Hammill Institute on Disabilities

Journal of Learning Disabilities
1–16
© Hammill Institute on Disabilities 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0022219414559974
journaloflearningdisabilities.sagepub.com
SAGE

In Search of Matthew Effects in Reading

Athanassios Protopapas, PhD[1], Rauno Parrila, PhD[2], and Panagiotis G. Simos, PhD[3]

Abstract

The concept of Matthew effects in reading development refers to a longitudinally widening gap between high achievers and low achievers. Various statistical approaches have been proposed to examine this idea. However, little attention has been paid to psychometric issues of scaling. Specifically, interval-level data are required to compare performance differences across performance ranges, but only ordinal-level data are available with current literacy measures. To demonstrate the interpretability problems of contrasting growth slopes, we use data from a longitudinal study of literacy development. We explore the possibility of comparing across ages, matched for performance, and we examine the consequences of nonlinear growth, temporal lag estimates, and individual differences in developmental progression. We conclude that, although conceptually appealing, the widening gap prediction is not empirically testable.

Keywords

methodological issues, reading, longitudinal, research method

The concept of Matthew effects in reading development refers to a longitudinally widening gap between high achievers and low achievers (Stanovich, 1986; Walberg & Tsai, 1983). That is, children at a low initial level of performance for a given skill are hypothesized to remain at a relatively lower level in this skill and other related skills that depend on it. Moreover, their rate of development is slower than the rate of development of children with high initial levels of performance. If true, the ensuing performance divergence, as children with poorer skills fall increasingly behind, poses significant challenges for educators and educational systems.

The Matthew hypothesis invokes reciprocal causation among different variables that contribute to reading skills. For good readers, enhanced print exposure supports the consolidation of decoding and word recognition skills and helps them improve lexical knowledge underlying expert reading performance (Joshi, 2005; Stanovich, 1986). In contrast, students with reading difficulties are unlikely to accumulate comparable reading experiences and thereby to obtain similar benefits from exposure to print. Thus, the gap between initially low-performing and high-performing students gradually widens, leading to a “fan spread” effect (Aarnoutse & Van Leeuwe, 2000) and divergent performance among subgroups differing in starting skill levels (Bast & Reitsma, 1997; Stanovich, 1986, 2000). Poor readers are consequently more likely to show reduced rates of growth of word recognition and fluency skills and progressively higher risk to demonstrate deficient performance on increasingly more demanding reading comprehension tasks (compared to their peers).

In the present study we are not concerned with the reciprocal relations among different components of developing reading skills. Instead, we focus on the longitudinal examination of the (within-construct) widening gap among children of initially differing performance, which has been difficult to establish for various reading skills. Several longitudinal investigations have attempted to confirm the predicted empirical patterns arising from the theoretical framework of Matthew effects by comparing 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 range of sources, with inconsistent and often negative findings (e.g., Aarnoutse & Van Leeuwe, 2000; Bast & Reitsma, 1997, 1998; Cain & Oakhill, 2011; Huang, Moon, & Boren, 2014; Leppänen, Niemi, Aunola, & Nurmi, 2004; Luyten & ten Bruggencate, 2011; Morgan, Farkas, & Wu, 2011; Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005; Protopapas, Sideridis, Mouzaki, & Simos, 2011; Scarborough & Parker, 2003; B. A. Shaywitz et al., 1995; Stainthorp & Hughes, 2004; Thomson, 2003; see recent

[1]University of Athens, Greece
[2]University of Alberta, Edmonton, Canada
[3]University of Crete, Herakleion, Greece

Corresponding Author:
Athanassios Protopapas, Department of Philosophy & History of Science, University of Athens, MITHE, Ano Ilissia University Campus, GR-157 71 Zografos, Greece.
Email: aprotopapas@phs.uoa.gr

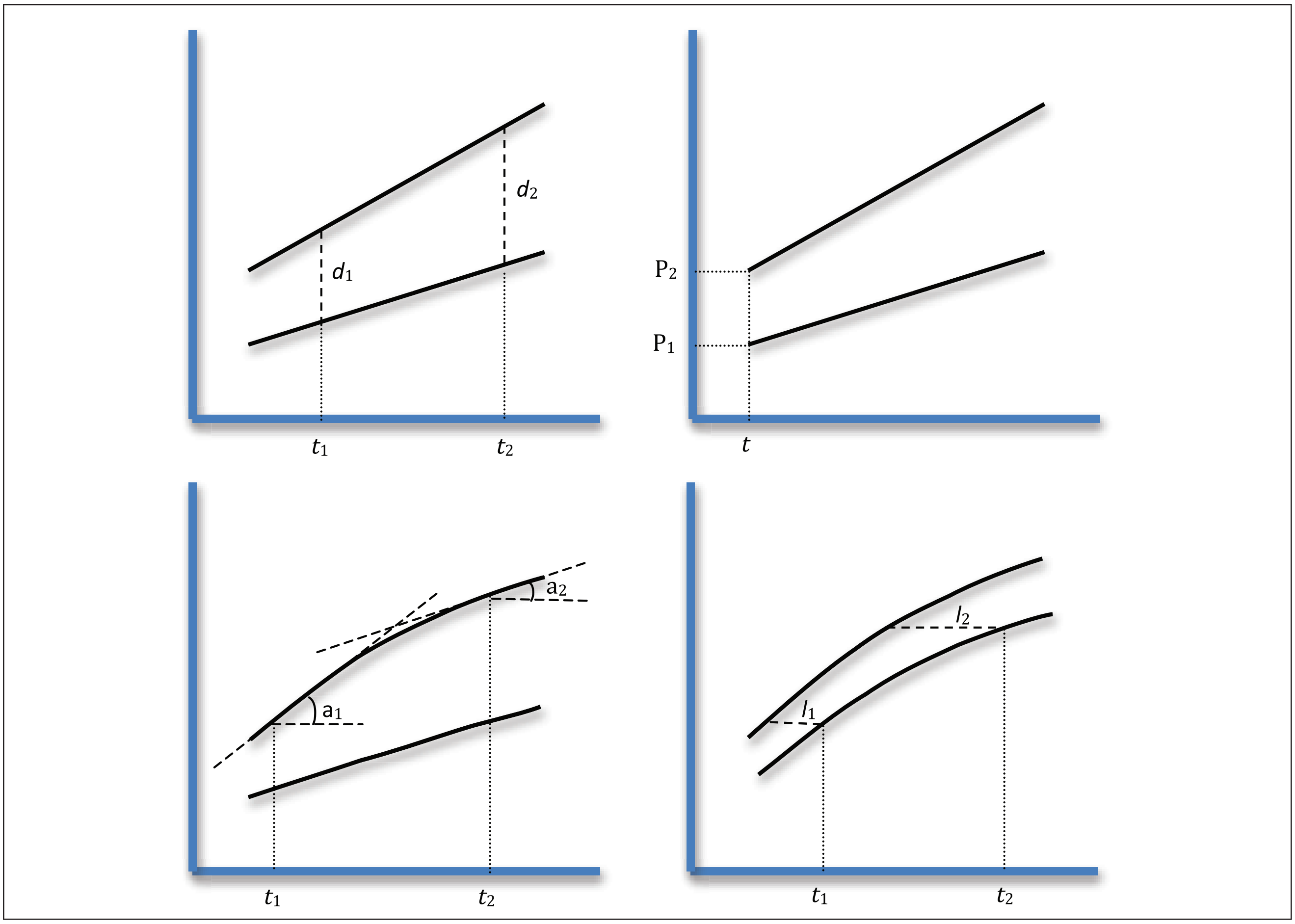

Figure 1. Diagrammatic illustration of growth comparisons that are relevant for the evaluation of Matthew effects.
Note. Each panel shows the hypothetical performance of a subgroup, high performing atop low performing. The horizontal axis corresponds to time (scaling in years or grades). The vertical axis corresponds to the performance measure under examination for potential Matthew effects. See text for explanation.

review in Pfost, Hattie, Dörfler, & Artelt, 2014). Overall, it seems fair to say that Matthew effects in reading development remain elusive, as their predicted long-term effects have proved remarkably difficult to establish empirically. Frequently, the data point in the opposite direction, that is, of children with low initial reading performance partially closing the gap between them and their higher-performing peers.

The question of how best to investigate the presence of Matthew effects in longitudinal data has received some discussion in the literature from a statistical point of view (e.g., Bast & Reitsma, 1997; Parrila et al., 2005). Various types of models and tests have been proposed and considered, yielding little practical difference in outcomes despite differing assumptions and operationalizations. An indicative sample of recent studies can be found in a special issue of this journal devoted to the topic of Matthew effects. Several authors applied varieties of multilevel growth modeling. For example, Protopapas et al. (2011) modeled the growth of raw scores and examined group differences in linear time slopes. Luyten and ten Bruggencate (2011) and Morgan et al. (2011) modeled the growth of item response theory–scaled scores; the former examined the covariance between person-level random intercepts and linear time slopes and the latter examined the effects of predictor variables on person-level intercepts and linear time slopes. Besides growth modeling, Protopapas et al. (2011) also examined variance differences in raw scores across time, and Cain and Oakhill (2011) examined group × time interactions in raw scores (see Note 1). All of these approaches capitalize on statistical comparisons between performance differences across time.

Consider the top left panel in Figure 1, which displays a rough schematic of the prototypical Matthew effect, through a linear modeling lens. Time flows along the horizontal axis; performance on some measure is referred to the vertical axis. Mean performance of two groups is plotted as two separate lines, indicating a linear increase in performance as a function of time. The two groups differ in performance:

their distance, in units of the specific measure scale, is d_1 at Time 1 (t_1) and d_2 at Time 2 (t_2).

The idea behind testing for Matthew effects is that if d_2 is greater than d_1 then this is evidence for divergence, that is, the lower-performing group is “falling behind” compared to the higher-performing group. When several longitudinal data points are available, permitting reliable estimation of growth, direct comparisons between distances can be replaced by tests of interaction between slopes (that is, the linear effect of time) and groups. This approach offers more power, provided that the developmental progression is reasonably approximated by straight lines.

The question, however, remains: Can any of these statistically sophisticated approaches provide evidence relevant to the Matthew effect framework? Here we suggest that a very important issue has received less attention than it deserves. It seems that a critical obstacle in establishing the purported Matthew effect may not be statistical in nature but, rather, psychometric. Specifically, we argue that the typically used measurement instruments fail to establish a metric scale on which differences can be meaningfully compared across distant performance levels.

Scaling of behavioral measurements is understood to be less than interval, generally conforming to an ordinal scale (Cliff & Keats, 2003). Typical practice, as expressed in introductory statistics and assessment textbooks, involves acknowledgement of psychometric scales as ordinal, followed by total disregard of the theoretical and practical implications. Occasionally, “approximately interval” or “plastic interval” scaling is defined (e.g., Coolican, 1994, p. 193) to justify numerical calculations and quantitative statistical comparisons. There is a long-standing debate, including very strong criticism, concerning the nature of measurement in psychometrics (e.g., Michell, 1997, 2008b, 2009). This may strike practicing psychologists and educators as overly philosophical, not really affecting their applied concerns and methods. However, when it comes to longitudinal comparisons of differences across performance ranges, the scaling problems manifest themselves in pervasive ways that cannot be dismissed.

In the present article we consider the extent to which the empirical difficulty in establishing divergence as hypothesized in the Matthew effects framework may be related to psychometric issues of measurement. To demonstrate the problems, we present and discuss a series of analyses, testing longitudinal performance patterns for differential development among core literacy skills. We discuss the effects of gradually diminishing improvement, typical of many psychoeducational scales, and we examine different ways of identifying divergence, focusing on the interpretation of the statistical findings. For simplicity, we consider only analyses of raw data; item response theory scaling is specifically addressed in the discussion. Our main question is this: To what extent is it possible to document the presence of Matthew effects as patterns of longitudinal divergence?

Description of Data

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 to 4 in the 2004–2005 school year, were followed through Grades 4 to 6 two years later. Details about the sample have been reported previously (Protopapas et al., 2011). The first assessment (Wave 1) was administered at the spring of the first study year, followed by two measurements per school year (fall and spring, at roughly 6-month intervals), for 2 more years, totaling five measurements per child.

Of the large battery of tests administered to each child, only four are used here (more details about each measure can be found in earlier reports, e.g., Protopapas et al., 2007, 2011): (a) *Spelling* was assessed by a 60-word list, dictated in order of increasing difficulty (see Mouzaki, Sideridis, Protopapas, & Simos, 2007, for the psychometric properties of this test). (b) *Word reading fluency* was assessed by a 112-high-frequency-word list, printed on a single sheet in order of increasing length, to be read aloud. (c) *Vocabulary* was assessed with an adaptation of the *Peabody Picture Vocabulary Test–Revised* (PPVT; Dunn & Dunn, 1981; see Simos, Sideridis, Protopapas, & Mouzaki, 2011, for details of the adaptation and psychometric analysis). Children were asked to identify one picture out of four that best represented the word pronounced by the examiner. (d) *Reading comprehension* was assessed with Subtest 13 of the *Test of Reading Performance* (Padeliadu & Sideridis, 2000; Sideridis & Padeliadu, 2000), which includes six passages of ascending length and difficulty, each followed by two to four multiple-choice questions.

In the following graphs and tables, raw performance on each measure is always reported as follows: number of correct words for spelling, number of correct words within 45 seconds for fluency, raw PPVT total score (i.e., number of correct choices plus baseline) for vocabulary, and number of correct choices (answers to questions) for comprehension. Raw data points are plotted in Figure 2. Table 1 lists the correlations between all measures at Wave 1 above the diagonal. Partial correlations, below the diagonal, control for grade.

The top and bottom quartiles of each measure at Wave 1 formed the corresponding “high-performing” and “low-performing” groups. That is, the low-performing group included the lowest 25% of children (i.e., those scoring below the 25th percentile), whereas the high-performing group included the top 25% (i.e., those scoring above the 75th percentile). Thus, in each analysis half of the data are used, excluding the middle 50% “average” performers (depending on the corresponding grouping variable). Comparison of the top against the bottom quartile compounds positive and negative Matthew effects, because any “rich-get-richer” effects of the top quartile relative to the average performance are effectively added to any “poor-get-poorer” effects of the bottom

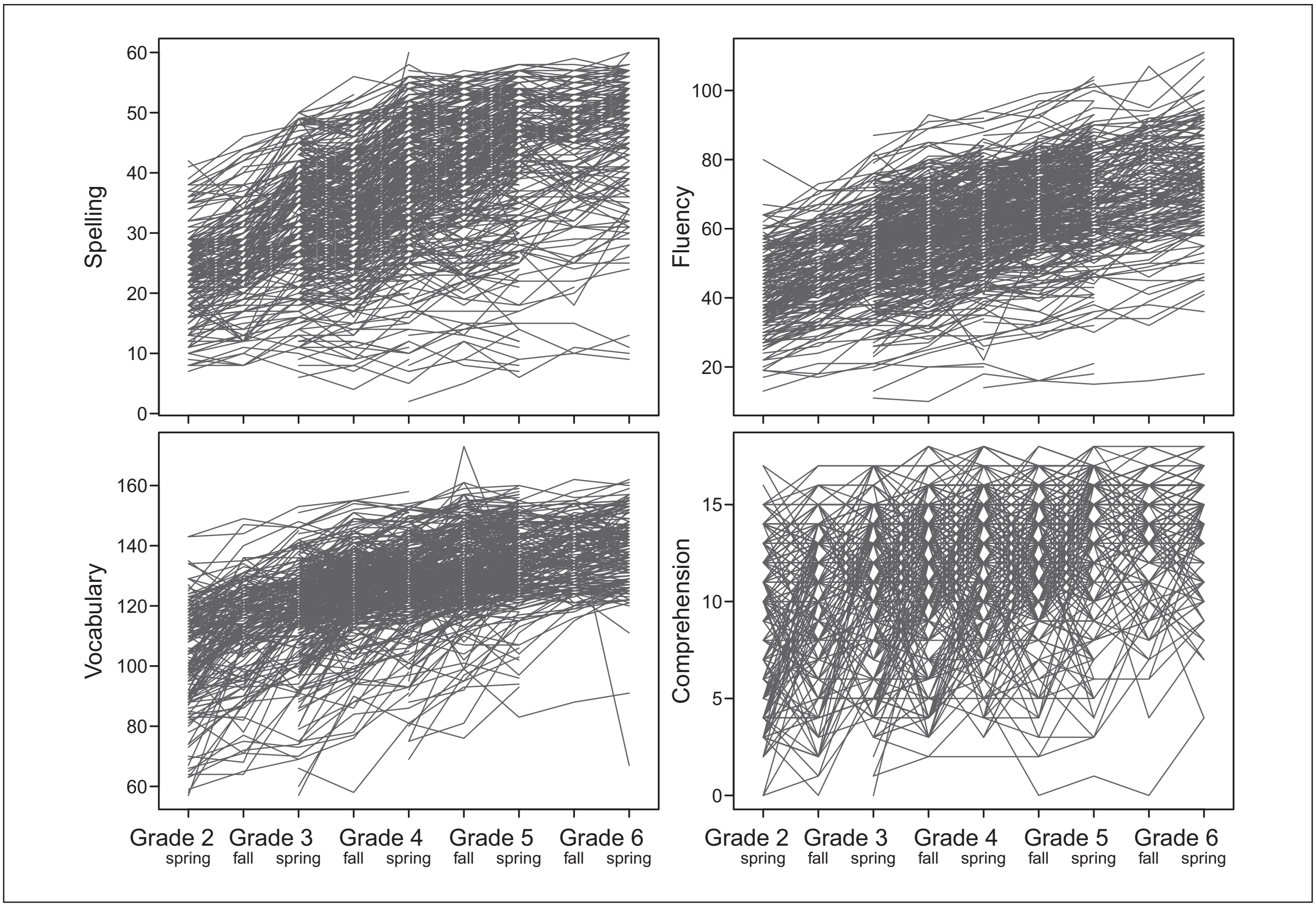

Figure 2. Raw data for the four measures.
Note. Each line connects the data points of a single child over the five measurement waves.

Table 1. Correlations Among Measures at First Measurement (Wave 1).

Measure	1	2	3	4
1. Word spelling		.81	.51	.47
2. Reading fluency	.73		.43	.41
3. Receptive vocabulary	.33	.23		.61
4. Reading comprehension	.36	.29	.54	

Note. Pearson correlation coefficient is above the diagonal. Partial correlation coefficient, controlling for grade, is below the diagonal.

quartile relative to the average performance. Therefore, our choice of subgroups maximizes the power to detect divergence patterns, if any exist.

It should be noted that, because the sample is a community cohort, low-ability subgroups are not necessarily learning disabled. However, a substantial proportion of children in the low-ability groups are at significant risk for learning disability. In particular, data from another Greek community sample suggest that low word reading fluency and especially poor reading comprehension are associated with other cognitive deficits such as processing speed, working memory, sustained attention, and executive skills, as well as with comorbid symptoms of attention deficit (Papaioannou et al., 2014).

To avoid problems related to regression to the mean, stemming from conflating grouping and dependent variables, initial-performance groups were formed on the basis of correlated but distinct measures. Because grouping was done on the basis of Wave 1 measures, only Wave 1 correlations are pertinent to variable selection. Based on the observed pattern of correlations, fluency and spelling served as grouping variables for each other, as did vocabulary and comprehension. The longitudinal performance on each measure by each cohort is plotted in Figure 3 by the dotted lines,

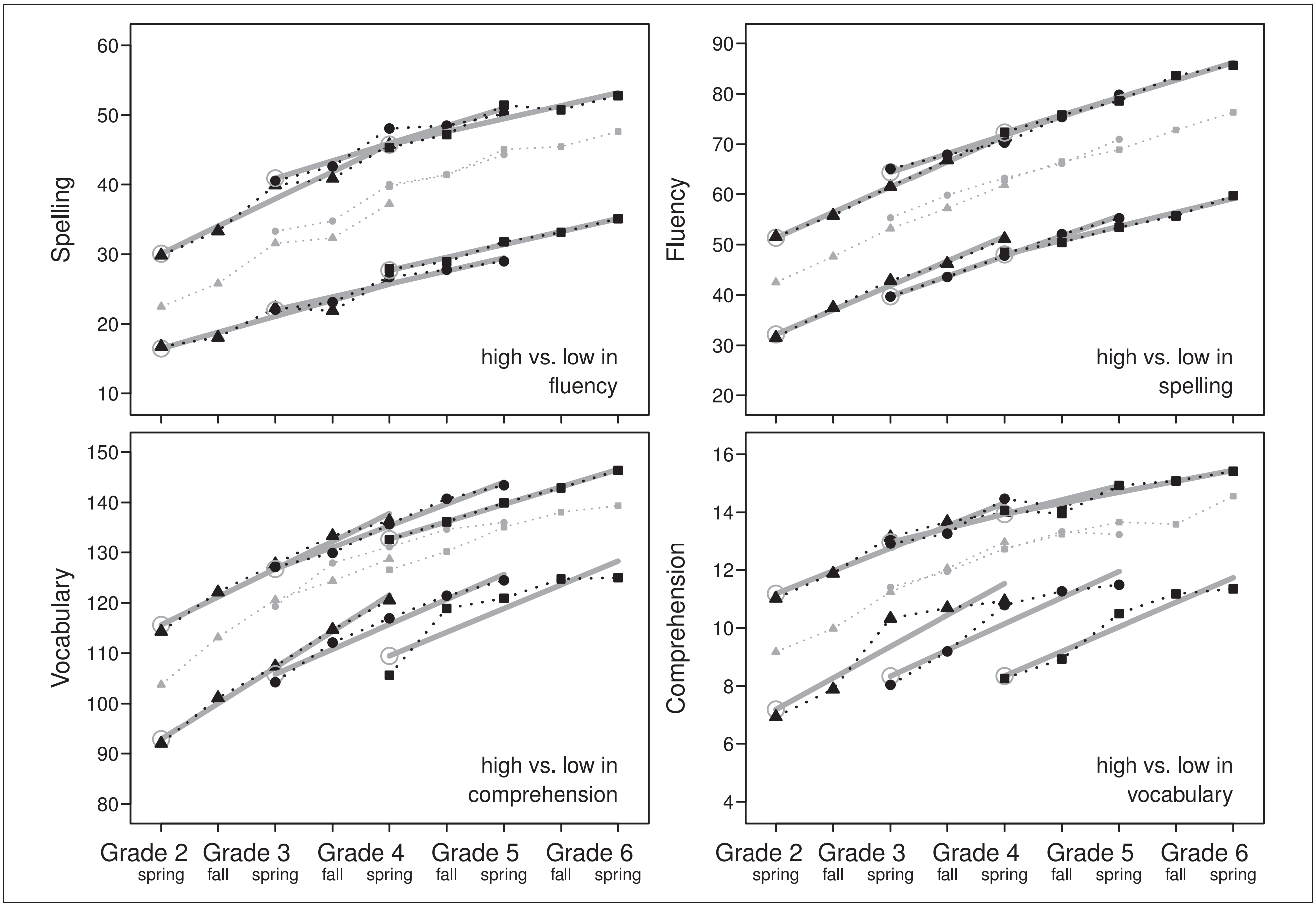

Figure 3. Condition means (unstandardized data; filled markers joined by dotted lines) and modeled longitudinal progression (gray solid lines) for each dependent variable (grouping variable indicated within panel), for the cohorts of Grades 2 (triangles), 3 (circles), and 4 (squares).

Note. Unfilled light gray circles plot modeled intercepts for each cohort subgroup. The light gray dotted lines with light gray markers in between the high- and low-performing subgroups display the performance of the middle 50% of the sample that is not included in the analyses.

grouped on the basis of high versus low performance on the corresponding selection measure. For example, in the top-left panel we see the spelling performance of the bottom fluency quartile versus the top fluency quartile. Children from each cohort are distinguished by marker shape: triangles, circles, and squares mark performance for children in the Grade 2, 3, and 4 cohorts, respectively. The performance of the middle 50% (average performers) is also plotted in each panel for reference, as a faint dotted trail between the top and bottom quartile (low and high performers).

Approaches to Establishing Matthew Effects

The data were modeled using linear mixed-effects models (see the appendix). Initially, the data shown in each panel of Figure 3 were fitted by a single model, in which the fixed effects of grade cohort (3 levels), group (high vs. low), and wave (linear) were allowed to interact. A significant triple interaction would suggest that group × wave interactions were not equal across cohorts. As it turns out, grade cohort interacted with longitudinal differences between performance groups only for spelling, $\chi^2(2) = 16.79$, $p < .0005$, indicating that growth differences between performance groups were not fully homogeneous across cohorts.

Slope Comparisons by Cohort

To examine differential growth directly within each cohort, grade cohorts were examined individually. Models including a group × wave interaction were compared to restricted models excluding the interaction (see the appendix). Table 2 lists the outcome of this test series, including modeled slopes (linear effects of wave) for each performance group. In these tests, significant negative interactions correspond to convergence of the two performance groups within each cohort (indicated with "C"), whereas positive ones correspond to divergence (indicated with "D"). There is scant

Table 2. Differential Growth Between High-/Low-Performing Groups .

		Grade 2 cohort					Grade 3 cohort					Grade 4 cohort				
		Growth β		Interaction			Growth β		Interaction			Growth β		Interaction		
Dependent variable	Grouping variable	High	Low	$\chi^2(1)$	p	C/D	High	Low	$\chi^2(1)$	p	C/D	High	Low	$\chi^2(1)$	p	C/D
Spelling	Fluency	3.9	2.3	33.71	.000	D	2.5	1.9	3.64	.056		1.9	1.8	.01	.928	
Fluency	Spelling	5.0	4.8	.34	.558		3.7	4.0	.50	.477		3.5	2.8	3.30	.069	
Vocabulary	Comprehension	5.6	7.1	6.63	.010	C	4.3	4.9	.83	.361		3.5	4.7	2.85	.091	
Comprehension	Vocabulary	0.8	1.1	2.64	.104		0.5	0.9	7.32	.007	C	0.4	0.8	10.13	.001	C
Spelling	Spelling	3.7	2.7	9.76	.002	D	2.1	2.0	0.12	.732		1.2	2.0	12.54	.000	C
Fluency	Fluency	4.8	4.6	.13	.719		3.2	2.6	1.04	.308		3.3	3.1	.31	.580	
Vocabulary	Vocabulary	3.8	9.7	92.52	.000	C	3.0	6.6	49.64	.000	C	1.7	5.3	23.94	.000	C
Comprehension	Comprehension	0.3	1.7	61.36	.000	C	0.0	1.2	40.26	.000	C	0.0	1.0	32.66	.000	C

Note. For each combination of dependent and grouping variable, group × wave interactions are examined in separate analyses for each grade cohort. C = convergence; D = divergence. Rows on the bottom part of the table list the results of analyses with grouping variable being the same as the dependent variable. These analyses are strongly subject to regression to the mean, evident in increased "convergence" outcomes.

evidence for either pattern of interaction but the little there is suggests that patterns of divergence may be discernible in spelling whereas patterns of convergence seem more likely in comprehension. Table 2 also lists the results of the same tests when the selection variable is the same as the outcome variable, to demonstrate the potential effect of regression to the mean, namely an increased rate of convergence.

The clarity of the conceptual picture reflected in these analyses belies a serious weakness arising from the nature of the measurement process. Consider the top right panel of Figure 1: At time *t*, the performance of the "low" group is P_1 and the performance of the "high" group is P_2. To base comparisons on the difference $P_2–P_1$, as if it were a distance, amounts to an implicit assumption of constant-interval scaling. That is, a 1-unit difference around P_1 is supposed to be equivalent to a 1-unit difference around P_2. Unless this is the case, direct numerical comparison between two slopes—or two differences—is meaningless. It is not, however, typically established whether scales used to evaluate reading development in general, and Matthew effects in particular, satisfy this criterion. It is unclear whether it may be possible in principle to establish such invariance, given the nature of the constructs under measurement and the scales used to assess them. The problem is not restricted to linear modeling approaches. It is equally necessary to establish metric equivalence of changes across performance levels, regardless of the chosen statistical model. Because the whole point is to document (or refute) a "widening gap," we need a valid measure of gap width. The same point applies to methods attempting to establish differences in measures of dispersion (e.g., the "fan spread" pattern); metric equivalence across performance levels is still required.

Consider a hypothetical case in which a group of low-achieving children score "two years behind" in some scale of interest, that is, a sizeable achievement gap. For example, a child in this group might answer correctly questions of difficulty up to "Grade 2" level, whereas children in the comparison group answer correctly questions of difficulty up to "Grade 4" level. Now, a "comparable rate of development" would be claimed if the low-achieving children improve by as many points as the higher-achieving children over the same period of time. Say, within 6 months the comparison group achieves a 2-point increase in raw score by answering correctly two questions beyond Grade 4 level. The comparable 2-point improvement for the low-achieving children would be to answer correctly two questions beyond Grade 2 level. But it is not clear that there is any sense in which two Grade 5–level questions can be considered equivalent to two Grade 3–level questions. In short, the meaning of interactions or slope comparisons is unclear unless they refer to largely overlapping performance ranges.

In terms of our data, the interpretability of the interactions (or lack thereof) indicated in Table 2 is severely compromised by the lack of an established metric equivalence between the performance levels of the "low" and "high" groups.

Slope Comparisons Across Cohorts

To address this grave shortcoming, a potential solution might be to consider growth slopes from a common reference point. By definition, Matthew effects refer to groups differing in initial performance, so this sounds prima facie contradictory. However, differential growth need not be tested at equal ages. Instead, we may identify appropriate ages in the high-performing and the low-performing subgroups, at which they have similar performance. Traditionally this approach is referred to as a reading-level match design. The age difference would be the developmental lag at that point. Comparing growth of the two

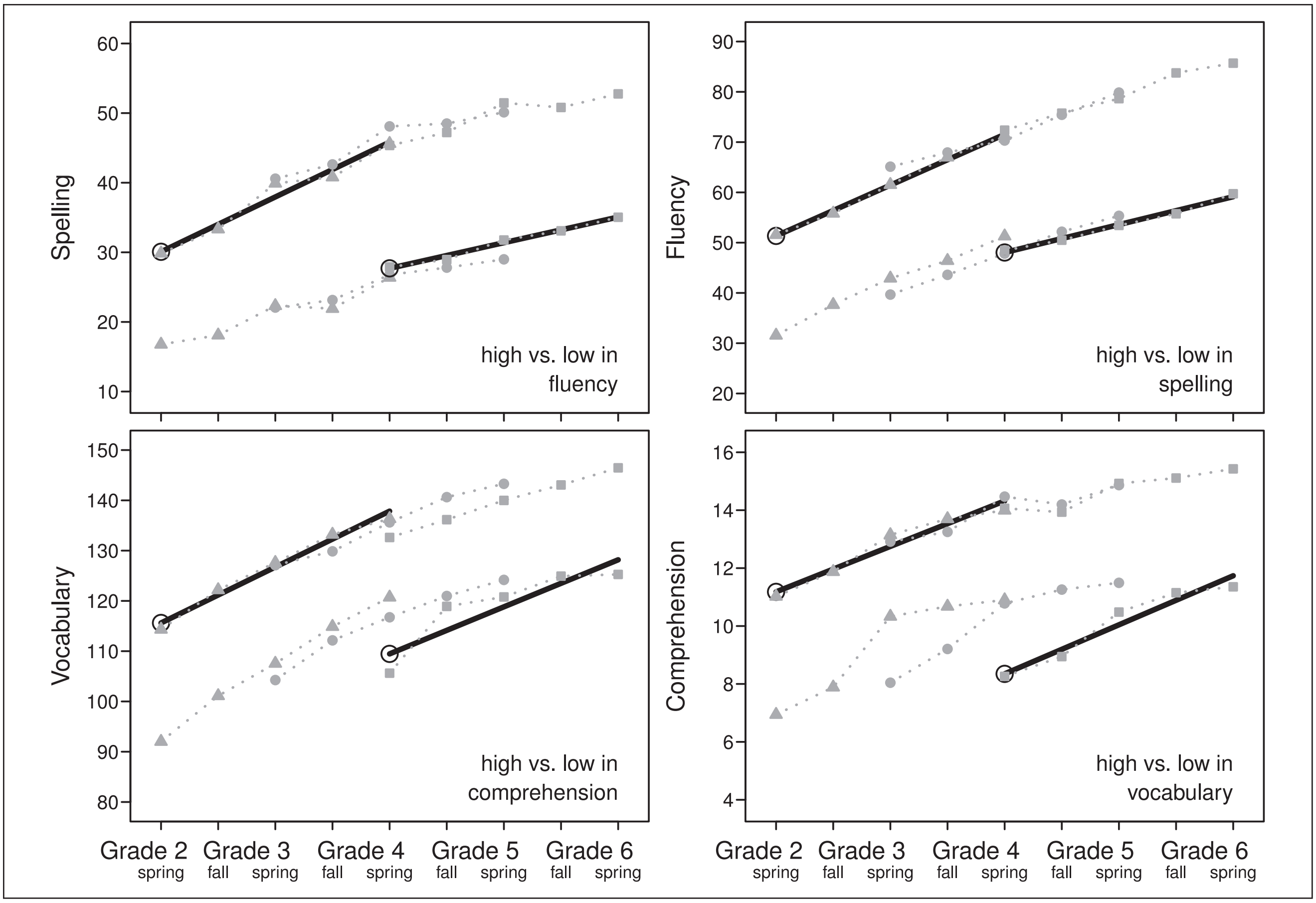

Figure 4. The same panels as in Figure 3, in which the modeled intercepts and slopes of only the high-performing subgroup of the Grade 2 cohort and the low-performing subgroup of the Grade 4 cohort are displayed (in black).
Note. Condition means for all subgroups are plotted in gray, joined by dotted lines, as in Figure 3.

groups from that point on might inform as to whether they are converging or further diverging. The common starting point, or at least nearby range along the critical dimension, obviates the scaling issue.

In Figure 4, the same panels as in Figure 3 are plotted again, highlighting performance of two subgroups: The low-performing subgroup of the Grade 4 cohort is displayed against the high-performing subgroup of the Grade 2 cohort. It is seen that, in each case, performance of these two subgroups is comparable, if not fully overlapping. Since the initial performance levels for these two subgroups are about the same, ordinal relations suffice for the comparison and so their developmental trajectories can be directly tested on the comparable intervals around their common intercept.

The modeled slopes of the growth trajectories highlighted in Figure 4 are included in Table 2. The interaction between wave and group for these pairs of growth lines, tested via model comparison, indicated significant divergence for spelling, $\chi^2(1) = 46.36$, $p<.0005$, and fluency, $\chi^2(1) = 29.67$, $p<.0005$, but no significant difference for vocabulary, $\chi^2(1) = 1.54$, $p = .215$, or comprehension, $\chi^2(1) = 0.13$, $p = .720$. These findings suggest that spelling and fluency in the lowest-performing subgroup develop at lower rates than in the highest-performing subgroup, consistent with a Matthew effect. As with the preceding analysis, this approach yields no evidence for Matthew effects for vocabulary and reading comprehension.

Developmental Slope Invariance

Comparisons across cohorts are vulnerable to changes in children's background experience, educational practices and materials, or other social and situational factors ("cohort effects"; see, e.g., Coolican, 1994, pp. 160–161). Moreover, the aforementioned analyses hinge on the stability of the growth rate of the high-performing subgroup, which serves as reference against which to evaluate the relative growth of the low-performing subgroup. If the high performers did not improve at a reasonably stable rate then the question of whether the low performers improve at a similar or lower rate would be difficult to examine using these data, because there would not be a constant reference to compare to. So,

Table 3. Slope Invariance Across Grades.

						Quadratic effect of wave								
		Cohort × wave interaction				Grade 2 cohort			Grade 3 cohort			Grade 4 cohort		
Dependent variable	Grouping variable	β_{2-3}	β_{3-4}	$\chi^2(2)$	p	β	$\chi^2(1)$	p	β	$\chi^2(1)$	p	β	$\chi^2(1)$	p
Spelling	Fluency	−1.45	−0.68	53.49	.000	−0.22	4.40	.036	−0.41	10.12	.001	−0.34	8.58	.003
Fluency	Spelling	−1.36	−0.23	16.36	.000	−0.01	0.00	.966	0.43	4.03	.045	−0.04	0.03	.870
Vocabulary	Comprehension	−1.23	−0.87	17.20	.000	−0.68	5.79	.016	−0.05	0.03	.860	−0.07	0.08	.784
Comprehension	Vocabulary	−0.29	−0.10	8.28	.016	−0.12	2.20	.138	−0.06	0.73	.394	0.01	0.03	.853

Note. Slope comparisons are displayed for the high-performing groups only. Left: Interaction between cohort and the linear effect of wave in analyses including all 3 grade cohorts. Right: Quadratic effects of wave in separate analyses for each grade cohort.

this question reasonably arises: Do the data actually justify the assumption of constant improvement for the high-performing groups, against which the improvement of the low-performing groups can be judged? Although a consistent developmental trajectory would not conclusively establish the required psychometric reference discussed above, it might nevertheless serve to partially alleviate the scaling concerns, in conjunction with the analysis of slopes in overlapping performance ranges (across cohorts).

Consider the bottom left panel of Figure 1. The performance of the high-performing group is plotted with a decelerating slope, consistent with approaching "plateau" performance. The slope of growth for this group is given by angle a_1 at time t_1 and angle a_2 at time t_2. A sizeable initial rate of increasing performance appears substantially reduced later on—subject to the interpretational difficulties concerning nonoverlapping performance ranges, as discussed above. The diminishing slope would invalidate any attempt purported to assess the growth of the high-performing group, since a reliable estimate of slope presupposes stable linearity (see Note 2).

The stability of growth rates of the high-performing subgroups across the age range considered in this dataset was examined in two ways: First is *between* cohorts, by testing the interaction of cohort with the linear effect of wave. In this test, a nonsignificant interaction would indicate statistically equivalent slopes of linear growth for the different grade cohorts, that is, constant rate of growth throughout the entire age range in the data. Second is *within* cohorts, by testing the quadratic effect of wave on top of the significant linear effect. In this test, a nonsignificant quadratic term would indicate approximately linear growth across the five measurement points covering a 2-year interval. The former test is more demanding because it concerns a larger age range. The latter test is an absolute prerequisite to the interpretation of any slope comparisons because linear slope differences would be impossible to interpret in the presence of quadratic effects.

Table 3 shows the results of both tests. It is clear that the more demanding criterion is not met as the high-performing groups develop at significantly less steep rates in higher grades (the corresponding slopes are listed in Table 2 and displayed graphically in Figure 3). This finding raises concerns regarding the interpretability of the slope comparison across cohorts.

The results of the within-cohort tests are only slightly more reassuring, as several of the growth curves apparently exhibit significant quadratic slopes. Negative coefficients predominate, consistent with diminishing improvement across waves. Vocabulary and spelling exhibit consistent decelerating growth across cohorts, occasionally reaching significance within cohorts as well. Fluency stands out insofar as there is no evidence of within-cohort growth deceleration, with quadratic effect estimates being nonsignificant or positive. Still, even for fluency there is a significant interaction of wave by cohort, that is, between-cohort growth deceleration.

In sum, none of our measures exhibited stable linear growth for the high-performing group, even though fluency came close, at least within-cohort. It seems that the gradually diminishing rate of improvement for the high-performing group fails to support its status as a fixed reference against which the low-performing group can be gauged. If the performance of a high-performing subgroup of children does not improve at a constant rate, this renders questionable any interpretation given to relative improvements of lower-performing subgroups in the same or different cohorts.

Moreover, a more insidious problem becomes evident, given the discussion of ordinal scaling. The very notion of linear growth hinges on the assumption of equal amounts of improvement over equal time intervals, as this is what "linear" means. If performance is not measured on a constant-interval scale then amounts of improvement cannot be meaningfully compared across performance ranges. Thus it is in principle impossible to test whether amounts of improvement are equal, and therefore it is impossible to justify the assumption of linearity. In other words, the scaling issue seems to undermine the notion of growth curve analysis more generally.

Age Analysis by Performance Level

To circumvent the problem of performance scaling, an alternative suggestion might be to establish a temporal reference instead. That is, instead of comparing performance levels at any given age, one might compare ages at fixed performance levels. Absolute lags could then be established at predetermined time points, on the basis of which to base assertions regarding "catching up" or "falling behind." The bottom right panel of Figure 1 illustrates this principle. At time t_1, the low-performing group exhibits a certain performance. We may calculate the corresponding time at which the high-performing group exhibits this mean performance. The difference between the two times is Lag 1 (l_1 on the diagram). Similarly, at time t_2 there is a temporal Lag 2 (l_2 on the diagram) in the attainment of the new performance level. These lags are affected by the decelerating slopes of the growth curves as much as by their differing angles, or even more, but they remain interpretable nevertheless, as they are expressed in absolute time (or grade) units. If Lag 2 is significantly longer than Lag 1 then the low performers are falling behind, that is, there are Matthew effects. If Lag 2 is significantly shorter than Lag 1 then they are catching up. All we would need to test the significance of this difference would be an estimate of variability along time, to define confidence intervals for the lags.

Unfortunately, it would be very difficult to apply this idea in realistic situations. Datasets are typically collected to cover a range of ages, by sampling children at predetermined time (or grade) points. The distribution of scores for each time point is then properly calculated and compared to other time points. To perform the reverse analysis we would need to sample performance levels and record all ages that may achieve them, a distinctly unachievable feat. Figure 5 illustrates the problem. We have plotted mean grade levels for a range of performance by averaging the estimated time points of the children achieving this performance. The time points were estimated by first computing a linear model for each child and then interpolating along the time dimension within the range of times and scores recorded (no extrapolation outside the observed range). We then calculated the mean time at which each score was achieved, based on the children models that included this score in their observed ranges. These means are in meaningful time (i.e., grade) units.

However, as seen in Figure 5, the rate of increase in mean times is much lower than the rate observed in the underlying data. The reason is that average scores were recorded in a much larger range of ages than more extreme scores. In particular, estimates for low and high scores are obtained from nonrepresentative age samples. For example, consider spelling (top left panel). Raw scores around 30 were recorded by many children in every grade, whereas scores less than 20 were only recorded by a few children, mostly from the Grade 2 cohort, and scores above 40 were recorded mainly by children from the Grade 4 cohort. Taking stock of the overall trends in the raw data (cf. Figure 2), it seems reasonable to expect that many more low scores would likely be recorded in earlier grades and high scores in later grades, had we sampled them. The lack of data points before the spring of Grade 2 and after the spring of Grade 6 severely limits the range of times (ages) contributing to the extreme score distributions much more than it limits the range of times (ages) contributing to average score distributions. This distorts the estimated means, rendering the resulting progression meaningless and therefore useless for the estimation of score-referenced lags.

Would it be possible to achieve the desired objective by sampling a wider range of ages? Although this might help diminish the sampling problem it would hardly eliminate it, due to inherent floor and ceiling effects: On one hand children do not produce meaningful spelling scores before Grade 1. On the other hand it is impossible to improve beyond spelling every word correctly, which is the level approached by the better spellers (not uncommon in relatively transparent orthographies). So the extreme scores will always be undersampled and hence the computed curves will be of doubtful reference value. Even though the temporal reference appears theoretically attractive, it does not seem workable in practice, with actual tests and realistic data distributions.

General Discussion

The results of our search for Matthew effects remain largely equivocal. Consistent with most previous studies (e.g., Aarnoutse & Van Leeuwe, 2000; Bast & Reitsma, 1997, 1998; McCoach, O'Connell, Reis, & Levitt, 2006; Parrila et al., 2005; Scarborough & Parker, 2003; Thomson, 2003; but cf. Cain & Oakhill, 2011; Hart & Risley, 1995), we might conclude in favor of convergence, rather than divergence, for vocabulary and comprehension, the two least constrained of our measures (cf. Paris, 2005). In contrast, we might note some evidence for divergence in fluency and, especially, spelling (both being relatively constrained skills in the sense of Paris, 2005), again consistent with some previous reports (e.g., Bast & Reitsma, 1997, 1998; but cf. Aarnoutse & Van Leeuwe, 2000; Thomson, 2003). Overall, our results are consistent with the literature in producing little and inconsistent evidence for Matthew effects.

What do these findings mean for our understanding of literacy development? Having gone through all these analyses, can we draw any conclusions with confidence? The aforementioned discussion suggests that we cannot. Specifically, the interpretability of the findings remains questionable, primarily due to pervasive scaling issues. The problem does not appear amenable to rectification by any sort of statistical procedure, as the scaling issues inherently

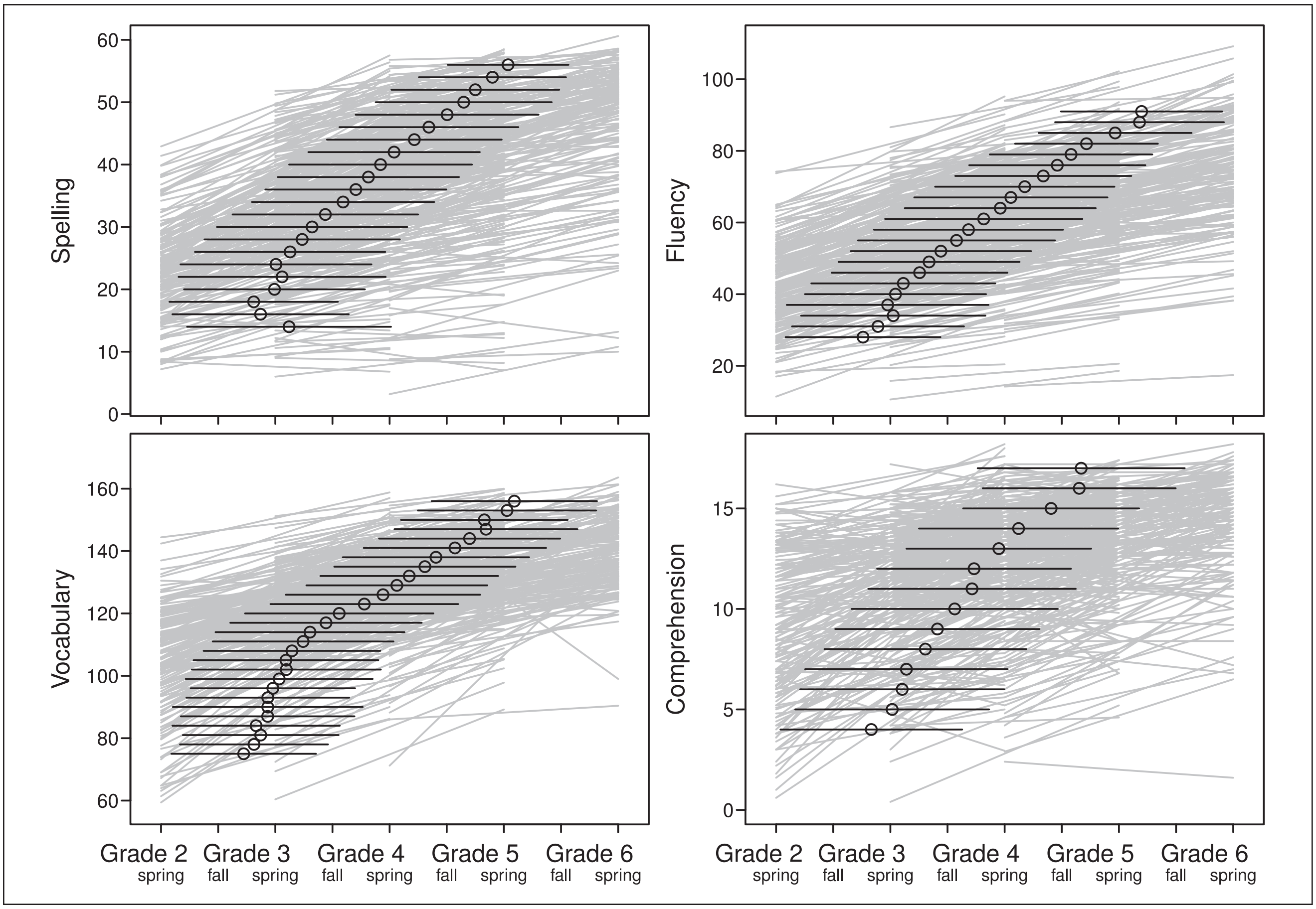

Figure 5. Distribution of ages (in grade units) per score for each measure.
Note. Gray lines in the background display individual modeled linear growth (one line per child). Unfilled black circles show mean grade at the displayed score; horizontal black lines plot corresponding standard deviation.

plague most, if not all, available psychoeducational measures. It seems that the main obstacle in establishing the purported Matthew effects may be psychometric in that our measurement instruments fail to establish a metric scale on which differences can be meaningfully compared across performance levels.

There is a clear sense in which a child measuring 150 centimeters is 50% taller than a child measuring 100 cm: One can use a single 50-cm long stick, which will fit exactly 3 times in the former case and 2 times in the latter. Unfortunately, there is no sense in which a child achieving a score of 150 on the PPVT is 50% more "vocabular" than a child achieving 100. The numbers are deceiving because the vocabulary construct is not a true interval scale. All we can be sure is that the higher score corresponds to higher vocabulary skill in the sense that the former child surpasses more same-age children than the latter child—the exact proportion depending on the specific score distribution for the norming group. The problem is that our measures are constructed and calibrated to yield quantitative estimates referenced against the norming population only. Thus, we can be reasonably confident in the percentage of children in the reference sample that perform better than a given score but we cannot ascribe any further quantitative properties to this score. Yet our psychometric scales provide numerical scores, and this tricks us into the impression that they are actual numbers, with quantitative structure (Michell, 2009), that directly map a constant-interval scale onto the theoretical constructs under study. This is far from being the case.

The scaling issue is inherent in the types of measures used in psychoeducational assessment and is not a particular flaw of classical psychometrics that can be alleviated through item response theory (IRT) methods. IRT scaling results in a latent construct, corresponding to individual ability (θ), which is presumed to be constant-interval and is treated accordingly as a quantitative variable. This θ construct maps onto the mental construct of interest to the extent the measure is valid. Unfortunately, evidence of good model fit does not constitute evidence that the IRT latent construct is indeed constant-interval, as presumed, because ordinal relations among abilities and difficulties suffice for adequate model fit (Michell, 2008a, 2009). In other words, IRT scaling is thought to be constant-interval only because it is a priori assumed so.

Even if the IRT latent construct were indeed constant-interval, it is unknown (and unknowable) whether it would map *uniformly* (i.e., linearly) onto the mental construct of interest (cf. de Ayala, 2008). If it does not, then our measure is not on a constant-interval scale with respect to what we are trying to measure. Suppose there is a "reading ability" mental construct and a valid IRT-scaled instrument to assess it. At issue here is whether equal differences in θ correspond to equal differences in reading ability. Because the IRT model will fit as long as ordinal relations among selected items hold, interval scaling of θ—which is a priori assumed to hold and is a prerequisite to running the analysis—is uninformative with respect to the scaling of the mental construct.

Moreover, although IRT calibration is theoretically invariant, in practice it is sample-referenced and depends on culling of (a possible majority of) poorly fitting items and on assumptions about goodness of fit. As explained by Cliff and Keats (2003, p. 20), interval scaling is only possible when item characteristic curves never cross, a goal only attainable for narrowly constrained sets of items and populations, and when multiple-choice guessing rates are constant across ability, which is unlikely to be true. In general, differences in item difficulty map onto differences in probability of correct response by different-ability children, but ability is not a true interval-scale construct on which meaningful (noncircular) quantitative differences can be calculated. In other words, there is no absolute sense in which two children with $\theta_1 = -2$ and $\theta_2 = -1$ are "equally different" with respect to another pair of children with $\theta_3 = 1$ and $\theta_4 = 2$. All we may conclude is that there are (ordinal) differences in the sets of items that are likely to be answered correctly by each child, as in any other psychometric scale.

It follows from this analysis that the same criticism applies to any composite score derived from multiple measures (for example, combining word and nonword accuracy and fluency, or word decoding and comprehension), whether IRT scaled or not, because the arbitrary scale of the composite is removed from any grounding that might provide the necessary interval reference. All of this makes little difference in practice for most psychometric purposes, and is generally no cause for concern among educational psychologists. However, when it comes to comparing differences among nonoverlapping ranges of performance then interval scaling attains crucial importance. If psychometric scaling tricks inherently cannot address the core of the problem, is there anything that can be done?

We contend that the problem lies with the notion of "equal progress" itself, which is ill-conceived from an empirical standpoint. How is a given amount of progress defined across grades and performance levels? As noted above, a 2-point increase in raw score means different things at different performance levels (e.g., two items of different difficulty and possibly altogether different properties). There is no clear sense in which two different words may constitute equivalent "one-word differences." For fluency measures, this problem can be somewhat reduced, by using reading lists (or texts) containing only easy words, all at about the same level. In that case the reading rate could meaningfully be referenced to actual "words per minute" because similar words would be counted against absolute time. This, however, does not seem possible for the other measures.

Moreover, scaling is often entirely arbitrary, and different reasonable decisions may lead to different conclusions. For example, if a child reads 100 words per minute (wpm) at one time point, and then 120 wpm at a different time point, this can be expressed as a 20-wpm or as a 20% increase. If another child reads 50 wpm at the first time point and 60 wpm at the second time point, this can likewise be expressed as a 10-wpm or as a 20% increase. By raw counts, there is a Matthew effect, as the second child fails to improve as much as the first one, apparently "falling behind." However, by relative proportions, both children progressed equally, so there is no Matthew effect. This problem is not limited to fluency assessments, but will be apparent with any rescaling operation, such as transformations commonly (and reasonably) applied to bring the data in closer approximation to the normal distribution, or to express a score in a more easily understood unit or range. If an important outcome hinges on such arbitrary scaling decisions then the value of any general conclusions seems greatly undermined.

A partial remedy of the scaling problem may be provided by comparing development in overlapping ranges. It is even possible to approximate interval scaling based on ordinal data, when performance ranges are fully overlapping (Mehta, Neale, & Flay, 2004). In this case, diverging slopes would indeed be consistent with further "falling behind" of the low-performing group. However, this comparison presupposes stability of the development rate that serves as reference. Herein lies the second major difficulty in the examination of Matthew effects. Measurement scales are typically normed to conform to certain distributions *within* age groups, not *across* them. Because there is a certain amount of ground to cover in skill development, which is gradually attained, and because there is more ahead than behind at the earlier levels, psychoeducational measures typically level off somewhat at higher ages. This is not a flaw in our versions of the tests. For PPVT, the quadratic effect over age in our sample is similar to that observed in a large U.S. study (Farkas & Beron, 2004; see Simos et al., 2011, Table 4, for the comparison). The increasingly shallower slope of growth curves for the measures we employed (compounded by within-cohort quadratic effects in several cases) prevents safe conclusions regarding the cross-cohort slope comparisons. This problem may be handled by norming future instruments to produce strictly linearly increasing mean raw scores as a function of age. Although this approach would not solve the scaling issue, and would not allow comparisons of differences across performance levels, it would permit meaningful comparisons of growth slopes between groups with similar performance (see Note 3).

All of the aforementioned maneuvering around measures and scores would be gratuitous if we could directly examine the meaningful quantity of interest, which is the amount of time a given group may be said to lag behind, in relation to some—appropriately defined—reference group. Unfortunately, to compare time lags we need estimates of variation along the time dimension so that they can be used to determine effect sizes in the temporal dimension. If we could somehow measure lags l_1 and l_2 in the bottom right panel of Figure 1 we could confidently determine whether the low-performing group is catching up or falling behind. Unfortunately, reliable estimates of dispersion over the time dimension, for given scores, seem impossible to attain given (a) practical constraints of sampling and (b) the time-limited nature of the development of skills under study. This realization serves to remind us of the fragility of the constructs themselves and, in particular, of the Matthew effects framework as a potential empirical research tool rather than merely a conceptual device for studying skill growth and its educational implications.

It should be noted that none of the issues discussed above are specific to particular studies or particular languages. The methodological criticisms outlined here do not depend in any way on properties of specific languages or orthographies. Although the data used for illustration originated in a Greek sample, the problems arise because of psychometric issues that are present in psychoeducational tests in general, and would apply equally to English as well as other languages.

For example, a straightforward application of our arguments to the study of Protopapas et al. (2011) in Greek would indicate that the comparison of raw score differences across time points is based on an implicit assumption of constant interval scaling which, as we have argued, is not valid. Moreover, in the same study, modeling growth in reading comprehension using linear slopes fails on two counts: First, potentially differential curvilinear growth was ignored, thereby invalidating the between-groups comparison. And second, the notion of linearity itself cannot be established in the absence of constant interval scaling, thus invalidating the slope comparisons more generally. Similar arguments can be raised in relation to studies in other languages, such as that of Parrila et al. (2005), in which both Finnish-speaking and English-speaking children were examined, contrasting a highly transparent with an opaque orthography. Parrila et al. did model curvilinear growth via quadratic models, but they examined growth trajectories using latent growth modeling, in effect comparing rates of growth, i.e., score differences, across performance ranges. Moreover, they compared variances across measurement points. Both of these comparisons are undermined by the lack of interval scaling in the reading measures, in both Finnish and English, as the critical notion of "equal growth" cannot be established without a constant-interval scale. Therefore, neither "linear growth" nor "differences in growth" can be empirically demonstrated on the basis of the available data, thereby largely invalidating any conclusions drawn on the basis of the reported analyses, including the reported latent growth classes of Canadian and Finnish children.

To the extent that comparisons of performance across ranges cannot be used to establish an increase or decrease in achievement difference, meta-analytic approaches to studies of the purported widening gap are no more interpretable than the studies they are based on. Thus, comparisons between studies reporting increasing vs. decreasing achievement differences, such as that of Pfost et al. (2014), are subject to the same criticisms stemming from the lack of constant interval scaling. The suggestion that "highly constrained skills lead to a compensatory developmental pattern" (p. 30) is naturally limited by the psychometric scales used to assess the skills in question in the original studies. The overall pattern of substantial heterogeneity in the findings reported in this meta-analysis may be largely attributed to noise due to the inadequate scaling properties of the psychoeducational assessment instruments, which may cause unsystematic apparent differences in some performance ranges but not others.

Measurement problems are not our sole impediment to understanding relative rates of reading development and documenting longitudinal convergence or divergence. Further difficulties stem from a lack of conceptual clarity regarding developmental progression and insufficient empirical foundation of "typical" developmental curves. In most measurable (i.e., relatively simple) cognitive constructs a more-or-less standard pattern is observed: Within each individual an initial period of rapid growth is followed by a period of relatively diminished growth rate, a protracted period of relatively stable performance (possibly with slow growth or decline) and, finally, a period of decline (often associated with "old age"). Individual differences in performance are seen at all points along this progression except at the initial zero point. Now, if children start equal (at zero) yet exhibit different performance levels at some later point in time, this means that they must have progressed at a different developmental rate at least for some time. That is, individual differences in performance entail individual differences in rate of growth; specifically they

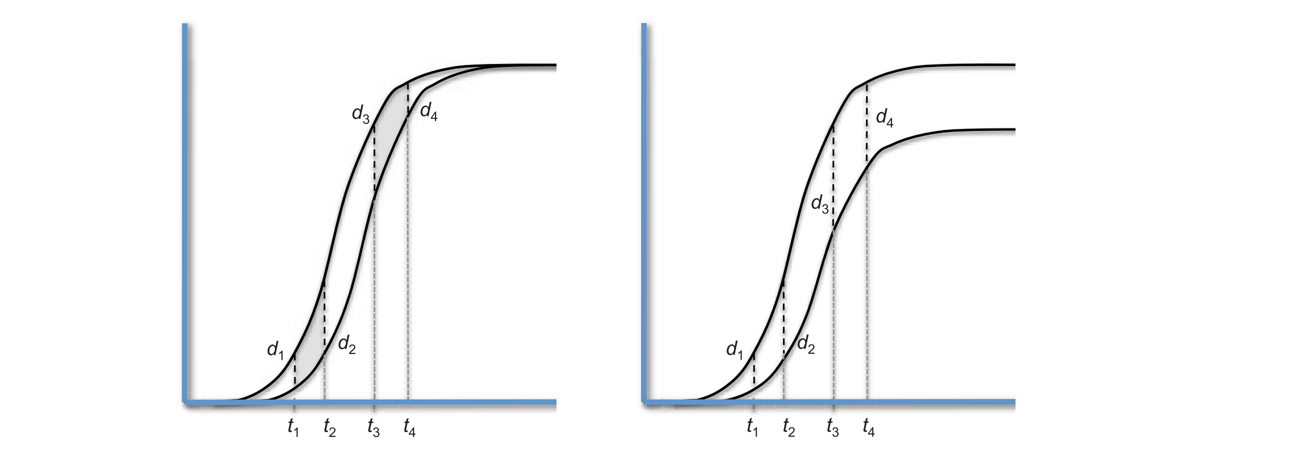

Figure 6. Diagrammatic illustration of developmental growth curves.
Note. Each panel shows the hypothetical performance of two individual children. Axes as in Figure 1. See text for explanation.

entail divergent performance at earlier times. Therefore, the "initial" differences postulated (or observed) in the Matthew effects framework are hardly initial as they must reflect divergent performance already. In this sense the Matthew effect can be said to be necessarily present in any skill where individual differences are present.

The ensuing phase, in which growth typically slows down, is the one usually examined for "Matthew effects," that is, for divergent growth. At this point, differences in performance no longer entail differences in growth because a gap is already established. But the way growth is modeled within this period can easily obscure informative individual differences in developmental patterns caused by relative timing rather than differential growth. For example, early deceleration of individuals who reached their level of stable performance earlier may cause other individuals to appear to be "catching up" even though they may only be slower to reach their own phase of stable performance. Conversely, late initial rapid growth will appear sigmoid-shaped compared to earlier-rising individual developmental curves. From the point of view of Matthew effects, such individual differences in developmental timing are reduced to questions of convergence or divergence, thus obscuring the crucial role of individual growth curves, which remain largely unstudied. To address this problem it will be necessary to understand individual skill development in finer temporal resolution and to determine appropriate models to track growth at the level of the individual.

In Figure 6 we plot hypothetical growth curves for two pairs of children, assuming for a moment that performance can be meaningfully measured on a constant-interval scale. The pair on the left exemplifies the "developmental lag" situation, in which the second child follows an identical developmental curve but is somewhat delayed in time. In contrast, for the pair on the right a more "genuine" performance difference is seen insofar as the lower-performing child never catches up to the level of the higher-performing one. Note that the pattern of growth differences along these distinct developmental paths hardly distinguishes among them. In both cases, an initial $d_1 > 0$ is consistent with early-phase divergence, followed by further divergence as $d_2 > d_1$, demonstrating full-blown "Matthew effects." Further on, in both cases $d_3 > d_4$, consistent with later-phase convergence, that is, "catching up," completely missing the crucial distinction between these developmental paths. In other words, even if the interval scaling problems were somehow addressed, a focus on quantification of the gap—either directly or via measures of local dispersion—would not achieve the desired goal of establishing how reciprocal relations among reading skills and reading practice may produce wide disparity in individuals' later reading outcomes.

Within the field of learning disabilities, the "Matthew" concept was initially introduced to partly account for the persistence and often apparent worsening of reading achievement in many students who present with early signs of difficulty in acquiring basic reading skills relative to their average-and above-average performing peers (Stanovich, 1986). The notion that the risk for meeting criteria for reading disability may increase over time in these students and may comprise more complex and increasingly challenging skills (such as reading comprehension) provided further support for the need for early interventions targeting a wider range of poor readers—including those who perform in the borderline range on phonological decoding and word recognition tasks in the early grades. Several studies have established that low performers are unlikely to show adequate progress in subsequent grades unless they receive systematic and targeted interventions (e.g., S. E. Shaywitz et al., 1999; Snowling, Muter, & Carroll, 2007). Early learning difficulties can have critical, negative long-term educational, occupational, and health consequences when not sufficiently addressed (e.g., Parsons & Bynner, 2005), but such negative outcomes may be preempted with adequate intervention (Bruck, 1987). Why two individuals reach the same end point via different routes and why some individuals end up having a lower stable performance level are, therefore, important questions with real-life instructional consequences. In this sense, the Matthew effects framework may be very useful conceptually, to think about reciprocal relations during skill development, and pragmatically, in highlighting the potentially grave long-term consequences of poor early performance.

However, our analysis suggests that the Matthew effects framework does not lend itself to empirically testable hypotheses, defined by specific performance comparisons on the basis of available psychoeducational measures. Of importance, the present study did not examine longitudinal changes in the likelihood that particular students met criteria for reading disability. Moreover, our analyses did not take into account emotional-motivational traits as they may be impacted by early reading-related capacities and in turn affect subsequent reading-related behaviors (e.g., classroom engagement, exposure to print). In view of these limitations, our results in no way negate the need for early interventions targeting basic reading skills in struggling readers. We suggest that, instead of trying to prove or disprove the presence of fan spread or performance divergence effects, developmental and learning disabilities researchers turn to the study of the original mechanisms behind the Matthew effect, that is, the reciprocal relations between skills and learning experiences. Future attempts to examine the presence and potential consequences of Matthew effects may perhaps more fruitfully focus directly on the development and strength of the reciprocal relationships that mediate performance divergence (Stanovich, 1986) rather than on the elusive “widening gap” itself.

Appendix

Notes on Statistical Analyses

Data were modeled in R (R Development Core Team, 2011) using linear mixed-effects models (package lme4; Bates, Maechler, & Bolker, 2011) including random slopes per participant. Initially, the data shown in each panel of Figure 3 were fitted by a single model including all 3 cohorts:

```
dv ~ cohort*group*wave+(1+wave|subj)
```

In this formula, dv stands for the dependent variable, the asterisk denotes an interaction among flanking terms, and the plus sign denotes additive (noninteracting) terms. The parenthesized factors denote random effects associated with participants, namely random intercepts and growth slopes. The wave factor was mean centered to minimize spurious random intercept–slope correlations (Baayen, 2008).

In this model, a significant triple interaction would suggest that group × wave interactions were not equal across cohorts. This was tested by a χ^2 test against a restricted model excluding the triple interaction:

```
dv ~ (cohort+group+wave)^2+(1+wave|subj)
```

In the restricted model, only up to second-order interactions were permitted, as denoted by the square term. To allow fixed-effects comparison by χ^2 test, maximum likelihood estimation was applied, instead of restricted maximum likelihood (Faraway, 2006, pp. 158–159).

Similarly, to examine differential growth directly within each cohort, grade cohorts were examined individually, using models of the form

```
dv ~ group*wave+(1+wave|subj)
```

compared (via χ^2 test) against restricted models excluding the interaction:

```
dv ~ group+wave+(1+wave|subj)
```

Acknowledgments

We are grateful to Georgios Sideridis, Ioannis Dimakos, John Kirby, and David Francis for constructive criticism and useful suggestions on a previous version of the article.

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.

Notes

1. In addition, McNamara, Scissons, and Gutknecth (2011) compared effect sizes of between-group differences in standardized scores across grades. Approaches employing standardized scores have been criticized for providing no insight into relative individual differences across times. Bast and Reitsma (1998) pointed out that standard score analyses can inform us only about changes in rank orderings over time. To compute standard scores, raw scores are transformed to distributions with equal variance. Scores for different ages are standardized by reference to distinct standardization samples. Individual differences at different ages are based on different subsets of items, because different items are of appropriate difficulty for different ages. As a result, quantitative differences in standard scores between children are not comparable across age groups and can be used only to compute percentile ranks. As noted by Stanovich (2000, p. 154), the only way for a low-performing child to fall behind in percentile rank is by “passing up” a lower-rank child, which, however, would also be subject to Matthew effects, to infinite regress. Thus the

process of standardization specifically prevents the detection of increasing differences. For this reason, in this article we do not consider approaches employing standard scores, even though they can be clinically relevant for identifying and following children with learning disabilities (for example, a practitioner may need to know how a student who was classified as a low achiever/at risk in one grade fares compared to the norm 1 or 2 years later).

2. Stable linearity is not strictly necessary if equality of curvature can be established instead. However, testing for a null difference in quadratic terms hardly justifies the assumption of equal curvature, as the power to detect such higher-order differences is typically very low.
3. Other formulations in the literature seem to simply presuppose interval scaling and deal with growth curve estimation as if measures were quantitative (e.g., Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; cf. Rogosa & Willett, 1985). These approaches will work in practice on overlapping ranges when there is a common reference, either as a starting point or as a no-change state (plateau), in which case ordinal relations suffice to determine the existence of differences. However, they have yet to be formulated in a way that can be applied to the investigation of Matthew effects. Even if they were so formulated they would be applicable only to comparisons between groups over fully overlapping ranges. Therefore they would not directly address the issue of the "widening gap" across a range of skill development, while also being subject to the criticisms of reading-level match designs (e.g., Coolican, 1994; Jackson & Butterfield, 1989; see also Van den Broeck & Geudens, 2012).

References

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

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

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.

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes* (R package version 0.999375-42). Retrieved from cran.R-project.org/package=lme4

Bruck, M. (1987). The adult outcomes of children with learning disabilities. *Annals of Dyslexia*, *37*, 252–263.

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

Cliff, N., & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.

Coolican, H. (1994). *Research methods and statistics in psychology* (2nd ed.). London, UK: Hodder & Stoughton.

de Ayala, R. J. (2008). A commentary on historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*, 209–212.

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

Faraway, J. J. (2006). *Extending the linear model with R: Generalized mixed effects and nonparametric regression models*. Boca Raton, FL: Chapman & Hall.

Farkas, G., & Beron, K. (2004). The detailed age trajectory of oral vocabulary knowledge: Differences by class and race. *Social Science Research*, *33*, 464–497.

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.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes.

Huang, F. L., Moon, T. R., & Boren, R. (2014). Are the reading rich getting richer? Testing for the presence of the Matthew effect. *Reading and Writing Quarterly*, *30*, 95–115.

Jackson, N. E., & Butterfield, E. C. (1989). Reading-level-match designs: Myths and realities. *Journal of Reading Behavior*, *21*, 387–411.

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

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

Luyten, H., & ten Bruggencate, G. (2011). The presence of Matthew effects in Dutch primary education, development of language skills over a six-year period. *Journal of Learning Disabilities*, *44*, 444–458.

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.

McNamara, J. K., Scissons, M., & Gutknecth, N. (2011). A longitudinal study of kindergarten children at risk for reading disabilities: The poor really are getting poorer. *Journal of Learning Disabilities*, *44*, 421–430.

Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: Latent growth curves with ordinal outcomes. *Psychological Methods*, *9*, 301–333.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355–383.

Michell, J. (2008a). Invariance, artifact, and the psychological setting of Rasch's model: Comments on Engelhard. *Measurement*, *6*, 205–209.

Michell, J. (2008b). Is psychometrics pathological science? *Measurement*, *6*, 7–24.

Michell, J. (2009). The psychometrician's fallacy: Too clever by half? *British Journal of Mathematical and Statistical Psychology*, *62*, 41–55.

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities*, *44*, 472–488.

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.

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.

Papaioannou, S., Mouzaki, A., Sideridis, G. D., Antoniou, F., Padeliadu, S., & Simos, P. G. (2014). Cognitive and academic abilities associated with symptoms of attention-deficit/hyperactivity disorder: A comparison between subtypes in a Greek non-clinical sample. *Educational Psychology*. Advance online publication. doi:10.1080/01443410.2014.915931

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, *40*, 184–202.

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.

Parsons, S., & Bynner, J. (2005). *Does numeracy matter more?* London, UK: University of London, Institute of Education, National Research and Development Centre for Adult Literacy and Numeracy.

Pfost, M., Hattie, J., Dörfler, T., & Artelt, C. (2014). Individual differences in reading development: A review of 25 years of empirical research on Matthew effects in reading. *Review of Educational Research*, *84*, 203–244.

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

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*, 203–228.

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.

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.

Shaywitz, S. E., Fletcher, J. M., Holahan, J. M., Shneider, A. E., Marchione, K. E., Stuebing, K. K., & Shaywitz, B. A. (1999). Persistence of dyslexia: The Connecticut Longitudinal Study at Adolescence. *Pediatrics*, *104*, 1351–1359.

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. G., Sideridis, G. D., Protopapas, A., & Mouzaki, A. (2011). Psychometric evaluation of a receptive vocabulary test for Greek elementary students. *Assessment for Effective Intervention*, *37*, 34–49.

Snowling, M. J., Muter, V., & Carroll, J. (2007). Children at family risk of dyslexia: A follow-up in early adolescence. *Journal of Child Psychology and Psychiatry*, *48*, 609–618.

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.

Van den Broeck, W., & Geudens, A. (2012). Old and new ways to study characteristics of reading disability: The case of the nonword-reading deficit. *Cognitive Psychology*, *65*, 414–456.

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