 15 seconds. A conservative estimate of reliability estimated from the correlation with the Denckla and Rudel test is .85.

Analyses

Subtype identification—The procedure for identifying subtypes was identical to our initial study (Peterson et al., 2013). Following Castles and Coltheart (1993), we applied the regression outlier procedure. First, we identified “pure” subtypes, whose performance was poorer than expected for age on one of the two subtype dimensions, but within normal limits on the second dimension. Pure phonological dyslexia was defined as a deficit in PC (-1.5 SD below the control group mean for age) and an age-corrected OC score within one standard deviation of the control group mean (i.e., -1 SD). Similarly, pure surface dyslexia was defined as a deficit in age-corrected OC (-1.5 SD) and an age-corrected PC score within one SD of the control group mean. Individuals who did not have a deficit in either PC or OC (i.e., both > -1.5 SD) were included in a “mild” subgroup based on the assumption that relatively spared PC and OC skills would relate to less pronounced real-world reading difficulty.

Individuals meeting a relative phonological, relative surface, or mixed dyslexia pattern were then identified among the remaining dyslexic participants. Following previous research, we used measures of PC and OC that were not age adjusted. First, PC was regressed on OC for control subjects, and the resulting regression equation was used to compute a standardized residual for all participants in the study. Low scores indicated poorer PC than expected given an individual's own OC abilities. Individuals with dyslexia who had scores of -1.5 or less were identified as having relative phonological dyslexia. Similarly, OC was regressed on PC for control subjects, and the equation used to calculate a standardized residual that indexed OC ability relative to PC ability. Participants with dyslexia who had scores of -1.5 or less were identified as having relative surface dyslexia. A mixed dyslexic subgroup was also identified in two ways. First there were individuals with dyslexia who did not meet criteria for any of the other dyslexic subtypes. Their age-corrected performance was at least 1 SD below the control group mean for both OC and PC, as well as at least 1.5 SD below the control group mean on at least one of the variables. Further, they did not meet regression-outlier criteria for either relative phonological or surface dyslexia, indicating they were not differentially impaired at either process. In addition, some individuals met criteria for *both* relative phonological and surface dyslexia and were included in the mixed subgroup. Though this seems paradoxical, it can occur when performance on both word types is extremely poor, since the predicted scores regress to the mean.

To maximize sample sizes, primary analyses in the current study divided the dyslexia sample into three subgroups: phonological dyslexia (combination of pure and relative), surface dyslexia (combination of pure and relative), and balanced dyslexia (combination of mild and mixed). When relevant, follow-up analyses based only on “pure” subtypes are reported.

Subtype stability—Stability of subtypes over time was estimated using kappa. This statistic is typically used as a measure of inter-rater reliability; in this case, time was considered the rater. First, we computed the stability of psychometric dyslexia itself, to provide an upper limit for expected stability for subtypes. For that analysis, we defined dyslexia based on the composite of the two single word reading measures falling at least 1.5 SD below age expectation based on the control group. Next, we computed kappa for the subtypes as a whole. We also considered kappa for those with phonological dyslexia compared to all others with dyslexia, and for those with surface dyslexia compared to all others with dyslexia.

Neuropsychological profiles—We tested whether the validity of the subtype distinction was supported by differing neuropsychological profiles on the variables of verbal IQ, performance IQ, PA, and RAN. At each time point, we used repeated-measures ANOVA to test for a construct by subtype interaction, and we qualitatively compared results across time.

Prognosis—We evaluated whether initial assessment subtype informed prognosis by comparing raw score growth in single word reading, spelling production, and reading comprehension across time. For each literacy variable, we tested for a time by subtype interaction in a repeated-measures ANOVA model.

Results

Four individuals who had had dyslexia at initial assessment showed good recovery of their reading problems at follow-up (single word recognition composite within .55 SD of control group mean). These individuals were excluded from further analyses of dyslexia subtypes, though they were of course retained in the analysis of the stability of psychometric dyslexia. Three of the four had been classified as being in the mild subgroup at initial assessment, and one had been in the pure phonological group at initial assessment. Table 2 shows the subtype breakdown for the 68 remaining participants with dyslexia.

Subtype Stability

The longitudinal stability of psychometric dyslexia was high ($\kappa = 0.83$, 95% CI = 0.73-0.92).

Table 3 shows the breakdown of the dyslexic sample according to their primary subtype classification at both time points. Overall longitudinal stability was only fair, and significantly weaker than the stability of psychometric dyslexia ($\kappa = 0.34$, 95% CI = 0.16-0.53). Comparing those with phonological dyslexia to all others with dyslexia improved stability to the moderate range ($\kappa = 0.49$, 95% CI = 0.27-0.70). In contrast, comparing those with surface dyslexia to all others with dyslexia revealed only fair stability ($\kappa = 0.26$, 95% CI = 0-0.58), though the confidence intervals for the phonological and surface subtype analyses were large and overlapping. In a follow-up analysis, we found that the stability of subtype classification for “pure” subtypes (i.e., pure phonological, pure surface, or mixed) over time was low ($\kappa = 0.18$, 95% CI = 0-0.43).

Neuropsychological profiles

Figures 1 and 2 show the pattern of performance for VIQ, PIA, PA and RA for the phonological, surface, and balanced dyslexia subtypes at times 1 and 2, respectively. Scores are reported as z-scores relative to the performance of controls. At initial assessment, there were main effects of both construct ($F(2.5, 169.2) = 10.91$, $p < .001$, $\eta^2_p = .14$) and subtype ($F(2, 67) = 7.28$, $p = .001$, $\eta^2_p = .18$), but these were modified by a significant subtype by construct interaction ($F(5.1, 169.2) = 10.34$, $p < .001$, $\eta^2_p = .24$). To better understand this interaction, we conducted follow-up univariate ANOVAs for each construct with between-subjects factor of subtype. There were no subgroup differences for VIQ, PIQ, or RAN. However, the three subtypes were not equivalent for PA ($F(2, 67) = 20.63$, $p < .001$, $\eta^2_p = 0.59$). Tukey post-hoc testing revealed that the phonological subgroup had poorer PA than either the surface dyslexia or balanced dyslexia subgroups, who did not differ from one another. The magnitude of the subgroup differences was large (phonological vs. surface: $d = 1.93$; phonological vs. balanced: $d = 1.48$). A very similar pattern was evident at follow-up. There was no main effect of subgroup ($p = .13$) but there was again a main effect of construct ($F(2.6, 133.3) = 13.01$, $p < .001$, $\eta^2_p = .21$), which was modified by a significant subtype by construct interaction ($F(5.2, 133.3) = 4.39$, $p = .001$, $\eta^2_p = .15$). Follow-up ANOVAs again revealed no subgroup differences for VIQ, PIQ, or RAN, but differences on PA were significant ($F(2, 65) = 10.03$, $p < .001$, $\eta^2_p = .24$). Tukey post-hoc tests revealed that the phonological subgroup had significantly poorer PA than the balanced subgroup (p

< .001, $d = 1.24$) and marginally poorer PA than the surface subgroup ($p = .06$, $d = .86$). The balanced and surface subgroups did not differ from one another ($p = .3$).

Prognosis

We conducted 4 repeated-measures ANOVA analyses to explore whether initial assessment subtype informed literacy prognosis. In each case, we analyzed raw scores on one of the literacy variables (PIAT Rec, WRAT Spelling, PIAT Comp, or TLWRT) with a within-subjects factor of time and between subjects factor of subtype (phonological, surface, or balanced) (see Table 4). In every case, there was, unsurprisingly, a large and significant main effect of time ($\eta^2_p = 0.85 - 0.94$) but no subtype by time interaction, thus providing no evidence that subtype membership was relevant to prognosis. A main effect of subtype was significant only for PIAT Rec ($F(2, 65) = 3.84$, $p = .027$, $\eta^2_p = .11$). Follow-up Tukey tests revealed that individuals with phonological dyslexia read fewer words on the PIAT than individuals with balanced dyslexia ($p = .03$).

Discussion

Following in the footsteps of adult cognitive neuropsychology, researchers have long sought to identify valid subtypes among children who struggle to learn to read (Coltheart, Masterson, Byng, Prior, & Riddoch, 1983; Olson et al., 1985; Temple & Marshall, 1983). Although a core phonological deficit theory was the predominant explanation of dyslexia for many years (Vellutino, Fletcher, Snowling, & Scanlon, 2004), mounting evidence suggests that a single phonological deficit cannot fully explain the dyslexia phenotype (Bishop, McDonald, Bird, & Hayiou-Thomas, 2009; Peterson, Pennington, Shriberg, & Boada, 2009; Snowling, Gallagher, & Frith, 2003). Some scientists remain convinced that the disorder is heterogeneous and composed of distinct subgroups (e.g., Friedmann & Lukov, 2008; Heim et al., 2008), although we think the data better support a dimensional view (McArthur et al., 2013; Olson et al., 1985; Pennington et al., 2012). A variety of subtyping schemes have been explored historically (Morris et al., 1998; Ramus, 2003; Stanovich, 1988; Wolf & Bowers, 1999) with no clear consensus emerging in the literature. The phonological/surface dyslexia distinction is among the most widely studied and has garnered support from both computational models and cross-linguistic studies. However, very limited evidence supports the external validity of the distinction. The present study evaluated the longitudinal stability of phonological and surface subtypes of dyslexia over 5 years, and also asked whether subtype related to neuropsychological profile or prognosis. While we found some modest support for the validity of the distinction, our results raise doubt that subtype membership is particularly meaningful.

We found that the overall longitudinal stability of the subtypes was only fair, and significantly weaker than the stability of dyslexia itself. Despite some previous evidence that identification of pure subtypes might be more reliable than identification of relative subtypes (Sprenger-Charolles et al., 2011), overall stability of “pure” subtypes only was poor. The pattern of results suggested that phonological dyslexia was likely more stable than surface dyslexia, though confidence intervals for the kappa estimates were large and overlapping. These results are broadly similar to those of Manis and colleagues, who conducted the only

previously published studies on the longitudinal stability of the phonological and surface dyslexia subtypes (Manis & Bailey, 2008; Manis et al., 1999). Our approach differed from that of Manis and colleagues in several ways. Our sample had a wider age range, and the follow-up interval was much longer. We also used different measures of PC and OC, and had a different method for identifying subtypes (regression outlier versus z-score subtraction). The general similarity in findings despite the substantial methodological differences strengthens our confidence that the pattern is real.

Why is the phonological profile more stable than the surface profile? This finding might initially appear to relate to differences in measurement reliability for PC and OC skills. Reliability was stronger for the PC composite than the OC composite, and the correlation of PC skills over time was greater than for that of OC skills (in part because one of the initial assessment OC measures was not readministered at follow-up). However, *all* subtype equations depend on *both* PC and OC, so differing reliabilities cannot explain the present results or those of Manis and colleagues. Those researchers suggested that children tended to move in and out of the surface category on the basis of environmental factors like exposure to print or instructional type (Manis & Bailey, 2008). This is an intriguing possibility, but to date, it lacks empirical support. Indeed, at initial assessment in this sample, we found no evidence for a link between the surface dyslexia profile and exposure to print (Peterson et al., 2013). An alternate possibility is that performance on OC tasks might be more influenced by factors external to the reading system and more likely to fluctuate over time (e.g., fatigue, attention). All the OC tasks we included (but not all the PC tasks) used a multiple-choice format, which could be more susceptible to such influences. On the other hand, Manis and colleagues included OC and PC tasks that were more similar in response format to one another (i.e., nonword reading and exception word reading).

If the subtype distinction is valid, then the subtypes should be associated with differing neurocognitive profiles. Several previous researchers have reported a link between phonological dyslexia and poor PA (Bowey & Rutherford, 2007; Castles et al., 1999; Manis et al., 1996 Stanovich et al., 1997; Sprenger-Charolles et al., 2011), and we replicated that finding at both time points. This pattern is essentially inevitable given the high correlation between PA and PC. In contrast, the surface dyslexia profile was not associated with differentially good or poor performance on any cognitive variable at either time point, although given the small size of the surface dyslexia group, our power to detect significant group differences was limited. In the larger initial study, we found preliminary evidence for a possible link between impaired RAN and surface dyslexia, but found no such relationship in this smaller sample. There are theoretical reasons to expect poor semantic skills would contribute to a surface dyslexia profile (Nation & Snowling, 1998; Plaut, McClelland, Seidenberg, & Patterson, 1996), but we did not find any differential lowering of VIQ, which is strongly correlated with semantic skills, in our surface dyslexia subgroup. Previous studies (including our own initial study) included both exception word reading and orthographic choice tasks, and demonstrated that when the surface dyslexia group is defined on the basis of one of those tasks, it also shows differentially poor performance on the second type (Manis & Bailey, 2008; Manis et al., 1996; Peterson et al., 2013). In some sense, this provides validation for the surface subtype. However, since both tasks are reading

tasks, this finding is probably more relevant to the question of whether orthographic coding has been measured reliably than whether surface dyslexia relates to a distinct neurocognitive profile.

Finally, we asked whether subtype at initial assessment affected literacy prognosis by examining raw score growth in word reading, spelling, and reading comprehension over time. Results were very clear: the average literacy levels of the subgroups were similar initially, and grew similarly over time. Thus, knowing a child's subtype at one time provided no information about his or her predicted progress in literacy over the coming years. Because the stability of psychometric dyslexia itself was so high, we know that the reading skills of the large majority of individuals with dyslexia remain far behind. This conclusion agrees with previous work based on data from this same sample across the full range of individual differences (Hulslander et al., 2010).

In summary, we found modest and asymmetric support for the validity of the phonological/surface dyslexia distinction. The phonological subtype showed moderate longitudinal stability and was associated with differentially poor PA. The surface dyslexia subtype was fairly rare in this sample. Its longitudinal stability was not very good, and we found no evidence for a distinct cognitive profile. Furthermore, subtype was irrelevant to prognosis. Taken together, we think these results question whether that the subtype distinction is clinically meaningful or important, though further research is needed to more definitively answer this question. Key unanswered questions include whether the subtypes are associated with differing responses to specialized treatments, and most critically, whether those responses transfer to real-world literacy tasks.

The current study has a number of important strengths, including its longitudinal design and relatively long follow-up interval. We are aware of no previous study that has statistically evaluated the stability of the phonological/surface subtypes over time, and compared those results to an appropriate benchmark (i.e., the stability of dyslexia itself). However, the work also includes a number of limitations. We were concerned only with the phonological and surface dyslexia subtypes, defined according to relative performance on specific word reading tasks, so our results do not speak to the validity of other proposed subtyping schemes for dyslexia. We had follow up data for only a minority of participants from the initial study, and so our overall sample size was relatively small (though similar to many other studies in the literature). The sample size of the pure surface dyslexia subgroup was too small to allow meaningful comparisons of the pure subtypes only. One subtyping measure was not re-administered at follow-up, and a second measure showed a ceiling effect, so orthographic coding skill was estimated with a single measure rather than a composite. Our assessment of neuropsychological constructs was limited to IQ, PA, and RAN, and did not cover all specific hypothesized reading component skills (McArthur et al., 2013; Ziegler et al., 2008). We did not account for variables like intervention history, exposure to print, or instructional type. A future treatment study should use an experimental design to manipulate such factors, which would more definitively address the clinical relevance of the phonological/surface distinction. Current results, in combination with previous literature, do not provide sufficient evidence for using this distinction to guide

educational or clinical practice at present, although it is possible that future research will support a different conclusion.

Acknowledgments

This work was supported by grants from the National Institutes of Health (HD-27802 and DC-05190).

References

- Bednarek D, Saldana D, Garcia I. Visual versus phonological abilities in Spanish dyslexia boys and girls. *Brain and Cognition*. 2009; 70:273–278. [PubMed: 19332361]
- Bishop DVM, McDonald D, Bird S, Hayiou-Thomas ME. Children Who Read Words Accurately Despite Language Impairment: Who Are They and How Do They Do It? *Child Development*. 2009; 80:593–605. [PubMed: 19467013]
- Bosse ML, Tainturier MJ, Valdois S. Developmental dyslexia: The visual attention span deficit hypothesis. *Cognition*. 2007; 194:198–230. [PubMed: 16859667]
- Bowey JA, Rutherford J. Imbalanced word-reading profiles in eighth-graders. *Journal of Experimental Child Psychology*. 2007; 96:169. [PubMed: 17266981]
- Castles A, Coltheart M. Varieties of developmental dyslexia. *Cognition*. 1993; 47:149–180. [PubMed: 8324999]
- Castles A, Datta H, Gayan J, Olson RK. Varieties of developmental reading disorder: Genetic and environmental influences. *Journal of Experimental Child Psychology*. 1999; 72:73–94. [PubMed: 9927524]
- Coltheart M, Masterson J, Byng S, Prior M, Riddoch J. Surface dyslexia. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*. 1983; 35:469–495.
- Coltheart M, Rastle K, Perry C, Langdon R, Ziegler J. DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*. 2001; 108:204–256. [PubMed: 11212628]
- Decker SN. Cognitive processing rates among disabled and normal young adults: A nine year follow-up study. *Reading and Writing: An Interdisciplinary Journal*. 1989; 2:123–134.
- DeFries, J. Colorado Reading Project. In: Gray, D.; Kavanaugh, J., editors. *Biobehavioral measures of dyslexia*. Parkton, MD: York Press; 1985. p. 107-122.
- DeFries JC, Filipek PA, Fulker DW, Olson RK, Pennington BF, Smith SD, Wise BW. Colorado Learning Disabilities Research Center. *Learning Disability Quarterly*. 1997; 7-8:19.
- Denckla MB, Rudel RG. Rapid 'automatized' naming (R.A.N.): Dyslexia differentiated from other learning disabilities. *Neuropsychologia*. 1976; 14:471–479. [PubMed: 995240]
- Dunn, LM.; Markwardt, FC. Examiner's manual: *Peabody Individual Achievement Test*. Circle Pines, MN: American Guidance Service; 1970.
- Ellis, AW.; Young, A. *Human Cognitive Neuropsychology*. London: Erlbaum; 1988.
- Friedmann N, Lukov L. Developmental surface dyslexias. *Cortex*. 2008; 44:1146–1160. [PubMed: 18761129]
- Griffiths YM, Snowling MJ. Predictors of exception word and nonword reading in dyslexic children: The severity hypothesis. *Journal of Educational Psychology*. 2002; 94:34–43.
- Gustafson S, Ferreira J, Roennberg J. Phonological or orthographic training for children with phonological or orthographic deficits. *Dyslexia*. 2007; 13:211–229. [PubMed: 17624906]
- Harm MW, Seidenberg MS. Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*. 1999; 106:491–528. [PubMed: 10467896]
- Harm MW, Seidenberg MS. Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. *Psychological Review*. 2004; 111:662–720. [PubMed: 15250780]
- Heim S, Tschierse J, Amunts K, Wilms M, Vossel S, Willmes K, et al. Huber W. Cognitive subtypes of dyslexia. *Acta Neurobiologiae Experimentalis*. 2008; 68:73–82. [PubMed: 18389017]

- Hulslander J, Olson RK, Willcutt EG, Wadsworth SJ. Longitudinal Stability of Reading-Related Skills and Their Prediction of Reading Development. *Scientific Studies of Reading*. 2010; 14:111–136. [PubMed: 20563241]
- Jastak, S.; Wilkinson, GS. *The Wide Range Achievement Test--Revised: Administration manual*. Wilmington, DE: Jastake Associates, Inc; 1984.
- Jimenez JE, Rodriguez C, Ramirez G. Spanish developmental dyslexia: Prevalence, cognitive profile, and home literacy experiences. *Journal of Experimental Child Psychology*. 2009; 103:167–185. [PubMed: 19321176]
- Manis, FR.; Bailey, CE. Exploring heterogeneity in developmental dyslexia: A longitudinal investigation. In: Reid, G.; Fawcett, LS.; Siegel, LS.; Manis, FR., editors. *The Sage Handbook of Dyslexia*. London: Sage Publications Ltd; 2008. p. 149-173.
- Manis FR, Seidenberg MS, Doi LM, McBride-Chang C. On the bases of two subtypes of development dyslexia. *Cognition*. 1996; 58:157–195. [PubMed: 8820386]
- Manis FR, Seidenberg MS, Stallings L, Joanisse M, Bailey C, Freedman L, et al. Keating P. Development of dyslexic subgroups: A one-year follow up. *Annals of Dyslexia*. 1999; 49:105–134.
- Markwardt, FC, Jr. *Peabody Individual Achievement Test-Revised*. Circle Pines, MN: American Guidance Service; 1989.
- McArthur G, Kohnen S, Larsen L, Jones K, Anandakumar T, Banales E, Castles A. Getting to grips with the heterogeneity of developmental dyslexia. *Cognitive Neuropsychology*. 2013; 30:1–24. [PubMed: 23614389]
- Morris RD, Stuebing KK, Fletcher JM, Shaywitz SE, Lyon GR, Shankweiler DP, et al. Shaywitz BA. Subtypes of reading disability: Variability around a phonological core. *Journal of Educational Psychology*. 1998; 90:347–373.
- Nation K, Snowling MJ. Semantic processing and the development of word-recognition skills: Evidence from children with reading comprehension difficulties. *Journal of Memory and Language*. 1998; 39:85.
- Olson, R. Genetic and environmental influences on phonological abilities and reading achievement. In: Brady, SA.; Braze, D.; Fowler, CA., editors. *Explaining individual differences in reading: Theory and evidence*. New York, NY US: Psychology Press; 2011. p. 197-216.
- Olson, R.; Forsberg, H.; Wise, B.; Rack, J. Measurement of word recognition, orthographic, and phonological skills. In: Lyon, GR., editor. *Frames of Reference for the Assessment of Learning Disabilities: New Views on Measurement Issues*. Baltimore, MD: Paul H. Brookes Publishing; 1994. p. 243-278.
- Olson, RK.; Forsberg, H.; Wise, B. Genes, environment, and the development of orthographic skills. In: Berniger, VW., editor. *The varieties of orthographic knowledge I: Theoretical and developmental issues*. Dordrecht, the Netherlands: Kluwer; 1994. p. 27-71.
- Olson, RK.; Kliegl, R.; Davidson, BJ.; Foltz, G. Individual and developmental differences in reading disability. In: MacKinnon, GE.; Waller, TG., editors. *Reading research: Advances in theory and practice*. Vol. 4. New York: Academic Press; 1985. p. 1-64.
- Olson RK, Wise B, Connors F, Rack J, Fulker D. Specific deficits in component reading and language skills: Genetic and environmental influences. *Journal of Learning Disabilities*. 1989; 22:339–348. [PubMed: 2738467]
- Pennington BF, Santerre-Lemmon L, Rosenberg J, MacDonald B, Boada R, Friend A, et al. Olson RK. Individual prediction of dyslexia by single versus multiple deficit models. *Journal of Abnormal Psychology*. 2012; 121:212–224. [PubMed: 22022952]
- Peterson RL, Pennington BF, Olson RK. Subtypes of developmental dyslexia: Testing the predictions of the dual-route and connectionist frameworks. *Cognition*. 2013; 126:20–38. [PubMed: 23010562]
- Peterson RL, Pennington BF, Shriberg LD, Boada R. What influences literacy outcome in children with Speech Sound Disorder? *Journal of Speech Language and Hearing Research*. 2009; 52:1175–1188.

- Plaut DC, McClelland JL, Seidenberg MS, Patterson K. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*. 1996; 103:56–115. [PubMed: 8650300]
- Ramus F. Developmental dyslexia: Specific phonological deficit or general sensorimotor dysfunction? *Current Opinion in Neurobiology*. 2003; 13:212–218. [PubMed: 12744976]
- Seidenberg MS, McClelland JL. A distributed, developmental model of word recognition and naming. *Psychological Review*. 1989; 96:523–568. [PubMed: 2798649]
- Snowling MJ, Gallagher A, Frith U. Family risk of dyslexia is continuous: Individual differences in the precursors of reading skill. *Child Development*. 2003; 74:358–373. [PubMed: 12705560]
- Sprenger-Charolles L, Cole P, Lacert P, Serniclaes W. On subtypes of developmental dyslexia: Evidence from processing time and accuracy scores. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie experimentale*. 2000; 54:87.
- Sprenger-Charolles L, Siegel LS, Jimenez JE, Ziegler JC. Prevalence and Reliability of Phonological, Surface, and Mixed Profiles in Dyslexia: A Review of Studies Conducted in Languages Varying in Orthographic Depth. *Scientific Studies of Reading*. 2011; 15:498–521.
- Stanovich KE. Explaining the differences between the dyslexic and the garden variety poor reader: The phonological-core variable-difference model. *Journal of Learning Disabilities*. 1988; 21:590–604. [PubMed: 2465364]
- Stanovich KE, Siegel LS, Gottardo A. Converging evidence for phonological and surface subtypes of reading disability. *Journal of Educational Psychology*. 1997; 89:114–127.
- Temple CM, Marshall JC. A case study of developmental phonological dyslexia. *British Journal of Psychology*. 1983; 74:517–533. [PubMed: 6640236]
- Vellutino FR, Fletcher JM, Snowling MJ, Scanlon DM. Specific reading disability (dyslexia): what have we learned in the past four decades? *Journal of Child Psychology and Psychiatry*. 2004; 45:2–40. [PubMed: 14959801]
- Wadsworth SJ, DeFries JC, Olson RK, Willcutt EG. Colorado longitudinal twin study of reading disability. *Annals of Dyslexia*. 2007; 57:139–160. [PubMed: 18060583]
- Wechsler, D. Wechsler Intelligence Scale for Children--Revised. New York: The Psychological Corporation; 1974.
- Wechsler, D. Wechsler Intelligence Scale for Children--Third Edition. San Antonio, TX: The Psychological Corporation; 1991.
- Wechsler, D. Wechsler Intelligence Scale-Third Edition. San Antonio, TX: Psychological Corporation; 1993.
- Wilkinson, GS. Wide Range Achievement Test-3. Wilmington, DE: Jastak Associates-Wide Range Inc; 1993.
- Wolf M, Bowers PG. The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology*. 1999; 91:415–438.
- Ziegler JC, Castel C, Pech-Georgel C, George F, Alario FX, Perry C. Developmental dyslexia and the dual route model of reading: Simulating individual differences and subtypes. *Cognition*. 2008; 107:151. [PubMed: 17959161]

Figure 1.
Performance of dyslexia subgroups on cognitive tasks at initial assessment.
* phonological dyslexia < balanced dyslexia = surface dyslexia

Figure 2.
Performance of dyslexia subgroups on cognitive tasks at follow-up.
* phonological dyslexia < balanced dyslexia

Table 1
Sample Characteristics for the Dyslexic and Control Groups at Follow-up

	Dyslexia group	Control group	
N	72	65	
Age in years: mean (SD)	15.19 (1.88)	15.49 (1.67)	
Gender: % male	48.6	49.2	
Race: % Caucasian	80.7	84.3	
Word recognition z-score: mean (SD)	-3.71 (2.07)	0.00 (1.00)	$t(135) = 13.11, p < .001$
PIAT-R Reading Recognition percentile	27 th	81 st	$t(113.9) = 12.54, p < .001$
PIAT-R Reading Comprehension percentile	30 th	73 rd	$t(118.8) = 9.57, p < .001$
WRAT-III Spelling percentile	27 th	72 nd	$t(135) = 13.61, p < .001$
Full Scale IQ: mean (SD)	93.17 (10.74)	108.55 (10.16)	$t(135) = 8.59, p < .001$

Table 2
Proportion of Dyslexia Group Meeting Various Subtype Criteria at Follow-up

Primary Subtype	n (% of dyslexia group)	Detailed Subtype	n (% of dyslexia group)
Phonological	44 (64.7)	Pure Phonological	16 (23.5)
		Relative Phonological	28 (41.2)
Surface	8 (11.8)	Pure Surface	2 (2.9)
		Relative Surface	6 (8.8)
Balanced	16 (23.5)	Mild	6 (8.8)
		Mixed	10 (14.7)

Table 3
Characterization of the Dyslexia Group According to Subtype across Time

Initial assessment Subtype				
		Phonological	Surface	Balanced
Follow-up Subtype	Phonological	36	4	4
	Surface	1	3	4
	Balanced	7	3	6

Table 4
Literacy Growth over Time for the Phonological, Surface, and Balanced Dyslexia Subgroups

Measure	Initial subgroup	Time	Raw Score: mean (SD)	Time $\times$ subtype interaction
PIAT Rec	Phonological	1	31.70 (7.82)	$p = .65, \eta^2_p = .01$
		2	73.00 (12.50)	
	Surface	1	34.60 (6.85)	
		2	78.80 (10.58)	
	Balanced	1	38.21 (6.02)	
		2	79.93 (6.21)	
TLWRT	Phonological	1	52.98 (27.78)	$p = .11, \eta^2_p = .07$
		2	102.66 (31.23)	
	Surface	1	53.30 (26.55)	
		2	117.20 (25.41)	
	Balanced	1	64.71 (17.85)	
		2	123.29 (14.55)	
PIAT Comp	Phonological	1	32.21 (8.95)	$p = .92, \eta^2_p = .00$
		2	73.98 (11.63)	
	Surface	1	33.90 (9.66)	
		2	75.50 (11.98)	
	Balanced	1	37.71 (7.49)	
		2	78.00 (8.72)	
WRAT Spell	Phonological	1	14.73 (4.98)	$p = .40, \eta^2_p = .03$
		2	31.23 (5.43)	
	Surface	1	14.30 (3.43)	
		2	32.60 (3.95)	
	Balanced	1	15.57 (4.57)	
		2	33.57 (3.59)	