Individual Differences in Reading Development: A Review of 25 Years of Empirical Research on Matthew Effects in Reading

Maximilian Pfost, John Hattie, Tobias Dörfler & Cordula Artelt

Abstract: The idea of Matthew effects in reading – the widening achievement gap between good and poor readers – has attracted considerable attention in education research in the past 25 years. Despite the popularity of the topic, however, empirical studies that have analyzed the core assumption of Matthew effects in reading have produced inconsistent results. This review summarizes the empirical findings on the development of early interindividual differences in reading. We did not find strong support for the general validity of a pattern of widening achievement differences or for a pattern of decreasing achievement differences in reading. The inclusion of moderating variables, however, allowed a clearer picture to be painted. Matthew effects were more likely to occur for measures of decoding efficiency, vocabulary, and composite reading scores when the achievement tests were not affected by deficits in measurement precision. Further, moderators such as the applied analytic method or the orthographic consistency of the language were of less importance for the emergence of Matthew effects in reading. An additional meta-analysis of studies reporting correlations between a baseline level and a growth parameter yielded a small, negative mean correlation (r = -.214), which again was moderated by properties of the measures. Possible explanations for the reported findings are discussed.

Keywords: Matthew effects, reading development, primary school, reading difficulties

Participation in social, economic, and cultural life requires successful handling of written information, as written text contains not only facts and information, but also ideas, values, and cultural content (Artelt & Dörfler, 2010; Organisation for Economic Co-Operation and Development [OECD], 2003). As a consequence, teaching children to read is typically seen as one of the most significant accomplishments of primary education. The prerequisites needed for a successful acquisition of literacy competences, however, are not equally distributed across children when they enter the primary school system. Already prior to the formal act of reading instruction in school, there are significant individual differences in the comprehension of oral language, as there are differences in vocabulary or the students' awareness of the phonological and sound structure of spoken language (Bradley & Bryant, 1983; Burns & Kidd, 2010; Dickinson, McCabe, Anastasopoulos, Peisner-Feinberg, & Poe, 2003; Scarborough, 2002). As a consequence, at the beginning of formal education, some children are better equipped with the resources to optimize the impact of reading instruction and have rapid success in mastering these demands. Other children may, already at this early phase, lack the basic skills that are needed for fluent reading and are at risk of becoming problem readers (Scarborough, 1990; Snowling, 2001; Vellutino, Fletcher, Snowling, & Scanlon, 2004).

The results of international large-scale studies such as the Progress in Reading Literacy Study (PIRLS) 2006 have indicated that by Grade 4, which in several countries marks the end of primary schooling, there are huge individual differences in students' reading literacy. Whereas students reaching the top end of the PIRLS reading scale could not only locate significant details embedded in texts but could also integrate and interpret ideas across texts, students at the lower end of this scale had severe deficits in recognizing, locating, and reproducing explicitly stated information from the text (Mullis, Martin, Kennedy, & Foy, 2007). Cross-sectional comparisons alone cannot explain how, on average, the gap develops between those students who read well and those who do not, and cannot determine whether students who perform well when schooling begins perform even better at later time points in comparison to those students who begin school with fewer resources or who fall behind at some point in time. There is a need for longitudinal studies that track poor and good readers over the first few years of school to determine whether interindividual differences that occur early increase or decrease with time.

Interindividual Differences in the Development of Reading Literacy

When investigating the development of reading literacy from a longitudinal point of view, a high interindividual rank-order stability seems to be the norm, especially as children grow older (Boland, 1993; Butler, Marsh,

Accepted Author's Manuscript. This article was published in: *Review of Educational Research, 84*(2), 203-244
 http://rre.aera.net. DOI: 10.3102/0034654313509492

Sheppard, & Sheppard, 1985; A. E. Cunningham & Stanovich, 1997; Juel, 1988; Phillips, Norris, Osmond, & Maynard, 2002). This stability may be attributable to several causes, beginning with genetic differences between children (Grigorenko, 2004; Harlaar, Spinath, Dale, & Plomin, 2005; Hohnen & Stevenson, 1999; Olson et al., 2011) in combination with continuous environmental influences from the preschool, school, and family (Bradley, 1989; Hart & Risley, 1992; Rodríguez-Brown, 2010). Furthermore, when regarding the entire population of primary school students, a high normative change or growth in reading literacy, ranging from about a third to one standard deviation is to be expected annually (Bloom, Hill, Black, & Lipsey, 2008; Hill, Bloom, Black, & Lipsey, 2008).

The typical pattern after the beginning of formal education is, on average, higher learning gains in primary school and decreasing gains over the course of secondary schooling. For example, whereas between Grade 1 and Grade 2, average reading gains of about one standard deviation can be expected, in Grade 8, the normative annual learning gains shrink to a quarter of a standard deviation (Hill et al., 2008). It seems implausible, however, to assume that learning rates are equal for all students, raising the question of whether gains in learning are systematically related to the initial level of achievement. Thereby, three broad developmental patterns characterizing the relationship between initial reading level and successive reading literacy improvement can be distinguished (Figure 1). The first developmental pattern assumes a positive relationship between students' initial reading level and their intraindividual reading gains, leading to increasing differences in reading proficiency over the course of students' ontogenesis (Figure 1, Pattern A). This pattern of relationship is commonly referred to as a fan-spread or the Matthew effect in reading (Stanovich, 1986, 2000; Walberg & Tsai, 1983). Furthermore, within this positive correlation between initial competence level and individual development, absolute and relative Matthew effects can be distinguished (Rigney, 2010). An absolute Matthew effect describes a developmental pattern in which the students who read better show further positive reading literacy gains, whereas the students who read worse show negative gains. On the other hand, a relative Matthew effect assumes higher reading literacy gains for better readers, whereas the poor readers have flatter or only marginally increasing gains.

A second developmental pattern that characterizes the relationship between initial reading level and successive reading literacy change, by contrast, assumes a negative relationship between students' initial reading level and the developmental gains leading to decreasing differences in reading over the course of development (Figure 1, Pattern B), a relationship generally referred to as a compensatory model or a developmental-lag model of reading development (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005; Rourke, 1976). Again, two subtypes can be distinguished. Students performing well on a reading pretest might further increase their reading literacy but to a lower extent than students who perform worse on the pretest, leading to a relative compensation. On the other hand, the reading proficiency of the more proficient students might decrease, whereas the reading proficiency of the initially less proficient students might increase. This produces a developmental pattern that can be seen as the opposite of an absolute Matthew effect.

Finally, a third developmental pattern can be described, assuming stable proficiency differences between the high and low performing students (Figure 1, Pattern C). In this case, over the course of development, neither increasing nor decreasing differences in reading literacy are expected. Taken together, there are three broad patterns that characterize the development of interindividual differences in reading. Furthermore, as different patterns should reflect different underlying mechanisms, leading to contrasting expectations about developmental trends, some prominent explanations that support either increases or decreases in interindividual differences in reading literacy development will be illustrated in the following sections.

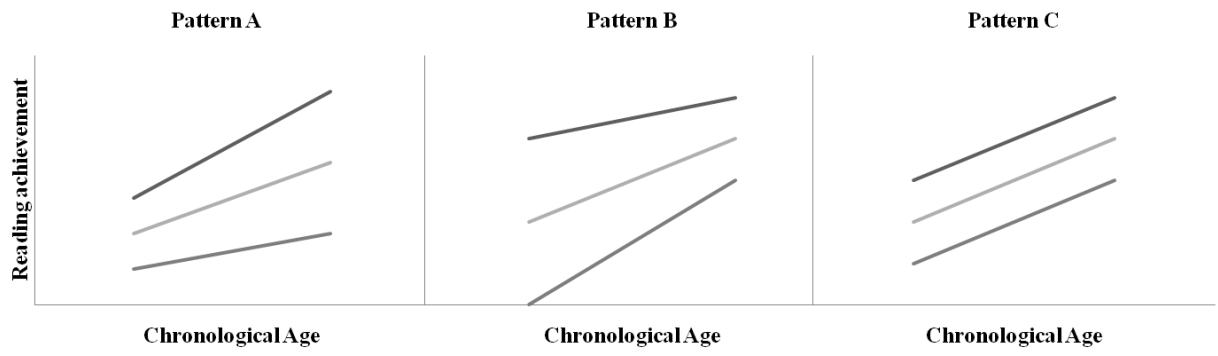

Figure 1. Three developmental patterns describing the development of interindividual achievement differences in reading. Pattern A depicts an increasing achievement gap/Matthew effect. Pattern B shows a decreasing achievement gap (compensatory model). And pattern C indicates a stable achievement gap.

The Concept of Matthew Effects in Reading

The concept of Matthew effects refers to the parable of talents (Matthew, 25:29) and describes a model of cumulative advantages of educational outcomes in which "those who score higher than others on pretests or other desirable attributes relevant to a treatment at the beginning of an experiment gain absolutely and relatively more than others from the same experience" (Walberg & Tsai, 1983, p. 360). Or, in other words, an initial advantage in a certain outcome tends to beget further advantages, whereas an initial disadvantage begets further disadvantages (Cook & Campbell, 1979; Rigney, 2010), creating in the long run, a widening gap or "fan-spread" between those who initially have more and those who initially have less.

In 1986, Stanovich transferred the concept to a model describing the development of interindividual differences in reading. The core of his model is the assumption of increasing interindividual differences in reading literacy due to self-reinforcing reciprocal causal mechanisms that connect reading literacy to factors that foster reading literacy development. Thereby, in addition to factors such as the genotype-environment correlation – meaning that environmental opportunities have a tendency to be compatible with the genotype (Gilger, Ho, Whipple, & Spitz, 2001; Harlaar, Dale, & Plomin, 2007; Scarr, 1992; Scarr & McCartney, 1983) – the role of reading behavior and practice is stressed (Stanovich, 1986, 2000): Better readers seem to be more motivated to read and hence read more. Free reading for these students is a major factor for the development of vocabulary; this in turn facilitates reading comprehension, and hence, as reading becomes more efficient, reading volume increases further.

This outcome has been called the *virtuous circle of reading* or, turning it the other way round, as the *vicious circle of nonreading* (Aunola, Leskinen, Onatsu-Arvilommi, & Nurmi, 2002; Pfost, Dörfler, & Artelt, 2010). We should note that these cumulative cycles are not a peculiarity found in the domain of reading or education. Rather, reciprocal cumulative processes seem to be typical mechanisms often found in the development of many attributes of psychological functioning and ill-functioning (Caspi, Bem, & Elder, 1989; Wachtel, 1994). Concerning the empirical support for these reciprocal relationships among reading motivation, reading behavior, and reading literacy, strong support has been found in observational studies (Bast & Reitsma, 1998; Harlaar et al., 2007; McElvany, Kortenbruck, & Becker, 2008; Morgan & Fuchs, 2007; Pfost et al., 2010), but the results of experimental studies have been less convincing (National Institute of Child Health and Human Development, 2000). Finally, meta-analytic results presented by Mol and Bus (2011) showed increasing correlations with age between measures of print exposure and oral language as well as technical reading skills. This finding is consistent with a developmental model of reciprocal causation of reading behavior and reading achievement.

A second explanation for the emergence of Matthew effects in reading has been found in research concerning specific reading disabilities. According to the cognitive deficit models of reading disability, deficient readers may be handicapped in their acquisition of reading skills due to underlying endogenous cognitive conditions or deficits that cannot be overcome (Pennington, 2006; Scarborough, 2002). For example, language-based deficits that might have some neurobiological, genetic, or congenital foundation seem of special importance (Shaywitz, Morris, & Shaywitz, 2008; Vellutino et al., 2004). In addition, persisting exogenous conditions such as adverse child-rearing patterns may cause early competence differences or symptoms that interfere with later reading development (Scarborough, 1990, 2002). Thus, taken together, as long as these conditions prevail, whether they are endogenous or exogenous, these students might be constantly impaired in the acquisition of their reading skills, which might lead to an increasing gap over the course of schooling. Note that although this explanation is often found in the literature on reading disabilities, the concept of persistent endogenous and exogenous conditions influencing the development of reading competencies also applies to the whole population of students. Conditions like parents reading practices and beliefs may support and motivate beginning readers as well as older and more advanced readers, and consequently promote reading development within and across different developmental periods (Baker, Scher, & Mackler, 1997; Klauda, 2009).

Finally, there is a third factor with regard to the timing of the emergence of Matthew effects. Previous results focusing on the development of social disparities in academic achievement have provided strong evidence that widening achievement scores are highly likely to occur during noninstructional periods (*summer setback;* cf., Entwisle & Alexander, 1992). By contrast, during the school season, students' socioeconomic status seems less important, suggesting that schools seem to have an equalizing effect. This relationship has been studied extensively by using data from the Early Childhood Longitudinal Study (ECLS-K; cf., Downey, von Hippel, & Broh, 2004; McCoach, O'Connell, Reis, & Levitt, 2006; Reardon, 2003). It seems as if the development of reading competence may follow a Matthew-effect pattern in summer and a compensatory developmental pattern during instructional periods (McCoach et al., 2006). It thus seems that factors leading to Matthew effects in reading are less tied to characteristics of schools (at least in primary schools with less strict forms of tracking), and may be better explained by attributes of the students and their families themselves (McCoach et al., 2006).

Compensatory and Developmental-Lag Models of Reading

In contrast to the presence of a Matthew effect in reading, it has been noted that there are alternative

positions that assume decreases in interindividual achievement differences over time. Compensatory models of reading literacy development assume that students with relatively low levels of cognitive resources at an early point in time will more or less automatically catch up in their proficiency level as these students grow older. Thereby, a developmental lag can be assumed: The rate of acquisition of developmental skills is delayed rather than impaired, which may be due to differences in brain maturation (Figure 2; Francis et al., 1996; Rourke, 1976). This means that, after the appearance of an early increasing competence gap, at a certain developmental age, the rate of competence development of the skilled readers tends to flatten or even reach a stable proficiency level. In comparison to the *early starters,* less skilled or developmentally lagged readers will begin their skill acquisition later. These *late starters* may however follow the same learning trend or curve as the early starters, just with a certain time delay. Consequently, as learning rates follow a negative accelerated trend (cf., Bloom et al., 2008; Hill et al., 2008), at a certain age, interindividual differences between the early and the late starters will decrease when the negative acceleration trend becomes higher for the better readers (Scarborough, 2002).

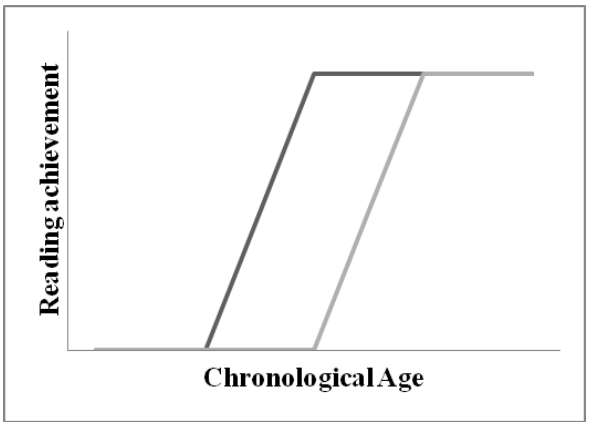

Figure 2. Developmental lag model of reading development.

The developmental lag model can easily be explained in the domain of reading when considering, for example, the learning of the names and sounds of the letters of the alphabet. Already in preschool, some children are familiar with many letters (e.g., when writing their names; Hildreth, 1936; Welsch, Sullivan, & Justice, 2003), whereas other children do not start learning to use letters until elementary school. As the number of elements that need to be mastered is finite and quite small (e.g., the German, like the English, alphabet contains 26 letters), we would expect that children who are familiar with some letters already prior to schooling will manage it quite well and in particular will be quite fast to recognize or name all the letters of the alphabet. Students who start elementary school with less letter knowledge might, however, have a lower learning rate for letters at first. Despite these initial problems, at a certain age, these later starting students will also manage this special task of naming all the letters of the alphabet quite well, which means that they are able to catch up on this task.

The developmental pattern for these tasks thus leads to a Matthew effect at the beginning of formal schooling and a compensatory pattern for the time to follow. Or to put this in more general terms: Developmental periods characterized by high negative accelerated growth trends (e.g., because students reach a proficiency level that cannot be increased anymore), seem to provide good conditions for the emergence of a compensatory developmental pattern. According to Paris (2005), we can expect this compensatory developmental pattern, in particular, for reading skills that are highly constrained, meaning that these skills are universally mastered in a relatively brief developmental time span, such as letter knowledge, phonics, or concepts of print. The reason, as stated in the example above, is that the opportunities for growth, especially for good readers, are quite limited within these constrained skills. Less constrained skills, such as vocabulary and comprehension, by contrast, might evoke a different pattern of the development of interindividual differences, as these skills might also provide additional possibilities for growth for the best readers. As a consequence, the observed pattern of development of interindividual differences in reading might be a function of time and the task or skill under consideration.

Furthermore, indirect empirical support for the adequacy of a developmental lag model has been provided by research that has addressed the question of how later reading is affected by the age at which reading instruction begins (cf., Cunningham & Carroll, 2011; Suggate, 2009). Suggate, Schaughency, and Reese (2013), for example, compared the development of reading skills of students in common state schools in New Zealand to students attending Steiner schools, who begin formal reading instruction about one and a half years later. According to these longitudinal results, the students from the Steiner schools who begin reading later finally catch up at about age 11 to the students who begin earlier, therefore closing the gap that emerged as a result of early reading instruction. Therefore, reading advantages of very young students might "wash out," as older students seem to be more efficient at acquiring these early reading skills (Suggate, 2012). However, we should be cautious when transferring this result of instruction that occurs at a different point in time to the question of the development of individual differences in the primary grades when students attain a comparable amount of instruction.

Taken together, the development of reading competence may follow a cumulative, stable, or compensatory developmental model, and differences in the observed developmental pattern may be due to factors such as the nature of the skill, the age of the students, the exposure to instruction, and so forth. Given the practical and theoretical significance of knowledge concerning the expected development of students with regard to their reading proficiency level, there is a need for research that integrates empirical

results from longitudinal studies in the domain of reading to address the hypothesis of Matthew effects in reading.

The purpose of this study was to review the available literature that has addressed developmental patterns of reading skills over the course of primary education. Although the concept of a Matthew effect in reading has often been integrated into a more general theory of individual differences in reading development (cf., Stanovich, 1986, 2000), the present literature review focuses exclusively on the question of whether it is adequate to assume divergent achievement trajectories of students with different initial skill levels. Concerning the supposed reciprocal mechanism underlying Matthew effects in reading, we refer to the studies cited above, especially the meta-analysis by Mol and Bus (2011) as well as the more narrative review by Morgan and Fuchs (2007).

Method

In order to provide an overview concerning the question of whether reading development in primary school can be best described in terms of Matthew effects or not, we conducted a broad literature search that aimed to identify all available published empirical studies on this topic. In a first step, we conducted an electronic search making use of databases such as PsycINFO, ERIC, and GoogleScholar. Terms such as *Matthew(-)effect, inter(-)individual differences, cumulative and effect, cumulative and deficit, developmental lag, and compensatory model* were used in combination with terms such as *reading, literacy,* and *vocabulary*. Due to a large number of hits, further terms such as *elementary* or *primary* and *longitudinal* were applied to restrict the number of references. Then, all papers listed in PsycINFO that used the term *Matthew* in their title were screened. In addition, articles citing the work by Stanovich (1986) were screened within PsycINFO and articles citing the work by Walberg and Tsai (1983) were screened within GoogleScholar. Only studies published since 1986 were included in this literature review as the notion of a Matthew effect in reading became quite popular with the 1986 Stanovich paper. Furthermore, in addition to the electronic search, titles of research papers in the following journals were screened manually: *Journal of Educational Psychology, Journal of Research in Reading, Reading and Writing, Reading Psychology,* and *Reading Research Quarterly*. Finally, citations in the included articles were examined. All literature searches were conducted in the summer of 2011. The search was updated in the summer of 2012.

Inclusion Criteria

To be selected for the review, the identified articles had to meet the following inclusion criteria:

1. The study focused on students of primary school age. Because the time of school enrollment as well as the number of years students spend in primary school varies between different countries and educational systems, an age criterion was also applied. Therefore, to be included in the review, the average age of the sample had to be in the range of 5:0 up to 11:11 years, basically reflecting the first to the sixth grade.
2. Analyses had to be based on longitudinal data; therefore, the data of the same students had to be available for at least two points of measurement that fell into our range of student age. Results referring to measurement points that were beyond our range of student age were not considered; however, in three papers (Jacobsen & Lundberg, 2000; Klicpera, Schabmann, & Gasteiger-Klicpera, 1993; Williamson, Appelbaum, & Epanchin, 1991), we could not separate growth rates for primary and secondary education as, for example, only one single parameter was estimated or reported across the entire time period. In these three cases, the available parameter was used as the best estimator of the relationship of interest.
3. Student outcomes had to contain measures of reading achievement (e.g., word, sentence, or text comprehension, decoding speed or accuracy, vocabulary, etc.).
4. The measured outcomes had to be on the same or at least a comparable metric across the different points of measurement. This criterion was fulfilled, for example, when the reading outcomes were measured twice using the same scale, a parallel version of the test with comparable test properties, or when the different outcomes measures were transferred to a common metric using some sort of test equating or linking.
5. The study was, at least partially, dedicated to analyzing the question of individual differences in the development of reading achievement dependent upon a prior reading achievement level (the Matthew effect hypothesis). Analyses focusing on longitudinal relations between different reading or precursor abilities were not included in the review (e.g., Frost, 2001; Greenfield Spira, Storch Bracken, & Fischel, 2005; Juel, Griffith, & Gough, 1986; Schneider, 2009).
6. The methods applied needed to allow some inferences about the existence of Matthew effects in reading. Analyses that allow the reporting of a covariance or correlation between a baseline level and a growth component were appropriate as were simplex or autoregressive models. Furthermore, studies grouping students into different proficiency groups by using one or several baseline-level reading measures and comparing their reading progress were adequate.
7. Students had to have attended the regular education system. Studies that focused on the effect of some sort of intervention, especially if the intervention was implemented by the researcher himself, were excluded from this review. A detailed discussion concerning the role of interventions for the emergence of Matthew effects is provided by Ceci and Papierno (2005). Nevertheless, as long as the intervention was part of the regular education

system itself (e.g., extra teaching), these studies were not excluded from the current review (e.g., Good, Baker, & Peyton, 2009; Wang, Algozzine, Ma, & Porfeli, 2011). Finally, the study by Carreker et al. (2007) was included in the review as the intervention took place prior to the analyzed time period and was not the sole focus of the analysis. Furthermore, the reported intercept-slope covariance in their study referred to all students independent of the treatment status.

8. Studies that focused on the reading development of students with diagnosed dyslexia or other psychiatric illness were not considered in the present analysis; however, analyses that focused on the development of deficient readers who did not suffer from a psychiatric illness or who showed just subclinical symptoms were included in the present review (e.g., Francis et al., 1996; Jacobsen & Lundberg, 2000). The rationale behind this criterion was twofold: On the one hand, a generalization of the finding would become increasingly difficult if the study contained different subpopulations; on the other hand, the underlying mechanism directing the development of achievement gaps might be different for students with and without clinically relevant symptoms.
9. Articles using basically the same dataset with the same methods and outcomes were included only once in the review (e.g., Aarnoutse, van Leeuwe, Voeten, & Oud, 2001; Bast & Reitsma, 1998; B. A. Shaywitz et al., 1995). Concerning the Early Childhood Longitudinal Study – Kindergarten Cohort (ECLS-K), we considered the papers by Foster and Miller (2007), Silberglitt and Hintze (2007), as well as Sonnenschein, Stapleton, and Benson (2010), due to their unambiguous interpretability with regard to the developmental pattern of reading achievement.
10. The article needed to be published in German or English.

Coding and Procedure

Every study that we included was reviewed with regard to the developmental pattern of reading achievement. Study results were coded as indicating a *Matthew effect* or *widening achievement gap* in reading when the study reported (a) a positive covariance or correlation between a baseline level or intercept and the (linear-) growth component, (b) a simplex model indicating a high rank-order stability in combination with an increasing (latent) variance in the outcome measures or an autoregressive coefficient greater than one, or (c) an increasing achievement difference with time for different proficiency groups constructed by using one or several baseline-level cognitive measures. The results were coded as indicating a *compensatory developmental pattern* or *decreasing achievement gap* when the analysis indicated (a) a negative covariance or correlation between a baseline level or intercept and the (linear-) growth component, (b) a simplex model indicating a decreasing (latent) variance in the outcome measures or an autoregressive coefficient below one, or (c) decreasing achievement differences with time for different proficiency groups constructed by using one or several baseline-level cognitive measures.

Third, results were coded as indicating a pattern of *stable achievement differences* when results showed (a) a near zero or nonsignificant covariance or correlation between a baseline level or intercept and the (linear-) growth component, (b) a simplex model indicating a constant or not significantly changing (latent) variance in the outcome measures or an autoregressive coefficient of one, or (c) constant or not significantly changing achievement differences with time for different proficiency groups. Finally, after a first screening of the reviewed studies, two further categories describing interindividual differences in reading achievement were included (Figure 3). Results were further coded as indicating a *pattern of delayed compensation* when results within one outcome variable first showed increasing achievement differences (e.g., increasing latent variance, positive intercept-slope covariance) and subsequently showed decreasing achievement differences (e.g., decreasing latent variance, negative intercept-slope covariance). Also, a *crossing fan spread pattern* was coded when simplex models reported an increasing latent variance but low rank-order stability (cf. Pattern E, Figure 3). In this case, individual differences increased with time, but poor readers did not necessarily remain poor readers in relation to the other students.

With regard to the outcome variables of reading achievement, these were assigned to some superordinate categories. The term *reading comprehension* was used whenever processes that focused in particular on aspects of comprehension were tested. The term *decoding skills* was used for measures of basic reading skills. Within this category, in a second run, we further differentiated *decoding speed* whenever the speed/amount of reading per time was the specific focus of the study. The term *decoding accuracy* was used whenever the correctness of the reading/reading errors were of interest. Finally, we applied the term, decoding efficiency, when the outcome combined speed and accuracy components. The term *prereading skills* summarized measures of letter-sound correspondences and letter naming tasks. *Vocabulary* referred to different kinds of measures of children's vocabulary, including expressive and receptive vocabulary measures. Finally, the term *reading proficiency* was used if the authors used a mixed or composite reading achievement score of different categories, for example, a combined score of vocabulary and reading comprehension. In addition to the mean reliability of the achievement measures, all measures were rated for whether floor or ceiling effects emerged.

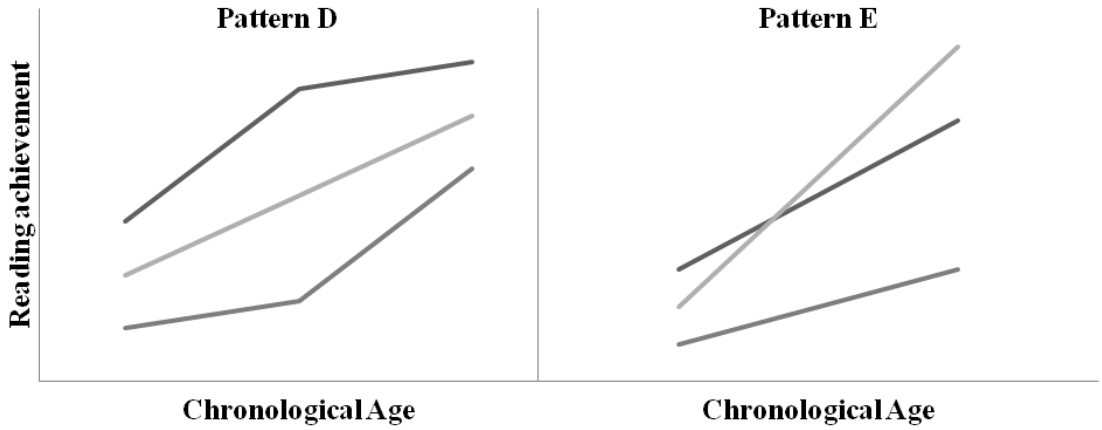

Figure 3. Two additionally reported developmental patterns describing the development of interindividual achievement differences in reading. Pattern D describes a developmental model of delayed compensation. Pattern E depicts a crossing fan-spread effect.

Measures were rated as to whether or not they were affected by *floor or ceiling effects*. This rating occurred, first, when a substantial proportion of students achieved the minimum or maximum scores of the scale, leading to an extremely skewed achievement distribution and a reduced variance of that distribution, or second, when the authors themselves reported the emergence of floor or ceiling effects. Furthermore, the reliability of the outcome variable was coded. When the reliability of the outcome variable varied between measurement points, the arithmetic mean was calculated. In order to create a good predictor for indicating problems with regard to the psychometric properties of the instruments, the two ratings were subsequently combined into a binary indicator: Measures were coded as *affected by low measurement precision* when they were rated as showing floor or ceiling effects or when the reported reliability was below .85; measures were coded as *not affected by low measurement precision* when no such indication was apparent.

Furthermore, grade level of the first wave of measurement, the number of points of measurement, and the main language of the students were rated. Languages were subsequently grouped according to their orthographic consistency (cf., Seymor, Aro, & Erskine, 2003; Ziegler et al., 2010; Ziegler & Goswami, 2006) in languages with relatively consistent grapheme-phoneme relations (e.g., Dutch, German, Finnish, and Swedish) and with relatively inconsistent grapheme-phoneme relations (e.g., English). All studies were coded by the first author, and agreed in discussion with the second author. In order to check for interrater agreement, 12 studies (41%) were coded further by the third author. Interrater agreement for the developmental patterns of reading achievement and the high inference ratings (measured construct, floor or ceiling effects) ranged from $\kappa = .79$ to $\kappa = .81$. Discrepancies between the raters were resolved by discussion.

The coded characteristics of the studies were related to the different patterns of reading development. When possible, categorical moderator variables were tested for significance by applying a χ^2 test of independence. However, inadequate expected cell counts (below 5) were a problem, so significance tests were often not able to be computed. Interval-scaled variables were related to the different patterns of reading development by testing for equal means using analysis of variance (ANOVA). Again, results should be interpreted with caution due to small number of studies. Nevertheless, the vote-counting method described above also has severe limitations, especially as the counted results do not consider the magnitude of the relationship between the baseline level and the growth component.

Studies reporting small positive or negative baseline levels-growth relationships were treated equal as studies reporting large positive or negative relationships. We tried to compensate for this drawback by meta-analyzing the subset of studies which report correlations between a baseline measure and a growth parameter separately. In cases that just a covariance in combination with the corresponding standard deviations was reported, a correlation was computed by dividing the estimated covariance through the product of the standard deviations. In a next step, the individual effect sizes respectively correlations were transformed using Fisher's transformation formula (cf., Card, 2012; Lipsey & Wilson, 2001). In the end, the effect sizes were weighted by the inverse of the variance of the effect size estimate and combined within a random effects model using Wilson's macros (2005; Lipsey & Wilson, 2001).

Results

Over the course of our electronic literature search, more than 4,000 references were screened (some references were duplicates due to overlapping searches and the use of different databases). Together with the manual screening of relevant journals and further citations found in the papers that were included in this review, 87 studies were scrutinized in detail. Finally, 28 studies that met the above criteria for analyzing the

development of Matthew effects in reading for primary school students were included in this review. All studies are listed in Table 1.

General Characteristics of the Studies and Samples

Within these 28 articles, 78 separate results about the development of interindividual differences in reading were reported. The number of results exceeded the number of studies for three reasons. First, some papers reported results from more than one sample, for example, Good et al. (2009) as well as Parrila et al. (2005), who reported distinct results from two separate samples. In addition, the papers by Rescorla and Rosenthal (2004) and Protopapas, Sideridis, Mouzaki, and Simos (2011) in each case reported results from three different cohorts of students, and Silberglitt and Hintze (2007) reported distinct results of partially different students in different grades. Second, 11 papers reported separate results for different reading measures based on the same sample (e.g., Kempe, Eriksson-Gustavsson, & Samuelsson, 2011, reporting results for four different reading measures).

Third, six articles reported separate results for different applied analytic methods based on the same outcome and sample (e.g., Aunola et al., 2002; Bast & Reitsma, 1997; Baumert, Nagy, & Lehmann, 2012; Leppänen, Niemi, Aunola, & Nurmi, 2004; Parrila et al., 2005; Protopapas et al., 2011). The sample sizes of the studies ranged from 28 to 358,032, with a median size of 243. The mean sample size was 11,835, indicating a right skewed sample size distribution across the reviewed studies. On average, the first point of measurement was between the first and second grades, and the final point of measurement was, on average, in fourth grade. The number of points of measurement ranged from 2 to 11 with a mean of 4.5 time points per reported result.

Most of the reported results were from samples hailing from the United States and Canada (n = 31, 39.7%) followed by Finland (n = 10, 12.8%), Greece (n = 9, 11.5%), and the Netherlands (n = 7, 9.0%). Other results were based on samples from the United Kingdom (n = 6, 7.7%), Sweden (n = 6, 7.7%), Austria (n = 5, 6.4%), and Germany (n = 4, 5.1%). Consequently, almost half of the results (n = 37, 47.4%) referred to samples of English-language students.

The most frequent reading measures were reading comprehension (n = 24, 30.8%), decoding efficiency (n = 22, 28.2%), and composite reading proficiency scores (n = 13, 16.7%). There were also measures of decoding accuracy (n = 10, 12.8%), decoding speed (n = 5, 6.4%), vocabulary (n = 3, 3.8%), and prereading skills (n = 1, 1.3%). Most of the studies did not report any constraints concerning the reading measures (n =59, 75.6%). Nineteen results (24.4%) were based on instruments that showed floor and/or ceiling effects. The average reliabilities of scores on the outcome measures ranged from .66 to .96, with a mean of .85.

Matthew Effect in Reading: General Findings

Of the reported 78 results, 33 (42.3%) indicated a decreasing achievement gap or compensatory pattern, 20 (25.6%) indicated stable achievement differences, and 18 (23.1%) indicated an increasing achievement gap or Matthew effect. Furthermore, six (7.7%) results indicated a pattern of delayed compensation, which means that these studies first found increasing and subsequently found decreasing achievement differences. One study reported a so-called crossing fan-spread pattern, meaning that despite an increasing latent variance in the outcome variable, low rank-order stability was present (Bast & Reitsma, 1997). The ratio of studies that reported an increasing to decreasing achievement gap was 0.55, indicating more studies that reported a compensatory developmental pattern than studies that reported a Matthew effect in reading. A χ^2 test for deviations from an equal distribution of the expected results between the three most prominent patterns (decreasing-stable-increasing achievement differences) was not significant (χ^2 = 5.606, [df] = 2, ns), indicating that none of these three developmental patterns was overrepresented in the total results.

We further need to mention that the reported descriptive finding that the number of studies that supported a compensatory developmental pattern exceeded the number of studies that indicated a Matthew-effect pattern almost by a factor of two should not be overemphasized for two reasons. On the one hand, the study by Protopapas et al. (2011) may substantially distort this general picture by reporting a compensatory developmental pattern nine times as the authors separately reported findings of three cohorts of students using three analytic approaches each time.

On the other hand, and of even more importance, there was substantial heterogeneity in the reported findings; thus, suggesting potential moderators that may determine the emergence of different developmental patterns should be of primary interest. Therefore, our main concern was to define the conditions under which the probability of the occurrence of Matthew effects in reading as opposed to compensatory developmental patterns was higher.

Moderator Analyses

In searching for potential moderators for the emergence of Matthew effects in reading, several properties of the studies were rated (Tables 2 and 3). In the subsequent section, first, patterns of reading development will be related to the measured construct, the language, and the grade level at the first point of measurement. Then, the role of the analytic model and the number of points of measurement will be explored. Finally, the properties of the outcome measures that were used will be taken into account.

Table 1

Summary of Reviewed Studies (in Alphabetical Order)

Author	Country	N	Study description	Analytic model	Construct, mean reliability	Findings	Comments
Aarnoutse & van Leeuwe (2000)	Netherlands	556	Grade 1 – Grade 6, 11 points of measurement	Formation of proficiency groups; Comparison of subsequent growth rates	Decoding efficiency (reading aloud wordlists; R = .85); Reading comprehension (R = .81); Vocabulary (R = .85)	o : Decoding efficiency - : Reading comprehension - : Vocabulary	No significance test statistics provided; Authors report an emergence of ceiling effects
Aunola, Leskinen, Onatsu-Arvilommi, & Nurmi (2002)	Finland	105	Grade 1, 3 points of measurement	LGC analysis; Simplex models	Reading proficiency (syllable recognition, word-to-picture matching, sentence and passage comprehension; R = .68)	- : Reading proficiency (LGC) -: Reading proficiency (Simplex)	In parts, no significance test statistics provided
Bast & Reitsma (1997)	Netherlands	235	Grade 1 – Grade 3, 6 points of measurement (decoding); 4 points of measurement (reading comprehension)	LGC analysis; Simplex models	Decoding efficiency (reading aloud wordlists); Reading comprehension	+ : Decoding efficiency (LGC) c : Decoding efficiency (Simplex) o : Reading comprehension (LGC) # : Reading comprehension (Simplex)	Simplex models for decoding show changing rank orders (crossing fan-spread pattern); Simplex models for reading comprehension show increasing (T1-T2) and decreasing (T2-T4) variance
Baumert, Nagy, & Lehmann (2012)	Germany	3167	Grade 4 – Grade 6, 3 points of measurement	LGC analysis; Simplex models	Reading comprehension (R = .86)	- : Reading comprehension (LGC) - : Reading comprehension (Simplex)	
Cain & Oakhill (2011)	Great Britain	31	Grade 3 – Grade 6, 2 points of measurement	Formation of proficiency groups; ANOVA	Decoding accuracy (word recognition in context; R = .84); Reading comprehension (R = .94); Vocabulary(R = .93)	o : Decoding accuracy o : Reading comprehension + : Vocabulary	

Table 1 (continued)

Author	Country	N	Study description	Analytic model	Construct, mean reliability	Findings	Comments
Carreker, Neuhaus, Swank, Johnson, Monfils, & Montemayor (2007)	USA	536	Grade 3 – Grade 5, 3 points of measurement	Growth models	Reading comprehension	+ : Reading comprehension	Some students were taught in Grade 1 and Grade 2 by specially trained teachers
Compton (2000)	USA	75	Grade 1, 7 points of measurement	Growth models	Decoding accuracy (reading wordlists aloud; reading aloud a list of nonwords)	o : Decoding accuracy (words) o : Decoding accuracy (nonwords)	Floor effects at the first point of measurement; Partial ceiling effects at the last point of measurement
Ditton & Krüsken (2009)	Germany	1,201	Grade 2 – Grade 4, 3 points of measurement	Correlation initial reading level – reading growth; Formation of proficiency groups	Reading comprehension (R = .66)	- : Reading comprehension	
Foster & Miller (2007)	USA	12,261	Kindergarten – Grade 3, 4 points of measurement	Formation of proficiency groups; Comparison of subsequent growth rates	Prereading skills (concept about print, letter – sound associations; R = .93); Reading proficiency (word identification, reading comprehension; R = .93)	- : Prereading skills # : Reading proficiency	Ceiling effects for prereading skills (Kindergarten – Grade 3) and for reading proficiency (Grade 3); No significance test statistics provided
Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher (1996)	USA	403	Grade 1 – Grade 9, 9 points of measurement	Formation of proficiency groups; Growth models	Reading proficiency (word identification, word attack, and passage comprehension)	- : Reading proficiency	
Good, Baker, & Peyton, (2009)	USA, Canada	a) 2,172 b) 358,032	Grade 1, 2 points of measurement	Formation of proficiency groups; Comparison of subsequent growth rates	Decoding efficiency (reading nonwords aloud; R = .83b)	a) o : Decoding efficiency b) - : Decoding efficiency	a) Oregon reading first sample b) DIBELS data system sample No significance test statistics provided

Table 1 (continued)

Author	Country	N	Study description	Analytic model	Construct, mean reliability	Findings	Comments
Jacobsen (1999); Jacobsen & Lundberg (2000)a	Sweden	171	Grade 2 – Grade 9, 3 points of measurement	Formation of proficiency groups; Comparison of subsequent growth rates	Decoding efficiency (wordchains; R = .85)	+ : Decoding efficiency (boys) o : Decoding efficiency (girls)	Some of the poor readers might show clinically relevant symptoms
Judge & Bell (2011)	USA	10,096	Kindergarten –Grade 5, 5 points of measurement	Growth models;	Reading proficiency (R = .94)	+ : Reading proficiency	
Juel (1988)	USA	54	Grade 1- Grade 4, 4 points of measurement	Formation of proficiency groups; Comparison of subsequent growth rates	Decoding accuracy (reading pseudo-words aloud; R = .93); Decoding accuracy (reading real words aloud; R = .96)	- : Decoding accuracy (pseudo-words) + : Decoding accuracy (real words)	No significance test statistics provided
Kempe, Eriksson-Gustavsson, & Samuelsson (2011)	Sweden	134	Grade 1 – Grade 3, 3 points of measurement	Formation of proficiency groups; Comparison of subsequent growth rates	Decoding efficiency (reading nonwords aloud; R = .86); Decoding efficiency (reading wordlists aloud; R = .82); Reading comprehension; Vocabulary	o : Decoding efficiency (nonwords) o : Decoding efficiency (wordlists) + : Reading comprehension + : Vocabulary	Floor effects for readers with reading difficulties in Grade 1 (decoding efficiency, reading comprehension); No significance test statistics provided
Kim, Petscher, Schatschneider, & Foorman (2010)	USA	12,536	Grade 1 – Grade 3, 3-4 annual points of measurement	Growth models	Decoding efficiency (reading nonwords aloud; R = .83); Decoding efficiency (reading connected text aloud; R = .94)	- : Decoding efficiency (nonwords) # : Decoding efficiency (connected text)	Intercept-slope correlation for decoding of connected text is positive in Grade 1 and negative in Grade 2 & 3
Klicpera, Schabmann, & Gasteiger-Klicpera (1993)	Austria	283	Grade 2 – Grade 8, 5 points of measurement	Formation of proficiency groups; Comparison of subsequent growth rates	Decoding speed (number of words read aloud); Decoding accuracy (percent of reading errors)	+ : Decoding speed - : Decoding accuracy	Ceiling effect for decoding accuracy for good readers

Table 1 (continued)

Author	Country	N	Study description	Analytic model	Construct, mean reliability	Findings	Comments
Klicpera, Schabmann, & Gasteiger-Klicpera (2006)	Austria	733	Grade 1 – Grade 4, 3 points of measurement	Formation of proficiency groups; Comparison of subsequent growth rates	Decoding speed (seconds per word read aloud); Decoding accuracy (percent of words read correctly); Decoding efficiency (total number of words read correctly)	- : Decoding speed - : Decoding accuracy + : Decoding efficiency	Ceiling effect for decoding accuracy for good readers
Leppänen, Niemi, Aunola, & Nurmi (2004)	Finland	196	Preschool – Grade 1, 4 points of measurement	Simplex models; Two-piece growth model	Reading proficiency (word/sentence reading aloud, sentence comprehension)	# : Reading proficiency (Simplex) # : Reading proficiency (LGC)	Simplex models first indicate increasing (T1-T3) and then decreasing (T3-T4) latent variances; Piecewise growth models indicate a positive covariance for the first and a negative covariance for the second growth component
Parrila, Aunola, Leskinen, Nurmi, & Kirby (2005)	a) Canada b) Finland	a) 198 b) 197	a) Grade 1 – Grade 5, 5 points of measurement b) Grade 1 – Grade 2, 4 points of measurement	LGC analysis; Simplex models	a) Decoding accuracy (Word attack; R = .94); Decoding speed (reading isolated words aloud; R = .94); Reading comprehension (R = .84) b) Decoding efficiency (reading a short story aloud); Decoding efficiency (wordchains); Reading comprehension	a) o : Decoding accuracy (LGC) # : Decoding accuracy (Simplex) - : Decoding speed (LGC) - : Decoding speed (Simplex) - : Reading comprehension (LGC) - : Reading comprehension (Simplex) b) o: Decoding efficiency (LGC) o : Decoding efficiency (Simplex) + : Decoding efficiency (Wordchains. LGC) + : Decoding efficiency (Wordchains, Simplex) - : Reading comprehension (LGC) o : Reading comprehension (Simplex)	a) Simplex models for decoding indicate first increasing (T1-T2) and then decreasing (T3-T5) latent variances; Floor effects reported for Grade 1 b) Simplex models for reading comprehension show significant variation in the latent variances between but no clear developmental trend

Table 1 (continued)

Author	Country	N	Study description	Analytic model	Construct, mean reliability	Findings	Comments
Pfost, Dörfler, & Artelt (2012)	Germany	1,124	Grade 3 – Grade 4, 3 points of measurement	Formation of proficiency groups; LGC analysis	Reading comprehension (R = .88)	+ : Reading comprehension	
Protopapas, Sideridis, Mouzaki, & Simos (2011)	Greek	a) 208 b) 192 c)187	a) Grade 2 – Grade 4, 5 points of measurement b) Grade 3 – Grade 5, 5 points of measurement c) Grade 4 – Grade 6, 5 points of measurement	Growth models; Log-linear multilevel model; Formation of proficiency groups	Reading comprehension (R = .78)	- : Reading comprehension (Subsample a), Growth model) - : Reading comprehension (Subsample a), Multilevel model) - : Reading comprehension (Subsample a), Proficiency groups) - : Reading comprehension (Subsample b), Growth model) - : Reading comprehension (Subsample b), Multilevel model) - : Reading comprehension (Subsample b), Proficiency groups) - : Reading comprehension (Subsample c), Growth model) - : Reading comprehension (Subsample c), Multilevel model) - : Reading comprehension (Subsample c), Proficiency groups)	
Rescorla & Rosenthal (2004)	USA	328	Grade 3 – Grade 10, 4 points of measurement	Formation of proficiency groups; ANOVA	Reading proficiency (vocabulary and reading comprehension; R = .80)	- : Reading proficiency(Cohort A) - : Reading proficiency (Cohort B) - : Reading proficiency (Cohort C)	

Table 1 (continued)

Author	Country	N	Study description	Analytic model	Construct, mean reliability	Findings	Comments
Silberglitt & Hintze (2007)	USA	7544	Grade 2 – Grade 6, 3 annual points of measurement	Formation of proficiency groups; Comparison of subsequent growth rates	Decoding efficiency (reading a passage aloud; R = .92)	+ : Decoding efficiency (Grade 2) + : Decoding efficiency (Grade 3) o : Decoding efficiency (Grade 4) o : Decoding efficiency (Grade 5) o : Decoding efficiency (Grade 6)	Lowest and highest performing students show lower growth rates in Grades 4, 5 & 6; No significance test statistic provided
Sonnenschein, Stapleton, & Benson (2010)	USA	6381	Kindergarten – Grade 5 5 points of measurement	LGC;	Reading proficiency (R = .93)	o : Reading proficiency	
Stainthorp & Hughes (2004)	Great Britain	24	Age 6 – Age 11, 3 points of measurement	Formation of proficiency groups; Comparison of subsequent growth rates	Decoding accuracy (word reading accuracy); Decoding speed (single word reading speed); Reading comprehension	- : Decoding accuracy o : Decoding speed o : Reading comprehension	Ceiling effect for decoding accuracy for precocious readers; No significance test statistics provided
Wang, Algozzine, Ma, & Porfeli (2011)	USA	5796	Grade 2, 3 points of measurement	Multilevel growth models	Decoding efficiency (reading connected text aloud; R = .95c)	+ : Decoding efficiency (Individual level)	
Williamson, Appelbaum, & Epanchin (1991)	USA	529	Grade 1 – Grade 8, 8 points of measurement	Growth models	Reading proficiency	+ : Reading proficiency (boys) + : Reading proficiency (girls)	No significant test statistic provided

Note. Findings were coded in the following way: + = Increasing interindividual differences/Matthew effect; o = stable interindividual differences; - decreasing interindividual differences; # = Delayed compensation; c = crossing fan-spread pattern. LGC = Latent Growth Curve, R = Mean reliability of the outcome.

aBoth studies refer to the same dataset. Findings for female students were taken from Jacobsen (1999) and for male students from Jacobsen and Lundberg (2000).

bInformation retrieved from https://dibels.uoregon.edu/market/assessment/measures/nwf.php [04 November 2013].

cInformation retrieved from https://dibels.uoregon.edu/market/assessment/measures/orf.php [04 November 2013].

Reading construct and language.

With regard to the construct of the outcome, measures of decoding speed and decoding efficiency were put together into one category because of the large amount of overlap in their task requirements. Tests of statistical significance were not conducted due to the problem of low expected cell frequencies. However, descriptive results showed an underrepresentation of results reporting Matthew effects in comparison to results reporting compensatory patterns for measures of reading comprehension and reading proficiency (Table 2). For measures of decoding efficiency or speed, a Matthew-effect pattern as well as a pattern of stable achievement differences was overrepresented in comparison to a compensatory pattern. With regard to vocabulary, two studies indicated a Matthew-effect pattern and one study indicated a compensatory pattern. For measures of decoding accuracy, one result indicated a Matthew effect in comparison to four results that indicated stable achievement differences and four results that indicated decreasing achievement differences. Finally, for prereading skills, one result indicated a compensatory developmental pattern.

Developmental patterns were related to the different languages. Languages were categorized into two broad categories reflecting differences in the orthographic consistency of these languages. In the first category, results from mainly English-speaking countries (anglophone regions of Canada, Great Britain, USA) were grouped together. In the second category, languages that are more consistent than English were grouped together (Dutch, German, Greek, Finnish, Swedish). Descriptive results did not show remarkable differences in the reported developmental patterns between these two language categories. This was supported by results of χ^2 statistics that indicated nonsignificant relations between the two language categories and the three main patterns of reading development, $\chi^2 = 2.162$, $df = 2$, *ns*; patterns of delayed compensation and the crossing fan-spread effect were excluded due to the low expected cell frequencies.

Characteristics of the study and the sample.

There seemed to be no clear relationship between students' age or grade level at the first point of measurement and the three main different developmental patterns, $F(2, 65) = 0.466$, *ns*. Results that indicated a pattern of delayed compensation, however, had a tendency to be based on younger students. However, due to a low amount of studies with such a developmental pattern, this finding was not tested for significance. Next, the role of the analytic model and the number of points of measurement was investigated. For the analytic model, there seemed to be a tendency in the direction that simplex models seemed to favor results supporting compensatory models of reading development or a pattern of delayed compensation (χ^2 tests were not conducted due to low expected cell frequencies). Concerning the total number of points of measurement, again, no clear relationship with the reported the achievement patterns was found, $F(2, 68) = 0.529$, *ns*.

Psychometric properties of applied reading measures.

Finally, the measurement properties of the applied instruments were related to the different patterns of reading development. Achievement tests were rated with regard to their difficulty. Test difficulty was assumed to be inappropriate when there was a clear tendency for the emergence of floor or ceiling effects. Descriptive analyses indicated that results based on measures that showed floor or ceiling effects were much more likely to report a decreasing achievement gap in comparison to results based on measures used within an appropriate range of test difficulty (Table 2). Whereas studies that used measures that were not rated as showing floor or ceiling effects had patterns with almost equal amounts of increasing and decreasing achievement differences, results based on measures that showed floor or ceiling effects indicated a pattern of decreasing achievement differences10 times more often than they indicated increasing achievement differences ($\chi^2 = 7.176$, $df = 2$, $p < .05$, for a test of equal cell frequencies within studies reporting floor/ceiling effects; $\chi^2 = 4.478$, $df = 2$, *ns*, for a test of independence of instrument properties and developmental patterns; however, two out of six cells had an expected cell frequency below five. Patterns of delayed compensation and the crossing fan-spread effect were not considered for either test due to low expected cell frequencies.

Second, the mean reliability of the achievement measures' scores was taken into account. Again, a clear pattern occurred. In comparison to results that reported a stable or increasing achievement gap, results that reported decreasing achievement differences were on average based on scores with lower reliability, $F(2, 44) = 12.283$, $p < .001$. The reported average reliabilities of the applied measures' scores were .89 ($SD = .047$, $SE = .014$, $n = 12$) for studies that reported stable achievement differences and .92 ($SD = .037$, $SE = .013$, $n = 8$) for studies that reported increasing achievement differences or a Matthew effect.

However, results indicating a compensatory developmental pattern were based on instruments with a mean reliability of .81 ($SD = .073$, $SE = .014$, $n = 27$). Furthermore, there seemed to be noticeably small overlap in the reported average reliabilities between the different developmental patterns (Figure 4): Whereas the lowest reliability for scores in studies reporting a Matthew effect in reading was .85, almost three quarter ($n = 20$, 74.1%) of the studies reporting a compensatory developmental pattern used instruments with reliabilities of scores of .84 or below.

Table 2

Frequencies of Different Developmental Patterns of Reading Achievement (Total and by Different Moderators)

	Achievement gap					
	Decreasing *n* (%)	Stable *n* (%)	Increasing *n* (%)	Delayed compensation *n* (%)	Crossing fan-spread effect *n* (%)	Ratio Increasing/ Decreasing
Total						
All results (*n* = 78)	33 (42.3%)	20 (25.6%)	18 (23.1%)	6 (7.7%)	1 (1.3%)	0.55
Results by construct						
Reading comprehension (*n* = 24)	16 (66.7%)	4 (16.7%)	3 (12.5%)	1 (4.2%)	0 (0.0%)	0.19
Reading proficiency (*n* = 13)	6 (46.2%)	1 (7.7%)	3 (23.1%)	3 (23.1%)	0 (0.0%)	0.50
Vocabulary (*n* = 3)	1 (33.3%)	0 (0.0%)	2 (66.6%)	0 (0.0%)	0 (0.0%)	2.00
Decoding efficiency/speed (*n* = 27)	5 (18.5%)	11 (40.7%)	9 (33.3%)	1 (3.7%)	1 (3.7%)	1.80
Decoding accuracy (*n* = 10)	4 (40.0%)	4 (40.0%)	1 (10.0%)	1 (10.0%)	0 (0.0%)	0.25
Prereading skills (*n* = 1)	1 (100%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0.00
Results by language (categorized)						
English (*n* = 37)	13 (35.1%)	12 (32.4%)	9 (24.3%)	3 (8.1%)	0 (0.0%)	0.69
Dutch, German, Greek, Finnish, Swedish (*n* = 41)	20 (48.8%)	8 (19.5%)	9 (22.0%)	3 (7.3%)	1 (2.4%)	0.45
Procedure						
Simplex models (*n* = 14)	7 (50.0%)	2 (14.3%)	1 (7.1%)	3 (21.4%)	1 (7.1%)	0.14
Covariance/Correlation Baseline-Growth (*n* = 25)	10 (40.0%)	6 (24.0%)	7 (28.0%)	2 (8.0%)	0 (0.0%)	0.70
Comparison of growth rates between proficiency groups (*n* = 39)	16 (41.0%)	12 (30.8%)	10 (25.6%)	1 (2.6%)	0 (0.0%)	0.63
Instruments						
No constraints (*n* = 59)	23 (39.0%)	14 (23.7%)	17 (28.8%)	4 (6.8%)	1 (1.7%)	0.74
Floor/ceiling effects (*n* = 19)	10 (52.6%)	6 (31.6%)	1 (5.3%)	2 (10.5%)	0 (0.0%)	0.10
Documentation						
Reliability reported (*n* = 50)	27 (54.0 %)	12 (24.0%)	8 (16.0%)	3 (6.0%)	0 (0.0%)	0.30
Reliability not reported (*n* = 28)	6 (21.4%)	8 (28.6%)	10 (35.7%)	3 (10.7%)	1 (3.6%)	1.67

Note. *n* = number of results. The ratio increasing/decreasing was calculated by dividing the number of results reporting an increasing achievement gap by the number of results reporting a decreasing achievement gap within each row.

Table 3

Average Grade Level, Number of Measurement Points, and Mean Outcome Reliability by Developmental Pattern

	Achievement gap					
	Decreasing	Stable	Increasing	Delayed compensation	Crossing fan-spread	
	M (SE)	M (SE)	M (SE)	M (SE)	M (SE)	Significance[a]
Grade Time 1[b] (*n* = 75)	1.97 (0.21)	1.89 (0.39)	1.61 (0.22)	0.50 (0.22)	1 (-)	ns
Number of measurement points (*n* = 78)	4.61 (0.41)	4.10 (0.49)	4.06 (0.41)	5.67 (1.12)	6 (-)	ns
Mean reliability of outcomes (*n* = 50)	0.809 (0.014)	0.885 (0.014)	0.919 (0.013)	0.937 (0.003)	- (-)	$p < .01$

[a]Mean differences were tested using ANOVA; Only the first three columns (decreasing-stable-increasing achievement gap) were considered due to the low number of results in column four and five.

[b]The attendance of preschool/kindergarten was coded with zero.

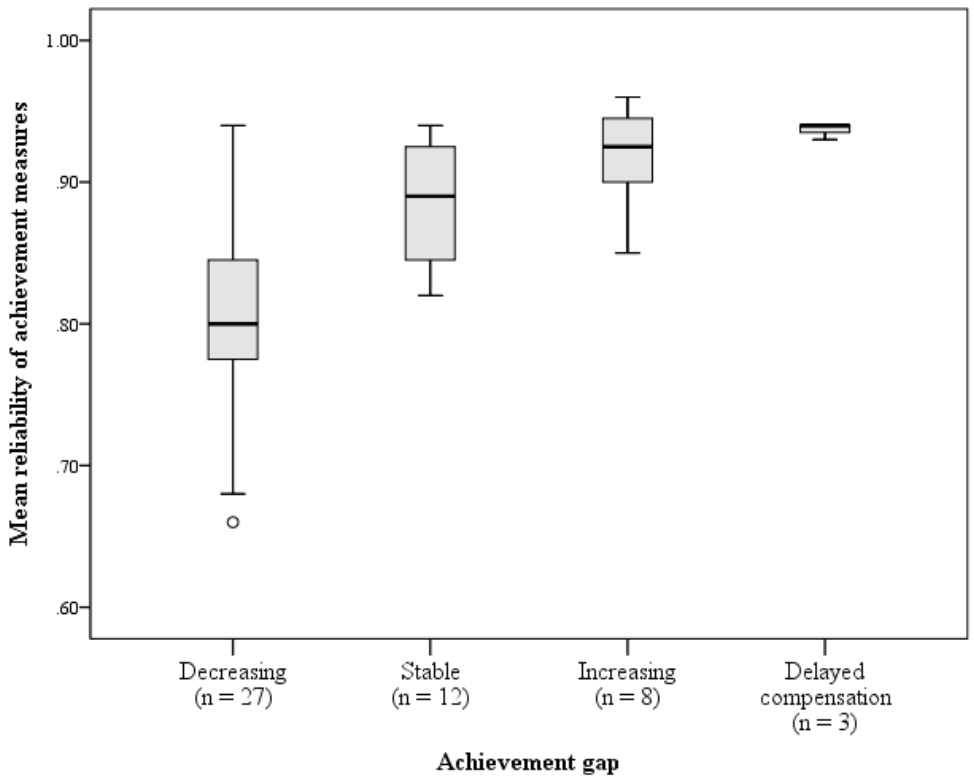

Figure 4. Mean reliabilities of the outcome measures by developmental pattern.

At this point however, we need to note that only 50 studies (64.1%) documented reliabilities of their measures' scores whereas 28 (35.9%) studies did not report any reliability estimates (Table 2). Additional analysis showed that studies not reporting the reliability of their measures' scores more often claimed a Matthew effect in reading than studies reporting reliabilities (χ^2 = 7.748, *df* = 2, $p < .05$; patterns of delayed compensation and the crossing fan-spread effect were excluded due to the low expected cell frequencies).

In a subsequent analysis, all measures were grouped into three categories: Measures affected by low measurement precision, for which either floor or ceiling effects were reported or for which the reported reliability was below .85, measures without such deficits but documented reliability of scores, and measures not reporting floor or ceiling effects and not documenting reliabilities. Relating this combined quality indicator of the achievement measures to the reported pattern of reading development showed a clear trend (Table 4): Measures affected by low measurement precision were strongly linked with a pattern of decreasing achievement differences. Results without such measurement flaws, by contrast, were linked with a pattern of stable and increasing achievement differences. Therefore, studies using measures free of such deficits were almost three times more likely to find Matthew effects in reading than a pattern of decreasing achievement differences, whereas results based on measures affected by low measurement precision reported a compensatory pattern in comparison to a Matthew-effect pattern 27 times more often.

Studies not documenting the reliability of scores on the used reading measures showed a comparable trend of favoring a Matthew effect pattern as did studies that were not affected by low measurement precision (Matthew effects were reported three times more often than decreasing achievement differences). Differences in developmental patterns between measures that were and that were not affected by these psychometric deficits as well as not reporting floor or ceiling effects or reliabilities were statistically significant, $\chi^2 = 28.421$, $df = 4$, $p < .001$; patterns of delayed compensation and the crossing fan-spread effect were not considered due to low expected cell frequencies.

Finally, the measured constructs were again related to the different developmental patterns, but only taking into account results that used measures without the reported measurement precision deficits. Tests of statistical significance were not conducted due to the problem of expected cell frequencies that were too small. However, descriptive results indicated that after taking the psychometric properties of the instruments into account, Matthew effects in reading were more often reported for measures of decoding efficiency or decoding speed and vocabulary than for measures of reading comprehension or decoding accuracy. Nevertheless, especially for measures of vocabulary and decoding accuracy, interpretations are preliminary due to the low number of available research findings.

Table 4

Frequencies of Different Developmental Patterns of Reading Achievement by Psychometric Properties of the Achievement Measures and Construct when using only Achievement Measures without Reported Deficits

	Achievement gap					
	Decreasing *n* (%)	Stable *n* (%)	Increasing *n* (%)	Delayed compensation *n* (%)	Crossing fan-spread effect *n* (%)	Ratio Increasing/ Decreasing
Measurement characteristics						
High precision, reliability documented (*n* = 18)	3 (16.7%)	6 (33.3%)	8 (44.4%)	1 (5.6%)	0 (0.0%)	2.67
No indication of floor/ceiling effects, reliability undocumented (*n* = 22)	3 (13.6%)	6 (27.3%)	9 (40.9%)	3 (13.6%)	1 (4.5%)	3.00
Low precision (*n* = 38)	27 (71.1%)	8 (21.1%)	1 (2.6%)	2 (5.3%)	0 (0.0%)	0.04
Results by construct using measures not affected by low measurement precision[a]						
Reading comprehension (*n* = 10)	3 (30.0%)	4 (40.0%)	2 (20.0%)	1 (10.0%)	0 (0.0%)	0.67
Reading proficiency (*n* = 7)	1 (14.3%)	1 (14.3%)	3 (42.9%)	2 (28.6%)	0 (0.0%)	3.00
Vocabulary (*n* = 2)	0 (0.0%)	0 (0.0%)	2 (100%)	0 (0.0%)	0 (0.0%)	(-)
Decoding efficiency/ speed (*n* = 19)	1 (5.3%)	7 (36.8%)	9 (47.4%)	1 (5.3%)	1 (5.3%)	9.00
Decoding accuracy (*n* = 2)	1 (50.0%)	0 (0.0%)	1 (50.0%)	0 (0.0%)	0 (0.0%)	1.00

Note. *n* = number of results; The ratio increasing/decreasing was calculated by dividing the number of results reporting an increasing achievement gap by the number of results reporting a decreasing achievement gap within each row.

[a]Studies documenting and not documenting the reliability of the used measures were combined.

The question whether two waves of data are sufficient for studying growth has been addressed several times in psychological and educational research (cf. Rogosa, Brandt, & Zimowski, 1982; Singer & Willett, 2003; Willett, 1982). In summary, there is unanimous consensus that more than two waves of data provide better conditions for studying change that studies relying on two data points. Multi-wave studies, in comparison to two-wave studies, allow us to make an inference about the shape of the growth function (e.g., whether the change is linear or accelerating with time). Furthermore, reliability of the estimated growth rate is increased with the number of waves of data. As shown by Willett (1982), growth-rate reliability can be increased considerably by adding further waves of data, especially if inter-individual heterogeneity in the true growth rate is low in comparison to the measurement error variance.

In the current review, two studies just relying on two waves of data were included. Cain and Oakhill (2011), using a series of ANOVAs, did not find a significant interaction between reading level and age for measures of word reading accuracy and reading comprehension between Grade 3 and Grade 6, supporting a pattern of stable achievement differences. For vocabulary, however, a significant interaction was found, indicating a Matthew effect. Good et al. (2009) estimated a slope of progress on measures of decoding efficiency using a nonsense word fluency test. Growth rate was estimated by using a difference score between students' reading performance at the beginning and at the middle of the first grade. In the smaller *Oregon reading first* sample, the highest average growth was found for students with some risk for reading failure. Students with the highest risk level as well as the lowest risk level showed a comparable reading progress rate, but slightly lower than students with some risk for reading failure. For the *DIBELS data system* sample, comprising data for more than 300,000 students, highest average reading growth was reported for students with the highest risk level. Students with lowest level of risk for reading failure showed the lowest reading progress, indicating decreasing individual differences in reading achievement.

In summary, disregarding the findings based on two data points, 32 findings support a pattern of decreasing achievement differences, 17 findings support a pattern of stable achievement differences and a further 17 findings support a Matthew effect pattern. The ratio of studies that reported an increasing to decreasing achievement gap was 0.53, $\chi^2 = 6.818$, $df = 2$, $p < .05$, and, therefore, was slightly lower than 0.55 estimated for all studies. However, the reported finding of marked differences in the reliability of scores on the measures between studies reporting a decreasing achievement gap ($M = 0.81$, $SD = 0.075$, $SE = 0.015$, $n = 26$), a stable achievement gap ($M = 0.89$, $SD = 0.044$, $SE = 0.015$, $n = 9$), or a Matthew effect ($M = 0.92$, $SD = 0.039$, $SE = 0.015$, $n = 7$) persists, $F(2, 39) = 10.572$, $p < .001$. Finally again, studies using measures that were affected by low measurement precision (reliability below .85 or reported floor or ceiling effects) were related to a compensatory developmental pattern whereas studies without such measurement flaws or studies not reporting psychometric properties of their instruments were linked with a pattern of stable and increasing achievement differences ($\chi^2 = 27.260$, $df = 4$, $p < .001$; patterns of delayed compensation and the crossing fan-spread effect were not considered due to low expected cell frequencies). Taken together, the number of measurement points seemed not a relevant technical variable that is linked to the emergence of Matthew effects in reading, neither when it was treated as a continuous variable (Table 3) nor when differentiating between studies relying on two or more than two data points.

Meta-Analytic Findings from Studies Reporting Correlations Between Baseline and Growth Parameters

Regarding the analytic model, 25 results were based on analyses using some form of covariance or correlation between a baseline level and a growth parameter (Table 2). The reported correlations ranged from -.959 to .708. Random-effects model indicate a mean correlation of -.214, $Z_r = -.217$, $SE = 0.091$, $k = 25$, $p < .05$, supporting a compensatory developmental pattern. However, there is strong heterogeneity of effect sizes as indicated by the Q statistic, $Q = 9{,}350.26$, $df = 24$, $p < .001$. Evaluating the influence of the reliability of the measures for the reported developmental pattern, a weighted regression analysis was conducted. Results showed a positive significant relationship of the reported reliability and the Fisher Z transformed correlation coefficient, $B = 4.954$, $SE = 0.914$, $k = 14$, $p < .001$. The model intercept was $B = -4.734$ ($SE = 0.779$, $p < .001$). Therefore, the expected baseline level – growth correlation given a mean reliability of .80 is -.648 ($Z_r = -0.771$). When using scores with a mean reliability of .90, however, the expected correlation is -.269 ($Z_r = -0.276$).

Notwithstanding, in addition to the just documented finding, again a bias within the subset of studies reporting baseline level-growth correlations seems apparent: studies not reporting any reliability of the measures' scores showed higher correlations ($r = .203$, $Z_r = .206$, $SE = 0.135$, k = 11, *ns*) than studies reporting reliabilities ($r = -.497$, $Z_r = -.545$, $SE = 0.118$, $k = 14$, $p < .001$). Finally, studies have again been categorized into studies affected by low measurement precision, indicating the usage of scores with a reliability below .85 or scores that were affected by floor or ceiling effects, studies not using measures without such measurement deficits but documented reliability, and studies not reporting floor or ceiling effects and not reporting reliabilities. The mean correlation was (a) -.581, $Z_r = -.665$, $SE = 0.125$, $k = 11$, $p < .001$ for studies affected by low measurement precision;(b) -.079, $Z_r = -.079$, $SE = 0.183$, k = 5, *ns*, for studies free of such deficits and with documented reliability; and (c) .241, $Z_r = .245$, $SE = 0.138$, $k = 9$, *ns*, for studies which did

not report floor or ceiling effects nor reliabilities. Significant heterogeneity was still present within studies affected by low measurement precision, Q_w = 25.02, df = 10, p < .01, but not within studies free of such deficits and documented reliability, Q_w = 4.83, df = 4, *ns*, or studies not reporting such characteristics, Q_w = 14.29, df = 8, *ns*.

Discussion

The concept of Matthew effects in reading has attracted considerable attention over the last 25 years. This review aimed to summarize empirical results that have been generated by a broad number of researchers dedicated to the analysis of the development of interindividual differences in reading. The first finding of our review was that there is no support for the overall validity of one special developmental pattern for all measures of reading skills in primary school. According to our review, there are slightly more studies that have reported a compensatory pattern of reading achievement development in comparison to a pattern of stable achievement differences as well as a pattern of increasing achievement differences between good and poor readers. However, the total distribution of these finding should not be overemphasized as there is substantial heterogeneity in the reported findings. Furthermore, an exclusion of the findings by Protopapas et al. (2011), who reported a compensatory reading achievement pattern nine times (three cohorts of students x three analytic methods), almost leads to a balance between the number of results supporting each developmental pattern. Because the included studies differed in several important characteristics, the question of moderating variables was raised. Important properties of the studies comprising the measured construct, the sample (age and language of the students, number of measurement points), the applied analytic strategy as well as the measures used (reliabilities and shortcomings of the test scores used) were rated and related to the reported developmental patterns of reading development.

Developmental Patterns and Different Reading Skills

Concerning the measured construct, results support the assumptions made by Paris (2005) that different developmental patterns may result as a function of the constraints of the measured skills. For highly constrained skills, in particular, a compensatory developmental pattern was assumed. Highly constrained skills are usually mastered completely and universally after a short critical developmental period. With regard to the present review, this particularly comprises measures of prereading skills such as letter naming tasks, concepts about print, and so forth, as well as measures of decoding accuracy due to their measurement restrictions (e.g., the percentage of reading errors is a highly skewed distribution with many students achieving zero errors). The results of our review provide much support for the assumption that highly constrained skills lead to a compensatory developmental pattern. Just one single result was found showing a Matthew effect in reading when it was based on measures of decoding accuracy or prereading skills, whereas a compensatory pattern was found five times. Nevertheless, individual differences in these highly constrained skills might contribute to Matthew effects in other less constrained reading skills by influencing the acquisition of these skills (Melby-Lervåg, Lyster, & Hulme, 2012; Scarborough, 2002).

Orthographic Consistency and Developmental Differences

In our second moderator analysis, studies were grouped according to the language of the test. It seemed worthwhile to ask whether differences in the orthographic consistency of languages (cf., Seymor et al., 2003; Ziegler et al., 2010; Ziegler & Goswami, 2006) would moderate the emergence of Matthew effects in reading, as these orthographic differences have consequences for students' acquisition of reading skills (e.g., Georgiou, Parrila, & Papadopoulos, 2008; Spencer & Hanley, 2003). According to the results of this review, however, this does not seem to be the case. The observed pattern of reading development does not seem to be dependent on this underlying feature of the language. Nevertheless, there may be higher order interactions that have not been explored in this review, and the findings should not be generalized to the development of reading in non-European languages.

Effects of the Procedure and Design of the Study

Furthermore, technical aspects of the different studies were taken into account. The question of which analytic method (simplex models, Latent Growth Curve analysis, formation of proficiency groups) seemed most appropriate for analyzing Matthew effects in reading had already been raised by several authors (Aunola et al., 2002; Bast & Reitsma, 1997; Baumert et al., 2012; Leppänen et al., 2004; Parrila et al., 2005). Most studies that specifically compared simplex models and LGC models using the same set of student data reported comparable results across the two approaches (Aunola et al., 2002; Baumert et al., 2012; Leppänen et al., 2004; Parrila et al., 2005) with two exceptions. First, analyzing the development of decoding efficiency, Bast and Reitsma (1997) reported a Matthew effect when LGC models were used and a crossing fan-spread pattern when simplex models were used. Second, Bast and Reitsma (1997), for measures of reading comprehension, and Parrila et al. (2005), for measures of decoding accuracy, reported a pattern of stable achievement differences when LGC models were used and a pattern of delayed compensation when simplex models were used. Overall, however, the results of the current review confirmed the results of Aunola et al. (2002), Baumert et al. (2012), Leppänen et al. (2004), and Parrila et al. (2005), as there did not seem to be a strong effect of the applied analytic method on the observed developmental reading pattern.

A small effect was found, however, whereby studies that used LGC models reported a comparable distribution of developmental patterns similar to studies

that used the approach of forming proficiency groups. Studies that used simplex models, by contrast, seemed to have a tendency to less often produce a Matthew-effect pattern, showing a pattern of delayed compensation instead. One explanation for this preliminary result may be that simplex models produce several results that need to be ordered for interpretation (e.g., a pattern of monotonically increasing variance) and hence include a random or error component several times, whereas LGC models produce one single correlation between the intercept and slope that can be easily assigned to the different developmental pattern (see Bast & Reitsma, 1997, for further discussion of differences between LGC and simplex models).

A second aspect that – from our perspective – seemed important to take into account was the number of measurement points. Again, there was no strong effect. Four to five points of measurement were the rule. Results leading to a pattern of delayed compensation differed slightly from the average as they were based on data with a mean of five to six measurement points. However the standard error was quite high, so this result should not be overemphasized.

Psychometric Properties of the Reading Measures and Reading Development

Last, psychometric properties of the measures that were used in the studies were considered as potential moderators. First, the instruments were rated as to whether they were used within their appropriate range of difficulty or, alternatively, whether they were affected by floor or ceiling effects. Results based on measures showing floor (e.g., Kempe et al., 2011; Parrila et al., 2005), ceiling (e.g., Aarnoutse & van Leeuwe, 2000; Klicpera et al., 1993; Klicpera, Schabmann, & Gasteiger-Klicpera, 2006; Stainthorp & Hughes, 2004), or floor and ceiling effects (e.g., Compton, 2000; Foster & Miller, 2007) significantly more often produced a compensatory pattern of reading development than did studies without these restrictions. There may be two explanations for this finding. The first explanation is that floor and ceiling effects go along with a skewed rather than a normal bell-shaped ability distribution as well as a decreased variance. This situation leads to reduced coefficients of rank-order stability as students are no longer as discernible from each other. Variance restrictions can lead to lower correlation coefficients, increasing the problem of regression toward the mean (Campbell & Kenny, 1999; see below for further discussion). The second explanation is that, whenever ceiling effects are apparent, independent of the question of whether they are task appropriate (e.g., letter naming) or purely due to a limited difficulty range of the measures, the best readers have no chance for further growth, whereas readers who have not yet reached the maximum task score (or minimum score, e.g., for reading errors) will be able to show a further improvement in their skills.

As a second moderator of the properties of the instruments, the average reliability of scores on the achievement measures was taken into account. Again, a significant trend was apparent. Results showing a compensatory developmental pattern were, on average, obtained from measures with lower reliabilities in comparison to results showing a Matthew-effect pattern. There may be one simple explanation for this finding: regression toward the mean. In its simplest form, regression toward the mean refers to the problem or phenomenon that students scoring very high at the first wave of measurement tend not to have as extreme scores at the second wave of measurement (Furby, 1973; Preacher, Rucker, MacCallum, & Nicewander, 2005). Furthermore, regression toward the mean increases as the correlation between two observations decreases (Campbell & Kenny, 1999). Measurement error deflates correlations. Consequently, the lower the reliability of the reading measures' scores, the higher were the effects of regression towards the mean in these studies, and the more likely these studies were to report a decreasing achievement gap pattern. Nevertheless, it might be that the size of reliability coefficients as well as the level of reporting of such properties goes along with further characteristics of the studies that have not been coded and therefore cannot be excluded as an alternative explanation for the reported findings.

Combining estimates of the reliability of the outcomes with the details about the emergence of floor and ceiling effects led to clear results about the presence of Matthew effects. Whereas results based on measures reporting floor or ceiling effects or scores with a low average reliability indicated a compensatory pattern of reading development 27 times more than they indicated a Matthew-effect pattern, results using measures without such deficits indicated a Matthew-effect pattern almost three times more often than they reported a compensatory pattern. Furthermore, Matthew effects were highly likely to occur for measures of decoding efficiency and vocabulary in these studies, but less likely for measures of reading comprehension. However, these are only preliminary findings given the small number of results within each cell, indicating the need for further research and empirical results using measures of reading comprehension and vocabulary.

Meta-analytic Results of Correlation Studies

In order to further underpin our findings, all studies reporting a correlation between a baseline measure and a growth parameter were quantitatively integrated. The findings provided further evidence for the critical role of the properties of the applied reading achievement measures for the emergence of Matthew effects. The mean correlation of all studies was slightly but significantly negative indicating decreasing achievement differences. However, heterogeneity of the correlations was very strong, negating the assumption that differences are solely due to sampling error. Again, score reliability and psychometric deficits of the measures were highly linked to the reported correlations. A strong negative mean correlation supporting a compensatory developmental pattern was found in studies using scores with low reliability or in studies with measures that were affected by floor or ceiling effects. Studies using measures

without such measurement deficits showed a mean correlation near zero supporting a pattern of stable achievement differences with time. Studies that did not document the reliability of their measures' scores and that did not report floor or ceiling effects also showed a mean correlation near zero although slightly positive in descriptive terms.

In summary, the meta-analytic findings of the subset of studies reporting correlations favor a model of stable achievement differences with time under the condition of sufficient reliability of the measures' scores. Therefore, under these conditions and within this subset of studies, neither increasing nor decreasing reading achievement differences are expected. Nevertheless, more than one third of the studies did not report the reliability of scores on the reading measures used, so results need to be generalized cautiously.

Limitations

The small number of high-quality empirical studies dedicated to the analysis of Matthew effects in reading is a serious limitation of this review. This dearth often led to low cell frequencies when we looked for moderating effects and limited the possibility of finding further fine-grained differences among the different studies. Further, each individual estimation or result was used as the unit of analysis, thereby violating the assumption of independent estimates: Studies based on the same sample of students using different outcomes or analytic procedures were treated as equal as were studies based on different samples. Thus, the more results that were reported on the same sample, the higher the impact of this sample. Weighting criteria were not applied because we expected the major sources of variation to occur between the single results, especially between the different measures, in addition to variation due to sample error. This expectation was confirmed as several studies reported different developmental patterns for separate reading measures within the same sample (e.g., Juel, 1988; Parrila et al., 2005). Then, although we were able to define criteria that differentiated the developmental patterns from each other for every type of applied analytic method, we were not able to further map the size of the effect for every single study on a single continuum across the different methods.

Thus, although all applied analytic approaches seemed to allow us to draw valid inferences about the existence of Matthew effects in reading, the reported results that often lacked detail were not transformable to a common metric across the different approaches. Consequently, the application of more sophisticated meta-analytic methods was just feasible within studies reporting correlations. Finally, we need to consider that not all studies, although treated equally throughout our analyses, provide equal evidence with regard to the question of Matthew-effects in reading. Some studies use a more appropriate design (e.g., Bast & Reitsma, 1997, 1998) or sampling plan (e.g., Baumert et al., 2012) than other studies, a fact that has not fully been taken in account within this study and a topic that needs further research.

Conclusion and Implications

The main question of this literature review was whether there is an empirical foundation for the assumption of a widening achievement gap in reading for primary school students. Although our results revealed no simple answer to this question, we were able to clearly describe conditions under which (relative) Matthew effects for reading are likely to occur and conditions under which a compensatory developmental model seems more appropriate. First, when describing the development of inter-individual differences for highly constrained skills, a stable or compensatory developmental model seems most appropriate. Second, with regard to less constrained measures of decoding efficiency, a Matthew-effect pattern or a pattern of stable achievement differences seems to best describe the development of these skills for primary school students. A widening achievement gap seems appropriate for describing the development of students' composite reading scores, although composite reading scores are not easy to interpret because they combine measures of higher and lower level reading skills.

Furthermore, to detect Matthew effects in reading, it is necessary that scores of the applied measures have a high reliability and lack any floor or ceiling effects. This constraint further underpins the importance of developing and using precise, high quality measures for future studies analyzing inter-individual differences in reading development. Finally, with regard to the development of reading comprehension, no unambiguous trend was found. This outcome might be attributable to the relevance of further moderators that have not been taken sufficiently into account and that need further exploration in future research. In summary, subsequent studies on Matthew-effects in reading should pay more attention to the psychometric properties of the measures, use three or more waves of data, and ask for further conditions increasing or decreasing the likelihood of detecting different developmental patterns in reading. In addition, studying Matthew-effects in reading by using a cross-national longitudinal dataset seems of special interest to us. Characteristics of the education system themselves, typically varying across different countries, might be related to the observed patterns of individual differences in the development of reading achievement, a finding that can only be revealed using cross-national datasets.

References

Aarnoutse, C., & van Leeuwe, J. (2000). Development of poor and better readers during the elementary school. *Educational Research and Evaluation, 6*, 251-278. doi:10.1076/1380-3611(200009)6:3;1-A;FT251

Aarnoutse, C., van Leeuwe, J., Voeten, M., & Oud, H. (2001). Development of decoding, reading comprehension, vocabulary and spelling during the elementary school years. *Reading and Writing, 14*, 61-89. doi:10.1023/A:1008128417862

Artelt, C., & Dörfler, T. (2010). Förderung der

Lesekompetenz als Aufgabe aller Fächer. Forschungsergebnisse und Anregungen für die Praxis [Fostering reading literacy is a mission of a school subjects. Research results and practical implications]. In Bayerisches Staatsministerium für Unterricht und Kultus (Ed.), *ProLesen. Auf dem Weg zur Leseschule - Leseförderung in den gesellschaftswissenschaftlichen Fächern* (pp. 13-36). Donauwörth, Germany: Auer.

Aunola, K., Leskinen, E., Onatsu-Arvilommi, T., & Nurmi, J.-E. (2002). Three methods for studying developmental change: A case of reading skills and self-concept. *British Journal of Educational Psychology, 72*, 343-364. doi:10.1348/000709902320634447

Baker, L., Scher, D., & Mackler, K. (1997). Home and family influences on motivations for reading. *Educational Psychologist, 32*, 69-82. doi:10.1207/s15326985ep3202_2

Bast, J., & Reitsma, P. (1997). Mathew effects in reading: A comparison of latent growth curve models and simplex models with structured means. *Multivariate Behavioral Research, 32*, 135-167. doi:10.1207/s15327906mbr3202_3

Bast, J., & Reitsma, P. (1998). Analyzing the development of individual differences in terms of matthew effects in reading: Results from a dutch longitudinal study. *Developmental Psychology, 34*, 1373-1399. doi:10.1037/0012-1649.34.6.1373

Baumert, J., Nagy, G., & Lehmann, R. H. (2012). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary schools? *Child Development, 83*, 1347-1367. doi:10.1111/j.1467-8624.2012.01779.x

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*, 289-328. doi:10.1080/19345740802400072

Boland, T. (1993). The importance of being literate: Reading development in primary school and its consequences for the school career in secondary education. *European Journal of Psychology of Education, 8*, 289-305. doi:10.1007/BF03174083

Bradley, L. (1989). The use of the HOME inventory in longitudinal studies of child development. In M. H. Bornstein & N. A. Krasnegor (Eds.), *Stability and continuity in mental development* (pp. 191-215). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read - a causal connection. *Nature, 301*, 419-421. doi:10.1038/301419a0

Burns, M. S., & Kidd, J. K. (2010). Learning to read. In P. Peterson, E. Baker & B. McGaw (Eds.), *International encyclopedia of education* (3 ed., pp. 394-400). Oxford, United Kingdom: Elsevier.

Butler, S. R., Marsh, H. W., Sheppard, M. J., & Sheppard, J. L. (1985). Seven-year longitudinal study of the early prediction of reading achievement. *Journal of Educational Psychology, 77*, 349-361. doi:10.1037/0022-0663.77.3.349

Cain, K., & Oakhill, J. (2011). Matthew-effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities, 44*, 431-443. doi:10.1177/0022219411410042

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York, NY: The Guilford Press.

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, NY: The Guilford Press.

Carreker, S. H., Neuhaus, G. F., Swank, P. R., Johnson, P., Monfils, M. J., & Montemayor, M. L. (2007). Teachers with linguistically informed knowledge of reading subskills are associated with a Matthew effect in reading comprehension for monolingual and bilingual students. *Reading Psychology, 28*, 187-212. doi:10.1080/02702710601186456

Caspi, A., Bem, D. J., & Elder, G. H. (1989). Continuities and consequences of interactional styles across the life course. *Journal of Personality, 57*, 375-406. doi:10.1111/1467-6494.ep8972739

Ceci, S. J., & Papierno, P. B. (2005). The rhetoric and reality of gap closing. When the "have nots" gain but the "haves" gain even more. *American Psychologist, 60*, 149-160. doi:10.1037/0003-066X.60.2.149

Compton, D. L. (2000). Modeling the growth of decoding skills in first-grade children. *Scientific Studies of Reading, 4*, 219-259.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design & analysis issues for field settings.* Chicago, IL: Rand McNally College Publishing.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*, 934-945. doi:10.1037/0012-1649.33.6.934

Cunningham, A. J., & Carroll, J. M. (2011). Reading-related skills in earlier- and later-schooled children. *Scientific Studies of Reading, 15*, 244-266. doi:10.1080/10888431003706309

Dickinson, D. K., McCabe, A., Anastasopoulos, L., Peisner-Feinberg, E. S., & Poe, M. D. (2003). The comprehensive language approach to early literacy: The interrelationships among vocabulary, phonological sensitivity, and print knowledge among preschool-aged children. *Journal of Educational Psychology, 95*, 465-481. doi:10.1037/0022-0663.95.3.465

Ditton, H., & Krüsken, J. (2009). Denn wer hat, dem wird gegeben werden? Eine Längsschnittstudie zur Entwicklung schulischer Leistungen und den Effekten der sozialen Herkunft in der Grundschulzeit [To those who have, will more be given? A longitudinal study concerning the

development of school achievement and the effects of social background during primary school]. *Journal for Educational Research Online, 1*, 33-61.

Downey, D. B., von Hippel, P. T., & Broh, B. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review, 69*, 613-635. doi:10.1177/000312240406900501

Entwisle, D. R., & Alexander, K. L. (1992). Race, poverty, school composition, and mathematics achievement in the first two years of school. *American Sociological Review, 57*, 72-84.

Foster, W. A., & Miller, M. (2007). Development of the literacy achievement gap: A longitudinal study of kindergarten through third grade. *Language, Speech, and Hearing Services in Schools, 38*, 173-181. doi:10.1044/0161-1461(2007/018)

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*, 3-17. doi:10.1037/0022-0663.88.1.3

Frost, J. (2001). Phonemic awareness, spontaneous writing, and reading and spelling development from a preventive perspective. *Reading and Writing, 14*, 487-513. doi:10.1023/A:1011143002068

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology, 8*, 172-179.

Georgiou, G. K., Parrila, R., & Papadopoulos, T. C. (2008). Predictors of word decoding and reading fluency across languages varying in orthograhpic consistency. *Journal of Educational Psychology, 100*, 566-580. doi:10.1037/0022-0663.100.3.566

Gilger, J. W., Ho, H.-Z., Whipple, A. D., & Spitz, R. (2001). Genotype-environment correlations für language-related abilities: Implications for typical and atypical learners. *Journal of Learning Disabilities, 34*, 492-502. doi:10.1177/002221940103400602

Good, R. H., Baker, S. K., & Peyton, J. A. (2009). Making sense of nonsense word fluency: Determining adequate progress in early first-grade reading. *Reading & Writing Quarterly, 25*, 33-56. doi:10.1080/10573560802491224

Greenfield Spira, E., Storch Bracken, S., & Fischel, J. E. (2005). Predicting improvement after first-grade reading difficulties: The effects of oral language, emergent literacy, and behavior skills. *Developmental Psychology, 41*, 225-234. doi:10.1037/0012-1649.41.1.225

Grigorenko, E. L. (2004). Genetic bases of developmental dyslexia: A capsule review of heritability estimates. *Enfance, 3*, 273-288. doi:10.3917/enf.563.0273

Harlaar, N., Dale, P., & Plomin, R. (2007). Reading exposure: A (largely) environmental risk factor with environmentally-mediated effects on reading performance in the primary school years. *Journal of Child Psychology and Psychiatry, 48*, 1192-1199. doi:10.1111/j.1469-7610.2007.01798.x

Harlaar, N., Spinath, F. M., Dale, P., & Plomin, R. (2005). Genetic influences on early word recognition abilities and disabilities: A study of 7-year-old twins. *Journal of Child Psychology and Psychiatry, 46*, 373-384. doi:10.1111/j.1469-7610.2004.00358.x

Hart, B., & Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology, 28*, 1096-1105. doi:10.1037/0012-1649.28.6.1096

Hildreth, G. (1936). Developmental sequences in name writing. *Child Development, 7*, 291-303.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177. doi:10.1111/j.1750-8606.2008.00061.x

Hohnen, B., & Stevenson, J. (1999). The structure of genetic influences on general cognitive, language, phonological, and reading abilities. *Developmental Psychology, 35*, 590-603. doi:10.1037/0012-1649.35.2.590

Jacobsen, C. (1999). How persistent is reading disability? Individual growth curves in reading. *Dyslexia, 5*, 78-93. doi:10.1002/(SICI)1099-0909(199906)5:2<78::AID-DYS127>3.0.CO;2-8

Jacobsen, C., & Lundberg, I. (2000). Early prediction of individual growth in reading. *Reading and Writing, 13*, 273-296. doi:10.1023/A:1026476712452

Judge, S., & Bell, S. M. (2011). Reading achievement trajectories for students with learning disabilities during the elementary school years. *Reading & Writing Quarterly, 27*, 153-178. doi:10.1080/10573569.2011.532722

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437-447. doi:10.1037/0022-0663.80.4.437

Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology, 78*, 243-255. doi:10.1037/0022-0663.78.4.243

Kempe, C., Eriksson-Gustavsson, A.-L., & Samuelsson, S. (2011). Are there any Matthew effects in literacy and cognitive development. *Scandinavian Journal of Educational Research, 55*, 181-196. doi:10.1080/00313831.2011.554699

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology, 102*, 652-667. doi:10.1037/a0019643

Klauda, S. L. (2009). The role of parents in adolescents' reading motivation and activity. *Educational Psychology Review, 21*, 325-363. doi:10.1007/s10648-009-9112-0

Klicpera, C., Schabmann, A., & Gasteiger-Klicpera, B. (1993). Lesen- und Schreibenlernen während der Pflichtschulzeit: Eine Längsschnittuntersuchung über die Häufigkeit und Stabilität von Lese- und Rechtschreibschwierigkeiten in einem Wiener Schulbezirk [Learning to read and write in school: A longitudinal study on the prevalence and stability of reading and writing difficulties in a Vienna school district]. *Zeitschrift für Kinder- und Jugendpsychiatrie, 21*, 214-225.

Klicpera, C., Schabmann, A., & Gasteiger-Klicpera, B. (2006). Die mittelfristige Entwicklung von Schülern mit Teilleistungsschwierigkeiten im Bereich der Lese- und Rechtschreibschwierigkeiten [The development of reading and spelling skills of students with learning disabilities during elementary school]. *Kindheit und Entwicklung, 15*, 216-227. doi:10.1026/0942-5403.15.4.216

Leppänen, U., Niemi, P., Aunola, K., & Nurmi, J.-E. (2004). Development of reading skills among preschool and primary school pupils. *Reading Research Quarterly, 39*, 72-93. doi:10.1598/RRQ.39.1.5

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.

McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierachical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology, 98*, 14-28. doi:10.1037/0022-0663.98.1.14

McElvany, N., Kortenbruck, M., & Becker, M. (2008). Lesekompetenz und Lesemotivation. Entwicklung und Mediation des Zusammenhangs durch Leseverhalten [Reading literacy and reading motivation: Their development and the mediation of the relationship by reading behavior]. *Zeitschrift für Pädagogische Psychologie, 22*, 207-219. doi:10.1024/1010-0652.22.34.207

Melby-Lervåg, M., Lyster, S.-A. H., & 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

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137*, 267-296. doi:10.1037/a0021890

Morgan, P. L., & Fuchs, D. (2007). Is there a bidirectional relationship between children's reading skills and reading motivation? *Exceptional children, 73*, 165-183.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 International report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timss.bc.edu/pirls2006/intl_rpt.html

National Institute of Child Health and Human Development (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Washington, DC: U.S. Government Printing Office. Retrieved from https://www.nichd.nih.gov/publications/pubs/nrp/pages/smallbook.aspx

Organisation for Economic Co-Operation and Development. (2003). *Literacy skills for the world of tomorrow - Further results from PISA 2000*. PISA: OECD Publishing. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2000/literacyskillsfortheworldoftomorrowfurtherresultsfrompisa2000-publications2000.htm

Olson, R. K., Keenan, J. M., Byrne, B., Samuelsson, S., Coventry, W. L., Corley, R., . . . Hulslander, J. (2011). Genetic and environmental influences on vocabulary and reading development. *Scientific Studies of Reading, 15*, 26-46. doi:10.1080/10888438.2011.536128

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*, 184-202. doi:10.1598/RRQ.40.2.3

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J.-E., & Kirby, J. (2005). Development of individual differences in reading: Results from longitudinal studies in English and Finnish. *Journal of Educational Psychology, 97*, 299-319. doi:10.1037/0022-0663.97.3.299

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

Pfost, M., Dörfler, T., & Artelt, C. (2010). Der Zusammenhang zwischen außerschulischem Lesen und Lesekompetenz. Ergebnisse einer Längsschnittstudie am Übergang von der Grund- in die weiterführende Schule [The relation between extra-curricular reading behavior and reading competence. Results from a longitudinal study at the transition from primary to secondary school]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42*, 167-176. doi:10.1026/0049-8637/a000017

Pfost, M., Dörfler, T., & Artelt, C. (2012). Reading competence development of poor readers in a German elementary school sample. An empirical examination of the Matthew effect model. *Journal of Research in Reading, 35*, 411-426. doi:10.1111/j.1467-9817.2010.01478.x

Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grade. *Journal of Educational Psychology, 94*, 3-13. doi:10.1037/0022-0663.94.1.3

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new

recommendations. *Psychological Methods, 10*, 178-192. doi:10.1037/1082-989X.10.2.178

Protopapas, A., Sideridis, G. D., Mouzaki, A., & Simos, P. G. (2011). Matthew effects in reading comprehension: Myth or reality? *Journal of Learning Disabilities, 44*, 402-420. doi:10.1177/0022219411417568

Reardon, S. F. (2003). *Sources of educational inequality: The growth of racial/ ethnic and socioeconomic test score gaps in kindergarten and first grade*. (Working Paper 03-05R). Population Research Institute, The Pennsylvania State University.

Rescorla, L., & Rosenthal, A. S. (2004). Growth in standardized ability and achievement test scores from 3rd to 10th grade. *Journal of Educational Psychology, 96*, 85-96. doi:10.1037/0022-0663.96.1.85

Rigney, D. (2010). *The Matthew effect. How advantage begets further advantage*. New York, NY: Columbia University Press.

Rodríguez-Brown, F. V. (2010). Family literacy. A current view of research on parents and children learning together. In M. L. Kamil, P. D. Pearson, E. Birr Moje & P. P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 726-753). New York, NY: Tayler & Francis.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726-748. doi:10.1037/0033-2909.92.3.726

Rourke, B. P. (1976). Reading retardation in children: developmental lag or deficit? In R. M. Knights & D. J. Bakker (Eds.), *The neuropsychology of learning disorders: theoretical approaches* (pp. 125-137). Baltimore, MD: University Park Press.

Scarborough, H. S. (1990). Very early language deficits in dyslexic children. *Child Development, 61*, 1728-1743. doi:10.1111/1467-8624.ep9103040636

Scarborough, H. S. (2002). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 97-110). New York, NY: The Guilford Press.

Scarr, S. (1992). Developmental theories for the 1990s: Development and individual differences. *Child Development, 63*, 1-19.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype-environment effects. *Child Development, 54*, 424-435.

Schneider, W. (2009). The development of reading and spelling. Relevant precursors, developmental changes, and individual differences. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood. Findings from a 20 year longitudinal study* (pp. 199-220). New York, NY: Taylor & Francis.

Seymor, P. H. H., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143-174.

Shaywitz, B. A., Holford, T. R., Holahan, J. M., Fletcher, J. M., Stuebing, K. K., Francis, D. J., & Shaywitz, S. E. (1995). A Matthew effect for IQ but not for reading: Results from longitudinal study. *Reading Research Quarterly, 30*, 894-906. doi:10.2307/748203

Shaywitz, S. E., Morris, R., & Shaywitz, B. A. (2008). The education of dyslexic children from childhood to young adulthood. *Annual Review of Psychology, 59*, 451-475. doi:10.1146/annurev.psych.59.103006.093633

Silberglitt, B., & Hintze, J. M. (2007). How much growth can we expect? A conditional analysis of R-CBM growth rates by level of performance. *Exceptional children, 74*, 71-84.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis. Modeling change and event occurence*. New York, NY: Oxford University Press.

Snowling, M. J. (2001). From language to reading and dyslexia. *Dyslexia, 7*, 37-46. doi:10.1002/dys.185

Sonnenschein, S., Stapleton, L. M., & Benson, A. (2010). The relation between the type and amount of instruction and growth in children's reading competencies. *American Educational Research Journal, 47*, 358-389. doi:10.3102/0002831209349215

Spencer, L. H., & Hanley, J. R. (2003). Effects of orthographic transparency on reading and phoneme awareness in children learning to read in Wales. *British Journal of Psychology, 94*, 1-28. doi:10.1348/000712603762842075

Stainthorp, R., & Hughes, D. (2004). What happens to precocious readers' performance by the age of eleven? *Journal of Research in Reading, 27*, 357-372. doi:10.1111/j.1467-9817.2004.00239.x

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407. doi:10.1598/RRQ.21.4.1

Stanovich, K. E. (2000). *Progress in understanding reading. Scientific foundations and new frontiers*. New York, NY: Guilford Press.

Suggate, S. P. (2009). School entry age and reading achievement in the 2006 Programme for International Student Assessment (PISA). *International Journal of Educational Research, 48*, 151-161. doi:10.1016/j.ijer.2009.05.001

Suggate, S. P. (2012). Watering the garden before a rainstorm. The case of early reading instruction. In S. Suggate & E. Reese (Eds.), *Contemporary debates in childhood education and development* (pp. 181-190). London, United Kingdom: Routledge.

Suggate, S. P., Schaughency, E. A., & Reese, E. (2013). Children learning to read later catch up to children reading earlier. *Early Childhood Research Quarterly, 28*, 33-48. doi:10.1016/j.ecresq.2012.04.004

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

Wachtel, P. L. (1994). Cyclical processes in personality and psychopathology. *Journal of Abnormal Psychology, 103*, 51-54.

Walberg, H. J., & Tsai, S.-L. (1983). Matthew effects in education. *American Educational Research Journal, 20*, 359-373. doi:10.3102/00028312020003359

Wang, C., Algozzine, B., Ma, W., & Porfeli, E. (2011). Oral reading rates and second-grade students. *Journal of Educational Psychology, 103*, 442-454. doi:10.1037/a0023029

Welsch, J. G., Sullivan, A., & Justice, L. M. (2003). That's my letter! What preschoolers' name writing representations tell us about emergent literacy knowledge. *Journal of Literacy Research, 35*, 757-776. doi:10.1207/s15548430jlr3502_4

Willett, J. B. (1982). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49*, 587-602. doi:10.1177/001316448904900309

Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal analyses of academic achievement. *Journal of Educational Measurement, 28*, 61-76. doi:10.1111/j.1745-3984.1991.tb00344.x

Wilson, D. B. (2005). Meta-analysis macros for SAS, SPSS, and Stata (Version 2005.05.23). Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., . . . Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science, 21*, 551-559. doi:10.1177/0956797610363406

Ziegler, J. C., & Goswami, U. (2006). Becoming literate in different languages: similar problems, different solutions. *Developmental Science, 9*, 429-453. doi:10.1111/j.1467-7687.2006.00509.x

Authors' Note

We would like to thank John Kirby (Queens University, Kingston, Canada), Tom Nicholson (Massey University Auckland, New Zealand), and Scott G. Paris (Educational Testing Service, Princeton, USA) for their helpful comments on an earlier draft of this article.

We further thank Ingrid Eder and Julia Scholz (both University of Bamberg, Germany) for their assistance.