# A Meta-Analysis of Reading-Level Match Dyslexia Studies in Consistent Alphabetic Orthographies

Rauno Parrila
Macquarie University
Dean Dudley
Macquarie University
Shuang Song
Capital Normal University
&
George K. Georgiou
University of Alberta

Author Note

Rauno Parrila, Department of Educational Studies, Macquarie University, Australia; Dean Dudley, Department of Educational Studies, Macquarie University, Australia (dean.dudley@mq.edu.au); Shuang Song, College of Teacher Education, Capital Normal University, Beijing, China (songsh326@gmail.com); George K. Georgiou, Department of Educational Psychology, University of Alberta, Canada (georgiou@ualberta.ca).

Correspondence concerning this article should be addressed to Dr. Rauno Parrila at rauno.parrila@mq.edu.au. Address: Department of Educational Studies, Macquarie University, Sydney NSW 2109, Australia. Phone +61 9850 6823.

ORCID Parrila: 0000-0003-4250-8980

## Abstract

We provide a meta-analytic review of all group-comparison studies that used reading-level match design, were conducted in highly consistent European orthographies, included children with dyslexia younger than 13-years of age as participants, and included measures of one or more of the potential causes of dyslexia. We identified 21 studies meeting these criteria that examined one or more of phonological awareness, rapid naming, verbal short-term memory, or auditory temporal processing. A random-effects model analyses showed first that the groups were matched imperfectly and they differed significantly in word reading measures not used for matching. Second, there were no significant differences between the individuals with dyslexia and their reading-level match controls in rapid naming, phonological memory, and auditory temporal processing. Finally, the analyses for phonological awareness showed a significant effect for comparisons that involved manipulating phonemes but not for tasks that involved manipulating syllables. The results are compatible with phonological deficit theories of dyslexia, but this conclusion is qualified by observed differences in reading skills and sample selection concerns.


*Keywords*: consistent orthographies; developmental dyslexia; meta-analysis; reading-level match design

# A Meta-Analysis of Reading-Level Match Dyslexia Studies in Consistent Alphabetic Orthographies

Developmental dyslexia, broadly defined as a difficulty in learning to read words despite conventional instruction, adequate intelligence, and sociocultural opportunity, is one of the most common childhood learning disabilities (Caravolas, 2005; Smythe & Everatt, 2004). Despite intensive research efforts across languages, there is substantial debate regarding the causes of developmental dyslexia (hereafter called dyslexia) in different orthographies (Caravolas, 2005; Elliot & Grigorenko, 2015; Goswami, 2002; McBride-Chang, Tong, & Mo, 2015). Several core deficits of dyslexia have been proposed (see e.g., Elliot & Grigorenko, 2015, and Parrila & Protopapas, 2017, for recent reviews), but the evidence for each of them has been inconsistent, particularly in highly consistent orthographies, such as Greek and Finnish, in which every grapheme roughly corresponds to one phoneme. As a result, we have a relatively long list of candidate causes of dyslexia only some of which have been extensively studied using methods that would allow causal arguments, such as longitudinal studies, randomized controlled interventions, or reading-level match designs (e.g., Backman, Mamen, & Ferguson, 1984; Bradley & Bryant, 1978; Goswami & Bryant, 1989). The purpose of this paper is to provide a meta-analysis of the studies in highly consistent alphabetic orthographies that used the reading-level (RL) match design to examine between-group differences in multiple constructs associated with and possibly causally linked to dyslexia.

Theoretically, examination of the status of multiple different causes addresses the tension between core deficit theories and probabilistic multifactorial models. A traditional theoretical approach to developmental dyslexia has been to posit a specific deficit in some cognitive or perceptual processes to account for word reading difficulties. The currently dominant theory of developmental dyslexia suggests that all or a large majority of children with dyslexia have a phonological deficit (e.g., Bishop & Snowling, 2004; Stanovich, 1988; Vellutino, Fletcher, Snowling, & Scanlon, 2004), as evidenced in poor performance on any task that requires phonological processing (for reviews, see Elliot & Grigorenko, 2015; Protopapas, 2014; Snowling, 2000; Vellutino et al., 2004), such as phonological awareness and phonological short-term memory tasks. Wolf and Bowers (1999) suggested that rapid automatized naming speed (RAN) constitutes a second core deficit in dyslexia that can exist either jointly or independently from the phonological deficit. According to the naming-deficit hypothesis, children with dyslexia show impairments in naming visually-presented familiar symbols, such as objects, colours, digits, and letters (Badian, Duffy, Als, & McAnulty, 1991; Holopainen, Ahonen, & Lyytinen, 2001; Wimmer, 1993), and possibly exhibit specific orthographic learning issues (see reviews in Georgiou & Parrila, 2013, and Kirby, Georgiou, Martinussen, & Parrila, 2010).

The additional proposed core deficit theories generally fall in one of two categories. In the first, alternative – possibly domain general – impairments are posited to explain word reading problems of different types. Theories in this category have, for example, proposed magnocellular dysfunctions (e.g., Stein & Walsh, 1997), visual-spatial attention difficulties

(Vidyasagar & Pammer, 2010), shorter visual attention span (Bosse, Tainturier, & Valdois, 2007; Zoubrinetzky, Bielle, & Valdois, 2014), sluggish visual attention (Franceschini, Gori, Ruffino, Pedrolli, & Facoetti, 2012), perceptual anchoring deficit (Ahissar, 2007), procedural memory deficits (see review in Lum, Ullman, & Conti-Ramsden, 2013), deficits in skill automatization (e.g., Nicolson, Fawcett, & Dean, 2001), or compromised temporal acuity (Laasonen, Service, & Virsu, 2002) as the core deficit underlying poor word reading. In the second category, the phonological deficit itself is attributed to a more general or lower-level dysfunction. The most prominent of these theories are the various auditory temporal processing theories (e.g., Gaab, Gabrieli, Deutsch, Tallal, & Temple, 2007; Goswami, 2011, 2015; Noordenbos, Segers, Serniclaes, & Verhoeven, 2013; Serniclaes et al., 2004; Tallal, 1980) that focus on identifying specific speech perception or processing deficits with various signature tasks (see e.g., Protopapas, 2014, for a review).

The theoretical assumption that behaviourally-defined developmental disorders such as dyslexia can have a single (or double) core cause has been challenged by multifactorial etiological models (e.g., Lyytinen et al., 1998; Pennington, 2006; van Bergen, van der Leij, & de Jong, 2014). In a seminal analysis of research in developmental disorders, Pennington (2006) concluded that converging evidence of dyslexia studies precipitates a major reconceptualization of the existing theoretical models. He proposed that probabilistic multiple deficit models (PMDM) are needed to provide a more realistic account of developmental disorders, their comorbidity, and the nondeterministic relationships between disorders and their presumed causes. Pennington (2006) suggested further that such PMDMs must include

protective and risk factors, multiple levels of analysis, bidirectional connections between constructs within each level (horizontal or intralevel interactions), and bidirectional connections between levels (vertical or interlevel interactions) to account for interactions between protective and risk factors functioning at different levels of analysis (see also, Ford & Lerner, 1992; Gottlieb, 1983, 1997; Gottlieb, Wahlsten, & Lickliter, 2006; Lyytinen et al., 1998). In terms of cognitive level predictors examined in this meta-analysis, PMDMs would suggest that while a significant deficit in any one of them could be associated with a reading deficit (possibly in absence of protective factors in other levels), none are necessary and a reading deficit could also result from an interaction of multiple subclinical deficits that jointly contribute to a clinically identifiable reading deficit. While models of this kind are not testable theories, theories derived from them can accommodate a wider range of results than core deficit theories, including diverging empirical results in cognitive (see below), neurological (see e.g., Protopapas & Parrila, 2018, for a discussion), and genetic (e.g., Carrion-Castillo, Franke, & Fisher, 2013) levels.

In terms of the core deficits, the empirical support for the phonological deficit hypothesis is generally stronger than the support for the RAN deficit in English-speaking individuals with dyslexia (see e.g., Pennington et al., 2012), whereas the same may not be true for more consistent alphabetic orthographies. There may be multiple explanations for the differences. First, several studies have found that children with dyslexia in consistent orthographies experience a reading speed difficulty and not necessarily a reading accuracy difficulty (e.g., Landerl, Wimmer, & Frith, 1997; Serrano & Defior, 2008; Wimmer, 1993).

This, in turn, has implications for the assumed causes of dyslexia. For example, Brizzolara et al. (2006) have documented that, in Italian, rapid naming speed rather than phonological awareness deficits represent the main cognitive marker of reading speed problems. Similar findings have been reported for Greek (Georgiou, Protopapas, Papadopoulos, Skaloumbakas, & Parrila, 2010), Spanish (Davies, Cuetos, & Glez-Seijas, 2007), Finnish (Torppa et al., 2013), and German (Landerl & Wimmer, 2000). Researchers have hypothesized that the joint effect of consistent grapheme-phoneme correspondences and systematic phonics instruction is sufficiently powerful to secure children's phonological skills after a few months of reading experience, regardless of their pre-reading levels of phonological awareness (Landerl & Wimmer, 2000; Papadopoulos, Georgiou, & Kendeou, 2009; Serrano & Defior, 2008). However, it is also possible that lack of significant results reflects task insensitivity and that even in consistent orthographies, children with dyslexia experience phonological awareness difficulties when phonological awareness is measured with sufficiently challenging tasks (i.e., phoneme elision, spoonerisms; Caravolas, Vólin, & Hulme, 2005; Constantinidou & Stainthorp, 2009; Nikolopoulos, Goulandris, & Snowling, 2003) and, therefore, phonological awareness should be considered a universal cause of reading difficulties (Katzir, Shaul, Breznitz, & Wolf, 2004; Paulesu et al., 2001; Ziegler et al., 2003).

Another possible reason for the inconsistent findings may be the methodological designs employed in different studies. Bryant and Goswami (1986) argued that for a processing skill to be considered a cause of reading difficulties it is not sufficient to show that a group of children with dyslexia perform worse than their general ability and age matched

peers, because the difference may simply reflect differences in reading experience (see Bradley & Bryant, 1978, for the original formulation of this argument). A group of children that is matched to children with dyslexia on both general cognitive and reading ability should also be included in studies examining dyslexia. According to Bradley and Bryant (1978), if "the two groups have reached the same reading level, and yet the backward readers are worse on a perceptual task, the fact that the two groups have the same reading ability as one another rules out the possibility that the backward readers' perceptual failure is merely the result of a lack of reading experience." (p. 746). This formulation has since been widely adopted (but see Mamen, Ferguson, & Backman, 1986, for an alternative formulation) and RL-match designs are commonly used to test assumptions of causality following the logic that if the poor readers perform poorer than RL-matched controls on task A assessing construct B, then construct B is a potential cause for dyslexia. However, if the poor readers only differ from their better reading chronological-age matched peers, the results are considered inconclusive.

Since the early studies (e.g., Backman, 1983; Bradley & Bryant, 1978; Bryant & Bradley, 1980; Seymour & Porpodas, 1980) the use of reading/spelling-level match design has become common in group-comparison studies in English. For example, a meta-analysis by Melby-Lervåg, Lyster, and Hulme (2012) included 27 studies with English-speaking children and 10 from languages other than English that had used RL-match design to examine differences in phonemic awareness. Interestingly, and in contrast to the discussion above, their results indicated that the English studies showed a smaller overall phonemic awareness difference ($d$ = -0.49) between children with dyslexia and RL control children than the non-

English studies ($d$ = -0.83), a result we suspect was largely driven by one outlier study by Constantinidou and Stainthorp (2009; see Figure 5 in Melby-Lervåg et al., 2012, p. 335). Melby-Lervåg et al. also found 24 RL-match design studies, 21 of which were in English, that included verbal short-term memory measures. The studies produced an overall nonsignificant effect size ($d$ = -0.09) and the three non-English studies showed an average effect size of -0.12. It is important to note that the non-English studies in Melby-Lervåg et al. (2012) includes a wider variety of languages than this review, including French and Danish that are often not classified as orthographically consistent.

While not as extensively as in English, the RL-match design has now been used with children learning to read highly consistent European orthographies and no reviews or meta-analyses are available attempting to summarize the findings. Given the divergent findings from individual studies and varying theoretical accounts for the causes of dyslexia, the purpose of the present study was to provide a meta-analytic review of the findings of all group-comparison studies examining the potential causes of dyslexia in highly consistent European orthographies (see below) and that include a reading-level matched control group. We reasoned that a meta-analysis (1) focusing on studies using RL-match design with younger readers (see below) and (2) examining the effect of potential moderators of effects (age and type of tasks) could provide the clearest evidence so far for each of the examined constructs. In terms of possible causes of dyslexia, we identified a sufficient number (three or more) of studies meeting these criteria that examined one or more of phonological awareness, rapid naming, verbal short-term memory, or auditory temporal processing. As we were not

able to locate at least three studies with RL-match comparison group examining magnocellular deficits, visual-spatial attention, visual attention, visual attention span, perceptual anchoring, temporal acuity, or skill automatization deficits, these are not considered further in this meta-analysis.

We report below on three separate comparisons. The main comparison for the purpose of this study is the comparison between the children with dyslexia and the reading-level matched group. The comparisons involving the chronological-age matched peers are always expected to show differences in favour of them and may be considered as sample selection and task validity checks for the current purposes.

## Method

### Data Collection

The inclusion, search, and coding procedures are detailed in Figure 1. The present meta-analysis is limited to studies conducted in European alphabetic orthographies that are considered to be highly consistent (Seymour, Aro, & Erskine, 2003). Accordingly, only dyslexia studies conducted in Finnish, Greek, Spanish, Hungarian, Icelandic, Italian, Swedish, Norwegian, and German were searched for the review. Studies conducted in English, French, Danish, Portuguese, and Dutch were excluded. To select the studies, we first searched in computerized databases (ERIC, Medline, PsycINFO, PubMed, ProQuest, and Google Scholar) for studies published in English from January 1990 to February 2019 (the final search was conducted in March 2019). The following descriptors were used in our search: Set 1 - Finnish, Greek, Spanish, Hungarian, Icelandic, Italian, Swedish, Norwegian, and German,

paired with Set 2 - phon* awareness, phonological processing, RAN, rapid naming, naming speed, verbal short-term memory, working memory, temporal processing, speed of processing, visual attention, perceptual anchoring, temporal acuity, articulation, and Set 3 - dyslexia, reading deficits, decoding deficits, word recognition deficits, learning disabilities, and reading-level control. Within each set the OR command was used and between sets the AND command was used.

Abstracts of peer-reviewed studies, dissertations, and book chapters were subsequently examined by the first and the second author, who examined all the full texts and extracted the information analysed here. Further, aware of the complexities involved in using RL- match design with older participants (see e.g., Deacon, Parrila, & Kirby, 2008, for a discussion), we included only studies in which the children with dyslexia/reading deficit (RD) were younger than 13 years of age. The requirement of a younger RL control group with established reading skills resulted in RD samples necessarily being in Grade 3 or above and the youngest RD sample included had a mean age of 8.5 years.

The final sample included 21 studies, involving 604 children with RD, 605 reading-level control children, and 608 chronological-age matched control children. Table 1 lists the included studies, their sample sizes and the ages of children in different groups, the selection criteria for the RD group as explained by the authors of the papers, the tasks used to match RD and RL groups, other matching variables if reported, and what analysed construct each study contributed information to. In order to prevent violation of independence of observations (i.e., including multiple reports using data from the same sample), reports by the

same author were examined in order to detect duplicate samples. When the same data was used in several publications, the most recent publication was included.

For the reading level constructs (word reading accuracy/efficiency, nonword reading accuracy/efficiency) and cognitive level constructs (phonological awareness, rapid naming, verbal short-term memory, and auditory temporal processing) examined below, operational criteria were first established in order to determine the indicators of each construct. For *word reading accuracy/efficiency*, we included results from tasks that required participants to read real words aloud without assessing comprehension. We included all reported measures, including those used for matching the groups, but analysed matching variables separately (see below). Four studies used sentence level reading tasks for matching and these were included under word reading. For *nonword reading accuracy/efficiency* task, tasks involved decoding of pronounceable nonwords. When the number of correct or incorrect items was scored, it was considered an accuracy task; when correct responses within a time limit, reading speed, or task completion time was reported, the task was considered an efficiency task. If the study reported on several word reading or nonword reading measures (other than the matching tasks), accuracy measures were combined into one estimate and efficiency measures into a separate estimate. Thus, one study could contribute two estimates to word reading and nonword reading analyses.

A task was considered to be a measure of *phonological awareness* if it required participants to orally manipulate, generate, or judge a phoneme, onset, rhyme, or syllable in words. Tasks requiring judgment or manipulation of syllables or rimes were combined into

one estimate. Tasks focusing on phonemes were further divided into two categories: matching, blending, and segmentation tasks were combined to estimate differences in simple phonemic awareness, and deletion, substitution and spoonerism tasks were combined to estimate complex phonemic awareness. As a result, three studies contributed three estimates to phonological awareness calculations whereas the others contributed one or two estimates.

To be considered a measure of *rapid naming (RAN)*, the test needed to require quick naming of an array of objects, colours, letters, or digits. RAN-Letters and RAN-Digits were combined into an estimate of alphanumeric naming speed, and RAN-Colours and RAN-Objects into an estimate of non-alphanumeric naming speed.

Measures of *verbal short-term memory* (VSTM) included tasks in which children needed to repeat a spoken list of words, nonwords, digits, or sentences immediately after hearing them. VSTM tasks were further classified into two types: those that required repeating real words, digits or sentences, and those that required repeating nonwords.

*Auditory temporal processing* (ATP) tests assessed the ability of an individual to discriminate between sequentially presented auditory stimuli. The tasks were divided into two categories on the basis of whether they included two (AXB) or three (2IFC; see Goswami et al., 2010, for details) stimuli.

All studies were coded twice by trained coders and interrater reliability was estimated. The interrater correlation (Pearson's) for the reported continuous variables (mean scores, SD of scores, and sample sizes) was .99 and the agreement rate was 97%. In regards to the moderator variables (see below), the interrater correlation for age was .99 and the agreement

rate 96%. Cohen's K for the categorical moderator variables (type of task) was .97. The few disagreements in the coding were resolved by discussion between the coders after consultation with the first author.

**Moderator Variables**

**Age.** Mean age of each sample was coded. Mean ages of the RD sample were tested for the comparisons of RD vs CA. Mean ages of the RA sample were tested for the comparison of RD vs RA and CA vs RA.

**Type of tasks.** Word reading and nonword reading tasks were classified into reading accuracy and reading efficiency tasks (see above). Phonological awareness tasks were classified into three categories: syllables and rimes, simple, and complex (see above). RAN tasks were classified based on the stimuli as alphanumeric RAN tasks or as non-alphanumeric RAN tasks (see above). VSTM tasks were classified into two categories (real words/digits/sentences vs. nonwords). Because fewer than five studies reported data on a moderator for the auditory temporal processing tasks, the analysis for this moderator was not performed.

**Meta-analytic Procedures**

The computer program Comprehensive Meta-Analysis, version 3 (Borenstein, Hedges, Higgins, & Rothstein, 2005) was used for the majority of the analyses. The mean and standard deviation of test scores and sample size for each group, as well as information pertinent to group average age and test type were coded from the studies and entered into a

predefined data sheet in the computer program. The 21 studies included in our meta-analysis are listed in Table 1 and identified by an asterisk in the reference list.

Hedge's *g* effect sizes were calculated for each comparison as many of the samples were small. The Hedge's *g* statistic expresses the difference of the means in units of the pooled standard deviation, and it is interpreted similarly to Cohen's *d.* A 95% confidence interval (CI) was calculated for each effect size to examine whether the effect size was significantly different from zero. The overall *g* value was estimated by calculating a weighted average of the effect sizes from each study. We used a random-effects model, which rests on the assumption that variation between studies can be systematic, and not only due to random error. Whether or not the overall effect size differed from zero was tested with a *z* test. In order to examine whether the between studies variation in *g* value was significant, the *Q* test of homogeneity was used (Hedges & Olkin, 1985). A significant value on this test indicates reliable variability between the effect sizes in the sample of studies. $I^2$ was used in order to determine the magnitude of heterogeneity. $I^2$ is the proportion of total variation between effect sizes that is caused by real heterogeneity rather than chance. The Classic Fail-Safe N and Trim and Fill (Duval & Tweedie, 2000) methods were used for assessing publication bias. Given the number of comparisons (three per construct), forest plots are included as supplementary material.

Given the relatively small number of studies, we examined if any of the effect sizes were outliers, defined as more than 2 SDs from the mean. When outliers are identified, we report results both with and without the outlier values.

## Results

We report below on three separate comparisons. As the main comparisons of interest were the comparisons between the dyslexia/reading disabled (RD) group and the reading-level matched (RL) group, the full results for those are reported first in each section. Only the main results are reported for the comparisons involving the chronological-age matched (CA) group as these comparisons are of secondary interest here.

### Word Reading

First, we examined the between group differences in the matching variables, typically either word reading accuracy or efficiency (see Table 1). By definition, there should be no differences between the RD and RL groups, and substantial differences between these two and the CA group. Eighteen studies reported sufficient data on the matching variables to compare RD and RL groups. The overall mean effect size was $g$ = -0.132, 95% CI [-.258, -.006], $z(17) = -2.06$, $p = .039$, favouring the RL group. After removing one outlier, the overall mean effect size was significant, $g$ = -0.164, 95% CI [-.296, -.033], $z(16) = -2.454$, $p < .014$, favouring the RL group. Fourteen studies reported sufficient data on the matching variables to compare RD and RL groups to the CA groups. As expected, the mean effect sizes were large and significant for both comparisons favouring the CA groups: RD vs. CA $g$ = -2.569, 95% CI [-3.124, -2.014], $z(13) = -9.067$, $p < .001$, and RL vs. CA, $g$ = -2.764, 95% CI [-3.466, -2.061], $z(13) = -7.710$, $p < .001$. In sum, the CA comparison groups performed considerably better than the RD and RL groups, but the matching of the latter groups was not perfect with the RL groups showing better reading skills than the RD group.

Eleven studies reported on 16 comparisons of RD, RL, and CA groups in word reading measures not used for matching. The studies included 343 children in the RD groups (mean sample size = 31.18, *SD* = 20.71, range = 12–89, age ranged from 8.72 to 11.94 years), 367 children in the RL groups (mean sample size = 33.36, *SD* = 17.95, range = 12–67, age ranged from 6.83 to 8.89 years) and 368 children in the CA groups (mean sample size = 33.45, *SD* = 14.62, range = 12–64, age ranged from 8.90 to 11.98 years).

***Comparison of RD and RL.*** The overall mean effect size was large, $g$ = -0.996, 95% CI [-1.475, -0.516], $z(14) = -4.071$, $p < .001$, favouring the RL group. The between studies variation in $g$ values was significant, $Q(14) = 210.069$, $p < .001$, $I^2 = 92.859$. After removing the three outlier effect sizes (all favouring RL), $g$ dropped to -.486, 95% CI [-0.830, -0.142], $z(12) = -2.763$, $p = .006$. The funnel plot searching for publication bias indicated that two studies could be trimmed from the left, bringing $g$ to -1.199.

The mean effect size for the eight reading accuracy comparisons, $g$ = -1.152, 95% CI [-2.158, -0.146], was larger than the mean for the nine reading efficiency comparisons, $g$ = -0.894, 95% CI [-2.019, 0.231]. The difference was largely driven by two studies with $g$s > -3.4 and after the outliers were removed the $g$s were -0.759, 95% CI [-1.221, -0.298] for accuracy and -0.183, 95% CI [-0.596, 0.231] for efficiency. The correlation between the age of the RL group and the effect size was significant, $r = .501$, $p = .048$, indicating that when the RL children were older, the differences tended to be larger.

***Comparison of RD and CA.*** One study reported data for only the RD and RL groups, so the comparison of RD and CA was carried out with only ten studies reporting on 15

comparisons. The overall mean effect size was large and significant, $g$ = -1.948, 95% CI [-2.438, -1.457], $z(14)$ = -7.783, $p < .001$. The variation in $g$ values between studies was significant, $Q(14)$ = 171.211, $p < .001$, $I^2$ = 91.823. After removing two outliers, the overall effect size was -1.426, 95% CI [-1.726, -1.126].

***Comparison of CA and RL.*** One study reported data for only the RD and RL groups, so the comparison of CA and RL was carried out with only nine studies reporting 12 comparisons. The overall mean effect size was large and significant, $g$ = 1.067, 95% CI [0.619, 1.514], $z(11)$ = 4.668, $p < .001$, favouring the CA group. No outliers were identified. The variation in $g$ values between studies was significant, $Q(11)$ = 101.196, $p < .001$, $I^2$ =89.130.

**Nonword Reading**

Twelve studies reported comparisons of nonword reading tasks among the RD, RL, and CA groups. The studies included 353 children in the RD groups (mean sample size = 29.42, $SD$ = 20.67, range = 10–89, age ranged from 8.72 to 12.80 years), 345 children in the RL groups (mean sample size = 28.75, $SD$ = 15.30, range = 10–67, age ranged from 6.83 to 9.20 years), and 347 children in the CA groups (mean sample size = 28.92, $SD$ = 11.62, range = 10–47, age ranged from 8.90 to 12.90 years).

***Comparison of RD and RL.*** Twelve studies reported altogether 19 effect sizes for either nonword reading accuracy or efficiency. The overall mean effect size was significant, $g$ = -0.981, 95% CI [-1.353, -0.609], $z(18)$ = -5.169, $p < .001$, favouring the RL group. The variation in $g$ values was significant, $Q(18)$ = 174.992, $p < .001$, $I^2$ = 89.714. After removing

three outliers, the overall effect size was $g$ = -0.633, 95% CI [-.922, -.345], $z(15)$ = -4.301, $p$ < .001. The funnel plot indicated that two effect sizes from the left could be trimmed, resulting in $g$ of -1.156.

The mean effect size for the eight nonword reading accuracy comparisons, $g$ = -1.095, 95% CI [-1.896, -0.294], was significant and identical to the mean effect size for the eleven reading efficiency comparisons, $g$ = -1.077, 95% CI [-1.787, -0.366]. After removing one outlier from accuracy comparisons and two from efficiency comparisons, the mean effect sizes remained significant at -0.825, 95% CI [-1.404, -0.246] and -0.632, 95% CI [-0.950, -0.315], respectively. The effect sizes correlated -.04 with the RD groups' ages, .17 with RL ages, and -.24 with the difference between the two; none approached significance. Finally, given the observed differences in the matching variables, we examined what nonword repetition mean effect sizes would have been subtracting the effect sizes for the matching difference from the observed effect sizes. After removing the two outliers, the mean $g$ was significant at -.679, 95% CI [-1.056, -0.302].

***Comparison of RD and CA.*** Two study reported data for only the RD and RL groups, so the comparison of RD and CA was based on 10 studies including 17 effect sizes. The overall mean effect size was large and significant, $g$ = -2.325, 95% CI [-2.834, -1.816], $z(16)$ = -8.947, $p$ < .001. The variation in $g$ values between studies was significant, $Q(16)$ = 198.455, $p$ <.001, $I^2$ = 91.938. After removing three outliers, the overall effect size was $g$ = -1.655, 95% CI [-1.952, -1.358], $z(13)$ = -.10.910, $p$ < .001. The funnel plot indicated that three studies should be trimmed from the left, resulting in $g$ of -2.733.

***Comparison of CA and RL.*** Three studies did not include all the data needed to compare CA and RL groups, so the comparisons were based on nine studies reporting on 16 comparisons. The overall mean effect size was large and significant, $g = 1.173$, 95% CI [0.758, 1.589], $z(15) = 5.533$, $p < .001$. The between studies variation in $g$ values was significant, $Q(15) = 134.927$, $p < .001$, $I^2 = 88.883$. After removing one outlier, the overall effect size was $g = .962$, 95% CI [0.656, 1.269], z(14) = 6.155, $p < .001$. The funnel plot suggested trimming five effect sizes from the right to the mean, resulting in $g$ of 1.666.

**Phonological Awareness**

Nineteen studies compared performance on phonological awareness tasks among the RD, RL, and CA groups. The studies included 530 children in the RD groups (mean sample size = 27.89, *SD* = 17.38, range = 10–89, age ranged from 8.48 to 12.80 years), 558 children in the RL groups (mean sample size = 29.37, *SD* = 15.74, range = 10–67, age ranged from 6.48 to 9.68 years), and 558 children in the CA groups (mean sample size = 29.37, *SD* = 13.78, range = 10–64, age ranged from 8.36 to 12.90 years).

***Comparison of RD and RL.*** Nineteen studies contributed 33 comparisons to these analyses. The overall mean effect size was significant, $g = -0.426$, 95% CI [-0.636, -0.217], $z(31) = -3.995$, $p < .001$. The between studies variation in $g$ values was significant, $Q(31) = 244.215$, $p < .001$, $I^2 = 87.306$. After removing the one outlier, the effect size was $g = -0.339$, 95% CI [-0.518, -0.160], $z(30) = -3.715$, $p < .001$. The funnel plot indicated that we could trim 8 studies from the left, resulting in a significant $g$ of -0.699.

The effect sizes correlated significantly with the RL groups' ages, $r = .437$, $p = .020$, but not with RD age ($r = .311$) or the age difference between the groups ($r = .037$). Nine effect sizes assessed syllable or rime level manipulations (both accuracy and speed) and produced a nonsignificant mean *g* of -0.302, 95% CI [-0.778, 0.173]. Eight effect sizes from tasks classified as simple phonemic awareness tasks (see above) produced a significant *g* value of -0.366, 95% CI [-0.646, -0.087] favouring the RL group. Finally, 14 effect sizes from complex phonemic awareness tasks produced a nonsignificant *g* of -0.208, 95% CI [-0.611, 0.194]. Removing the one outlier (favouring RD) resulted in a significant *g* of -0.335, 95% CI [-0.657, -0.013]. Finally, when we controlled for the imperfect matching (see above) and removed the one outlier, the mean effect size for the remaining 26 comparisons remained significant, $g = -0.270$, 95% CI [-0.475, -0.064].

***Comparison of RD and CA.*** Nineteen studies contributed 33 comparisons to these analyses. The overall mean effect size was large and significant, $g = -1.114$, 95% CI [-1.302, -0.927], $z(32) = -11.666$, $p < .001$. The between studies variation in *g* values was not significant, $Q(32) = 194.115$, $p < .001$, $I^2 = 83.515$. After removing the one outlier, the effect size was $g = -1.029$, 95% CI [-1.171, -0.888], $z(31) = -14.258$, $p < .001$. The funnel plot indicated that eight effects to the left could be trimmed, resulting in a *g* of -1.322.

***Comparison of CA and RL.*** The overall mean effect size was large and significant, $g = 0.739$, 95% CI [0.609, 0.870], $z(31) = 11.093$, $p < .001$. The between studies variation in *g* values was significant and large, $Q(31) = 92.273$, $p < .001$, $I^2 = 66.404$. After removing one

outlier, the effect size was $g$ = 0.701, 95% CI [0.584, 0.819], $z(31) = 11.669$, $p < .001$. The funnel plot indicated that no studies could be trimmed.

**Rapid Automatized Naming**

Fourteen studies reported comparisons between the RD, CA, and RL groups on RAN. The studies included 312 children in the RD groups (mean sample size = 22.28, *SD* = 7.82, range = 11–36, age ranged from 9.23 to 11.98 years), 402 children in the RL groups (mean sample size = 28.71, *SD* = 17.52, range = 11–67, age ranged from 6.83 to 9.68 years), and 392 children in the CA groups (mean sample size = 28.00, *SD* = 14.48, range = 12–64, age ranged from 9.26 to 11.98 years).

***Comparison of RD and RL.*** The overall mean effect size was not significant, $g$ = -0.177, 95% CI [-0.437, 0.084], $z(15) = -1.330$, $p = .184$. The between studies variation in $d$ values was significant, $Q(15) = 59.269$, $p < .001$, $I^2 = 74.612$. After removing two outliers, the overall effect size was $g$ = -0.019, 95% CI [-0.154, 0.117], $z(13) = -0.274$, $p = .785$. The funnel plot suggested trimming four effects from the left, resulting in significant $g$ of -0.342 [-0.598, -0.085] favouring the RL group. This indicates that there may be a publication bias favouring the RD group in the original pool of studies.

Five studies reported on alphanumeric RAN tasks, with one study reporting an outlier value seven times higher than the next value. The four remaining values had a mean of 0.043. Nine studies reported on nonalphanumeric RAN tasks and had a mean of -0.08. The differences between the tasks were not examined further.

Age of the RL group correlated significantly ($r$ = .519, $p$ = .039) with the effect sizes, indicating that the older the RL participants, the more the results favoured the RD group. The age of the RD ($r$ = .342) and the age difference between the two groups ($r$ = -.129) did not correlate significantly with the effect sizes. Finally, controlling for the imperfect matching changed the direction of the effect, $g$ = 0.137, 95% CI [-0.065, 0.339] but did not affect the conclusion of nonsignificant differences.

***Comparison of RD and CA.*** The overall mean effect size was large and significant, $g$ = -1.400, 95% CI [-1.711, -1.090], $z(16)$ = -8.843, $p$ < .001. The between studies variation in $g$ values was significant, $Q(16)$ = 81.757, $p$ < .001, $I^2$ = 80.430. After removing the one outlier, the overall effect remained significant, $g$ = -1.238; 95% CI [-1.471, -1.005], $z(15)$ = -10.417, $p$ < .001. The funnel plot indicated no need for trimming.

***Comparison of CA and RL.*** The overall mean effect size was large and significant, $g$ = 1.427, 95% CI [1.070, 1.784], $z(14)$ = 7.830, $p$ < .001. The variation in $g$ values between studies was significant, $Q(14)$ = 81.065, $p$ < .001, $I^2$ = 82.730. After removing the one outlier, $g$ remained significant at 1.245, 95% CI [0.976, 1.514], $z(13)$ = 9.067, $p$ < .001. The funnel plot indicated no trimming needed.

**Verbal Short-term Memory**

Twelve studies reported comparisons between the RD, CA, and RL groups on verbal short-term memory (VSTM). The studies included 303 children in the RD groups (mean sample size = 25.25, $SD$ = 12.08, range = 11–53, age ranged from 9.23 to 11.98 years), 347 children in the RL groups (mean sample size = 28.92, $SD$ = 18.42, range = 112–67, age

ranged from 7.25 to 9.68 years), and 334 children in the CA groups (mean sample size = 27.83, *SD* = 14.43, range = 12–64, age ranged from 9.20 to 11.98 years).

***Comparison of RD and RL.*** The overall mean effect size was not significant, $g$ = -0.172, 95% CI [-0.545, 0.201], $z(13) = -0.905$, $p = .365$. The between studies variation in $g$ values was significant, $Q(13) = 91.057$, $p < .001$, $I^2 = 85.723$. After removing the one outlier, the overall effect size remained nonsignificant at $g = 0.042$, 95% CI [-0.181, 0.265], $z(12) = 0.366$, $p = .714$. The funnel plot indicated trimming of five studies to the left, resulting in a significant difference, $g = -0.537$ 95% CI [-0.948, -0.125], possibly indicating publication bias in favour of the RD group.

Five studies (with one extreme outlier that was removed) reported on VSTM tasks that required repetition of nonwords, $g = 0.068$, 95% CI [-.243, 0.379] and nine on tasks where the stimuli consisted of digits, words, or sentences, $g = -0.089$, 95% CI [-0.828, 0.651]. The effect sizes also did not correlate significantly with any of the age variables. Controlling for imperfect matching (and one outlier favouring RL) resulted in a nonsignificant $g$ value of 0.176 95% CI [-.0437, 0.399].

***Comparison of RD and CA.*** The overall mean effect size was significant, $g = -0.913$, 95% CI [-1.316, -0.511], $z(14) = -4.451$, $p < .001$. The between studies variation in $g$ values was not significant, $Q(14) = 97.652$, $p < .001$, $I^2 = 86.687$. After removing the one outlier, the overall effect size remained significant, $g = -0.653$, 95% CI [-0.871, -0.435], $z(13) = -5.870$, $p < .001$. The funnel plot indicated six effects could be trimmed from the left, resulting in $g$ value of -1.434, 95% CI [-1.908, -0.959].

***Comparison of CA and RL.*** The overall mean effect size was significant, $g = 0.638$, 95% CI [0.362, 0.914], $z(11) = 4.523$, $p < .001$. The between studies variation in $d$ values was significant, $Q(11) = 41.126$, $p < .001$, $I^2 = 73.253$. No outliers were identified and the funnel plot indicated no need for trimming.

**Auditory Temporal Processing**

Five studies reported comparisons between RD, CA, and RL groups on auditory temporal processing. The studies included 107 children in the RD groups (mean sample size = 21.40, $SD = 6.23$, range = 16–32, age ranged from 10.20 to 11.98 years), 111 children in the RL groups (mean sample size = 22.20, $SD = 6.57$, range = 14–32, age ranged from 8.10 to 9.68 years), and 111 children in the CA groups (mean sample size = 22.20, $SD = 6.83$, range = 13–32, age ranged from 10.00 to 11.98 years).

***Comparison of RD and RL.*** The overall mean effect size was not significant, $g = -0.191$, 95% CI [-0.735, 0.354], $z(6) = -0.686$, $p = .493$. The between studies variation in $d$ values was significant, $Q(6) = 41.536$, $p < .001$, $I^2 = 85.555$. After removing one outlier the effect size was $g = 0.106$, 95% CI [-0.129, 0.342], $z(5) = 0.885$, $p = .376$. The funnel plot indicated trimming two studies from the left, resulting in $g$ value of -0.499 95% CI [-1.078, 0.080]. Given the limited age range and number of effect sizes, no moderator analyses were completed.

***Comparison of RD and CA.*** The overall mean effect size was significant, $g = -0.633$, 95% CI [-1.068, -0.198], $z(6) = -2.849$, $p = .004$. The between studies variation in $d$ values was not significant, $Q(6) = 25.948$, $p < .001$, $I^2 = 76.877$. After removing one outlier, the $g$

was -0.436, 95% CI [-0.659, -0.214]. The funnel plot indicated trimming two studies from left, resulting in *g* value of -0.863, 95% CI [-1.307, -0.418].

***Comparison of CA and RL.*** **Five studies, seven effects** The overall mean effect size was significant, $g = 0.561$, 95% CI [-0.248, 1.371], $z(6) = 1.360$, $p = .174$. The between studies variation in *d* values was not significant, $Q(6) = 84.509$, $p < .001$, $I^2 = 92.900$. Removing two outliers resulted in *g* of 0.669, 95% CI [0.137, 1.201], $z(4) = 2.464$, $p = .014$. The funnel plot indicated trimming one study to the right, resulting in *g* value of 1.061, 95% CI [0.096, 2.027].

## Discussion

The purpose of the present study was to provide a meta-analytic review of reading-level match group-comparison studies conducted in highly consistent European alphabetic orthographies examining the potential causes of dyslexia. We located a sufficient number of studies, 21 altogether, on phonological awareness, rapid naming, verbal short-term memory, and auditory temporal processing to calculate effect sizes. Magnocellular deficits, visual-spatial attention, visual attention, visual attention span, perceptual anchoring, temporal acuity, and skill automatization deficits were not considered in this meta-analysis as we did not locate enough studies meeting our inclusion criteria. We focused on studies with younger students because of complexities involved in using reading-level match design with older participants (see Deacon et al., 2008, for a discussion) and the difficulty of inferring causes with older participants with substantial experience with literacy (e.g., Huettig, Lachmann, Reis, & Petersson, 2018). Further, we examined whether age of the participants, age

difference between the groups, and the type of task affected the results. Finally, we also compared the groups with reading disabilities (RD) and their reading-level (RL) matched controls to chronological-age (CA) control groups where possible to verify that the tasks used across studies and the samples selected showed the expected differences favouring the CA groups.

We examined first differences in reading words and nonwords. Given that we included only studies with a reading-level matched control group and that 18 of the 21 included studies reported matching the RD and RL groups on one or more word reading tasks, we expected no differences in word reading measures with the exception of possible rate differences when matching was based on accuracy alone. To our surprise, we first observed a small, but significant, difference ($g$ = -0.132) between the RD and the RL groups on the matching tasks, and a medium ($g$ = -0.486 after controlling for outliers) difference favouring the RL group on word reading tasks not used for matching. Contrary to our expectations, the differences between the two groups were larger in word reading accuracy tasks than in word reading efficiency tasks that did not show significant differences. Typically, children learning to read a highly consistent orthography reach a high level of accuracy very quickly even if they have a reading disability, as indicated by ceiling effects by Grade 2 in multiple studies (e.g., Babayiğit & Stainthorp, 2007; Georgiou, Parrila, & Papadopoulos, 2008; Georgiou, Torppa, Manolitsis, Lyytinen, & Parrila, 2012). A closer look at the word reading accuracy results in the eight studies involved indicated that in six of them the accuracy results for the RL group show a possible ceiling effect and, as a result, restricted variability that likely inflated the

effect size estimates for the accuracy tasks. As an extreme example, Jimenez et al. (2005) reported accuracy means (and SDs) of 28.6 (1.26), 29.6 (0.62), and 29.8 (0.52) for the dyslexic, RL matched, and CA matched groups on a task with a maximum score of 30. Some of the accuracy matching (see Table 1) was based on tasks with substantial ceiling effects as well. These observations highlight the perils of using a single task for matching, a procedure that is open not only to the regression-to-the-mean effect (see Lundberg & Høien,1990, for an extended demonstration with somewhat older readers), but also to task specific variability resulting from differences between items, instructions, choice of strategy, and task difficulty. Thus, our first conclusion is that, as a group, the included studies did not include a proper reading-level matched control group, but a younger control group that happened to perform similarly to the older children with dyslexia in one reading task at the time of testing while differing significantly on other reading tasks. In light of this finding, we advise caution in interpreting both the deficit estimates and the accuracy vs. efficiency comparison results – perhaps the method of calculating effect sizes is poorly suited for analysing word reading accuracy results across studies that include many examples of severe variability limitations. Further, the observed differences in word reading results complicate the interpretation of nonword reading and phonological awareness results in particular as both of these can be reciprocally related to word reading (e.g., Perfetti, Beck, & Bell, 1987) or impacted by differences in matching procedures (e.g., van den Broeck & Geudens, 2012).

Nonword results were in line with expectations and in spite of the highly consistent grapheme-phoneme correspondences children in these orthographies are required to learn,

they do make errors and are slower when reading nonwords than when reading words. In our meta-analysis, children with dyslexia were reliably poorer than their reading-level matched controls. Again, the accuracy difference was slightly larger than the efficiency difference, albeit not significant, likely indicating ceiling effects in accuracy for the comparison group. Our meta-analytic techniques do not allow examination of why and on what kind of items children with dyslexia made errors, but three of the reviewed studies offered an interpretation of the observed nonword reading deficits (see also Wimmer & Schurz, 2010). Serrano and Defior (2008) attributed the poorer nonword reading of their participants to insufficient grapheme-phoneme correspondence knowledge. Jimenez and Ramirez (2002) noted that their participants made a significantly higher number of reversal, lexicalization, and omission errors than the reading-level match controls, whereas the opposite was true for phonological errors, defined as misapplication of context dependent phonological rules or accent rules. Wimmer (1996), in turn, suggested that the observed nonword reading deficit resulted from impaired efficiency in generating pronunciations and not from deficient grapheme-phoneme correspondence knowledge. Wimmer (1996) reasoned that if "syllables are assembled at a comparatively slow speed, as presumably is the case for dyslexic children, then there is obviously a higher chance that earlier assembled syllables fade from short-term memory and that interference from most recent syllables occur" (p. 87). It seems clear from the above that nonword reading errors are still relatively poorly understood and more research is required to better understand the processes underlying them and whether they are the same across the orthographies. We should also note that the nonword reading results could reflect a selection

bias. Van den Broeck and Geudens (2012) argued that when groups are matched based on word reading, there will likely be a nonword reading difference as the older children's word reading score will reflect greater familiarity with words, whereas the younger RL groups' performance will reflect better decoding skills. We will discuss this further below.

In terms of the cognitive skills, phonological awareness is the most commonly cited cause of dyslexia, and perhaps also the one with the most support for a causal status in English (however, see Castles & Coltheart, 2004; Huettig et al., 2018, for an alternative interpretation). At the same time, it has not been always associated with dyslexia in consistent orthographies, and when it is, only some of the participants seem to be affected (Georgiou, Papadopoulos, Zarouna, & Parrila, 2012; Tobia & Marzocchi, 2014). In the current meta-analysis, 20 of the 33 effect sizes were smaller than -0.20 favouring the RL group, seven fell in between -0.20 and 0.20, and six favoured the RD group. The overall effect size of -0.399 compares well to what Melby-Lervåg et al. (2012) reported for English-speaking subjects ($d$ = -0.49), although it is considerably smaller than $d$ = -0.83 they reported for the non-English sample, possibly reflecting the use of a different pool of languages examined and the lack of controlling for the effects of outliers (see introduction). The effect sizes for tasks requiring manipulation of syllables or rimes, simple phonemic awareness tasks, and complex phonemic awareness tasks were comparable in size, but the first was not significantly different from zero due to large variability across studies. That simple phonemic tasks showed the most robust differences between the RD and RL groups is contrary to the existing literature suggesting that frequent null results with phonological awareness tasks are likely due to too

simple tasks (e.g., Caravolas et al., 2005; Constantinidou & Stainthorp, 2009). Examination of the actual tasks administered across the studies does not show any clear trend; the simple phonological awareness tasks included a variety of matching, blending and segmentation tasks that produced larger differences than the supposedly more complex deletion, substitution, and spoonerism tasks.

Rapid naming, verbal short-term memory, and auditory temporal processing analyses showed no reliable differences between the dyslexia groups and their reading-level match controls. They did, however, show reliable differences between the chronological age-matched controls and children with dyslexia, and between older controls and the younger reading-level matched children. Thus, it is difficult to argue that the tasks used across the studies would not have been sensitive to individual differences. One possible interpretation of the results is that all of these tasks capture developmental changes that may be associated with reading acquisition but are not affected by reading experience that the reading-level match design supposedly controls. Rapid naming and verbal short-term memory analyses indicated further that there is a possible publication bias that if corrected, would result in significant differences favouring the RL groups. We doubt, however, that the assumptions underlying these analyses are entirely applicable to the contexts examined here.

Theoretically, our results are at least partially compatible with the phonological-deficit hypothesis that posits a specific deficit in some aspect of phonological processing to account for word reading difficulties. Further, our results suggest that such a deficit is most likely located at the level of phonemic processing. This conclusion is qualified by two caveats that

future research has to address. First, we observed a half a standard deviation difference in word reading skills (after removing the outliers) that may have contributed to 1/3$^{rd}$ standard deviation difference in phonological awareness tasks due to reciprocal connections between reading and phonological awareness. At least some proportion of the word reading differences resulted in using accuracy tasks that artificially deflated the variability (see above), and the same or similar tasks were also used to match the RD and RL groups in seven studies. Thus, the second caveat we have is the method of matching RD and RL groups seems to have significant effect on the results. For example, the mean difference in the phonological awareness tasks, $g$ = -1.142, 95% CI [-1.954, -0.330] for the studies that matched the two groups on the basis of reading accuracy is significantly larger than the mean difference, $g$ = -0.030, 95% CI [-0.330, 0.271] for the 11 studies that matched the participants of the basis of reading efficiency. Further, the three studies that matched the RD and RL groups on nonword reading produced four phonological awareness effect size estimates that varied from 0.03 to 1.44 favouring the RD group, thus possibly extending van den Broeck and Geudens' (2012) selection bias argument to include phonological awareness.

Given the variability in phonological awareness results, it is up to future research to examine how exactly does the matching criteria affect the cognitive profiles of the resulting RL groups. As the criteria for dyslexia (see Table 1) did not include phonological deficit in any of the studies, and was remarkably consistent across the studies, the differences in matching likely resulted in RL samples that varied in their cognitive skills, including phonological awareness. More specifically, the current studies do not allow determining

under what conditions phonological awareness tasks can be reliably used to identify a phonological awareness deficit in consistent orthographies, in which, among other things, orthographic information can be more readily used to support performance. In sum, to paraphrase Wimmer and Shurtz (2010), the phonological deficit explanation fared better than the rival accounts we examined, but whether it fared well enough to qualify as the cause of dyslexia is open to debate.

In terms of the cognitive level predictors we examined in this meta-analysis, multifactorial models (e.g., Lyytinen et al., 1998; Pennington, 2006; van Bergen et al., 2014) would predict that a significant deficit in any one of them could be associated with a reading deficit (possibly in the absence of protective factors in other levels); however, none are necessary *per se* and a clinically identifiable reading deficit could also result from an interaction between multiple subclinical deficits. While models of this kind are not yet formulated as testable theories, they can accommodate a wider range of results than core deficit theories, including diverging empirical results in cognitive, neurological (e.g., Protopapas & Parrila, 2018, for a discussion), and genetic (e.g., Carrion-Castillo et al., 2013) levels. As developmental systems theories, multifactorial etiological models also inherit the idea that an explanation of dyslexia requires understanding of the developmental system with its interacting risk and protective factors. Thus, they avoid the causality arguments reading-level match design theorists (see e.g., Backman et al., 1984; Bryant & Goswami, 1986; Goswami & Bryant, 1989; Jackson & Butterfield, 1989; Mamen et al., 1986; Vellutino & Scanlon, 1989) have engaged in by conceptualizing development as a sequential emergence

of new structural and functional properties and competencies at all levels of analysis (Gottlieb et al., 2006). This implies that a causal explanation of dyslexia must describe the developmental system that led to the observed outcome over some meaningful period of time. Accordingly, we can study components of the system, such as phonological awareness and rapid naming or individual genes, in relative isolation, but individual components neither explain nor cause (normal or abnormal) development in any meaningful sense without an account of the cognitive, physical, biological, and social factors ("developmental niche") that interact with and shape the components of interest over time. While several prominent authors have recently acknowledged the limitations of traditional models (e.g., Catts, 2017; Huettig et al., 2018), systems approaches in general have had little traction in dyslexia research

(however, see Morrison & O'Connor, 2017, for an example of a systems approach to reading development), perhaps because systems approaches pose formidable empirical and theoretical challenges the answers to which poorly match our dominant research traditions and presuppositions.

To conclude, we found qualified support for the phonological deficit hypothesis of dyslexia. At the minimum, our results suggest that more research is needed to establish the precise nature of the phonological deficit and under what circumstances it is reliably associated with a reading deficit. This, of course, in no way affects the value of studying the cognitive skills included in this meta-analysis and the many that were excluded due to the lack of studies; however, it does suggest limits to the causal arguments made on the basis of the results of those studies.

References

Ahissar, M. (2007). Dyslexia and the anchoring-deficit hypothesis. *Trends in Cognitive Sciences, 11*(11), 458–465.

Babayiğit, S. & Stainthorp, R. (2007). Preliterate phonological awareness and early literacy skills in Turkish. *Journal of Research in Reading, 30*, 394–413. doi: 10.1111/j.1467-9817.2007.00350.x

Backman, J. (1983). The role of psycholinguistic skills in reading acquisition: A look at early readers. *Reading Research Quarterly, 18*, 466-479. doi: 10.2307/747381

Backman. J. E., Mamen, M., & Ferguson, H. B. (1984). Reading level design: conceptual and methodological issues in reading research. *Psychological Bulletin, 96*, 560-568. doi: 10.1037/0033-2909.96.3.560

Badian, N., Duffy, F. H., Als, H., & McAnulty, G. B. (1991). Linguistic profiles of dyslexics and good readers. *Annals of Dyslexia, 41*, 221-245. http://www.jstor.org/stable/23768526.

Bishop, D. V. & Snowling, M. J. (2004). Developmental dyslexia and specific language impairment: Same or different? *Psychological Bulletin, 130*, 858–886. doi:10.1037/0033-2909.130.6.858

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive Meta-Analysis Version 2 [Computer software]. Englewood, NJ: Biostat.

Bosse, M. L., Tainturier, M. J., & Valdois, S. (2007). Developmental dyslexia: The visual attention span deficit hypothesis. *Cognition, 104,* 198–230. doi: 10.1016/j.cognition.2006.05.009

Bradley, L. & Bryant, P. (1978). Difficulties in auditory organization as a possible cause of reading backwardness. *Nature*, *271*, 746–747.

Brizzolara, D., Chilosi, A., Cipriani, P., Di Filippo, G., Gasperini, F., Mazzotti, S., et al. (2006). Do phonologic and rapid automatized naming deficits differentially affect dyslexic children with and without a history of language delay? A study of Italian dyslexic children. *Cognitive and Behavioral Neurology, 19*, 141-149. doi: 10.1097/01.wnn.0000213902.59827.19

Bryant, P. E. & Bradley, L. (1980). Why children sometimes write words which they do not read. In U. Frith (ed.), *Cognitive processes in spelling* (pp. 355 – 370). London: Academic Press.

Bryant, P. & Goswami, U. (1986). Strengths and weaknesses of the reading level design: A comment on Backman, Mamen, and Ferguson. *Psychological Bulletin, 100*, 101-103. doi: 10.1037/0033-2909.100.1.101

Caravolas, M. (2005). The nature and causes of dyslexia in different languages. In M. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 336-356). London, UK: Blackwell Publishing.

Caravolas, M., Vólin, J., & Hulme, C. (2005). Phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies: Evidence from

Czech and English children. *Journal of Experimental Child Psychology, 92*, 107-139. doi: 10.1016/j.jecp.2005.04.003

Carrion-Castillo, A., Franke, B., & Fisher, S. E. (2013). Molecular genetics of dyslexia: An overview. *Dyslexia*, *19*, 214–240. doi: 10.1002/dys.1464

Castles, A. & Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? *Cognition, 91,* 77–111.

Catts, H. (2017). Early identification of reading disabilities. In K. Cain, D. Compton, & R. Parrila (eds.), *Theories of reading development* (pp. 311-331). Amsterdam: Benjamins. doi: 10.1075/swll.15.18cat

Connor, C. M., & Morrison, F. J. (2017). Child characteristics by instruction interactions, literacy, and implications for theory and practice. In K. Cain, D. Compton, & R. Parrila (eds.), *Theories of reading development* (pp. 507-524). Amsterdam: Benjamins. doi 10.1075/swll.15.27con

*Constantinidou, M. & Stainthorp, R. (2009). Phonological awareness and reading speed deficits in reading-disabled Greek-speaking children. *Educational Psychology, 29*, 171-186. doi: 10.1080/01443410802613483

*Cuetos, F., Martinez-Garcia, C., & Suarez-Coalla, P. (2018). Prosodic perception problems in Spanish dyslexia. *Scientific Studies of Reading, 21*, 41-54. doi: 10.1080/10888438.2017.1359273

Davies, R., Cuetos, F. & Glez-Seijas, R. M. (2007). Reading development and dyslexia in a transparent orthography: A survey of Spanish children. *Annals of Dyslexia, 57*, 179-198. doi: 10.1007/s11881-007-0010-1

Deacon, S. H., Parrila, R., & Kirby, J. R. (2008). A review of evidence on morphological processing in dyslexics and poor readers. In G. Reid, A. Fawcett, F. Manis, & L. Siegel (eds.), *The SAGE Handbook of Dyslexia* (pp. 212-237). London: Sage Publications.

*Diamanti, V., Goulandris, N., Campbell, R., & Protopapas, A. (2018). Dyslexia profiles across orthographies differing in transparency: An evaluation of theoretical predictions contrasting English and Greek. *Scientific Studies of Reading, 22*, 55–69. doi: 10.1080/10888438.2017.1338291

Duval, S. & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463. doi:10.1111/j.0006-341X.2000.00455.x

Elliott, J. G. & Grigorenko, E. L. (2014). *The dyslexia debate*. Cambridge University Press.

Ford, D. H. & Lerner, R. M. (1992). *Developmental systems theory: An integrative approach*. Newsbury Park, CA: Sage.

Franceschini, S., Gori, S., Ruffino, M., Pedrolli, K., & Facoetti, A. (2012). A causal link between visual spatial attention and reading acquisition. *Current Biology, 22,* 814–819. doi: 10.1016/j.cub.2012.03.013

Gaab, N., Gabrieli, J. D. E., Deutsch, G. K., Tallal, P., & Temple, E. (2007). Neural correlates of rapid auditory processing are disrupted in children with developmental dyslexia and ameliorated with training: An fMRI study. *Restorative Neurology and Neuroscience, 25,* 295–310.

*Georgiou, G. K., Papadopoulos, T. C., Zarouna, E., & Parrila, R. (2012). Are auditory and visual processing deficits related to developmental dyslexia? *Dyslexia, 18*, 110-129. doi: 10.1002/dys.1439

Georgiou, G. & Parrila, R. (2013). Rapid naming and reading. In L. Swanson, K. Harris & S. Graham (eds.), *Handbook of learning disabilities* (2nd ed., pp. 169-185). New York: Guilford.

Georgiou, G. K., Parrila, R., & Papadopoulos, T. C. (2008). Predictors of word decoding and reading fluency in English and Greek: A cross-linguistic comparison. *Journal of Educational Psychology, 100*, 566–580. doi: 10.1037/0022-0663.100.3.566

*Georgiou, G. K., Protopapas, A., Papadopoulos, T. C., Skaloumbakas, C., & Parrila, R. (2010). Auditory temporal processing and dyslexia in an orthographically consistent language. *Cortex, 46*, 1330-1344. doi: 10.1016/j.cortex.2010.06.006

Georgiou, G., Torppa, M., Manolitsis, G., Lyytinen, H., & Parrila, R. (2012). Longitudinal predictors of reading and spelling across languages varying in orthographic consistency. *Reading and Writing: An Interdisciplinary Journal, 25*, 321–346. doi: 10.1007/s11145-010-9271-x

Goswami, U. (2002). Phonology, reading development, and dyslexia: A cross-linguistic perspective. *Annals of Dyslexia, 52*, 141-163. doi: 10.1007/s11881-002-0010-0

Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. *Trends in Cognitive Sciences, 15,* 3–10. doi: 10.1016/j.tics.2010.10.001

Goswami, U. & Bryant, P. (1989). The interpretation of studies using the reading level design. *Journal of Reading Behavior, 21*, 413-424. doi: 10.1080/10862968909547687

*Goswami, U., Wang, S., Cruz, A., Fosker, T., Mead, N., & Huss, M. (2010). Language universal sensory deficits in developmental dyslexia: English, Spanish, and Chinese. *Journal of Cognitive Neuroscience, 23*, 325–337. doi: 10.1162/jocn.2010.21453

Gottlieb, G. (1983). The psychobiological approach to developmental issues. In M. M. Haith & J. J. Campos (Eds.), *Handbook of child psychology* (4th ed., vol. 2, pp. 1-26). New York: John Wiley & Sons.

Gottlieb, G. (1997). *Synthesizing nature-nurture: Prenatal roots of instinctive behavior*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gottlieb, G., Wahlsten, D., & Lickliter, R. (2006). The significance of biology for human development: A developmental psychobiological systems view. In R. M. Lerner & W. Damon (Eds.), *Handbook of child psychology* (6 ed., Vol. 1, pp. 210-257). Hoboken, NJ: John Wiley & Sons.

*Gustafson, S. (2001). Cognitive abilities and print exposure in surface and phonological types of reading disability. *Scientific Studies of Reading, 5,* 351-375. doi: 10.1207/S1532799XSSR0504_03

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Holopainen, L., Ahonen, T., & Lyytinen, H. (2001). Predicting delay in reading achievement in a highly transparent language. *Journal of Learning Disabilities, 34*, 401-413. doi: 10.1177/002221940103400502

Huettig, F., Lachmann, T., Reis, A., & Petersson, K. M. (2018). Distinguishing cause from effect – many deficits associated with developmental dyslexia may be a consequence of reduced and suboptimal reading experience. *Language, Cognition and Neuroscience, 33* (3), 333-350. doi: 10.1080/23273798.2017.1348528

Jackson, N. E. & Butterfield, E. C. (1989). Reading-level-match designs: Myths and realities. *Journal of Reading Behavior, 21*, 387-412. doi: 10.1080/10862968909547686

*Jiménez, J. E. (1997). A reading-level matched study of phonemic processes underlying reading disabilities in a transparent orthography. *Reading and Writing: An Interdisciplinary Journal, 9*, 23-40. doi: 10.1023/A:1007925424563

*Jiménez, J. E., García, E., Ortiz, R., Hernández-Valle, R., Guzmán, R., Rodrigo, M., et al. (2005). Is the deficit in phonological awareness better explained in terms of task differences or effects of syllable structure? *Applied Psycholinguistics, 26*, 267-283. doi: 10.1017/S0142716405050174

*Jiménez, J. E. & Ramírez, G. (2002). Identifying subtypes of reading disability in the Spanish language. *The Spanish Journal of Psychology, 5*, 3-19. doi: 10.1017/S1138741600005783

* Jiménez, J. E., Rodríguez, C., & Ramírez, G (2009). Spanish developmental dyslexia: Prevalence, cognitive profile, and home literacy experiences. *Journal of Experimental Child Psychology, 103*, 167-185. doi: 10.1016/j.jecp.2009.02.004

Katzir, T., Shaul, S., Breznitz, Z., & Wolf, M. (2004). The universal and the unique in dyslexia: A cross-linguistic investigation of reading and reading fluency of Hebrew- and English-speaking children with reading disorders. *Reading and Writing: An Interdisciplinary Journal, 17*, 739-768. doi: 10.1007/s11145-004-2655-z

Kirby, J. R., Georgiou, G. K., Martinussen, R., & Parrila, R. (2010). Naming speed and reading: From prediction to instruction. *Reading Research Quarterly, 45*, 341–362. doi: 10.1598/RRQ.45.3.4

Laasonen, M., Service, E., & Virsu, V. (2001). Crossmodal temporal order and processing acuity in developmentally dyslexic young adults. *Brain and Language, 80*, 340-354. doi:10.1006/brln.2001.2593

*Landerl, K. & Wimmer, H. (2000). Deficits in phoneme segmentation are not the core problem of dyslexia: Evidence from German and English children. *Applied Psycholinguistics, 21*, 243-262.

Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German-English comparison. *Cognition, 63*, 315-334. doi: 10.1016/S0010-0277(97)00005-X.

*Loizidou-Ieridou, N. (2011). Reading, spelling and phonological ability in 9 to 10 year old Greek-speaking dyslexic children. In conference proceeding Σχολές Επιστημών της

Αγωγής: ο ρόλος τους στις προκλήσεις της σύγχρονης κοινωνίας" (pp. 357-374). Frederick University, Nicosia, Cyprus.

Lum, J. A. G., Ullman, M. T., & Conti-Ramsden, G. (2013). Procedural learning is impaired in dyslexia: Evidence from a meta-analysis of serial reaction time studies. *Research in Developmental Disabilities, 34*, 3460-3476. doi: 10.1016/j.ridd.2013.07.017

Lundberg, I. & Høien, T. (1990) Patterns of information processing skills and word recognition strategies in developmental dyslexia. *Scandinavian Journal of Educational Research, 34* (3), 231-240. doi: 10.1080/0031383900340305

Lyytinen, H., Ahonen, T., Aro, M., Aro, T., Närhi, V., & Räsänen, P. (1998). Learning disabilities: A view of developmental neuropsychology. In R. Licht, A. Bouma, W. Slot, & W. Koops (Eds.), *Child neuropsychology: Reading disability and more*. Delft, NL: Eburon.

Mamen, M., Ferguson, H. B., & Backman, J. E. (1986). No difference represents a significance finding: The logic of the reading level design. A response to Bryant and Goswami. *Psychological Bulletin, 100*, 104-106. doi: 10.1037/0033-2909.100.1.104

McBride-Chang, C., Tong, X., & Mo, J. (2015). Developmental dyslexia in Chinese. In W. S.-Y. Wang & C. Sun (eds.)*, The Oxford handbook of Chinese linguistics*. doi: 10.1093/oxfordhb/9780199856336.001.0001

Melby-Lervåg, M., Lyster, S.-A., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin, 138*, 322-352. DOI 10.1037/a0026744

Nicolson, R. I., Fawcett, A. J., & Dean, P. (2001). Developmental dyslexia: The cerebellar deficit hypothesis. *Trends in Neurosciences, 24* (9), 508–511. doi:10.1016/S0166-2236(00)01896-8

*Nikolopoulos, D. (1999). Cognitive and linguistic predictors of literacy skills in the Greek language. The manifestation of reading and spelling difficulties in a regular orthography. PhD thesis, University College London.

*Nikolopoulos, D., Goulandris, N., & Snowling, M. (2003). Developmental dyslexia in Greek. In N. Goulandris (Ed.), *Dyslexia in different languages* (pp. 53–67). London: Whurr.

Noordenbos, M. W., Segers, E., Serniclaes, W., & Verhoeven, L. (2013). Neural evidence of the allophonic mode of speech perception in adults with dyslexia. *Clinical Neurophysiology, 124,* 1151–1162. doi: 10.1016/j.clinph.2012.12.044

Papadopoulos T. C., Georgiou, G. K., & Kendeou, P. (2009). Investigating the double-deficit hypothesis in Greek: Findings from a longitudinal study. *Journal of Learning Disabilities, 42*, 528-547. doi: 10.1177/0022219409338745

Parrila, R. & Protopapas, A. (2017). Dyslexia and word reading problems. In K. Cain, D. Compton, & R. Parrila (eds.), *Theories of reading development* (pp. 333-358). Amsterdam: Benjamins. doi 10.1075/swll.15.19par

Paulesu, E., Démonet, J.-F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., et al. (2001). Dyslexia: Cultural diversity and biological unity. *Science, 291*, 2165-2167. doi: 10.1126/science.1057179

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition, 101*, 385-413. doi: 10.1016/j.cognition.2006.04.008

Pennington, B. F., Santerre-Lemon, L., Rosenberg, J., McDonald, B., … & Olson, R. K. (2012). Individual prediction of dyslexia by single versus multiple deficit models. *Journal of Abnormal Psychology, 121*, 212-224. doi: 10.1037/a0025823

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

Protopapas, A. (2014) From temporal processing to developmental language disorders: Mind the gap. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369* (1634), 1-11. doi: 10.1098/rstb.2013.0090

Protopapas, A. & Parrila, R. (2018). Is dyslexia a brain disorder? *Brain Sciences, 8*, 61. doi:10.3390/brainsci8040061

Serniclaes, W., Van Heghe, S., Mousty, P., Carr., R., & Sprenger-Charolles, L. (2004). Allophonic mode of speech perception in dyslexia. *Journal of Experimental Child Psychology, 87*, 336–361. doi: 10.1016/j.jecp.2004.02.001

*Serrano, F. & Defior, S. (2008). Dyslexia speed problems in a transparent orthography. *Annals of Dyslexia, 58*, 81-95. doi: 10.1007/s11881-008-0013-6

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143 – 174. doi: 10.1348/000712603321661859

Seymour, P. H. K. & Porpodas, C. D. (1980). Lexical and non-lexical processing of spelling in dyslexia. In U. Frith (ed.), *Cognitive processes in spelling* (pp. 443 – 473). London: Academic Press.

Smythe, I. & Everatt, J. (2004). Dyslexia: A cross-linguistic framework. In I. Smythe, J. Everatt, & R. Salter (Eds.), *International Book of Dyslexia* (pp. 1-29). Chichester, UK: John Wiley & Sons.

Snowling, M. J. (2000). *Dyslexia* (2nd ed.). Oxford, UK: Blackwell.

*Soriano, M. & Miranda, A. (2010). Developmenal dyslexia in a transparent orthography: A study of Spanish dyslexic children. *Literacy and Learning, 23*, 95-114. doi:10.1108/S0735-004X(2010)0000023006

Stanovich, K. E. (1988). Explaining the differences between dyslexic and the garden variety poor readers: The phonological-core variable-difference model. *Journal of Learning Disabilities, 21*, 590–604, 612. doi: 10.1177/002221948802101003

Stein, J. & Walsh, V. (1997). To see but not to read; the magnocellular theory of dyslexia. *Trends in Neuroscience, 20,* 147-152. doi: 10.1016/S0166-2236(96)01005-3

*Surányi, Z., Csépe, V., Richardson, U., Thompson, J. M., Honbolygó, F., & Goswami, U. (2009). Sensitivity to rhythmic parameters in dyslexic children: A comparison of Hungarian and English. *Reading and Writing: An Interdisciplinary Journal, 22*, 41–56. doi: 10.1007/s11145-007-9102-x

Tallal, P. (1980). Auditory temporal perception, phonics, and reading disabilities in children. *Brain and Language, 9*, 182–198. doi: 10.1016/0093-934X(80)90139-X

*Talli, J., Sprenger-Charolles, L., & Stavrakaki, S. (2016). Specific language impairment and developmental dyslexia: What are the boundaries? Data from Greek children. *Research in Developmental Disabilities, 49-50*, 339-353. doi: 10.1016/j.ridd.2015.12.014

*Tobia, V. & Marzocchi, G. M. (2014). Cognitive profiles of Italian children with developmental dyslexia. *Reading Research Quarterly, 49*, 437-452. doi: 10.1002/rrq.77

Torppa, M., Parrila, R., Niemi, P., Poikkeus, A.-M., Lerkkanen, M.-K., & Nurmi, J.-E. (2013). The double deficit hypothesis in the transparent Finnish orthography: a longitudinal study from kindergarten to grade 2. *Reading and Writing: An Interdisciplinary Journal, 26,* 1353-1380. doi: 10.1007/s11145-012-9423-2

van Bergen, E., van der Leij, A. & de Jong, P. F. (2014). The intergenerational multiple deficit model and the case of dyslexia. *Frontiers in Human Neuroscience, 8*, 346. doi: 10.3389/fnhum.2014.00346

Van den Broeck, W. & Geudens, A. (2012). Old and new ways to study characteristics of reading disability: The case of the nonword-reading deficit. *Cognition, 65*, 414-456. doi: 10.1016/j.cogpsych.2012.06.003

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry, 45*, 2-40. doi: 10.1046/j.0021-9630.2003.00305.x

Vellutino, F. R. & Scanlon, D. M. (1989). Some prerequisites for interpreting results from reading level matched designs. *Journal of Reading Behaviour, 21*, 361-385. doi: 10.1080/10862968909547685

Vidyasagar, T. R. & Pammer, K. (2010). Dyslexia: A deficit in visuo-spatial attention, not in phonological processing. *Trends in Cognitive Sciences, 14*, 57–63. doi: 10.1016/j.tics.2009.12.003

*Wimmer, H. (1993). Characteristics of developmental dyslexia in a regular writing system. *Applied Psycholinguistic, 14*, 1-33. doi: 10.1017/S0142716400010122

*Wimmer, H. (1996). The nonword reading deficit in developmental dyslexia. *Journal of Experimental Child Psychology, 61*, 80-90. doi: 10.1006/jecp.1996.0004.

Wimmer, H. & Schurz, M. (2010). Dyslexia in regular orthographies: Manifestation and causation. *Dyslexia, 16*, 283-299. doi: 10.1002/dys.411

Wolf, M. & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology, 91*, 415–438. doi: 10.1037/0022-0663.91.3.415

*Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Körne, G. (2003). Developmental dyslexia in different languages: Language-specific or universal. *Journal of Experimental Child Psychology, 86*, 169-193. doi: 10.1016/S0022-0965(03)00139-5.

Zoubrinetzky, R., Bielle, F., & Valdois, S. (2014). New insights on developmental dyslexia subtypes: Heterogeneity of mixed reading profiles. *PloS one, 9*, e99337. doi: 10.1371/journal.pone.0099337