Hammill Institute on Disabilities

Journal of Learning Disabilities
44(5) 402–420

DOI: 10.1177/0022219411417568
http://jld.sagepub.com
SAGE

Matthew Effects in Reading Comprehension: Myth or Reality?

Athanassios Protopapas[1], Georgios D. Sideridis[2], Angeliki Mouzaki[2], and Panagiotis G. Simos[2]

Abstract

The presence of Matthew effects was tested in students of varying reading, spelling, and vocabulary skills. A cross-sequential design was implemented, following 587 Grade 2 through 4 students across five measurement points (waves) over 2 years. Students were administered standardized assessments of reading, spelling, and vocabulary. Results indicated that the hypothesized fan-spread pattern for Matthew effects was not evident. Low and high ability groups were formed based on 25th and 75th percentile cutoffs on initial measures of spelling, reading accuracy and fluency, vocabulary, and reading comprehension. Multilevel modeling suggested that low and high ability groups had significantly different starting points (intercepts) and their pattern of growth on passage comprehension did not indicate that the gap would increase over time. Instead, some analyses, especially of the youngest cohorts, showed significant convergence. However, there was no evidence of eventually closing the gap. Thus, although the poor students may not be getting poorer, they do not get sufficiently richer either.

Keywords

reading comprehension, Matthew, development, longitudinal, Greek

The ability to comprehend written material has been studied extensively in relation to several contributing factors such as language and cognitive skills, cultural opportunities and engagement in literacy activities, teaching methods and motivation to read, and exposure to reading. The independent or combined effects of the aforementioned factors have instigated various hypotheses regarding reading acquisition. Rapidly accumulating evidence has established strong links between prereading and early language skills and future reading problems, enabling prediction of reading development and identification of children at risk for failure in learning to read. Prevention of reading difficulties has been central in early identification and intervention efforts that aim to reduce special education referrals. Developmental findings have validated such approaches by highlighting the decisive role of low reading achievement at the beginning of school (Aarnoutse, Mommers, Smits, & Van Leeuwe, 1986; Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Juel, 1988; Juel, Griffith, & Gough, 1986; Torgesen & Burgess, 1998). Early identification and early intervention have been found to be more effective in reducing reading difficulties than remediation programs offered later in schools to students who have already presented with reading problems. Accordingly, the stability of reading performance across school years has become a very important issue for longitudinal research along with the implications and progression of differences between competent and poor readers.

An often discussed hypothesis in the reading literature concerning the development of individual differences in reading is known as the "Matthew effect," proposed by Walberg and Tsai (1983) and popularized by Stanovich (1986). According to this hypothesis, good readers improve their reading skills faster than do poor readers over the years by taking advantage of their fluent and unobstructed exposure to reading. In turn, enhanced print exposure supports the consolidation of decoding and word recognition skills and helps to improve lexical knowledge underlying expert reading performance (Joshi, 2005; Stanovich, 1986). Conversely, students experiencing difficulties in learning to read are less likely to have similar reading experiences and to enjoy similar benefits from exposure to print. Thus, the gap between the two groups of students gradually widens, leading to a "fan-spread" effect (Aarnoutse & Van Leeuwe, 2000). The overall conceptual framework of Matthew effects thus hinges on (a) stable rank orderings among students and (b) reciprocal causation among reading skills and reading practice, leading to (c) divergent performance among subgroups with differences

[1]University of Athens, Athens, Greece
[2]University of Crete, Rethimno Crete, Greece

Corresponding Author:
Georgios D. Sideridis, Department of Psychology, University of Crete, GR-741 00 Rethimno, Greece.
Email: sideridis@psy.soc.uoc.gr

in starting skill levels (Bast & Reitsma, 1997; Stanovich, 1986, 2000).

Empirical Investigation of Matthew Effects

Several longitudinal investigations have attempted to confirm the predicted empirical patterns arising from the theoretical framework of Matthew effects by comparing student groups of good and poor readers or by more sophisticated statistical modeling of individual variability across time. A variety of different techniques have been used to analyze longitudinal data from a variety of sources. Some studies have sought to confirm the reciprocal causation hypothesis, examining longitudinal correlations among reading skills and print exposure. These studies have generally reported findings in line with the Matthew framework (e.g., Cunningham & Stanovich, 1997, 1998; Harlaar, Dale, & Plomin, 2007; Mol & Bus, 2011; for a review and discussion of earlier findings see Stanovich, 2000). Immediate effects of practice also seem consistent with the Matthew framework. For example, highest ability children benefited most from story rereadings and retellings and from explanation of new vocabulary (Penno, Wilkinson, & Moore, 2002). Similarly, higher level fifth grade readers benefited more from spellings, and made increasingly larger gains in learning new vocabulary words, than lower level readers (Rosenthal & Ehri, 2008). Another set of studies has generally confirmed the stability of rank orderings among students, such that above average performers tend to remain above average (or at most slip down to average) whereas below average performers remain below average (or at most attain average performance). Some studies have reported extreme stability (e.g., Juel, 1988), whereas others have reported some mobility within a context of substantial stability (Phillips, Norris, Osmond, & Maynard, 2002).

However, a different picture has emerged from studies aiming to confirm the predicted fan-spread effect in reading performance, especially with regard to the most critical skill of passage comprehension. In an early attempt to examine Matthew effects, Shaywitz et al. (1995) followed 396 English-speaking children from kindergarten through Grade 6 and found no divergence in reading skills; however, they were criticized for using standard scores (Bast & Reitsma, 1998; Stanovich, 2000). Bast and Reitsma (1997, 1998) reported a fan-spread pattern for word decoding but not for reading comprehension, in Dutch children followed through Grades 1 through 3. However, Aarnoutse and Van Leeuwe (2000) found more often converging, rather than diverging, performance through Grades 1 through 6 among subgroups of Dutch children with different initial reading skills. Scarborough and Parker (2003) reviewed the literature up to that point and likewise found no negative consequences of diminished reading experience in 57 English-speaking children (including some learning disabled) followed from Grade 1 through Grade 8. Catching up, rather than further falling behind, was also reported by Thomson (2003) for 252 children in a special school for dyslexics. At the other end of the ability spectrum, Stainthorp and Hughes (2004) found that precocious readers simply maintained, rather than increased, their initial advantage in reading skill. Parrila, Aunola, Leskinen, Nurmi, and Kirby (2005) also concluded that individual differences remain stable, and more likely decrease than increase, in a sample of 198 English-speaking children in Ontario followed through Grades 1 through 5 and another sample of 197 children in Finland followed through Grades 1 and 2. Parrila et al. concluded that a compensatory, rather than a cumulative, model of reading development accounted best for the data.

Factors extrinsic to reading skill have recently been brought into the picture, as researchers have examined the potential role of social and demographic variables such as socioeconomic status (SES) and race. McCoach, O'Connell, Reis, and Levitt (2006) and Morgan, Farkas, and Hibel (2008) have followed several thousand children in the United States from kindergarten through Grades 1 and 3, respectively, and found that low-skill children in certain disadvantaged sociodemographic groups were more likely to lag behind, exhibiting lower growth rates. These effects, however established and important, do not constitute evidence for the aforementioned Matthew framework insofar as it is not the interrelation between initial skill and reading exposure per se that causes the performance divergence but, rather, cumulative risk effects of socioeconomic factors. Thus, McCoach et al. concluded against Matthew effects, as reading skills of high-level and low-level readers in their sample, "converged during the school year" (p. 25), whereas Morgan et al. summarized their findings as a "one-sided Matthew effect" (p. 196) for certain high-risk subgroups in the population. Therefore, overall and including the latest studies focusing on sociodemographics, it seems that evidence for a cumulative longitudinal effect of reading practice accentuating initial skill differences is lacking, despite the intuitive appeal of the Matthew framework and the supporting findings for reciprocal relations and stable rank ordering.

What Constitutes Evidence in Favor of Matthew Effects?

According to Bast and Reitsma (1997), the Matthew effect model can be described using a set of interrelated hypotheses. Focusing on reading comprehension, for the Matthew effect to be present, two assumptions must be met. The first assumption states that differences in the development of reading comprehension between low and high ability students will be demonstrated with divergent trajectories of growth, in the context of a stable rank ordering of individual student performance. The second assumption states that the observed differences in the development of reading comprehension are a function of other reading skills (e.g., decoding) or other

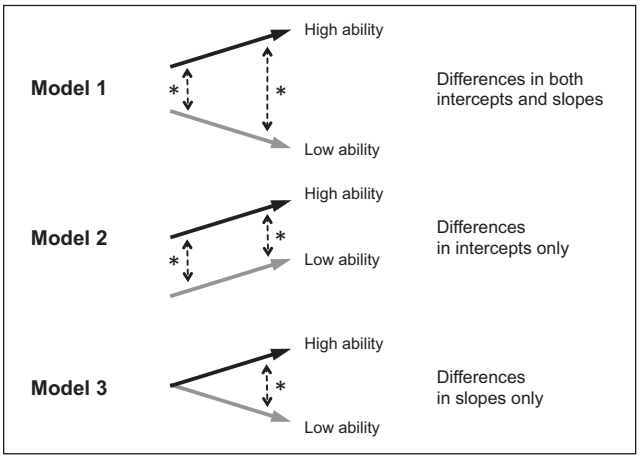

Figure 1. Graphical depiction of models that support (fully or partially) Matthew effects in reading comprehension

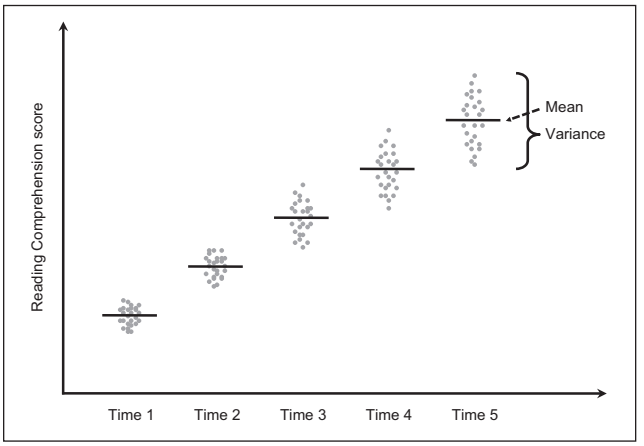

Figure 2. Graphical depiction of the fan-spread hypothesis indicating Matthew effects

cognitive skills (e.g., vocabulary; see Bast & Reitsma, 1998), in a relationship of reciprocal causation.

The Matthew effect model has been criticized as being too philosophical in nature and not amenable to testing by specific mathematical hypotheses (Bast & Reitsma, 1998). Although it is true that no specific interactions of exogenous predictors are posited, it nevertheless can be conceptualized as a growth model by attending to differences in (a) the means and (b) the slopes. A graphic outline of alternative growth models is presented in Figure 1. A qualitative interpretation of the gospel "the poor get poorer and the rich get richer" is presented as Model 1, in which evidence in favor of Matthew effects would be indicated by a significant interaction between means and slopes across the two ability groups. In other words, if both the mean and the slope of the high ability group are significantly different from the mean and slope of the low ability group, then that would be evidence not only that the low ability individuals are at a disadvantage at the start but also that the initial gap between the two groups has expanded over time. The pattern of findings in Model 1 describes a full ramification of the Matthew hypothesis.

To attempt a comprehensive interpretation of previous findings within the Matthew framework, we may consider two alternative patterns of results as providing partial validation of the framework. Model 2, in Figure 1, posits that the two ability groups display similar growth trajectories but different intercepts. Thus, the observed initial difference remains constant over time so that the low ability group will never catch up with the high ability group. Last, Model 3 suggests that although the two groups perform at the same level initially, their growth trajectories diverge over time. This pattern constitutes a partial manifestation of the Matthew effect because the two groups, defined as low and high ability on the basis of some other skill, such as decoding, may not differ in reading comprehension at the earliest learning stages.

Another hypothetical pattern of findings that would be indicative of a Matthew effect has been described as the fan-spread pattern (see Figure 2). This hypothesis states that overall variability in reading performance will increase over time as a result of increasing differences between ability groups. For example, the trajectories of growth for a subsample of high-performing students may be more consistent, occupying the upper part of the distribution, compared to the growth curves of students with lower scores, which may appear more variable and inconsistent. Such a pattern of increased variability can also be said to exhibit a partial Matthew effect (Shaywitz et al., 1995).

What Is the Best Statistical Model to Test These Predictions?

Bast and Reitsma (1997) suggested that growth models, linear and nonlinear, are likely appropriate for modeling the relationships posited by the Matthew model at either the latent or the measured variable level. However, they concluded that a time-series model may be more appropriate, compared to simple growth models, as it will likely model more efficiently the relationship between adjacent data points (or waves) by including an autocorrelation parameter. They recommended the first-order autoregressive model (AR1, or simplex model), which in essence posits that the correlation between measurement waves will likely decrease as the distance between time points increases. Thus, the autocorrelation function will likely decrease when contrasting data five time points apart (i.e., for Lag 5 estimates) compared to Lag 1 or Lag 2 estimates.

The definition of the AR1 process is that the magnitude of the autocorrelation function decays with lag. Although this may be true in many cases, there may be important exceptions. For example, the estimate of the autocorrelation function for decoding will likely be different compared to vocabulary. In

decoding one would expect that the autocorrelation function would be affected (inflated or deflated) because for older students a "plateau" may be reached (ceiling effects). Thus, the lack of variability of later scores, and subsequent smaller range of responses, will likely alter the magnitude of the autocorrelation function. The direction of the effect may go either way, but it is more likely for the estimate of correlation to decrease as variability in early measurements would be associated with "constant" responses at later time points. Thus, the AR1 pattern may not be evident with decoding but may be so with vocabulary, the scoring of which is likely not affected by ceiling effects, as vocabulary development is expected to continue over a wider age range.

The main point is that the first-order autoregressive model will not likely work universally across variables, although it is appealing in its implementation. On the other hand, the linear growth model also accounts for the observed correlational pattern of adjacent data (but not the unique AR1 pattern) and is particularly more appropriate for brief longitudinal data. In fact, the most important aspect of the autoregressive model, not discussed by Bast and Reitsma (1997), is that it requires a large number of repeated observations (20–30 or more) to model the specific time-series autocorrelation pattern. This requirement, however, is unlikely to be met in practice, as longitudinal studies in reading typically extend to 3 to 8 measurement points. With such an extremely small number of observations, and corresponding number of lags ranging between 2 and 7, it is almost impossible to even evaluate an autoregressive or other type of pattern in the data (i.e., moving average or integration or both).

The Present Study

The purpose of the present study was to evaluate the presence of Matthew effects in reading comprehension for Greek elementary school students. The pattern of growth in reading comprehension scores was evaluated as a function of different levels of initial ability. On a narrow reading of the Matthew effect, focusing on comprehension alone, this objective would entail examining the comprehension progress in groups initially differing in comprehension. However, this approach would be subject to regression to the mean, as the criterion variable would be identical to the outcome variable, potentially obscuring any divergent development. Moreover, such restricted focus on a single dimension might prevent discovery of more complex sets of interrelations among component and related skills. Therefore, in this study we examined growth in reading comprehension scores in groups differing in initial ability in spelling, word and pseudoword reading accuracy, vocabulary, and reading fluency, all of which are significant concurrent predictors of reading comprehension in Greek (Protopapas, Sideridis, Mouzaki, & Simos, 2007). We selected these skills based on the "simple view" of reading (Hoover & Gough, 1990), according to which reading comprehension is a function of print-dependent (e.g., decoding, fluency) and print-independent components (e.g., vocabulary, as a proxy of oral language skill; see Protopapas, Simos, Sideridis, & Mouzaki, in press). Fluency was considered important in the context of the "double-deficit" view of reading difficulties (Wolf & Bowers, 1999), which is specifically relevant for regular orthographies (Wimmer, Mayringer, & Landerl, 2000) and in particular for Greek (Papadopoulos, Georgiou, & Kendeou, 2009). We added spelling as a predictor because it requires phonological, orthographic, and semantic skills (Ehri & Wilce, 1979; Nagy & Scott, 2000) and as such constitutes a reliable index of "lexical quality" (Perfetti, 1992) and is thus a strong predictor of both reading itself and component processes (Rosenthal & Ehri, 2008).

Thus, the present study evaluated the two positions of the Matthew effect framework, namely, (a) the fan-spread effect and (b) the predictive ability of decoding, fluency, vocabulary, and spelling as longitudinal determinants of reading comprehension development. For this purpose, we employed a hierarchical linear modeling (HLM; Bryk & Raudenbush, 1992) approach to growth in reading comprehension over time, relative to potential causal predictors. We chose HLM (a) to model individual growth trajectories and thus test the fan-spread pattern of Matthew effects and (b) to examine the effects of specific predictors (such as vocabulary and reading fluency) on intercepts and growth parameters.

Method

Participants

The analysis employed data collected through the University of Crete longitudinal study on the development of reading skills, in which 587 students from 17 public elementary schools in Greece, attending Grades 2 through 4 in the 2004–2005 school year, were followed through Grades 4 through 6 two years later. Participating schools from different regions (Attica, Crete, and Ionian islands) included seven urban, seven semiurban, and three rural schools. Students were first randomly selected from each class. Then, the students whose parents consented to their participation were assessed by the research team. All students were fluent speakers of Greek (including 48 non-native speakers), had never been retained in the same grade, and did not suffer from any physical or mental handicaps necessitating enrollment in special education. The attrition rate was approximately 10% between the first and last assessment. Students who moved and changed school during the study were tracked down in subsequent testing periods and reassessed whenever possible.

For the present study, the full sample of 587 students were followed longitudinally across five consecutive waves (measurement points) spaced approximately 6 months apart. There were 208 second graders (101 boys and 107 girls), 192 third graders (92 boys and 100 girls), and 187 fourth graders

(90 boys and 97 girls) in the initial assessment (Wave 1). Because of missing data points, in the specific analyses reported below the sample size ranged between 464 and 587 students for the different combinations of variables and methods.

Low-and high-skill groups in reading comprehension, word reading accuracy, pseudoword reading accuracy, word reading efficiency (fluency), word spelling, and vocabulary were formed using 25th and 75th percentile cutoff scores. Thus, only half of the sample contributed data points to the analyses involving different ability groups. The choice of the 25th and 75th percentiles was made to maximize differences between low and high ability groups while retaining sufficient levels of statistical power (Cohen, 1992).

Procedures

Participating students were assessed individually in a quiet room at their school by qualified examiners. Examiner qualification was ensured via special training and certification procedures including one-on-one evaluation of testing skills and reliability of administration. Individual assessments were completed within two 45-min sessions, depending on age and individual differences. The children were assessed on a series of measures encompassing reading, cognitive, and behavioral domains. In this article we report findings based on measures of word and pseudoword reading accuracy, reading comprehension, word reading fluency, spelling, and receptive vocabulary.

Measures

Reading comprehension. Reading comprehension was assessed with Subtest 13 of the *Test of Reading Performance* (TORP; Padeliadu & Sideridis, 2000; Sideridis & Padeliadu, 2000). The test included six passages of ascending length (word counts per passage: 19, 26, 51, 65, 97, and 85) each followed by 2 to 4 multiple-choice questions (with four options each). Children were asked to read each passage aloud and then to read and answer all the questions following each passage. Passages and questions were presented on a test booklet and children were allowed to look at the passages while answering the questions. Passages (five narratives, one expository) became progressively more difficult by increasing vocabulary level and syntactic complexity. Most comprehension questions related to story characters and their actions, whereas a few of the later questions concerned story topic and main idea. The total number of questions for the six passages was 18, including 13 explicit, answered with information found directly in the passage, and 5 implicit, involving some "higher" thinking in terms of reader judgment based on the text information. Each was scored with 0 (*correct selection*) or 1 (*incorrect or no response*). Responses were scored during test administration to allow application of a floor-performance discontinuation criterion (when all questions following a passage were answered incorrectly), in which case questions to subsequent (not administered) passages were also scored with 0.

Word and pseudoword reading accuracy. The accuracy of word and pseudoword identification was assessed using Subtests 5 and 6 of the TORP. Subtests 5 and 6 included lists of 40 words and 19 pseudowords, respectively, in order of increasing difficulty, printed in two columns on a single sheet presented for the child to read aloud without time pressure. Words ranged in length from two to five syllables and pseudowords from two to three syllables. Responses were scored with 0 (*inaccurate item reading*), 1 (*correct phoneme sequence but incorrect stress*), or 2 (*phonologically accurate response, including correct stress*). In both subtests, administration was discontinued when students scored 0 on 6 consecutive items.

Word reading fluency. This task included a list of 112 high frequency words, printed on a single sheet in four columns in order of increasing length (1–6 syllables), presented for the child to read aloud in 45 s, as fast as possible without making errors, starting at the top of each column. Words were initially selected on the basis of frequency of appearance in the Hellenic National Corpus (Hatzigeorgiu et al., 2000; http://hnc.ilsp.gr), a corpus of (at the time) approximately 34 million words (tokens) compiled from a wide selection of texts (mainly popular Greek books published after 1990 and daily newspapers). All 112 items in the word list were among the 1,000 most frequent word forms in the corpus. To further ensure that a sufficient number of words visually familiar to the youngest students in the study were included in the list, 30 items were among those appearing in the basic vocabulary selection of the second grade reading textbook used nationwide.

Spelling. Orthographic ability was assessed by spelling to dictation 60 words selected from the basic reading vocabulary for Grade 1 through 6 textbooks. The words were arranged in increasing order of difficulty based on their grade level appearance, confirmed by teacher ratings. The examiner pronounced each word, first in isolation and then in sentence context to demonstrate its use. Children wrote the words in a numbered form after the examiner repeated the word in isolation. Each word was scored with 1 point for accurate spelling, ignoring stress errors (typically omissions of the stress diacritic, which are quite frequent). The selection of the words ensured representation of key instructional units of grammar and spelling rules taught in each grade (i.e., derivation, verb conjugation, and noun or adjective declension suffixes). Testing was discontinued when students scored 0 on 6 consecutive items. A psychometric analysis of this test can be found in Mouzaki, Sideridis, Protopapas, and Simos (2007).

Vocabulary. Receptive vocabulary was assessed by a Greek adaptation of the *Peabody Picture Vocabulary Test–Revised* (PPVT-R; Dunn & Dunn, 1981), in which changes were made in the order of appearance of some items or words and in the items or target words featured in some templates, based

on pilot assessment data tested with the original materials. In this test, each child was asked to identify one picture out of four that best represented the word pronounced by the examiner. Response accuracy was scored with 1 or 0. The test was discontinued after 8 incorrect answers within 10 consecutive questions. Further details on the adaptation and psychometric properties of this Greek version are reported elsewhere (Simos, Sideridis, Protopapas, & Mouzaki, in press).

Research Design

A cross-sequential research design was implemented, in which cohorts of students from the elementary Grades 2, 3, and 4 were followed over a period of 2 years. There was an initial (intercept) estimate of the total raw score on the passage comprehension test at the first assessment (Wave 1; spring of first year) and two measurements per school year (fall and spring, at roughly 6-month intervals) following that, for a total of five measurements per child. This design is more powerful, compared to a strictly longitudinal design in which a single age group is followed over time, because different cohorts are evaluated over time and thus, different populations are assessed for their mean ability levels and also their growth pattern over time.

Data Analyses

Grade-adjusted standard scores for the reading accuracy, fluency, comprehension, and spelling measures and age-adjusted standard scores for PPVT-R vocabulary were used to form ability groups. The total raw passage comprehension score was used in the analyses as a dependent variable.

Stability of parameter estimates. To ensure that the estimated parameters (means) were free of bias, we reestimated the sample means using robust methods (Efron, 1979, 1982; Efron & Tibshirani, 1993). For each sample mean we created 1,000 resamples via sampling without replacement from the original data and estimated the mean of each bootstrap distribution using the following formula,

$$M_{boot} = \frac{1}{k}\sum m^*,$$

with M_{boot} representing the mean of the bootstrap distribution and m^* the mean of each bootstrap sample (k denoting the number of replications, set to the customary 1,000; Chernick, 2007). We estimated the bias of the mean, expressed as the difference between the estimates provided by the sample data and those by the bootstrap distribution. The purpose of this method is to simulate the sampling distribution of a parameter (the mean in our case) and not rely on the estimates of the sample only (which may be biased). A bias of less than 1 unstandardized unit was considered negligible.

Evaluating the fan-spread hypothesis. The sign of the correlation between the intercept and slope of growth in our multilevel models served as the main indicator for the presence of the fan-spread pattern described previously. A negative correlation between an intercept and corresponding slope would indicate that individuals with high initial scores (i.e., intercepts, at Wave 1) tended to have shallower slopes, indicating relatively slower improvement over time, whereas individuals with low initial scores (intercepts) showed steeper slopes, indicative of faster rates of development. Thus, a negative correlation would suggest that lower ability children caught up with children of initially higher ability, arguing against the presence of a Matthew effect. In contrast, a positive correlation would suggest that students with high initial scores tended to have steeper slopes compared to individuals with low initial scores, further diverging over time. Thus, the presence of positive correlations between intercepts and slopes would be indicative of the fan-spread effect consistent with the presence of a Matthew effect (MacCallum, Kim, Malarkey, & Glaser, 1997). A complementary method for evaluating the presence of the fan-spread effect was through a log-linear multilevel model in which the variability at later time points was modeled (instead of fixed). Thus, two multilevel models were simultaneously fitted to the data, one allowing heterogeneous variances at Level 1 and one assuming equal variances. The two models were compared by a chi-square difference test. A significant chi-square in the predicted direction (more variability at later time points) would be indicative of the fan-spread pattern.

Evaluating the differential growth hypothesis. Multilevel modeling was employed to assess the trajectories of growth for each age group on reading comprehension. This method essentially models data from nested structures (Bryk & Raudenbush, 1992; Roberts, 2004). Examples of such structures include observations nested within students, students nested within classrooms, classrooms nested within schools, and so on. In the present study, student observations over time (waves) were nested within student characteristics. This is a "within-between" design, in which the time series constitutes the Level 1 unit and "within-student" information associated with it (i.e., variability in the means and slopes of individual children) is attributed to "between-student" factors (i.e., different ability levels). All analyses were conducted using the HLM 6.1 software (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). The intraindividual level variance at the first level of the analysis (Level 1) was assessed by the following model,

$$Y = \beta_{0j} + \beta_{1ij}X_{1j} + r_{ij},$$

in which the β parameters represent Level 1 coefficients (intercepts and slopes), X_{ij} is the Level 1 predictor for case i belonging to group j, and r_{ij} is the Level 1 random effect (residual),

Table 1. Presence of Bias Between Sample Estimates of Mean and Those of the Bootstrap Distribution for the Mean of the Independent Variables

Independent Variable	Sample Mean	Bias	SEM_{Boot}	95% CI of Mean
Grade 2				
Spelling	22.932	−0.008	0.513	21.894−23.922
Word reading accuracy	70.189	0.005	0.550	69.073−71.243
Pseudoword reading accuracy	25.553	0.161	0.444	24.680−26.485
Vocabulary	103.049	−0.001	1.230	100.504−105.547
Reading fluency	42.296	0.006	0.819	40.627−43.848
Grade 3				
Spelling	32.173	−0.021	0.687	30.743−33.461
Word reading accuracy	74.031	−0.005	0.469	73.126−74.911
Pseudoword reading accuracy	29.372	0.018	0.493	28.388−30.387
Vocabulary	117.445	0.025	1.145	115.221−119.791
Reading fluency	53.859	−0.031	0.967	51.802−55.752
Grade 4				
Spelling	38.102	−0.033	0.754	36.554−39.494
Word reading accuracy	76.247	−0.006	0.326	75.640−76.860
Pseudoword reading accuracy	30.995	0.005	0.482	30.113−31.914
Vocabulary	124.199	−0.051	1.043	122.065−126.263
Reading fluency	61.726	0.033	0.895	59.921−63.466

Note: The confidence intervals around the mean (bootstrap) are the bias corrected accelerated (BCA) intervals (*T* intervals were deemed inappropriate because of the likelihood that some distributions deviated from normality). In all cases the magnitude of bias was negligible. *SEM* = standard error of the mean; CI = confidence interval.

assumed to be normally distributed as $N(0, \sigma^2)$. The slope β_1 models the change in the dependent variable for one unit of change in the X_{ij} predictor (De Leeuw & Hox, 2003; Kreft & de Leeuw, 1998; Shin, Espin, Deno, & McConnell, 2004).

At the interindividual level of analysis (Level 2), which reflects the prediction of Level 1 intercepts and slopes of students belonging to different ability groups, we fit the following model to the data,

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1} W_{1j} + u_{ij},$$

in which the β parameter represents intercepts or slopes at Level 1 that are now modeled as dependent variables for Group *j*. These dependent variables (intercepts and slopes) are predicted at Level 2 by an "s" number of independent variables W_{js}, which represent the groupings (low or high ability students in various subjects). The γ_q coefficient represents the *q* number of slopes at Level 2, and u_{ij} expresses the random error term at this level (normally distributed residual).

What constitutes evidence for the presence of Matthew effects in reading comprehension? As mentioned earlier, there are several findings that would be consistent with a Matthew effect (see Figure 1). The presence of significant between-group differences in intercepts and in slopes (Model 1) would suggest a full expression of the Matthew effect model, with observed initial gaps *expanding* over time. The presence of significant differences in intercepts but not in slopes would indicate that the observed gap at Time 1 is *maintained* over time (Model 2). Alternatively, the presence of significant differences in slopes but not in intercepts would point to the development of an initially nonexistent gap in reading ability over time (Model 3). Among these three models, only the first fully expresses the theoretical consequences of the Matthew framework, whereas the other two may be considered only partial manifestations of the model.

Results

Prerequisite Analyses: Bootstrapping Point Estimates and Levels of Variability

Initially a set of descriptive analyses were conducted to assess the stability of our point estimates (means) and their validity in representing the population. In the presence of large bias parameters the representativeness of our sample would be questionable. The results from the simulation indicated that the mean bias was negligible (see Table 1 for estimates). In only one occasion was the point estimate larger than one tenth of one raw unit. Thus, the bias was almost zero across all tests.

Before testing any formal model it is important to establish that ample levels of variability are present around the parameters of interest. Thus, the following model was fit to the data to assess the variance around the point estimate (mean) of reading comprehension along with its growth parameter:

Table 2. Intercorrelations Between Measured Variables by Grade

Variable	1	2	3	4	5
Grade 2					
1. Reading comprehension	1				
2. Spelling	.079	1			
3. Word reading accuracy	.218**	.543**	1		
4. Pseudoword reading accuracy	.138*	.363**	.491**	1	
5. Vocabulary	.295**	.183**	.252**	.168*	1
6. Reading fluency	.097	.701**	.538**	.339**	.167*
Grade 3					
1. Reading comprehension	1				
2. Spelling	.141	1			
3. Word reading accuracy	.109	.609**	1		
4. Pseudoword reading accuracy	.107	.573**	.565**	1	
5. Vocabulary	.272**	.367**	.404**	.179*	1
6. Reading fluency	.095	.759**	.564**	.515**	.194*
Grade 4					
1. Reading comprehension	1				
2. Spelling	.278**	1			
3. Word reading accuracy	.273**	.636**	1		
4. Pseudoword reading accuracy	.157*	.601**	.564**	1	
5. Vocabulary	.226**	.452**	.444**	.316**	1
6. Reading fluency	.180	.709**	.502**	.474**	.312**

Note: Calculated with data from Wave 1 (first time point).
*$p < .05$. **$p < .01$.

$$\text{LEVEL 1: } RC_i = \beta_{0i} + e_{ti}$$

$$\text{LEVEL 2 : } \beta_{0i} = \gamma_{00} + r_{0i}$$

Fit results indicated a nonzero grand mean of reading comprehension ($M = 10.14$, $p < .001$) and nonzero slope equal to 1.14 units ($p < .001$), confirming that there is information (nonzero) to be modeled. The main interest, however, lies in the variances. Examination of the random effects revealed that 53.4% of reading comprehension variance was at the between-student level, reflecting the intraclass correlation coefficient. The remaining 46.6% of reading comprehension variance was at the within-student level. Thus, there was ample variability around the mean of reading comprehension between as well as within students (across the five measurements). Last, the reliability of the mean reading comprehension score was .775, indicating high consistency of individual student scores around the mean estimate (or in other words homogeneity in individuals' estimates of ability).

To establish that the linear slope (growth of reading comprehension) was nonzero and contained enough variability, the following model was fit to the data:

$$\text{LEVEL 1: } RC_i = \beta_{0i} + \beta_{1i}(Slope_i) + e_{ti}$$

$$\text{LEVEL 2: } \beta_{0i} = \gamma_{00} + r_{0i}$$

$$\beta_{1i} = \gamma_{10} + r_{1i}$$

The results indicated that the within-person variance was reduced by 25.7% by fitting the slope parameter (linear growth of reading comprehension). Thus, 25.7% of the within-student information can be attributed to a linear slope. This amount was both significant and substantial, considering that the total within-student information was 46.6% in the previous model. Although the reliability estimate of the linear slope was low ($\rho = .123$), the model is not invalidated, as estimates would have to fall below .100 to warrant alternative action (such as to fix the parameter rather than to leave it free to vary).

Intercorrelations Between Measured Variables

Table 2 presents the intercorrelations between variables for each age group, revealing that the pattern of correlations was quite stable across grades. This suggests that the *process of reading* is independent of age and rather invariant (see Protopapas et al., 2007, for component skills analysis of the Wave 1 data consistent with invariance). This stability in

the functioning of the variables across age groups is important for generalization and theory development.

Matthew Hypothesis 1: Is There a Fan-Spread Pattern?

As depicted in Figure 2, the fan-spread hypothesis entails that variability between individuals increases over time because of expansion of individual differences in reading ability. As high-achieving students improve faster than low-achieving students, the change in means is accompanied by larger individual differences (i.e., between-student variance). Statistically, this pattern would be manifested as a positive correlation between individual intercepts and slopes in a growth curve model. To estimate this correlation, we fit the following model to the data for each grade:

Model 1

LEVEL 1:

$$RC_i = \beta_{0i} + \beta_{1i}(Slope_i) + e_{ti}$$

LEVEL 2:

$$\beta_{0i} = \gamma_{00} + r_{0i}$$

$$\beta_{1i} = \gamma_{10} + r_{1i}$$

Results indicated that the correlation between intercepts β_0 and corresponding slopes β_1 was strongly negative across grade groups. Specifically, the correlation coefficient was −.756 in Grade 2, −.891 in Grade 3, and −.959 in Grade 4, suggesting the opposite pattern of what a Matthew effect would predict, and thus providing no evidence for a Matthew effect.

In a complementary approach, the fan-spread pattern was evaluated by use of box plots for each grade ($N_1 = 3$) across the five waves ($N_2 = 5$). These findings are presented in Figure 3, demonstrating the decreasing spread. To confirm the finding statistically, the following multilevel log-linear model was fit to the data to test for equality of Level 1 variances (i.e., variance of means across waves).

LEVEL 1:

$$RC_i = \beta_{0i} + \beta_{1i}(Slope_i) + e_{ti}$$

$$Var(r_{ij}) = \sigma^2_{t\xi} \text{ AND } \text{LOG}(\sigma^2_{t\xi}) = \alpha_0 + \alpha_1(Slope_{ij})$$

LEVEL 2:

$$\beta_{0i} = \gamma_{00} + r_{0i}$$

$$\beta_{1i} = \gamma_{10}$$

When fitting this model to the data for each grade group, results indicated heterogeneity of variance in reading comprehension scores, but again in the opposite direction from the one predicted by the fan-spread hypothesis. Across grades, there was a tendency for reading comprehension scores to increase while variance estimates decreased. Thus, the overall conclusion is that the fan-spread pattern was not present in the current study.

Matthew Hypothesis 2: Are the Trajectories of Growth Different Between Different Ability Groups?

To test this hypothesis, several multilevel models were constructed. Initially, the following linear growth model was fit to the data:

Model 2

LEVEL 1:

$$RC_i = \beta_{0i} + \beta_{1i}(Slope_i) + e_{ti}$$

LEVEL 2:

$$\beta_{0i} = \gamma_{01} HiGrp_i + \gamma_{02} LoGrp_i + r_{0i}$$

$$\beta_{1i} = \gamma_{10} HiGrp_i + \gamma_{11} LoGrp_i + r_{1i}$$

The Level 1 equation simply describes a regression analysis problem with β_1 being the growth parameter and β_0 being the intercept. In the Level 2 equations, however, there is no overall intercept. Thus, the terms γ_{01}/γ_{02} and γ_{10}/γ_{11} correspond to the two ability groups, coded as 1 versus 0 (defining group membership vs. absence, respectively). Specifically, γ_{01} equals the mean of the high ability group at Wave 1 and γ_{02} the mean of the low ability group at the same time point, as group-specific intercepts. The coefficients γ_{10} and γ_{11} reflect the predicted slopes of the high and low ability groups, respectively. The purpose of this specification was to be able to directly compare coefficients with each other (between-group comparisons) to decompose interaction effects. In the case of a significant interaction (i.e., different intercepts and slopes for the two groups) a series of simple effects was tested, based on the following model:

Model 3

LEVEL 1:

$$\begin{aligned} RC_1 = {} & \beta_{1i}(\text{TIME1}_i) + \beta_{2i}(\text{TIME2}_i) \\ & + \beta_{3i}(\text{TIME3}_i) + \beta_{4i}(\text{TIME4}_i) \\ & + \beta_{5i}(\text{TIME5}_i) + e_i \end{aligned}$$

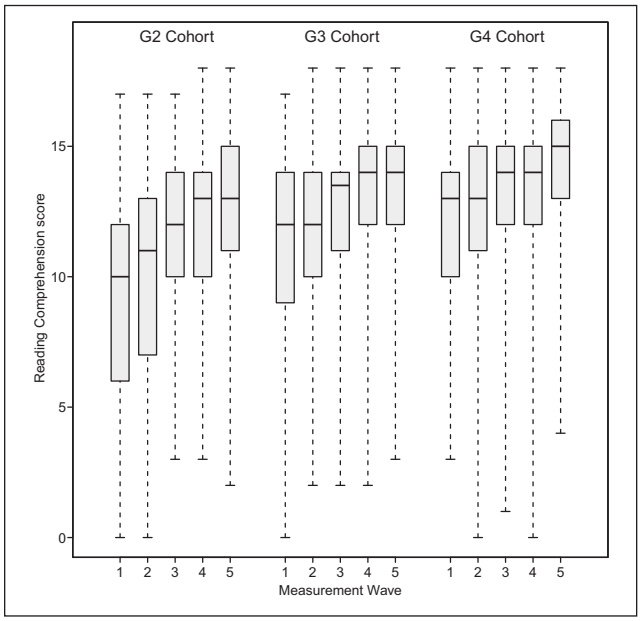

Figure 3. Reading comprehension performance distribution over time for each grade cohort
Note: Boxes enclose 50% of the data (25th–75th percentiles); lines indicate medians; bars extend to the full range. G = grade

LEVEL 2:

$$\beta_{1i} = \gamma_{11} HiGrp_i + \gamma_{12} LoGrp_i + r_{1i}$$

$$\beta_{2i} = \gamma_{21} HiGrp_i + \gamma_{22} LoGrp_i$$

$$\beta_{3i} = \gamma_{31} HiGrp_i + \gamma_{32} LoGrp_i$$

$$\beta_{4i} = \gamma_{41} HiGrp_i + \gamma_{42} LoGrp_i$$

$$\beta_{5i} = \gamma_{51} HiGrp_i + \gamma_{52} LoGrp_i$$

The preceding model in essence represents a repeated-measures ANOVA with a within-person factor having five levels (i.e., the waves) and two between-person grouping variables that each define one ability group (in the absence of an intercept). Thus, differences between the two ability groups can be separately examined for each wave (simple effects). All comparisons were conducted using the multivariate chi-square test with 1 degree of freedom, which has a critical value of 3.84 units at ($p < .05$).

Growth in reading comprehension as a function of initial spelling ability. The modeling of relationships was organized according to grade level. Thus, Model 1 was initially fit to the data of Grade 2 students. Results indicated that both intercepts, $t(110) = 24.961$, $p < .05$, and slopes, $t(483) = 11.639$, $p < .05$, were nonzero, suggesting that the mean of our first measurement had ample variance and that there was a significant trend in the data. We now turn to the focal research question, which is whether the trajectories of growth for Grade 2 students were different between the two ability groups, namely, low and high spelling ability. After fitting Model 2 to the data, intercepts and slopes were tested for invariance between the two groups using chi-square tests of difference. Specifically, an interaction contrast was tested with the intercept and slope of Group 1 (modeled with coefficients 1 and –1) compared with the intercept and slope of Group 2 (modeled with coefficients −1 and 1). This interaction was significant, $\chi^2(2) = 490.861$, $p < .001$, consistent with a difference between the growth trajectories of the two ability groups. In the presence of a significant interaction, several *simple effects* should be tested to clarify the type of interaction (ordinal, disordinal, etc.). One type of comparison involves examination of between-group differences in intercepts and slopes (cf. Figure 1). Results indicated that the intercepts were significantly different between the two groups, $\chi^2(1) = 35.889$, $p < .001$, using robust estimates; the growth slopes, however, were not, $\chi^2(1) = 2.242$, *ns*. Thus, this analysis suggests that the between-group differences observed at the first measurement point (significant effect) were maintained over time (the effect did not change in relation to the intercept). Figure 4a shows the modeled initial ability intercepts and growth slopes for these groups.

An alternative evaluation of this type of interaction involves decomposing the trend by evaluating simple effects. Thus, when fitting Model 3 to the data, post hoc tests allowed a comparison of means in reading comprehension across different time points (similar to an independent samples *t* test across each wave). Significant differences between spelling ability groups were found, favoring the high ability group, in Wave 1, $\chi^2(1) = 20.751$, $p < .05$, Wave 2, $\chi^2(1) = 36.771$, $p < .05$, Wave 3, $\chi^2(1) = 21.253$, $p < .05$, Wave 4, $\chi^2(1) = 15.335$, $p < .001$, and Wave 5, $\chi^2(1) = 20.940$, $p < .001$. This finding suggests that the observed initial differences were maintained over time. Figure 5a shows the estimated group means per measurement wave for this comparison.

Similar findings emerged for Grade 3 students. Specifically, there was a significant interaction between the two groups in their trajectories of growth, $\chi^2(2) = 1028.841$, $p < .001$. When broken down into simple effects, significant intercept differences were found, $\chi^2(1) = 24.981$, $p < .001$, but no growth slope differences, $\chi^2(1) = 1.869$, *ns*. Nevertheless, following the significant interaction, differences in means were tested at each time point with significant differences emerging across all time points: Wave 1: $\chi^2(1) = 21.010$, $p < .001$; Wave 2: $\chi^2(1) = 21.026$, $p < .001$; Wave 3: $\chi^2(1) = 22.428$, $p < .001$; Wave 4: $\chi^2(1) = 15.620$, $p < .001$; Wave 5: $\chi^2(1) = 23.376$, $p < .001$. This suggests that the observed initial spread was maintained over time.

Last, with regard to the Grade 4 cohort, there was again a significant interaction between groups in their growth pattern, $\chi^2(2) = 2132.371$, $p < .001$, because of significant group differences in intercepts, $\chi^2(1) = 24.342$, $p < .001$, but not in

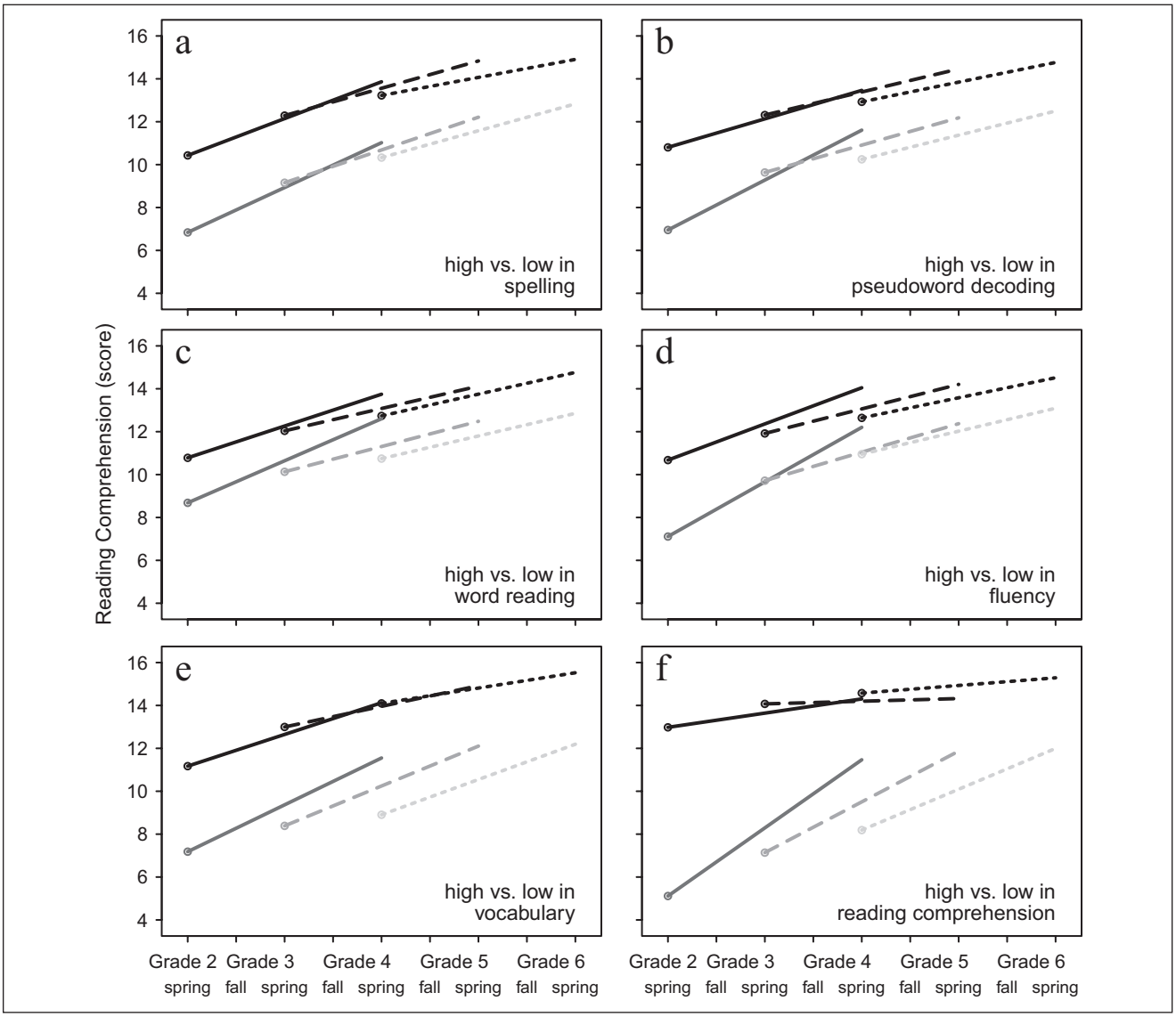

Figure 4. Modeled trajectories of growth (predicted intercepts and slopes) in reading comprehension total scores for low ability (gray) and high ability (black) groups across grade cohorts

Note: Solid, dashed, and dotted lines indicate cohorts first assessed in spring of Grades 2, 3, and 4, respectively. The ability groups were formed with regard to students' scores on the indicated variables.

growth slopes, $\chi^2(1) = 2.162$, *ns*. Post hoc tests for simple effects indicated significant differences at every time point: Wave 1: $\chi^2(1) = 15.798$, $p < .001$; Wave 2: $\chi^2(1) = 23.958$, $p < .05$; Wave 3: $\chi^2(1) = 18.269$, $p < .05$; Wave 4: $\chi^2(1) = 11.247$, $p < .05$; Wave 5: $\chi^2(1) = 11.800$, $p < .05$.

Growth in reading comprehension as a function of initial pseudoword decoding. With high and low ability groups formed on the basis of pseudoword reading accuracy at Wave 1, a significant growth × group interaction (Model 2) for Grade 2 students, $\chi^2(2) = 641.291$, $p < .001$; indicated that the trajectories of growth differed between the two ability groups. Chi-square difference tests indicated significant differences in intercepts, $\chi^2(1) = 42.958$, $p = .05$, and slopes, $\chi^2(1) = 11.814$, $p = .001$, consistent with converging, rather than diverging, comprehension performance (see predicted trajectories of growth in Figure 4b). The analysis of simple effects showed that the two groups were significantly different at Wave 1, $\chi^2(1) = 39.466$, $p < .05$; Wave 2, $\chi^2(1) = 26.850$, $p < .05$; Wave 3, $\chi^2(1) = 17.596$, $p < .05$; and Wave 5, $\chi^2(1) = 21.343$, $p < .05$ (see estimated group means per wave in Figure 5b). For the Grade 3 cohort, the trajectories of growth were different for the two ability groups, as manifested by a significant interaction, $\chi^2(2) = 1376.594$, $p < .001$, but only the differences in intercepts exceeded levels of significance, $\chi^2(1) = 19.165$,

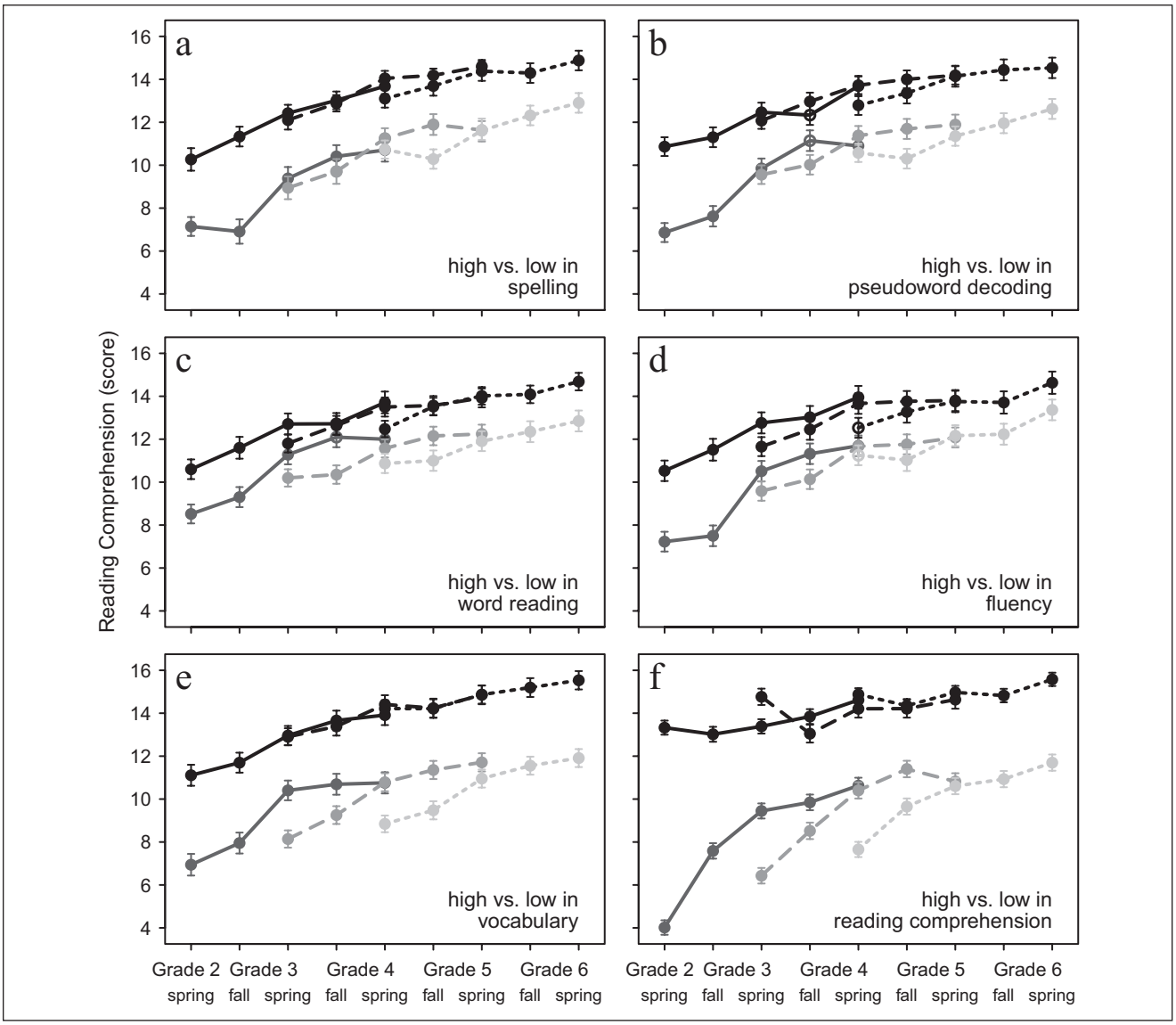

Figure 5. Modeled group means in reading comprehension per measurement point for low ability (gray) and high ability (black) groups across grade cohorts

Note: Solid, dashed, and dotted lines indicate cohorts first assessed in spring of Grades 2, 3, and 4, respectively. Open circles indicate not significantly different group means. Ability groups formed as indicated in Figure 4 note.

$p < .001$. The analysis of simple effects showed that the two groups were significantly different at Wave 1, $\chi^2(1) = 13.515$ $p < .001$; Wave 2, $\chi^2(1) = 19.534$; $p < .001$; Wave 3, $\chi^2(1) = 18.066$, $p < .001$; Wave 4, $\chi^2(1) = 15.695$, $p < .001$; and Wave 5, $\chi^2(1) = 13.665$, $p < .001$. Last, for the Grade 4 cohort, the interaction between growth and group was again significant, $\chi^2(2) = 1083.751$, $p < .001$, showing a significant difference in intercepts only, $\chi^2(1) = 17.853$, $p < .001$. Interaction decomposition indicated that mean differences were significant at Wave 1, $\chi^2(1) = 11.313$, $p < .001$; Wave 2, $\chi^2(1) = 18.086$, $p < .001$; Wave 3, $\chi^2(1) = 18.281$, $p < .001$; Wave 4, $\chi^2(1) = 19.864$, $p < .001$; and Wave 5, $\chi^2(1) = 11.010$, $p < .01$. Therefore, when considering groups varying on decoding ability there was evidence for convergence of reading comprehension slopes for Grade 2 students, in the context of otherwise stable between-group differences.

Growth in reading comprehension as a function of initial word reading accuracy. The same multilevel models were fit with regard to high and low ability groups based on the Wave 1 measure of word reading accuracy. For the Grade 2 cohort, a significant growth × group interaction was observed, $\chi^2(2) = 662.813$, $p < .001$ (Figure 4c). This interaction was decomposable into a significant difference of intercepts only, $\chi^2(1) = 12.080$, $p < .001$, although the slopes were different

by one-tailed test (again consistent with convergence rather than divergence). Most simple effects exceeded significance, specifically at Wave 1, $\chi^2(1) = 9.568$, $p < .001$; Wave 2, $\chi^2(1) = 9.427$, $p < .01$; Wave 3, $\chi^2(1) = 6.149$, $p < .05$; and Wave 5, $\chi^2(1) = 7.696$, $p < .05$ (Figure 5c).

For the Grade 3 cohort, there was again a significant interaction, $\chi^2(2) = 1113.308$, $p < .001$, followed by significant differences in intercepts only, $\chi^2(1) = 12.077$, $p < .01$; slope differences were not significant. Thus, initial between-group differences were maintained over time. Following up on the interaction, tests of simple effects indicated that low-and high-skilled readers were different across all time points: Wave 1: $\chi^2(1) = 6.013$, $p < .05$; Wave 2: $\chi^2(1) = 12.110$, $p < .001$; Wave 3: $\chi^2(1) = 13.839$, $p < .001$; Wave 4: $\chi^2(1) = 7.312$, $p < .01$; Wave 5: $\chi^2(1) = 7.862$, $p < .01$.

Grade 4 findings were in agreement with Grade 3 results. A significant interaction between ability group and slope of reading comprehension, $\chi^2(2) = 1489.690$, $p < .001$, was followed up by a significant difference in intercepts only, $\chi^2(1) = 9.682$, $p < .01$. Simple effects indicated between-group differences at Wave 1, $\chi^2(1) = 5.843$, $p < .05$; Wave 2, $\chi^2(1) = 12.012$, $p < .001$; Wave 3, $\chi^2(1) = 10.183$, $p < .01$; Wave 4, $\chi^2(1) = 7.660$, $p < .01$; and Wave 5, $\chi^2(1) = 11.152$, $p < .01$.

Growth in reading comprehension as a function of initial reading fluency. Forming ability groups on the basis of Wave 1 word reading fluency, a significant interaction was found for Grade 2 students, $\chi^2(2) = 389.045$, $p < .001$, with significant differences in both intercept, $\chi^2(1) = 27.866$, $p < .001$; and slope, $\chi^2(1) = 6.610$, $p < .01$; again consistent with convergence (Figure 4d). Significant between-group differences were observed at Wave 1, $\chi^2(1) = 19.796$, $p < .001$; Wave 2, $\chi^2(1) = 27.165$, $p < .001$; Wave 3, $\chi^2(1) = 13.157$, $p < .01$; Wave 4, $\chi^2(1) = 6.925$, $p < .01$; and Wave 5, $\chi^2(1) = 15.044$, $p < .001$ (Figure 5d). For Grade 3 students, differential Level 1 effects, $\chi^2(2) = 857.088$, $p < .001$; were the result of trajectories of growth being significantly different at the intercept level only, $\chi^2(1) = 11.616$, $p < .01$. Decomposition of the significant interaction indicated that there were between-group differences at Wave 1, $\chi^2(1) = 8.900$, $p < .01$; Wave 2, $\chi^2(1) = 11.704$, $p < .01$; Wave 3, $\chi^2(1) = 11.009$, $p < .01$; Wave 4, $\chi^2(1) = 11.696$, $p < .001$; and Wave 5, $\chi^2(1) = 7.035$, $p < .01$. Last, for Grade 4 students, the significant overall interaction, $\chi^2(2) = 1345.081$, $p < .001$, was decomposed into significant differences in intercepts only, $\chi^2(1) = 7.420$, $p < .05$, consistent with a stable difference between the two ability groups over time. Simple effects tests showed significant differences at Wave 2, $\chi^2(1) = 9.748$, $p < .05$; Wave 3, $\chi^2(1) = 5.046$, $p < .05$; Wave 4, $\chi^2(1) = 4.885$, $p < .05$; and Wave 5, $\chi^2(1) = 4.694$, $p < .05$.

Growth in reading comprehension as a function of initial vocabulary. Defining low and high ability groups on the basis of receptive vocabulary, a somewhat different picture emerged. Specifically, for Grade 2 students, the significant Group × Growth interaction, $\chi^2(2) = 486.973$, $p < .05$, was decomposed into significant differences in both intercepts, $\chi^2(1) = 37.988$, $p < .001$, and slopes, $\chi^2(1) = 4.150$, $p < .05$, consistent with convergent, rather than divergent, performance. The predicted slopes are depicted in Figure 4e. The simple effects tests for each wave showed that the low vocabulary group did not entirely close the gap, as the significant differences were maintained at Wave 1, $\chi^2(1) = 35.579$, $p < .001$; Wave 2, $\chi^2(1) = 24.236$, $p < .001$; Wave 3, $\chi^2(1) = 16.109$, $p < .001$; Wave 4, $\chi^2(1) = 25.019$, $p < .001$; and Wave 5, $\chi^2(1) = 22.966$, $p < .001$ (Figure 5e). For Grade 3 students, the overall growth by group interaction was again significant, $\chi^2(2) = 1416.487$, $p < .001$, arising from a difference in both intercepts, $\chi^2(1) = 66.828$, $p < .001$, and slopes, $\chi^2(1) = 13.499$, $p < .001$. As with Grade 2 students, differences in reading comprehension between low and high vocabulary students were observed across all time points: Wave 1: $\chi^2(1) = 60.621$, $p < .001$; Wave 2: $\chi^2(1) = 42.773$, $p < .001$; Wave 3: $\chi^2(1) = 41.262$, $p < .001$; Wave 4: $\chi^2(1) = 25.934$, $p < .001$; Wave 5: $\chi^2(1) = 36.681$, $p < .001$. Thus, again, the low vocabulary group gradually converged but never caught up to the high vocabulary group in reading comprehension during this period. Last, for Grade 4 students the results were similar, with a significant interaction between group and growth, $\chi^2(2) = 2083.886$ $p < .001$, again decomposed into significant differences between both intercepts, $\chi^2(1) = 105.563$, $p < .001$, and slopes, $\chi^2(1) = 16.666$, $p < .001$. Thus, in this cohort as well, the two ability groups started off at different levels and then converged somewhat, as the lower ability group increasingly approached the higher ability group, without reaching it. Specifically, there were significant between-group differences across all time points: Wave 1: $\chi^2(1) = 100.615$, $p < .001$; Wave 2: $\chi^2(1) = 54.280$, $p < .001$; Wave 3: $\chi^2(1) = 37.909$, $p < .001$; Wave 4: $\chi^2(1) = 44.883$, $p < .001$; Wave 5: $\chi^2(1) = 47.817$, $p < .001$.

Growth in reading comprehension as a function of initial reading comprehension. The expectation in this analysis was that group differences would be more pronounced as the independent and dependent variables originate in the same measurements. The grouping variable reflected the 25th versus 75th percentile dichotomy in reading comprehension, and the dependent variable was the continuous variable of reading comprehension. With regard to Grade 2 data, the interaction between group and growth was significant, $\chi^2(2) = 4054.338$, $p < .001$, with between-group differences in both intercept, $\chi^2(1) = 663.825$, $p < .001$, and slope, $\chi^2(1) = 93.245$, $p < .001$, consistent with convergent growth (Figure 4f). Simple effects analyses showed significant differences between groups across all time points: Wave 1: $\chi^2(1) = 1428.075$, $p < .001$; Wave 2: $\chi^2(1) = 82.373$, $p < .001$; Wave 3: $\chi^2(1) = 60.733$, $p < .001$; Wave 4: $\chi^2(1) = 47.375$, $p < .001$; Wave 5: $\chi^2(1) = 56.499$, $p < .001$ (Figure 5f). Similar findings were observed for Grade 3 students, with the significant interaction, $\chi^2(2) = 6861.135$, $p < .001$, decomposed into significant differences in both intercept, $\chi^2(1) = 341.718$, $p < .001$, and slope, $\chi^2(1) = 89.038$, $p < .001$. Again, all simple effects comparisons were

significant: Wave 1: $\chi^2(1) = 551.283$, $p < .001$; Wave 2: $\chi^2(1) = 47.891$, $p < .001$; Wave 3: $\chi^2(1) = 54.180$, $p < .001$; Wave 4: $\chi^2(1) = 21.786$, $p < .001$; Wave 5: $\chi^2(1) = 47.636$, $p < .001$. Results for Grade 4 students displayed a similar pattern. The significant interaction, $\chi^2(2) = 7398.519$, $p < .001$, was attributed to group differences in both intercept, $\chi^2(1) = 265.982$, $p < .001$, and slope, $\chi^2(1) = 53.121$, $p < .001$. Furthermore, significant between-group differences were observed across all time points: Wave 1: $\chi^2(1) = 393.996$, $p < .001$; Wave 2: $\chi^2(1) = 54.186$, $p < .001$; Wave 3: $\chi^2(1) = 54.530$, $p < .001$; Wave 4: $\chi^2(1) = 48.610$, $p < .001$; Wave 5: $\chi^2(1) = 58.361$, $p < .001$.

Discussion

The purpose of this study was to evaluate the presence of Matthew effects in reading comprehension in a Greek sample of elementary school students. More specifically, the pattern of growth in reading comprehension scores was evaluated as a function of different starting levels of ability in spelling, word and pseudoword reading accuracy, receptive vocabulary, and word reading fluency, as well as in reading comprehension itself. By fitting a series of linear growth curve models and examining specific hypotheses, results failed to provide support for the Matthew framework in the original sense of diverging performance. Instead, evidence in support of a weak version of Matthew effects was obtained in some analyses, in that performance differences between groups differing widely in initial ability were maintained across grades. Lower ability students do not seem to catch up in their reading comprehension, compared to higher ability students, in most cases.

Is the Fan-Spread Pattern of the Matthew Framework Evident?

Not only was the fan-spread pattern not evident across different ability groups, but in fact the opposite was the case. That is, at later time points the scores of the students were more internally consistent and displayed less spread, compared to their scores at earlier time points. Specifically, we found a sizeable negative correlation between intercepts and slopes, indicative of increasing homogeneity rather than heterogeneity of the reading comprehension scores over time. Thus, the fan-spread hypothesis is clearly not supported by the data, in agreement with the findings of Bast and Reitsma (1997, 1998) and Scarborough and Parker (2003) for reading comprehension at comparable ages. Specifically, Bast and Reitsma (1997) found no fan-spread pattern in their linear latent growth curve model, as "the correlation between intercept and growth did not significantly differ from zero" and "no significant variation in linear growth was found" (pp. 154–155). Similarly, an autoregressive latent variable (quasi simplex) model indicated stable individual differences and decreasing latent variance after the first few months of Grade 1. Lack of divergence in comprehension was replicated in a subsequent longitudinal study, using a simplex growth model to examine causal interrelations among several reading-related variables more systematically (Bast & Reitsma, 1998). In contrast, Bast and Reitsma (1997, 1998) found evidence for the fan-spread pattern in the development of decoding skills, indicating divergence between ability groups in word recognition through Grades 1 through 3. This partial dissociation between the development of decoding and comprehension warrants further investigation.

Are There Differences in Growth Trajectories Between Low-and High-Skill Groups?

Differences in intercepts between ability groups suggested that students scoring below the 25th percentile on various reading-related skills achieved lower passage comprehension scores at the first time point (Wave 1), across grades, compared to students scoring above the 75th percentile. This is a way to state the interrelations among reading skills that allow specific comparisons of growth between different skill measures. The initial differences in reading comprehension among the subgroups simply reflect the predictive power of different grouping skills with respect to concurrent reading comprehension (Protopapas et al., 2007).

Table 3 summarizes the slope differences for each grade cohort and grouping variable. When students were assigned to ability groups with regard to high versus low initial performance in spelling, word or pseudoword reading accuracy, or word reading fluency, significant interactions were found between ability group and growth trajectory. Evaluation of the simple effects indicated significant differences in mean reading comprehension score between the two groups across time points. Slope comparisons for Grade 3 and 4 cohorts resulted in no significant differences, indicating stable differences between subgroups at these ages for the duration of the study. Thus, the evidence is consistent with an interpretation that poor spellers or readers generally neither fall behind nor catch up with better spellers or readers in their ability to extract meaning from written text after Grade 2. Exceptions to this pattern were seen when Grade 2 students were grouped on the basis of word and pseudoword reading accuracy and fluency. In these analyses performance convergence was found in reading comprehension between the high and low ability groups (albeit marginally for word reading accuracy), in contradiction of the Matthew effects hypothesis. Thus, when grouping in terms of print-dependent component skills, no differential development of reading comprehension was evident specifically hampering children at the lower end of the ability spectrum. Rather, slight convergence was seen among the younger students, as low reading ability children initially attending Grade 2 caught up somewhat with their higher ability peers. However, this convergence was far from complete, and there was no indication that the gap might be closed in the future.

Table 3. Differences Between Ability Groups in the Trajectories of Reading Comprehension Growth (Slopes Only)

	Cohort		
Ability grouping variable	Grade 2	Grade 3	Grade 4
Spelling	No difference	No difference	No difference
Pseudoword reading accuracy	Convergence	No difference	No difference
Word reading accuracy	No difference[a]	No difference	No difference
Reading fluency	Convergence	No difference	No difference
Vocabulary	Convergence	Convergence	Convergence
Reading comprehension	Convergence	Convergence	Convergence

Note: Significance evaluated by chi-square difference test at $p < .05$, two-tailed.
[a]Convergence by one-tailed test.

The role of vocabulary has been undoubtedly very important for the prediction of reading comprehension (Gough & Tunmer, 1986; Stahl & Fairbanks, 1986). Lexical skills occupy a prominent position in the Matthew framework (Joshi, 2005; Stanovich, 2000). As Sénéchal, Ouellette, and Rodney (2006) note, vocabulary provides the building blocks for reading comprehension. Short-term cumulative effects from vocabulary, consistent with the Matthew framework, have been reported by Penno et al. (2002), who found greater benefits for high vocabulary students from repeated readings and from explanation of new words (see also Robbins & Ehri, 1994). In our data, vocabulary was the strongest concurrent and longitudinal predictor of reading comprehension (Protopapas et al., 2007; Protopapas, Mouzaki, Sideridis, Kotsolakou, & Simos, in press), potentially serving as a proxy for comprehension itself when an independent measure becomes necessary, to guard against regression to the mean. Thus, it is important that in the present study vocabulary grouping was consistently associated with convergence in reading comprehension scores in every grade cohort, in direct contradiction of the Matthew effects framework. This finding indicates that at least some of the observed convergence in the reading comprehension grouping was not entirely due to regression to the mean but may reflect higher rates of improvement for lower ability children.

Thus, our findings are consistent with the conclusions of Aarnoutse and Van Leeuwe (2000), Scarborough and Parker (2003), Thomson (2003), Parrila et al. (2005), and McCoach et al. (2006) in favoring a convergence or compensation rather than a divergence account, despite substantial and reliable differences in study-initial reading comprehension itself or in other print-dependent or print-independent skills.

Reciprocal Causation Without Divergence?

Our study joins a growing list of reports failing to observe significant patterns of divergence among children with differing initial levels of ability. Yet besides being intuitively appealing, the Matthew framework seems well supported in its predictions for reciprocal causation among reading skills, print exposure and reading practice, and cognitive development (Stanovich, 2000). As noted in the introduction, studies have shown moderate reliable correlations between reading comprehension and print exposure measures over a great age range (through adulthood; see the recent meta-analysis by Mol & Bus, 2011). Moreover, higher ability students routinely outperform lower ability peers in benefiting from practice and experience in settings of short-term experimental manipulation. Therefore, there seems to be a contradiction in that the causal story of the Matthew framework appears largely well founded but its predicted consequences prove difficult to establish and may be altogether absent. Given the numbers of children studied, over a range of ages, languages, orthographic systems, and abilities, and the increasingly robust and powerful statistical techniques applied to this question, it seems unlikely that major divergence patterns among differently abled students have been overlooked. Some other explanation is needed to resolve the paradox.

We propose that one possibility may be that the reciprocal causation model, with time-limited relations as posited by Stanovich (1986), may be largely valid, yet the pattern of diverging performance may not follow from it because of *diminishing returns*. That is, lower skill children may have such a large horizon for improvement ahead of them that it is relatively easy to make substantial gains. In contrast, higher skill children are already quite efficient, so additional improvements are relatively more difficult to attain. The outcome of this interplay among level and potential may be that higher ability children indeed make greater gains, but they are not greater in quantity, only in quality, because they result from covering more demanding ground. Therefore, such gains will not be evident in data from psychometric assessment scales. Assuming an equal "learning rate" throughout (cf. the Rescorla–Wagner theory for simple learning; Rescorla & Wagner, 1972, as cited in Anderson, 2000), the ability scale would lead us to predict relatively smaller gains for higher skill children. Yet those children achieve as large gains as lower skill children. Therefore, their "learning

rate" (i.e., the extent to which they benefit from individual learning instances) must be greater, as hypothesized by the Matthew framework.

A complementary form for the diminishing returns argument applies specifically to vocabulary. As is well established, most communication depends on a relatively small set of high frequency words. Most words are low frequency; hence, learning of new words is likely learning of increasingly rarer words. Therefore, for a child with low vocabulary, learning a new word may be quite useful, in that the new word may be more likely to be useful often, whereas for a child with high vocabulary, learning a new word will afford rare opportunity for actually using the new word. Thus, the observable benefit from a single new word will be larger for the low-vocabulary child than for the high-vocabulary child. As a result, the high vocabulary child would be expected to show less relative improvement in vocabulary and in the closely associated skill of reading comprehension. That the relative improvement among high and low vocabulary is evidently equivalent, therefore, may constitute evidence for the high-vocabulary child actually learning many more words for each word learned by the low-vocabulary child. Given the rapid fall off in word frequency of use as rank increases (Zipf, 1949; confirmed for Greek by Hatzigeorgiu, Mikros, & Carayannis, 2001), this might come to an appreciable difference in absolute number of lexical items in favor of the high-skilled children.

Even if this idea is on the right track, the actual operation of causation remains to be established in the long term and for specific hypothesized processes and measures. We may not overlook the relative influence of heritability versus environmental effects by restricting our longitudinal studies to patterns of correlation. For example, it is generally assumed that the long-term correlation among print exposure indices and reading skill reflects a bidirectional causal process (Mol & Bus, 2011). Yet the estimated genetic influences on reading comprehension far outstrip those on print exposure: Harlaar et al. (2007) reported a^2 around .66 for word reading efficiency in 7- to 12-year-olds, compared to .10 for author recognition among 10-year-olds. More impressively, Olson et al. (2011) reported $a^2 = .86$ for Grade 4 reading comprehension, compared to .77 for word recognition and .44 for vocabulary. Although it is not yet obvious how best to interpret these results, one plausible interpretation is that a genetically determined potential for comprehension develops largely through minor environmental interaction while leading to strong activity preferences, that is, children who are better at learning to understand text choose to read more. Whether this reading has a large direct causal effect on their future reading comprehension performance remains to be determined.

Limitations Resulting From Measurement and Analysis Issues

The detection of Matthew effects as patterns of diverging performance is hampered by a number of statistical difficulties. An obvious one relates to the scaling of the critical outcome variable, the rate of development of which is analyzed. Ideally, this should be an interval scale such that amounts of growth are quantitatively comparable at different regions of the scale. In this type of data a 2-point difference around a mean raw score of 10 would be equivalent to a 2-point difference around a mean raw score of 20. Can this ever be achieved for a reading comprehension scale? This is not an issue of standardization. As has long been forcefully argued (e.g., Bast & Reitsma, 1998), the use of standard scores is inappropriate for the study of relative growth because absolutely converging or diverging subgroups of children may nevertheless retain their relative rankings across time, rendering the divergence invisible on standard scores and discernible only in the overall group variance.

The measurement problem is deeper and runs into the definition of the outcome construct itself and its stability throughout the range of observed scores. Two complementary approaches are typically employed: One is to use age-equivalent scores, turning the metric into proportions of average yearly increase. This may initially appear as a valid transformation, but it can never be established by what criteria one year of average growth at, say, Grade 2, is comparable to one year of average growth at Grade 5. The second approach, adopted in our study and others, is to use absolute (raw) scores, ensuring that equal amounts of progress correspond to specific quantifiable indices (e.g., same number of comprehension questions answered correctly). This approach is not without limitations, as it is unclear whether the specific items composing an achievement scale (even one with established convergent and divergent validity) are truly equivalent in some metric sense. Therefore, it is possible to be absolutely confident of a divergence only when the initial performance is indistinguishable. In that case, progress from the same initial point on the scale is estimated for alternative groups (or individuals) and any interval difference is guaranteed to be meaningful. In the case of Matthew effects, we are interested by definition in initially differing groups; therefore, we can never be entirely certain of the meaning of equivalence or difference of growth trajectory slopes.

Another issue potentially affecting studies of comparative growth concerns regression to the mean. This is a well-understood problem occurring whenever a selection criterion is applied on the outcome variable itself. In the case of Matthew effects, this is an issue when the outcome variable is also the grouping variable, and when this has been necessitated by the research question a variety of statistical maneuvers have been applied to remediate, or at least assess, the extent to which the problem may actually occur. In our case, regression to the mean could not have been a determinant of our findings, because in most of our analyses the criterion variable (forming the low and high ability groups) was different from the outcome variable (viz., reading comprehension).

Finally, a host of statistical issues arise when considering how best to investigate the qualitative predictions of the

Matthew framework in quantitative stochastic terms. Bast and Reitsma (1997, 1998) have discussed alternative approaches at length and have advocated the use of a simplex model that includes autoregressive effects. This model, although likely more sensitive to carryover effects, could not be implemented in the present study with only five time points. Instead, we attempted to decompose any trend in the data by testing adjacent time points for changes in their point estimates. Although our approach may have failed to model properly the instability in rank order among individuals, it is probably the most appropriate one for evaluating directly the correlations and interactions between intercepts and slopes (Bast & Reitsma, 1997). Moreover, neither Bast and Reitsma's (1997, 1998) nor Parrila et al.'s (2005) use of a variety of alternative statistical approaches led to any substantial divergence in findings or conflicting interpretations. We are therefore reasonably confident in the reliability of our findings for the age range and skills tested.

Educational Implications and Future Directions

A significant increasing difference in the trajectories of growth over time would suggest a full manifestation of the Matthew model. That was not the case in the present study. Instead, we found that the gap observed at the initial measurement point either decreased or remained stable during the study period. In any case, there was no indication of eventual closing of the gap, even when statistically significant convergence was observed. In our view, this *partial* manifestation of the model is severe enough to warrant remedial action. The fact that low skilled groups may never catch up to the levels of their higher ability peers has grave educational implications. Obviously, the most immediate implication is the need to provide early interventions for reading. These interventions may be more important for early reading skills, such as decoding, that provide the prerequisites for subsequent learning. In the absence of basic reading skills, no further knowledge can be built.

In the future it will be interesting to explore multivariate models that combine cognitive and noncognitive factors as predictors of reading ability. For example, how home variables predict reading levels (Bakermans-Kranenburg, Bradley, & IJzendoorn, 2005) and how cognitive and noncognitive variables interact in the prediction of reading comprehension have not yet been ascertained in the context of Matthew effects. Also, the exploration of dynamic multivariate models (e.g., latent class models) may further aid our understanding of such complex phenomena and the patterns of variability displayed by different subgroups of students. After all, the variability and lack of homogeneity among individuals with challenging learning styles (such as students with learning disabilities) have been with the field of learning disabilities since its inception.

Acknowledgments

We are grateful to Areti Kotsolakou for help with the literature review.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Aarnoutse, C. A. J., Mommers, M. J. C., Smits, B. W. G. M., & Van Leeuwe, J. F. J. (1986). De ontwikkeling en samenhang van technisch lezen, begrijpend lezen en spellen [Development and relation between decoding, reading comprehension and spelling]. *Pedagogische Studiën*, *63*, 97–110.

Aarnoutse, C., & Van Leeuwe, J. (2000). Development of poor and better readers during the elementary school. *Educational Research and Evaluation*, *6*, 251–278.

Anderson, J. R. (2000). *Learning and memory: An integrated approach*. New York, NY: John Wiley.

Bakermans-Kranenburg, M., Bradley, R., & IJzendoorn, M. (2005). Those who have, receive: The Matthew effect in early childhood intervention in the home environment. *Review of Educational Research*, *75*, 1–26.

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

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.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Chernick, M. (2007). *Bootstrap methods: A guide for practitioners and researchers*. New York, NY: John Wiley.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

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.

Cunningham, A. E., & Stanovich, K. E. (1998). What reading does for the mind. *American Educator*, *22*, 8–15.

De Leeuw, E. D., & Hox, J. J. (2003). The use of meta-analysis in cross-national studies. In J. A. Harkness, F. J. R. van de Vijver & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 329–345). New York, NY: John Wiley.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test–Revised*. Circle Pines, MN: American Guidance Service.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B. (1982). *The jackknife, bootstrap, and other resampling plans*. Philadelphia, PA: SIAM.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Ehri, L., & Wilce, L. (1979). The mnemonic value of orthography among beginning readers. *Journal of Educational Psychology*, *71*, 26–40.

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.

Gough, P., & Tunmer, W. (1986). Decoding, reading and reading disability. *Remedial and Special Education*, *7*, 6–10.

Harlaar, N., Dale, P. S., & 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.

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., & Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari, E., Papageorgiou, H., Demiros, I. (2000, May–June). Design and implementation of the online ILSP corpus. In M. Gavrilidou, G. Carayannis, S. Markantonatu, S. Piperidis, & G. Stainhaouer (Eds.). *Proceedings of the Second International Conference of Language Resources and Evaluation (LREC)* (Vol. 3, pp. 1737–1740). Athens, Greece.

Hatzigeorgiu, N., Mikros, G., & Carayannis, G. (2001). Word length, word frequencies and Zipf's law in the Greek language. *Journal of Quantitative Linguistics*, *8*, 175–185.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, *2*, 127–160.

Joshi, R. M. (2005). Vocabulary: A critical component of comprehension. *Reading and Writing Quarterly*, *21*, 209–219.

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.

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.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

MacCallum, R., Kim, C., Malarkey, W., & Glaser, J. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*, 215–253.

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

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.

Morgan, P. L., Farkas, G., & Hibel, J. (2008). Matthew effects for whom? *Learning Disability Quarterly*, *31*, 187–198.

Mouzaki, A., Sideridis, G., Protopapas, A., & Simos, P. (2007). Διερεύνηση των ψυχομετρικών χαρακτηριστικών μιας δοκιμασίας ορθογραφικής δεξιότητας μαθητών της Β′, Γ′, Δ′, και Ε′ τάξης του δημοτικού σχολείου [Investigation of the psychometric characteristics of a spelling skill test for students of elementary school Grades 2, 3, 4, and 5]. *Epistimes tis Agogis*, *1*, 129–146.

Nagy, W., & Scott, J. (2000). Vocabulary processes. In M. Kamil, P. Mossentha, P. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 269–284). Mahwah, NJ: Lawrence Erlbaum.

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.

Padeliadu, S., & Sideridis, G. D. (2000). Discriminant validation of the Test of Reading Performance (TORP) for identification of children with reading difficulties. *European Journal of Psychological Assessment*, *16*, 139–146.

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

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

Penno, J. F., Wilkinson, I., & Moore, D. W. (2002). Vocabulary acquisition from teacher explanation and repeated listening to stories: Do they overcome the Matthew effect? *Journal of Educational Psychology*, *94*, 23–33.

Perfetti, C. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145–174). Hillsdale, NJ: Lawrence Erlbaum.

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 grades. *Journal of Educational Psychology*, *94*, 3–13.

Protopapas, A., Mouzaki, A., Sideridis, G. D., Kotsolakou, A., & Simos, P. G. (in press). The role of vocabulary in the context of the simple view of reading. *Reading and Writing Quarterly*.

Protopapas, A., Sideridis, G. D., Mouzaki, A., & Simos, P. G. (2007). Development of lexical mediation in the relation between reading comprehension and word reading skills in Greek. *Scientific Studies of Reading*, *11*, 165–197.

Protopapas, A., Simos, P. G., Sideridis, G. D., & Mouzaki, A. (in press). The components of the simple view of reading: A confirmatory factor analysis. *Reading Psychology*.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Robbins, C., & Ehri, L. C. (1994). Reading storybooks to kindergartners helps them learn new vocabulary words. *Journal of Educational Psychology*, *86*, 54–64.

Roberts, K. J. (2004). An introductory primer on multilevel and hierarchical linear modeling. *Learning Disabilities: A Contemporary Journal*, *2*, 30–38.

Rosenthal, J., & Ehri, L. C. (2008). The mnemonic value of orthography for vocabulary learning. *Journal of Educational Psychology*, *100*, 175–191.

Scarborough, H. S., & Parker, J. D. (2003). Matthew effects in children with learning disabilities: Development of reading, IQ, psychosocial problems from Grade 2 to Grade 8. *Annals of Dyslexia*, *53*, 47–71.

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In S. Newman & D. Dickinson (Eds.), *Handbook of early literacy research* (pp. 173–182). New York, NY: Guilford.

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

Shin, J., Espin, C. A., Deno, S. L., & McConnell, S. (2004). Use of hierarchical linear modeling and curriculum-based measurement for assessing academic growth and instructional factors for students with learning difficulties. *Asia Pacific Education Review*, *5*, 136–148.

Sideridis, G. D., & Padeliadu, S. (2000). An examination of the psychometric properties of the Test of Reading Performance (TORP) with elementary school students. *Psychological Reports*, *86*, 789–802.

Simos, P., Sideridis, G. D., Protopapas, A., & Mouzaki, A. (in press). Psychometric evaluation of a receptive vocabulary test for Greek elementary students. *Assessment for Effective Intervention.*

Stahl, S., & Fairbanks, M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*, 72–110.

Stainthorp, R., & Hughes, D. (2004). What happens to precocious readers' performance by the age of eleven? *Journal of Research in Reading*, *27*, 357–372.

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

Stanovich, K. E. (2000). *Progress in understanding reading*. New York, NY: Guilford.

Thomson, M. (2003). Monitoring dyslexics' intelligence and attainments: A follow-up study. *Dyslexia*, *9*, 3–17.

Torgesen, J. K., & Burgess, S. R. (1998). Consistency of reading-related phonological processes throughout early childhood: Evidence from longitudinal-correlational and instructional studies. In J. Metsala & L. Ehri (Eds.), *Word recognition in beginning reading* (pp. 161–188). Hillsdale, NJ: Lawrence Erlbaum.

Walberg, H. J., & Tsai, S. (1983). Matthew effects in education. *American Educational Research Journal*, *20*, 359–373.

Wimmer, H., Mayringer, H., & Landerl, K. (2000). The double-deficit hypothesis and difficulties in learning to read a regular orthography. *Journal of Educational Psychology*, *92*, 668–680.

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

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

About the Authors

Athanassios Protopapas, PhD, is an associate professor of cognitive science at the University of Athens, working on the representation and processing of spoken and written words. As a principal researcher at the Institute for Speech and Language Processing until 2010 he worked on psychoeducational assessment and computer-based screening.

Georgios D. Sideridis, PhD, is associate professor of research methods and applied statistics at the Department of Psychology, University of Crete. His research interests lie in the area of motivation and underachievement in students with and without learning disabilities.

Angeliki Mouzaki, PhD, is lecturer of special education and language disorders at the Department of Primary Education, University of Crete. Her research interests include the development of reading and spelling skills and related disorders and the identification and interventions for struggling readers in primary grades.

Panagiotis G. Simos, PhD, is professor of developmental neuropsychology at the Department of Psychology, University of Crete. His research focuses on neuropsychological and brain imaging studies of language, reading, and memory, exploring psychoeducational and neurophysiological profiles associated with specific reading disability and ADHD.