©1994 International Reading Association

Marinus H. van IJzendoorn
Adriana G. Bus
Leiden University, The Netherlands

Meta-analytic confirmation of the nonword reading deficit in developmental dyslexia

In a recent narrative review, Rack, Snowling, and Olson (1992) concluded that strong evidence exists for the phonological deficit hypothesis in explaining severe word reading and spelling problems that cannot be accounted for by sensory or neurological damage, lack of educational opportunity, or low intelligence. The phonological deficit hypothesis states that in these instances of dyslexia there is a highly specific deficit in the phonological language domain, which ultimately leads to problems in reading and spelling. Dyslexics are supposed to differ from normal readers in those qualitative aspects of reading that emphasize phonological processes. An alternative interpretation is the developmental lag or delay hypothesis. According to this hypothesis, normal and dyslexic readers differ only in the speed of development, and are equal in terms of qualitative aspects of reading style. The developmental lag hypothesis implies that dyslexics will perform poorly on phonological reading tests, but not more so than younger readers at the same reading stage who develop in a normal way. The phonological deficit hypothesis, on the other hand, predicts that dyslexics and (younger) normal readers may have the same word recognition ability but will differ strongly in phonological skills. Rack et al. (1992) describe and analyze a series of studies of nonword reading in dyslexics and reading-level-matched normal readers that may be considered as crucial tests for the validity of the deficit and the delay hypotheses.

In these studies, Snowling's (1980, 1981) paradigm of the nonword reading task has been applied in a variety of ways to assess phonological skill relatively independently of reading ability. Dyslexics are matched with normal readers in terms of reading level. The reading-level-matched design is used to compare dyslexics' performance on a nonword task with younger normal readers' performance on the same phonological skill measure. The design controls for differences in reading abilities that might influence the children's performance on the nonword task. The effectiveness of the design in reaching this goal depends, of course, on the adequacy of the matching procedure. In most studies the matching of dyslexics and normal readers is checked by a word recognition test that should show only minimal differences between normal and dyslexic readers. Rack et al. (1992) scrutinized all pertinent published studies using the nonword paradigm in the context of the reading-level-match design. Because the majority of studies showed (a) significant differences in nonword processing between dyslexics and normal readers against the background of (b) equivalence of word recognition abilities in both groups, the authors were convinced that there is "extremely strong evidence for the phonological deficit hypothesis" (p. 49). Furthermore, they analyzed in depth the causes of absence of phonological skill differences in about a third of the studies that seemed to contradict the deficit hypothesis, and pointed to several alternative hypotheses in terms of measures, designs,

and subjects characteristics Their work is a sublime example of a thorough narrative review taking stock of a decade of research on an important dimension of dyslexia

Although the authors exhaust the possibilities of the narrative review in an excellent way, a quantitative meta-analysis may supplement their approach for the following reasons First, a meta-analysis allows for a quantitative estimate of the overall effect size of a series of studies In our case, we may be able to quantify precise ly what the difference in phonological skills between dyslexics and normal readers is as well as test the adequacy of the reading-level matching procedure These are crucial statistics not only for testing the phonological deficit hypothesis, but also for determining how much we still do not know Second, a meta-analysis allows for a quantitative estimate of the stability of the combined probability level Rack et al (1992) rely on published studies, and the meta-analysis yields an estimate of the hypothetical number of unpublished studies with null results necessary to undermine the overall outcome Third, a meta-analysis describes the variability in study results, and tests for homogeneity of the set of pertinent studies Rack et al (1992) discriminate between the subset of studies finding significant nonword reading deficits in dyslexic readers and the subset of studies not finding significant differences, a meta-analysis might formally test whether the two subsets of studies have indeed been taken from different populations Fourth, whether or not a particular study showed a significant outcome may depend more on (restricted) sample size and chance than on reality From a meta-analytic perspective, studies showing an (insignificant) trend in the expected direction add to the combined probability level and effect size Fifth, a review should focus on inconsistent results and should suggest alternative hypotheses for unexpected outcomes In a narrative review, however, only speculations about factors explaining differences in results between studies are possible In a meta-analysis, alternative hypotheses can be tested in the formal sense Meta-analysis allows for testing the factors supposed to contribute to the variability of effect sizes in separate studies, on the basis of characteristics of those studies In this sense, a meta-analysis provides exactly the formal hypothesis testing that Rack et al (1992, p 49) explicitly asked for, and at the same time makes use of the large database on hand

In our meta-analysis, we will test the following hypotheses, all of which are derived from the Rack et al (1992) review

1 Do dyslexics and normal readers differ in terms of phonological skill despite equivalent word recognition abilities, and, if so, how large is the difference?

2 Does age—in particular, age of the matched normal readers—explain why some studies did not reveal any difference in phonological skill between dyslexics and normal readers? Rack et al (1992) hypothesized that 7-year-old readers might be prematurely exposed to tests for decoding unfamiliar letter strings, and therefore experience the developmentally normal difficulty with nonwords diminishing the nonword difference between normal and dyslexic readers

3 Is the kind of nonwords used to assess phonological skill related to the outcome of the studies? If nonwords are phonologically simple (e g monosyllabic) and if nonwords are highly visually similar to real words, they might not tap the phonological processing as sensitively as they would for more complex and dissimilar nonwords Studies using more extreme nonwords might yield larger differences between normal and dyslexic readers

4 In reading level-match designs the type of reading test used to match dyslexics and normal readers might explain variability between studies Tests involving oral reading of connected text for example might be measuring comprehension level instead of word recognition level and therefore obfuscate potential phonological differences between dyslexics and normal readers

5 In the comparison between dyslexics and normal readers differences in verbal intelligence should be minimal The phonological deficit hypothesis emphasizes the specificity of the reading deficit The adequacy and type of intelligence match between dyslexics and normal readers might therefore be important In particular, it is hypothesized that a close match on verbal intelligence is related to larger nonword reading differences

6 Phonological skill should not be considered to be a stable trait, and its sensitivity to special remediation has been established According to Rack et al (1992) more experience with special reading programs might lead to less obvious differences between trained dyslexics and normal readers in nonword reading ability

We tested these hypotheses by a quantitative meta-analysis of the same studies on phonological skill differences that Rack et al (1992) selected for their narrative review In this respect, our meta-analysis can be considered as a replication and extension of their seminal narrative review

Method

Database

The studies included in this meta-analysis were taken from Rack et al s (1992) review Two selection criteria were applied (a) Nonword reading had to have

Table 1 Characteristics of the studies on the nonword reading deficit

Study	Percentage of dyslexics	Age of dyslexics (months)	Age of normals (months)	7 year olds included	Reading test used	IQ test used	Nonwords simple (one syllable)	Nonwords similar to real words	Special program[a]
Snowling (1980)	33	145	114	no	Schonell	PPVT	yes	no	yes
Snowling (1981)	48	161	106	no	Schonell	PPVT	no	no	no
Baddeley et al (1982)	50	154	119	no	Schonell	WISC R/ Terman[a]	yes	yes	no
Kochnower et al (1983)	50	123	96	no	DST GE	PPVT	no	yes	no
DiBenedetto et al (1983)	50	123	96	no	DST GE	PPVT	no	yes	no
Olson et al (1985)	50	184	122	no	PIAT	WISC R	yes	no	yes
Siegel & Ryan (1988)	39	150	108	yes	WRAT GE[b]	PPVT	yes	no	yes
Manis et al (1988)	56	142 5	104 5	no	WRMT[b]	WISC R[m]	no	no	no
Hollingsworth & Johnston (1988)	50	102	86	yes	BAS[n]	WISC R[m]	no	no	yes
Olson et al (1989)	50[p]	187	124	no	PIAT	WISC R	no	no	yes
Beech & Harding (1984)	56	119	86 5	yes	Schonell	Raven	no	no	yes
Frauenheim & Hirsch Pesck (1985)	50	141	102	no	WRMT	PPVT	no	yes	no
Vellutino & Scanlon (1987)	50	144	94	no	Gilmore[n]	Slosson[q]	no	no	yes
Szeszulski & Manis (1987)	73	124	86	yes	Gilmore	WISC R[m]	no	no	no
Szeszulski & Manis (1987)	43	158	107	no	Gilmore	WISC R[n]	no	no	no
Johnston et al (1987)	50	102	86	yes	BAS	BAS[o]	no	yes	yes
Johnston et al (1987)	50	134	106	no	BAS	BAS	no	yes	yes
Baddeley et al (1988)	48	143	103	no	n a	WISC/ RAVEN[w]	yes	yes	no

7 year old children included in the normal comparison sample Monosyllabic nonwords Nonwords highly similar to real words [a] Involvement of subjects in special reading programs explicitly indicated Schonell Graded Word Recognition Test Peabody Picture Vocabulary Test [e] Wechsler Intelligence Scales for Children Revised and Terman Intelligence Test Decoding Skills Test Grade Equivalent Peabody Individual Achievement Test Word Recognition Wechsler Intelligence Scales for Children Revised Wide Range of Achievement Test Word Recognition Grade Equivalent Woodcock Word Identification Test [m] WISC R short form [n] BAS Word Reading Test Raven's Coloured Progressive Matrices [p] Gilmore Oral Reading Test Grade Equivalent Slosson Intelligence Test Low reading age subgroup High reading age subgroup Younger group British Ability Scales Older group WISC for dyslexics RAVEN for normal groups

been used to assess phonological reading skill and (b) the studies had to be based on the reading-level match design The authors included only published papers and do argue against a publication bias or the file drawer problem (Rosenthal 1991) In this research domain, null results would be as valuable as significant results because null results support the alternative developmental delay hypothesis (p 40) For the 16 studies included, we retrieved the appropriate test statistics (such as p r t, F) in one of the following ways (a) The test statistic was explicitly reported in the study, (b) the study provided means and standard deviations for the nonword reading test and for the word recognition test and we computed the t-statistic from these data, or (c) the study only provided an estimate of the significance level (e g , the difference in the nonword test between dyslexics and normal readers was [not] significant), and we included a conservative estimate (no significant effect $p = .50$, significant effect $p = .05$) In Table 2, the superscripts a, b, c and d are used to indicate which method had to be applied In some studies (Siegel & Ryan, 1988 Vellutino & Scanlon, 1987), results were reported on the level of five subgroups In these cases, we performed separate meta-analyses on the subgroups within these studies to compute an overall probability level which was included in the final meta analysis In some cases, only two sub groups were described (Johnston, Rugg, & Scott, 1987, Szeszulski & Manis 1987), these were included separately in the meta analysis In these latter cases, information on predictor variables would have been deleted if subgroups had been combined in advance

Predictors

The following predictor variables were derived from the studies

Age This includes age of dyslexics, age of normal readers, and the age difference between the two groups, furthermore, we used a separate variable indicating whether or not a specific study included 7-year-old normal readers (Hypothesis 2)

Nonword test The nonword tests used in the studies were analyzed in two dimensions complexity and similarity of the nonwords included in the test Complexity was defined as the use of nonwords with more than one syllable, and similarity was defined as the visual correspondence with real words, in particular the change of one (similarity) or more (difference) letters in a real word to create a nonword (Hypothesis 3)

Table 2 Significances and effect sizes per (sub-)sample for the nonword reading tests

				Significance		Effect size			
Study	Statistic	(*df*)	*N*	*Z*	*p*	Z_r	*r*	*r*	*d*
Snowling (1980)	$t = 2.79^a$	(52)	54	2.68	.004	.38	.36	.13	.77
Snowling (1981)	$t = 7.61$	(38)	42	2.62	.004	.43	.41	.17	.90
Baddeley et al (1982)	$p = .02^a$		30	2.05	.020	.39	.38	.14	.81
Kochnower et al (1983)	$t = 3.17$	(38)	40	2.97	.002	.49	.46	.21	1.03
DiBenedetto et al (1983)	$t = 2.74^b$	(38)	40	2.60	.005	.43	.41	.16	.89
Olson et al (1985)	$p = .05$		100	1.65	.050	.17	.16	.03	.33
Siegel & Ryan (1988)	$p = .0001^b$		110	3.72	.0001	.37	.35	.13	.76
Manis et al (1988)	$p = .05$		90	1.65	.050	.18	.17	.03	.35
Holligan & Johnston (1988)	$t = 3.11$	(38)	40	2.92	.002	.49	.45	.20	1.01
Olson et al (1989)	$p = .001^a$		115	3.09	.001	.30	.29	.08	.60
Beech & Harding (1984)	$t = 1.51^b$	(99)	101	1.53	.063	.15	.15	.02	.31
Treiman & Hirsch Pasek (1985)	$t = .057$	(72)	74	.060	.480	.01	.01	.00	.01
Vellutino & Scanlon (1987)	$p = .202^d$		150	.830	.202	.07	.07	.00	.14
Szeszulski & Manis (1987)	$t = 2.69^a$	(49)	51	2.58	.005	.38	.36	.14	.77
Szeszulski & Manis (1987)[f]	$p = .50^c$		35	.000	.500	.00	.00	.00	.00
Johnston et al (1987)[e]	$t = .97$	(38)	40	.960	.170	.16	.16	.02	.31
Johnston et al (1987)[b]	$t = 2.79$	(38)	40	2.64	.004	.44	.41	.17	.91
Baddeley et al (1988)	$p = .50^c$		31	.000	.500	.00	.00	.00	.00

[a] Presented in the study [b] Derived from means and standard deviations in the study [c] Conservative estimates [d] Based on Fisher's meta-analysis on five subgroups [e] Low reading age [f] High reading age [g] Younger group [h] Older group

Type of reading test We divided the reading tests used to match dyslexics with normal readers into two groups: those studies using the Gilmore Oral Reading Test or the Woodcock Word Identification Test (WRMT) (supposed to be less adequate, Rack et al 1992), and those studies using another reading test (Hypothesis 4).

Type of intelligence test The application of a purely verbal intelligence test such as the Peabody Picture Vocabulary Test (PPVT) is supposed to create a better match between dyslexics and normal readers than mixed verbal/performance tests or tests measuring only performance. Therefore, the studies were divided into two groups: those applying and those not applying the PPVT. To measure the adequacy of the intelligence matching we also derived the mean difference in intelligence between the dyslexic and normal reader groups (Hypothesis 5).

Special program In some studies it was reported that dyslexic subjects were recruited from special programs or units, in other studies it was not reported whether dyslexic subjects attended special schools or not (Hypothesis 6).

Besides these theoretically derived predictors, we also included some common predictors, such as sample size and publication year. We also included a variable Table, indicating whether, according to Rack et al (1992), the study belonged to the group of studies confirming the deficit hypothesis or to the group of studies with a null result. In Table 1, most predictors are included in Table 2 sample size is presented along with basic meta analytic results.

Meta analytic procedures

The unit of analysis in a single primary-level study is the subject, the unit of analysis in a meta-analysis of several primary-level studies is the outcome of those studies. Because of this fundamental difference in unit of analysis, meta-analysis has to apply a different set of statistical techniques. These techniques should, for example, take into account the fact that data points in meta-analysis are usually based on different sample sizes, and therefore may lack the homogeneity of variance required for the conventional statistics (Hedges & Olkin, 1985, Mullen, 1989, Rosenthal 1991). In our meta-analysis, the statistical tests of the studies under consideration were transformed to a few common metrics: the standard normal deviate (Z) and probability value (p) for significance level, and the correlation coefficient (r) and Fisher's Z for effect size. The standardized difference between the means of two groups, in our case the dyslexic and the normal group, was also computed (d).

On the basis of these common metrics, the following meta-analytic procedures were applied (Mullen, 1989):

1. We combined significance levels and effect sizes with the weighted Stouffer (1949) method. The formula for combining significance levels is

$$Z = \frac{\Sigma\, w_j Z_j}{\sqrt{\Sigma\, w_j^2}}$$

where w_j = sample sizes of the studies, Z_j = Z associated with significance levels of the studies.

The formula for combining effect sizes is:

$$\text{Fisher } Z = \frac{\Sigma\, w_i \text{ Fisher } Z}{\Sigma\, w_i}$$

where w_i = sample sizes of the studies, Fisher Z = Fisher Z associated with the effect sizes of the studies.

2. Tests for homogeneity of study results show whether study results might have been sampled from different populations. First, a test for homogeneity of significance levels was applied, based on the following formula:

$$X^2_{(k-1)} = \Sigma\, (Z_j - Z)^2$$

Second, the following formula for the homogeneity test of effect sizes was used:

$$X^2_{(k-1)} = \Sigma\, (N_j - 3)\, (\text{Fisher } Z - \text{Fisher } \bar{Z})^2$$

where k = number of studies included in the meta-analysis.

Third, a disjoint cluster analysis of effect sizes (Hedges & Olkin, 1985) was carried out, based on the following statistic:

$$U = \left(\Sigma\, \frac{\sqrt{N_j - 3}}{k}\right) \quad \text{Fisher } Z_j$$

The differences between rank-ordered and adjacently ranked Us are then tested against a preset significance level (in our case α = .05), and the set of studies is divided into significantly different subsets.

3. To estimate the probability that the variability of the p-values of the included studies can be significantly explained by the predictor variables, we used the following formula:

$$Z = \frac{\Sigma\, \lambda_j Z_j}{\sqrt{\Sigma \lambda_j{}^2}}$$

where λ_j = contrast weight assigned to the results of study j.

For the prediction of variability in effect sizes the following formula was used:

$$Z = \frac{\Sigma\, \lambda_j \text{ Fisher } Z_j}{\sqrt{\Sigma\, \frac{(\lambda_j{}^2)}{(N_j - 3)}}}$$

We performed a meta-analysis on nonword reading ability and on word recognition ability. Two studies were excluded from the second analysis because of missing data (Snowling, 1981; Vellutino & Scanlon, 1987). In reporting the results of our meta-analysis we will emphasize the effect size (r, d, or Fisher Z) as the most important indicator of the outcome of the study. The limited set of studies did not allow for the testing of a multivariate model of the predictors' (interactive) effects on the outcome of the studies. We will, however, use a standard alpha level as well as a Bonferronized alpha level to protect against capitalizing on chance. Both approaches will be used in our analysis to avoid overly conservative analyses and to leave room for exploration of interesting trends. The analyses were performed using Mullen's (1989) statistical package *Advanced BASIC Meta-Analysis*.

Results

Combined significance levels and effect sizes

In Table 2, the basic meta-analytic statistics of the studies are described. The studies included 1,183 subjects, about half of whom were dyslexic individuals. The effect sizes on the nonword test ranged from d = .00 to d = 1.03, and no negative effect sizes, indicating that dyslexics perform better than normal readers on the nonword task, were reported. The overall combined effect size for nonword reading ability was d = .48, which is comparable to a Fisher Z = .24, and a correlation coefficient r = .24. The combined probability level was 5.557 E-13 (Z = 7.25). In other words, the difference in phonological skill between dyslexic individuals and matched normal readers amounted to half a standard deviation, which was a highly significant result. The number of unretrieved or future studies averaging null results required to bring the combined probability level down under α = .05 is 423. This number of studies is four times the tolerance level of $5k + 10$ (where k = the number of studies included in the meta-analytic database; Rosenthal, 1991).

Table 3 Categorical predictors and combined significance and effect sizes for nonword reading

	Significance		Effect size			Comparison	
	Z	p_1	Z_{Fisher}	r	d	Z	p_1
7-year-olds							
Included	5.07	000	.30	.29	.60		
Excluded	5 38	000	22	21	.44	76	22
Nonword test							
Simple (one syllable)	4 76	000	.28	.27	.56		
Complex	5.59	.000	.23	22	.46	.11	46
Nonword test							
Similar to real words	3.69	000	.25	.24	.50		
Different	6 33	000	24	23	48	14	44
Reading test							
Gilmore/WRMT	2 05	020	.11	.11	23		
Other	7.68	.000	.30	.29	.62	2 77	003
PPVT							
Not applied	4 97	000	.19	.18	.37		
Applied	6.27	.000	.41	39	84	2.76	.003
Special program							
Yes	5 00	.000	25	.25	.51		
No	5.44	.000	.23	.23	.47	03	.49

The overall combined effect size for word recognition was $d = -.02$ (Fisher $Z = -.01$; $r = -.01$), with a standard normal deviate $Z = .52$, $p_1 = .30$ for the combined probability levels. The dyslexic readers did not differ from the matched normal readers on word recognition ability.

Homogeneity

The homogeneity of the significance levels was tested: X^2 ($df = 17$) = 22.46, $p = .17$. The chi-square for the homogeneity test of the effect sizes was X^2 ($df = 17$) = 27.09, $p = .057$. The disjoint cluster analysis did not yield significantly separate clusters of studies ($\alpha = .05$). There is no reason to assume that studies were derived from different populations. A comparison of combined effect sizes for studies that found a nonword reading deficit in dyslexic readers versus studies that did not seem to find such a deficit (Rack et al., 1992, Table 2) showed a significant standard normal deviate, $Z = 3.09$ ($p_1 = .001$). Combined effect size for studies finding a deficit was $d = .66$; combined effect size for studies that were supposed not to have found a nonword reading deficit in dyslexics was: $d = .27$, with a combined probability level of .005. Even when separate studies do not find a significant phonological deficit, their meta-analytic combination shows this deficit to be present.

Prediction

Although the study results did not appear to be heterogeneous, the variability in effect sizes of the studies is large enough to warrant trying to explain this variability on the basis of the predictor variables. In Table 3, the relevant statistics for categorical predictors are presented. Statistics for continuous predictors are given in the text.

Age did not predict variability in study results. Whether 7-year-old normal readers were included or not did not make a significant difference for the combined effect sizes in the two subsets of studies ($p_1 = .22$) Furthermore, the continuous variable age of normal readers did not predict variability in effect sizes either ($Z = .03$; $p_1 = .49$). The difference in age between the dyslexic and the normal group, however, was significantly related to the effect sizes. The correlation of age difference with the Fisher Z of each study was –.34, indicating that larger age differences were associated with smaller effect sizes (the standard normal deviate for the effect size of age difference was $Z = 1.76$, $p_1 = .04$).

The type of nonword test did not make a difference for the effect sizes (see Table 3). Whether or not simple (monosyllabic) or complex nonwords were used, or whether or not nonwords similar to real words were applied, did not determine the size of the effects of the studies involved.

The type of reading test used to match the dyslexic subjects with normal readers, however, did make a significant difference (see Table 3). As expected by Rack et al. (1992), studies using the Gilmore (words in context) or WRMT (regular words) showed a much smaller combined effect size than studies using other reading tests ($p_1 = .003$). Comparing the dyslexic and younger normal reader groups on the word recognition test, we also found that if dyslexic subjects scored lower on the word recognition test than the matched normal readers group, the dyslexic individuals had a relatively low score on the nonword reading test as well. To quantify this relation, the Fisher Z scores for the word recognition difference were correlated with the Fisher Z scores for the nonword reading differences ($r = -.37$, $Z = 1.77$, $p_1 = .04$, $N = 16$).

The type of intelligence test used in the matching procedure was also related to the effect sizes (see Table 3). If the most adequate (verbal) intelligence test—the PPVT—was used, the combined effect size for the studies involved was much larger compared to studies in which the PPVT was not included ($p_1 = .003$). Furthermore, the difference in intelligence between the dyslexic and normal readers was related to effect size: correlation with Fisher Zs was –.31 ($Z = 1.69$, $p_1 = .05$). If the dyslexic group scored higher on the intelligence test than the normal readers group, the effect size of the nonword reading test indicating the difference between

the two groups on phonological skill appeared to be smaller

If dyslexic subjects were participating in special programs, units, or schools, they did not show more phonological skill than dyslexic subjects involved in regular programs (see Table 3) If reading practice may be supposed to increase with age, the amount of reading practice did not appear to be relevant either: The age of dyslexic subjects is not related to effect size on the nonword reading test ($r = -.20$, $Z = 1.07$, $p_1 = .14$).

Some formal characteristics of the studies were related to effect size as well. Studies with larger sample sizes showed smaller effect sizes ($r = -.31$, $Z = 1.80$, $p_1 = .04$) Studies published in the early 1980s showed larger effect sizes than studies published more recently ($r = -.35$, $Z = 1.71$, $p_1 = .04$) Because 13 analyses were performed across the same set of studies, and predictor variables might well be correlated, a conservative, Bonferronized alpha level would be .008 (one-tailed) Our most robust findings, therefore, concerned the type of reading test and IQ test used in matching the dyslexics and the normal readers

Discussion and conclusions

The meta-analysis clearly supports Rack et al 's (1992) main conclusion that there is extremely strong evidence for the phonological deficit hypothesis We did find about half a standard deviation difference on the nonword reading task between dyslexic subjects and the reading-level-matched comparison group. At the same time, we did not find a difference in word recognition ability between the two groups The developmental delay hypothesis has therefore become implausible. Because the meta-analysis is based on almost 1,200 subjects, and because our fail-safe analysis showed that 423 further studies finding no support for the phonological deficit hypothesis are needed to render this hypothesis implausible, we feel it is safe to consider the phonological deficit to be an established fact. The law of diminishing returns might be applicable to new studies in this area; that is, the contribution of new primary-level studies on the existence of a nonword reading deficit will only be marginal

The overall effect size of half a standard deviation difference between dyslexic subjects and matched normal readers can be seen as quite modest (Cohen, 1977, but see Rosenthal, 1991), and much remains to be explained. In fact, less than 6% of the variance is explained on the basis of the nonword reading deficit. Even when we consider only the studies with optimal design features (i.e., using the PPVT as well as reading tests other than the Gilmore or WRMT), the combined effect size of this set of optimal studies is Cohen's $d = .84$ This effect size is comparable to a mean $r = .386$, and the proportion of explained variance in developmental dyslexia is 15% Though by definition groups with severe word recognition problems were selected for the studies, the nonword reading deficit explains a surprisingly small portion of the differences between normal and dyslexic readers Factors other than a nonword reading deficit, such as orthographic processing skill (Stanovich, 1991) or even experiences in the early stages of becoming literate (Teale & Sulzby, 1986), may therefore also be important. Of course, we do not exclude the possibility that the phonological deficit is a primary factor and that other explanations for the reading and spelling problems are, in whole or in part, consequences of this deficit (Stanovich, 1986).

Some studies showed much higher effect sizes than others. The reading-level-match design is a quasi-experimental design (Cook & Campbell, 1979) in a domain in which randomization is impossible. The implementation of this design, however, is difficult because the matching procedure might at any time produce unexpected differences between the groups, related to their performance on the nonword reading task (Backman, Mamen, & Ferguson, 1984). In our meta-analysis we found that studies with more adequate matching procedures showed a larger phonological deficit in dyslexic readers. In particular, studies with a better match on age, on intelligence, on reading level, and on word recognition yielded more impressive differences on the nonword reading task. The Gilmore and WRMT reading tests appeared to be less adequate matching tests than reading tests focusing on reading of irregular words out of context. The use of a verbal intelligence test like the PPVT leads to a larger difference on nonword reading between dyslexics and matched normal readers. If dyslexics and normal readers are matched on performance IQ, the specific phonological deficit might be contaminated with a general language deficit. A larger age range is related to a smaller nonword reading deficit. Inspection of the data revealed that larger age ranges were associated with relatively older normal readers (> 8 years). The age difference measure is, however, not very reliable and we should refrain from far-reaching conclusions. Dyslexics who are somewhat more intelligent than the matched normal readers also obscure the nonword reading effect, suggesting the mitigating influence of general competence. The most important indicator of the reading level is the word recognition test used in most studies to check whether the matching procedure had been successful or not Larger differences on this word recognition test favoring the dyslexic subjects lead to

smaller differences on the nonword test suggesting a less severe phonological deficit

Contrary to Rack et al s (1992) suggestions we did not find a relation between the age of the normal readers and the size of the nonword reading deficit In particular, the inclusion of 7-year-old normal readers did not significantly decrease the difference with dyslexic subjects Furthermore, dyslexic subjects participating in special remediation programs did not perform better than dyslexic subjects in regular schools We have to emphasize however that several studies were quite vague about the recruitment of dyslexic subjects Our decision to include in the special program group only those subjects whose participation in such a program was explicitly stated, might in some cases have led to wrong classifications Lastly, we did not find any significant influence of the materials used in the nonword tests Our meta-analysis did not support Rack et al s (1992) speculation that complexity and similarity of the nonwords might affect the outcome of the study The major weakness of studies on the phonological deficit hypothesis does not appear to be the kind of nonword reading test used to measure phonological skill, but the matching procedure used to create comparable groups of dyslexic and normal readers

In addition to Rack et al s (1992) review we also checked whether the size of the phonological deficit found in the studies was dependent on the number of subjects involved and on the year of publication We did indeed find that the more recent studies showed a somewhat smaller deficit than the early studies Two trends might be involved here First, during the last decade special and regular schools might have become more sensitive to the importance of phonological skill training for slow readers Second, older studies may be more exact replications of Snowling's (1980, 1981) pioneering studies, whereas more recent studies may have more variations in design that reduce the nonword reading deficit In meta-analyses, the same association between publication year and effect size has often been found (Mullen, 1989, Rosenthal, 1991) The relation between sample size and effect size seems to point to the possibility of a publication bias Of studies showing relatively small effect sizes, those using larger samples may have more chance of being published than those using smaller samples However, a plot of effect sizes by sample sizes (a so-called funnel plot, Light & Pillemer, 1984) did look like an inverted funnel, and did not show a conspicuous absence of the "small sample–small effect–nonsignificant result' studies (Mullen, 1989) The funnel plot, therefore, confirms Rack et al s (1992) suggestion that in this field null results are as important as significant results, and a publication bias should not be expected Furthermore, the file-drawer problem cannot be considered acute in view of the fact that more than 400 studies with null results would have to be available (unpublished or in press) to bring the combined probability level down to insignificance Nevertheless it might be worthwhile to search systematically for unpublished papers and dissertations on the phonological deficit hypothesis in order to broaden the database for our estimate of the combined effect size

In the meta analytic literature the potential weaknesses of the traditional narrative review are elaborated quite extensively (Cooper 1984 Mullen 1989 Rosenthal 1991) whereas the strengths of the meta analytic approach are heavily emphasized Usually at least three major differences between the traditional and the meta-analytic review are outlined The meta analytic review is supposed to be more precise more objective and replicable We have shown however how strongly a narrative review and a meta analysis of the same set of studies may converge A careful and thoughtful narrative review is invaluable for generating ideas and interpretations of discrepancies between studies The meta analytic approach might add formal tests and qualifications as to the generalizability of the results The most informative and reliable review of a research domain is therefore a combination of a thorough narrative review and a systematic meta analysis

REFERENCES

BACKMAN J MAMEN M & FERGUSON H B (1984) Reading level design Conceptual and methodological issues in reading research *Psychological Bulletin* 96 560 568

BADDELEY A D ELLIS N C MILES T R & LEWIS V J (1982) Developmental and acquired dyslexia A comparison *Cognition* 11 185 196

BADDELEY A D LOGIE R H & ELLIS N C (1988) Characteristics of developmental dyslexia *Cognition* 29 197 228

BEECH J R & HARDING L M (1984) Phonemic processing and the poor reader from a developmental lag viewpoint *Reading Research Quarterly* 19 357 366

COHEN J (1977) *Statistical power analysis for the behavioral sciences* New York Academic Press

COOK T D & CAMPBELL D T (1979) *Quasi experimentation Design and analysis issues for field settings* Chicago Rand McNally

COOPER H M (1984) *The integrative research review A social science approach* Beverly Hills CA Sage

DIBENEDETTO B RICHARDSON E & KOCHNOWER J (1983) Vowel generation in normal and learning disabled readers *Journal of Educational Psychology* 75 576 582

HEDGES L V & OLKIN I (1985) *Statistical methods for meta analysis* San Diego CA Academic Press

HOLLIGAN C & JOHNSTON R (1988) The use of phonological information by good and poor readers in memory and reading tasks *Memory & Cognition* 16 522 532

JOHNSTON R J RUGG M D & SCOTT T (1987) The influence of phonology on good and poor readers when reading for meaning

Journal of Memory & Language, 26, 57-68.

KOCHNOWER, J., RICHARDSON, E., & DIBENEDETTO, B. (1983). A comparison of the phonic decoding ability of normal and learning disabled children. *Journal of Learning Disabilities, 16,* 348-351

LIGHT, R J., & PILLEMER, D.B (1984). *Summing up The science of reviewing research.* Cambridge, MA: Harvard University Press

MANIS, F.R., SZESZULSKI, P.A , HOLT, L.K., & GRAVES, K. (1988). A developmental perspective on dyslexic subtypes. *Annals of Dyslexia, 38,* 139-153

MULLEN, B. (1989) *Advanced BASIC meta-analysis.* Hillsdale, NJ: Erlbaum.

OLSON, R K , KLIEGEL, R., DAVIDSON, B.J., & FOLTZ, G. (1985). Individual and developmental differences in reading disability In G E MacKinnon & T G Waller (Eds.), *Reading research Advances in theory and practice* (Vol 4, pp. 1-64). New York: Academic Press

OLSON, R.K , WISE, B , CONNERS, F , RACK, J., & FULKER, D. (1989). Specific deficits in component reading and language skills: Genetic and environment influences. *Journal of Learning Disabilities, 22,* 339-348.

RACK, J P., SNOWLING, M.J., & OLSON, R.K. (1992). The nonword reading deficit in developmental dyslexia· A review. *Reading Research Quarterly, 27,* 29-53.

ROSENTHAL, R. (1991). *Meta-analytic procedures for social research* (rev ed.). Newbury Park, CA: Sage

SIEGEL, L S , & RYAN, E B (1988). Development of grammatical sensitivity, phonological, and short-term memory skills in normally achieving and learning disabled children. *Developmental Psychology, 24,* 28-37.

SNOWLING, M J. (1980). The development of grapheme-phoneme correspondence in normal and dyslexic readers. *Journal of Experimental Child Psychology, 29,* 294-305.

SNOWLING, M.J. (1981). Phonemic deficits in developmental dyslexia. *Psychological Research, 43,* 219-234.

STANOVICH, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360-407.

STANOVICH, K E. (1991) Word recognition. Changing perspectives. In R. Barr, M.L. Kamil, P.B. Mosenthal, & P D. Pearson (Eds.), *Handbook of reading research* (Vol. 2., pp. 418-452). White Plains, NY: Longman.

STOUFFER, S.A. (1949). *The American soldier Vol. 1. Adjustment during army life* Princeton, NJ: Princeton University Press.

SZESZULSKI, P.A., & MANIS, F.R. (1987). A comparison of word recognition processes in dyslexic and normal readers at two reading-age levels *Journal of Experimental Child Psychology, 44,* 364-376.

TEALE, W H., & SULZBY, E (1986). *Emergent literacy Writing and reading* Norwood, NJ: Ablex.

TREIMAN, R., & HIRSCH-PASEK, K. (1985) Are there qualitative differences in reading behavior between dyslexic and normal readers? *Memory & Cognition, 13,* 357-364.

VELLUTINO, F.R , & SCANLON, D.M. (1987). Phonological coding, phonological awareness, and reading ability: Evidence from a longitudinal and experimental study. *Merrill Palmer Quarterly, 33,* 321-364

Received January 12, 1993
Revision received September 7, 1993
Accepted November 17, 1993

AUTHOR NOTES

This study was supported by a PIONEER grant from the Netherlands Organization for Scientific Research (NWO grant no. PGS 95-256).

Correspondence should be addressed to Marinus van IJzendoorn, Center for Child and Family Studies, Leiden University, PO Box 9555, NL-2300 RB Leiden, The Netherlands.