Problems with Early Reading Tests: Confounded Statistical Analyses and Policy Recommendations

Scott Paris
National Institute of Education
1 Nanyang Walk
Singapore 637616
+65 9738-0007
scott.paris@nie.edu.sg

Tests of early reading skills do not distinguish among rapid vs slow developing skills, referred to as constrained vs unconstrained skills (Paris, 2005). This creates problems for measurement and interpretation of reading development, especially when based on tests such as DIBELS and EGRA. As an example, the National Early Literacy Panel Report (2008) in the USA identified early predictors of reading achievement as good targets for instruction, and many of those skills are related to decoding. I suggest that the developmental trajectories of rapidly developing skills pose problems for traditional statistical analyses and policy making. Rapidly developing skills yield correlations with later reading success that change with learning so the predictive strengths are temporary and unstable. The correlations are strong only briefly when children demonstrate partial learning of a skill that they will master completely later. Thus, correlations with rapidly developing skills exaggerate the strength of the relation to later achievement, ignore the transient developmental window, and inflate effect sizes of interventions. I explain why these measurement problems should temper the use of early reading data to make decisions about curricula, materials, instruction, and policies.

Key words: reading assessments, constrained skills, developmental analyses

Background

In the USA, the need to distill the scientific evidence about reading development led to the establishment of panels of experts to review reading research, e.g., *Preventing Reading Difficulties* (Snow, Burns, & Griffin, 1998), the Report of the National Reading Panel, NRP (2000), and *Developing Early Literacy: Report of the National Early Literacy Panel* (2008). Among many important findings, the reviews identified five essential components of reading development; the alphabetic principle, phonemic awareness, oral reading fluency, vocabulary, and comprehension. These components, although not a theory of reading development, have been used to make decisions about what skills to assess and teach young children. In particular, research has shown strong correlations (i.e., predictive validity) between skills related to the first three components that emphasize decoding skills and subsequent reading achievement, and those data have been used to justify primary emphases on these skills for instruction and assessment in early primary grades. Although the practical and political exigencies regarding reading education may be served by identification of the five essential components, there is a need to examine the research evidence and developmental claims about the components.

The "reading wars" between top-down and bottom-up views of reading development were about control of the instructional agenda and financial resources devoted to literacy and, in the last 20 years, scientific evidence has been used to bolster the approaches based on the primacy of decoding skills. However, part of the reason for the preponderance of scientifically-based evidence on decoding skills is the greater ease in measuring and quantifying skills such as letter knowledge, phonemic awareness, and oral reading fluency compared to quantitative measures of comprehension among beginning readers. Thus, there are many more published studies that meet the criteria of review panels such as the NRP and NELP, so there is an imbalance in the availability of evidence based on quantifiable data and experimental methods. Ease of measurement favors studies of decoding skills compared to comprehension and the effects of training favor fast developing skills over slow developing skills. Thus, there is an inherent bias in reviews of literature that favor easily measured and fast developing skills among young readers.

Assessments of Early Reading

There are four basic kinds of early reading assessments. One group of popular assessments is informal reading inventories (IRIs) that include measures of word knowledge, oral reading rate, oral reading accuracy and mistakes, retelling, and comprehension questions that can be given to children just learning to read as well as proficient readers. These performance-based measures provide a wealth of information for the experienced teacher who listens to children read and respond to text because they reveal good and bad strategies used by children. However, IRIs require time-intensive one-on-one testing and considerable teacher expertise in observing children and evaluating the data. IRIs are appropriate measures of individual reading strengths and weaknesses, but they usually do not provide a common scale to track progress over time or compare students in a classroom.

A second group of measures are based on knowledge about reading and language, for example, knowing the alphabet and letter sounds, concepts about print, word boundaries, differences among genres, functions of text, and so forth. Children need to understand such concepts about their native language, but the knowledge enables rather than causes reading to develop. The basic concepts about print (e.g., word boundaries, direction of reading) are usually acquired in the first two years that children learn to read. A third group of measures includes traditional reading comprehension tests that usually involve silently reading many short passages and either answering questions, filling in missing words (cloze passages), or writing short answers. These kinds of tests are difficult for beginning readers who may struggle to decode and understand the words, but as decoding becomes more skilled, comprehension tests reveal more about understanding and responding to ideas in text.

A fourth group of measures aims to assess automatic decoding skills, for example, how quickly children recognize and say letters, words, and nonsense syllables, and how quickly they can read connected text. The advantage of these kinds of reading fluency measures is that they are quick and quantitative. One-minutes tests of how many letters on a page can be named or how many words in a text can be read are simple data to collect and easy data to use for screening, tracking progress, and monitoring growth in fluent decoding. Assessments of decoding fluency are the core measures in the DIBELS or the Dynamic Indicators of Basic Early Literacy Skills (Good & Kaminski, 2002), a widely used battery of early reading assessments. Indeed, it has been modified for use with languages other than English in the EGRA Toolkit (2009) and is now used globally to measure early reading development in many languages.

Early reading assessments found in commercial materials, state-designed assessment batteries, and teachers' daily use include tasks from all four groups, but measures of fluency and decoding predominate in primary grades because (a) the skills provide a necessary foundation for reading development, (b) they develop rapidly during this time frame, and (c) they are quick and quantitative. The data are seductively simple but may lead to two problems. One problem is the overemphasis on skills that are easy to measure, in both assessment and instruction of early reading. A second problem is the misinterpretation of the data and exaggerated claims about the measures. Reading skills that change rapidly yield unstable data compared to other skills that may reflect more enduring abilities of children. Differences in the growth trajectories of reading skills can influence assessment data and can lead to exaggerated claims about rapidly developing skills.

Distinctions Among Early Readings Skills

The criticisms of early reading assessments are based on a conceptual approach to reading development called “constrained skills theory” or CST (cf., Paris, 2005). The basic claim of CST is that reading skills have different developmental trajectories with some skills learned quickly to mastery levels while others skills develop over the lifespan. For example, the names and sounds of the 26 letters in the English alphabet are learned and known by all skilled readers. During the period of rapid learning about letter names and sounds, usually between 4-6 years of age in the USA, the mean scores and variances will vary widely among children depending on the relative degree of skill mastery in the sample, but the scores will only approximate a normal distribution temporarily during rapid learning in any given sample. This means that skills related to learning the alphabetic principle go from floor to ceiling levels in a brief developmental time frame, and consequently variances will be smaller at initial acquisition and when mastery is approached. Unequal variances along rapidly developing trajectories of mastery are expected among children for decoding skills more than slower developing vocabulary and comprehension skills.

Consider the development of a constrained skill such as name writing. It can be acquired in a relatively brief time frame and goes quickly from scribbles to letters to correct letter order to upper and lower case and so forth until the child writes his or her name correctly. The skill may begin between 2-5 years, and depending on the complexity of the name, the intensity of the instruction, and the criteria for success, name writing may be mastered in a few months, not years. The onset, duration, and age of mastery vary among children, but all readers learn to write their names. The average age of mastery approximates a normal distribution only if the sample is selected to exclude children who are obviously at floor and ceiling levels, which is exactly what traditional research has done, but then the distribution and variance of the scores are entirely sample dependent. One can imagine a sample of preschoolers from advantaged backgrounds who learn to write their names at 3-4 years of age whereas children from less advantaged circumstances may not master the skill until 4-5 years of age. The skill varies in accuracy and completeness for only a brief time period while children are learning, so any inferences about the skill development must be interpreted relative to this brief time period as transient relations. However, the NELP Report does not limit the predictive power of constrained skills to these narrow windows of rapid learning.

Constrained skills include many of the skills commonly referred to as emergent literacy and decoding skills including alphabet knowledge, concepts about print, and phonemic awareness. Similar constraints operate in measures of oral reading fluency but over a longer time frame, perhaps 5-6 years before an asymptote is reached. Constrained skills develop in nonlinear trajectories thus violating the homogeneity of variance assumptions necessary for parametric data analyses like those used in the NELP Report (Paris & Paris, 2006). We examine the following three fundamental problems. First, data analyses in the NELP Report treated all reading variables as normalized scores. This is an unwarranted assumption because some early reading skills simply are not normally distributed. Second, the calculations of correlation coefficients, such as Pearson *r*, or group difference *d* in the original studies are inappropriate for constrained skills, except in a narrow age range of mid-mastery. The consistent pattern of results across studies may be an artifact because most researchers avoid floor and ceiling effects in their data by sampling children who are likely to have partial mastery of the target skills. Third, the underlying nonlinear distributions and unequal variances along the developmental trajectories of constrained skills are ignored in the calculation of the estimated effect sizes in the NELP Report by averaging the reported correlation coefficient *r* or standardized difference *d* across a number of studies. The obtained effect sizes for aggregated skills are confounded by the combination of the different types of skills and the degree of skill mastery of each skill. Therefore, the effect sizes in studies of early reading development may inflate both (a) the predictive relations between early literacy skills and later reading outcomes and (b) the experimental impact of various treatments.

Interpretive Problems with Data on Constrained Skills

Three problems are evident with analyses of constrained predictor variables. First, normally distributed data are dependent on the specific sample including only or mostly children who have partial but incomplete mastery. Name writing is not normally distributed in the general population of people or children, and it is clearly at floor for most 1-3 year olds and clearly at ceiling for most children older than 5 years. Thus, the normal distribution is an artifact manufactured by selective sampling of children who have partial mastery of the constrained skill. The predictive power of name writing is zero when name writing is at floor or ceiling levels and is only significant for a limited time of rapid learning.

Second, some studies create a normal distribution of constrained skills by transforming raw scores. Transformations may change scores slightly or considerably depending on the specific sample and the transformation statistic that is applied. Thus, it obscures the developmental nature of the skill as well as the relations in the data in ways that are unknown to the reader. Erceg-Hurn and Mirosevich (2008) said, “The use of transformations is problematic for several reasons, including (a) transformations often fail to restore normality and homeoscedasticity, (b) they do not deal with outliers, (c) they can reduce power, (d) they sometimes rearrange the order of means…(e) they make the interpretation of results difficult… “(p.594). Beyond the statistical problems of transforming skewed scores, the construct validity of constrained skills is contradicted by procedures that attempt to normalize scores that are not normally distributed most of the time.

Third, to increase the difficulty of the tasks in the assessment of constrained skills, some studies add a task demand for speedy performance. This changes constrained skills into unconstrained measures, but the construct is more closely tied to automatic performance than knowledge. For example, rapid automatic naming (RAN) of letters, digits, words, or pictures adds extra demands to recognition so the measure changes from a knowledge-based to a performance-based measure with less constrained development over a longer time frame and with more uniform variance. Manufactured normal distributions distort the data, invite inappropriate statistical analyses, and lead to interpretations generalized beyond the specific sample and beyond the narrow age range of the sample – all critical errors which are not considered in the NELP Report. In addition, the analyses often exaggerate the relations among variables by using predictors and outcomes that are either similar or close in time. The tests then become more like measures of reliability of the test instruments than tests of predictive relations to general literacy achievement.

Unfortunately, neglect of the different developmental trajectories of constrained and unconstrained skills exacerbates this problem because the same interval between two test times may reveal slight growth in unconstrained skills but rapid growth in constrained skills. The rapid growth represents a much larger percentage of total mastery and the total scale of measurement in constrained skills. Thus, the changes in mean levels of performance on constrained skills will be greater compared to less constrained skills over shorter time intervals, and they will contribute more to analyses of effect sizes. For the same reasons, constrained skills assessed during rapid learning are more likely to show larger effect sizes for interventions. More rapid changes in mean levels of performance and greater variances during rapid learning inflate the significance of constrained skills

The NELP Report invites interpretations of causal relations between predictor variables and literacy outcomes, and furthermore, it invites those causal relations beyond the narrow age range of rapid learning. Although this report also analyzed multivariate studies with more than one predictor, and controlled for some variables, such as age and IQ, we cannot conclude that these target early skills were the causes of the later conventional literacy skills. It is possible that individual differences in maturation, cognitive development, and opportunities to learn are indicated by success on the predictor variables. In other words, early literacy predictors are simply proxies for potentially many other kinds of developmental changes in the child and in the child's environment. The proxy argument is supported by data in Table 2.2 that show that reading comprehension is predicted best by readiness (*r* = .59), concepts about print (*r* = .54), alphabet knowledge (*r* = .48), print awareness (*r* = .48), and phonemic awareness (*r* = .44). All of these constrained skills indicate a developmental head start among peers, but all readers will master these skills to nearly identical levels in ensuing years so they are indicators or proxies or enabling conditions rather than causes of beginning reading. Furthermore, the significance of these predictive relations is transient and only evident during the period of rapid learning; it disappears as all children approach mastery of these decoding skills.

The "proxy effect" interpretation is supported by the significant correlations between non-literacy skills that are also moderately correlated with later conventional literacy skills: as shown in Table 2.1, 2.2, and 2.3, *r* = .45 between IQ and decoding, *r* = .30 between performance IQ and decoding, *r* = .45 between arithmetic and decoding; r= .34 between performance IQ and reading comprehension, *r* = .35 between arithmetic and reading comprehension; *r* = .54 between IQ and spelling, *r* =.29 between performance IQ and spelling, and *r* = .50 between arithmetic and spelling. Theoretically, it is reasonable to propose that the early skills are prerequisites of the later skills, and they may be necessary, but they are not sufficient conditions for literacy development, just good indicators of relative differences in development among peers. Some of these differences may be attributed to opportunities to learn that can be addressed by parents and educators, but some of the differences may simply reflect intellectual and maturational variation among young children.

The NELP Report used meta-analyses of research studies to calculate the success of various treatments, and comparisons of effect sizes were the bases for evaluating the relative strength of interventions. However, it must be noted that the calculation of effect sizes and estimates of the power of a statistical test depend critically on equal variances along the developmental trajectories of the skills. Both are estimates of the differences between true population means, and both depend on the size of the variances. For example, assessing alphabet knowledge among 6 year olds ignores the huge differences in knowledge and experience among children and assumes that the sample mean is a good estimate of the population mean. It is clear that the sample means and variances will depend on the relative level of expertise as well as size and homogeneity of the sample. Thus, treatments introduced when the mean level of alphabet knowledge is low will yield greater mean differences (between pre and post-test scores) than treatments introduced when mean levels are nearer to the asymptote.

Meta-analyses of treatment effects that ignore the differences between constrained and unconstrained skills should be re-interpreted. In the NELP Report, the most effective interventions were often mathematical artifacts of short-term gains in mastery of constrained skills that inflate the gains, variances, and effect sizes relative to unconstrained skills. Such interventions are like picking "low-hanging fruit" and there is no evidence of enduring or generalizeable effects of the interventions. Alternative techniques need to be considered to evaluate treatment effects on beginning reading skills.

Conclusions

The NELP Report, despite sophisticated analyses and conventional conclusions, misinterprets data on beginning reading skills by neglecting differences in the developmental trajectories of different reading skills. We have identified three distinct problems in the NELP Report; assuming or manufacturing normal distributions inappropriately for constrained skills, misleading interpretations of correlations involving constrained skills, and exaggerated effect sizes with rapidly developing skills. Together these problems undermine the claims made in the NELP Report about the strength of early predictors of reading proficiency.

It is possible that more appropriate statistical tests will confirm the importance of constrained skills for reading development, but we anticipate that they will reveal (a) brief developmental windows of strong relations, and (b) that constrained skills are necessary but not sufficient to enable fluent reading with good comprehension. Moreover, the importance of constrained skills as early predictors is transient because the differences in mastery of constrained skills are proxies for many possible differences among children in opportunities to learn as well as individual differences in abilities. Conclusions that decoding skills deserve greater or earlier instruction for beginning readers than unconstrained skills are not warranted, and there are liabilities for early reading pedagogy that over-emphasizes decoding skills at the expense of vocabulary, comprehension, oral language, writing, and critical analyses of literacy. Skills related to the decoding of printed text are necessary for skilled reading, but traditional research claims need to be re-analyzed and re-interpreted with regard to the different developmental trajectories of various reading skills so that decoding is not afforded a privileged or exclusive role in early reading instruction. When reading development is considered across a broader K-12 time frame and when outcomes go beyond decoding at early grades to include measures of comprehension, engagement, and use of literacy, then the value of broad and comprehensive reading pedagogies will be more evident.

References

Early Grade Reading Assessment (EGRA) toolkit. (2009). RTI International, Research Triangle Park, North Carolina.

Erceg-Hurn, D.M., & Mirosevich, V.M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*, 591-601.

Good, R.H., & Kaminski, R.A. (Eds.) (2002). *Dynamic Indicators of basic Early Literacy Skills* (6th Ed.). Eugene, OR: Institute for the Development of Educational Achievement.

National Early Literacy Panel. (2008). Developing early literacy: Report of the National early Literacy Panle. Washington DC: National Institute of Literacy. Available at http://www.nifl.gov/earlychildhood/NELP/NELPreport.html

Paris, S.G. (2005). Re-interpreting the development of reading skills. *Reading Research Quarterly, 40*(2), 184-202.

Paris, S.G., & Paris, A.H. (2006). The influence of developmental skill trajectories on assessments of children’s early reading. In W. Damon, R. Lerner, K.A. Renninger, & I.E. Siegel (Eds.) *Handbook of Child Psychology, Sixth Edition, Volume 4: Child Psychology in Practice* (pp.48-74). Hoboken, NJ: Wiley.

Report of the National Reading Panel (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Washington, DC: National Institute of Child Health and Human Development and US Department of Education.

Snow, C.E., Burns, M.S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.