Published in final edited form as:
Educ Res. 2010 May ; 39(4): 347–351. doi:10.3102/0013189X10369931.

Misunderstood statistical assumptions undermine criticism of the National Early Literacy Panel's Report

Christopher Schatschneider and **Christopher J. Lonigan**
Florida State University Florida Center for Reading Research

There are many points of agreement between the authors of the critiques, what was said in the National Early Literacy Panel (NELP) Report, and the perspectives of the Panel members. Several of the critics complain that we did not emphasize strongly enough the role of oral language in later reading and writing, providing too large an emphasis on the role of skills like alphabet knowledge and phonological awareness. In fact, we merely summarized the extant data. Some of the criticisms are framed in statistical or methodological terms. In this response, we examine the validity of these claims and consider how they affect the Report conclusions.

Constrained Skills Theory

Paris and Luo's critique of NELP represents a broad criticism of research conducted on early literacy skills. These criticisms are grounded in Paris' (2005) Constrained Skills Theory (CST). According to CST, "all readers learn and master some reading skills completely to the same criteria," and because all readers master these skills, the window of time during which they differ is small and the distribution of scores is sample dependent. These skills, such as alphabet knowledge, concepts about print, and phonological awareness, are considered "constrained" because the skill progresses from a floor when children have no amount of the skill, to a developmental window in which children vary in considerably in the skill, to a ceiling where mastery is attained by almost everyone. Based on CST, Paris and Luo claim that there are "serious confounding factors in the analyses that may invalidate some of the conclusions in the report."

They argue that these skills cannot be analyzed using correlations because, by their nature, these variables violate the assumption of normality and "homogeneity of variance" that are "necessary for parametric data analyses like those used in the NELP Report, and, in fact, all traditional analyses of constrained skills." The implications of the CST are far-reaching. If the CST is correct, it implies that not only is the NELP report invalid but so are the hundreds of quantitative studies conducted on early literacy development. This is the at least the third time that Paris has advanced his CST (Paris, 2005; Paris, Carpenter, Paris, & Hamilton, 2005); however, to date, CST has not been subject to published feedback. Because this issue is echoed in several of the critiques, we believe that the basic empirical, statistical, and conceptual assumptions of CST deserve careful scrutiny.

There are many things with which we could take exception in the Paris and Luo critique, from the erroneous accusation that we combined measures of constrained and unconstrained skills (we didn't) to the claim that we invited causal interpretations from correlational data--despite our explicit caution against doing so (again, we didn't). However, we will restrict our discussion to the empirical, statistical, and conceptual basis of CST, which we believe, as detailed below, is based on flawed understanding of statistics, an idiosyncratic notion of causation, and assertions that are not borne out by the evidence.

Paris and Luo argue that constrained skills "…develop in nonlinear trajectories [that] violate the homogeneity of variance assumption necessary for parametric data analysis…" and that the analyses in the NELP report ignored the "…underlying nonlinear distributions and unequal variances along the developmental trajectories of constrained skills…" when the effect sizes were estimated. From their description of constrained skills and their ideas about how students grow, we will assume that that by nonlinear trajectories, they mean something similar to a logistic, or s-shaped function where students start with little or no skill, then develop skill at a fairly linear growth rate, until they begin to slow down as they reach the ceiling or asymptote. This is certainly a plausible trajectory for any skill that has an expectation of mastery. Let's imagine a group of children growing over time along these varied s-shaped trajectories. Furthermore, imagine if we were able to take a snapshot (i.e., an assessment) of these skills at various points in time as they progress. What would we expect to find? At the earliest stages, this snapshot would reveal a skewed distribution of skill, with many children exhibiting little or no skill, but with a small number of students well on their way toward mastery. A snapshot taken during the long middle stretch would most likely reveal a normal distribution of scores, as children are at various stages of mastery. Finally, toward the end, there would be another skewed distribution with many students at mastery.

This idea of snapshots of trajectories is an important one in understanding our response to Paris and Luo's criticism. The NELP report did not aggregate studies of "nonlinear trajectories" or combine them with linear ones. The NELP report examined and aggregated results from studies that took snapshots of children at various points in time and correlated those snapshots with later snapshots of reading performance. Therefore, the statistical criticism of the NELP report (and almost all studies used in the NELP report) from a CST perspective rests on the impact of estimating means and correlations on variables that may, on occasion, be skewed and not normally distributed.

The impact of variables that are not normally distributed

CST's statistical argument against the use of data collected on constrained skills in a parametric analysis rests on a fundamental misunderstanding of means and correlations and their analysis. A correlation, like a mean, is a *descriptive statistic*, and there are no assumptions concerning the data to which it can be applied, as claimed by Paris and Luo. Correlations are a measure of the association between two variables, and they can be computed on any set of paired variables regardless of the distributions or variance properties of those variables. Hays (1963) stated this fact clearly when he said:

> It is not necessary to make any assumptions at all about the form of the distribution, the variability of Y scores within X columns or "arrays," or the level of measurement represented by the scores in order to employ linear regression and correlation indices to describe a given set of data (p. 510).

It's true. Any two sets of numbers with any type of distribution--normal or not--with any kind of variance can have the correlation between them computed, and that correlation is a valid description of the association between those two variables in that set of data. In fact, one can correlate two variables even when the values for each variable are only zero or one (i.e., binary variables). This correlation is known as a phi-coefficient (Guilford, 1936), and it is obtained using the same formula one would use for a regular correlation. Two binary variables are about as non-normally distributed as they come. Yet, the phi-coefficient has been used profitably for over 80 years. Non-normal sample distributions do not make correlations invalid.

Non-normal distributions have an impact on two things relevant to correlations: The possible range of values the correlations can have, and the inferential-statistical test that indicates whether those correlations are different from zero. The maximum correlation

between two variables can range between +1.0 and −1.0. This maximum and minimum range assumes a normal distribution. If the variables from which the correlation is computed are not normally distributed, this range will be *smaller*. For example, the range of correlations between a binary variable and a normally distributed variable is between +.798 and -.798 (Shih & Huang, 1993). This has implications for the NELP report and anyone that uses correlations. To the extent that variables deviate from normal, the correlations obtained will be *smaller* than they would have been had the variables been normally distributed. This seems exactly the opposite of Paris and Luo's contention that correlations for "constrained" variables, which are presumably non-normally distributed, would be inflated relative to correlations between "unconstrained" variables, which are presumably normally distributed.

The second area regarding correlations where the normality of distributions of variables can be an issue is when an inferential-statistical test is applied to the correlation. That is, in addition to obtaining a correlation between two variables from a sample, it may also be desirable to make a statement regarding the probability that that particular correlation value was obtained from a sample with a population correlation between those two variables that was zero. If this probability were sufficiently small--say, less than .05--the correlation would be declared "statistically significant." The inferential-statistical test applied usually employs the *t*-distribution, and to use this distribution, a number of assumptions need to be made, including the assumption of normality. Statistical assumptions like the ones used in testing for significance of a mean or a correlation vary in their importance and their impact if they are violated. So, what is the impact of violating the assumption of normality on the statistical test for correlation? Slight to nonexistent. The statistical tests for correlations (and means) are robust against violations of normality. Havlicek and Peterson (1977) conducted a study of the potential consequences of violating the assumption of normality on correlations, and they note:

> It appears that the Pearson *r* can be used in nearly all situations in which there is a need for a measure of relationship between two variables, regardless of the shape of distributions of scores or the type of scales being used. Researchers can compute Pearson *r*s between samples of scores that are skewed in the same or opposite directions, when one distribution is normal and the other non-normal, when the type of scale used for either or both measures is noninterval, or when any combination of these situations occurs. In each case, the researcher can be assured that the probability statements for the obtained *r* will be accurate. Results from this study clearly suggest that the *r* is robust against non-normality and is adequate to cope with weak measurements. (p. 376)

These facts leave the NELP statistical analyses on pretty solid ground. The NELP analyses rest upon well-developed and (near) universally accepted statistical techniques whose properties have been known for decades. Correlations can be computed on any two variables regardless of their distributions or variances, and the inferential-statistical tests applied to those correlations are robust to violations of the assumption of normality that underlie those inferential-statistical tests. Finally, when two variables deviate from a non-normal distribution, the correlation obtained from a sample will be smaller than had the distribution been normal.

Paris and Luo also suggest that constrained variables will lead to erroneous conclusions in the calculation of effect sizes from group-intervention studies. They argue that "constrained skills are more likely to show larger effect sizes." We can think of no reason that this would be true, and their explanation neither follows from actual data nor make sense conceptually. Paris and Luo claim that changes in mean performance on constrained skills will be greater than changes in mean performance on unconstrained skills. Yet, this is contrary to developmental facts. For instance, between the ages of 2 and 5 years, many children learn

the names of the letters of the alphabet (a “constrained skill”). Even starting with no letter-name knowledge, the maximum “change in mean performance” is 26. In contrast, children acquire more than a thousand new vocabulary words (an “unconstrained skill”) during this same developmental period, which would put the minimum “change in mean performance” at something higher than 26. Despite the failure of this aspect of their argument, we are perplexed that they seem to be suggesting that we advocated that educators look for the largest effect size and do only that, which bears no resemblance to anything we actually recommended.

In the context of an intervention study, there seems to be no systematic process that would result in larger effect sizes for constrained versus unconstrained skills. Effect sizes are a standardized metric of the difference between a treatment group and a comparison or control group; effect sizes typically are calculated as the difference in observed group means on a measure following treatment divided by a standard deviation (*SD*; a measure of variability) of the measure obtained from that sample (i.e., [Mean-1 - Mean-2]/*SD*). A larger difference between the group means or a smaller *SD* results in a larger effect size. Whereas instruction that resulted in significant growth in a constrained skill may move more children toward the ceiling on the measure of a skill and, thus, reduce variance (i.e., resulting in a larger effect size), moving one group of children toward the ceiling would restrict the size of the possible difference between the group means because scores at the ceiling cannot move further away from the scores of the comparison group (i.e., resulting in a smaller effect size). If the level of skills exhibited by children in the study does not approach the ceiling on the measure, the degree of non-normality in the distribution of scores will have the same effect on the effect size as was the case for correlations--it will reduce it.

A final issue regarding the “normality” of distributions of measures of constrained versus unconstrained skills concerns whether there really are differences in the distributional properties of the measures. Paris simply asserts that there are differences in the distributional properties of constrained and unconstrained skills. This may be the case for some measures; however, it is not a necessity. For instance, an examination of the distributions and item-level scores in the norming sample of the *Test of Preschool Early Literacy* (TOPEL; Lonigan, Wagner, Torgesen, & Rashotte, 2007), which included a representative group of over 800 3-, 4-, and 5-year-old children, revealed little evidence of significant skew, floor, or ceiling effects for raw scores on any of the three subtests (Print Knowledge, Phonological Awareness, Definitional Vocabulary), and the proportion of children with correct/incorrect answers to individual items was similar across each subtest. In fact, the subtest with the most skew and even a hint of a ceiling effect was Definitional Vocabulary, an “unconstrained skill.” Such findings are not surprising because measures are typically constructed to assess a skill at a certain age or developmental level. Clearly, subtest scores from 9-year-olds given the TOPEL would have very different distributions than the distributions from the age group for which the measure was designed; however, the distribution is a property of the measure and the sample, not of the skill being measured.

Conceptual issues regarding early literacy skills

As should be apparent, CST is, in part, based on a misunderstanding of statistics. The arguments concerning distributions of variables do not undermine any of the conclusions drawn on the basis of correlations between early literacy variables and later reading outcomes or on the basis of differences between groups of children exposed to different instructional activities designed to promote the skills these variables represent. As noted by Paris and Luo, *all readers* ultimately acquire near asymptotic performance for skills like alphabet knowledge. Such a relation indicates the likely *necessity* of such skills; however, the relation does not indicate *sufficiency*. (e.g., alphabet knowledge is necessary but not sufficient for the development of reading and writing skills).

CST implies that because the period of development for some early skills is relatively brief these skills are only important predictors during acquisition. Although we agree that assessing an adult's alphabet knowledge is unlikely to reveal much concerning her or his reading skills--unless that adult is a non-reader, in which case it could provide an important clue to the underlying problem--understanding how children differ on early literacy skills at specific points during development can provide important information concerning their later literacy development. These early skills are not just predictive of later literacy skills during the period of their acquisition; they predict reading and writing skills many years later (e.g., see Table 1 in the previous rejoinder). The argument that constrained skills should not be conceptualized as enduring individual difference variables does not take into account the dimension of time or acknowledge a developmental window in which performance on certain skills are informative. For example, a child's ability to name letters at a *particular moment in time* is an enduring individual difference variable. No matter what happens in the future, that child assessed at that point in time was able to name a certain number of letters. That performance in relation to his or her peers is a strong predictor of future reading skill. This child's performance could be an indicator of her or his home literacy environment (HLE), a proxy for SES, an enabling condition that helps the child "crack" the alphabetic code, or any combination of the three. These possibilities are, in part, what makes this area of research so interesting.

CST advocates that these skills should be ignored because they do not represent "causes" of beginning reading because individual differences in reading performance exist even after most children have mastered a constrained skill. It is beyond the scope of this article to discuss fully the philosophical issues concerning the definition of "cause." We think that it is worthwhile to note, however, that "causes" for most complex human behaviors can be described as *INUS conditions*, that is as an insufficient but nonredundant part of an unnecessary but sufficient condition (Shadish, Cook, & Campbell, 2002). Such variables satisfy the basic definition of cause – that which brings something else into being. However, most such causes are not deterministic but probabilistic in that they do not ensure an outcome; they only make it more likely. Some skills identified as constrained by CST meet stringent tests of causality. For instance, some interventions that promote phonological awareness have an impact on decoding skills that is causal and enduring (e.g., Byrne & Fielding-Barnsley, 1991; 1993; 1995).

We do not see the strategy of ignoring or minimizing the importance of skills related to reading development advancing knowledge in this area or helping children who are struggling with learning. Regardless of whether a skill is a direct cause of literacy learning, it serves as a proxy for some other condition (e.g., SES, HLE), or represents an enabling condition, understanding how that skill relates to later reading and writing advances an understanding of the development of literacy. Perhaps a skill like alphabet knowledge mediates the influence of HLE on later reading and writing skills. Which is the cause, the HLE or the alphabet knowledge? Are there other ways to acquire alphabet knowledge? Do these routes to acquisition of alphabet knowledge have the same influence on later reading and writing? These are valuable questions that help explain how children become skilled readers. We believe that ruling-out certain skills simply because they are acquired early, are finite in scope, or may be acquired relatively quickly, needlessly distracts from potentially useful areas of inquiry.

Direct and indirect effects

Dickinson, et al. argue that we underestimated the role of oral language in later reading and writing because we only summarized direct effects. They note that there is some evidence that oral language is related to phonological awareness and other early literacy skills, and

therefore, some of the influence of oral language on literacy is mediated by phonological awareness and these other skills. We agree that oral language may be one factor in the development of phonological awareness (Lonigan, 2006; 2007). Unfortunately, Dickinson et al. make a statistical error that causes them to over-estimate the potential early contributions of oral language to later literacy. Whereas it is true that the *total effect* of a variable is the sum of its *direct effect* and all *indirect effects*; this sum cannot exceed the *total effect* (Mueller, 1996). In the primary analyses of the NELP report, we analyzed total effects, not direct effects. A total effect is revealed by the zero-order correlation between a variable and an outcome. A direct effect is the regression path between a variable and an outcome when additional variables are in the model. Indirect effects are the products of the regression paths between the index variable and the additional variables in the model and the paths between these other variables and the outcome.

We have a high degree of empathy for Dickinson et al. for making this error. One of us made the same mistake while a postdoctoral fellow (Lonigan, 1994). Direct effects and total effects are not synonymous. The summary of predictive relations between potential early literacy skills and conventional literacy skills in the Report was based on zero-order correlations, which are total effects. The Report does not underestimate the role of oral language skills on later decoding, reading comprehension, and spelling by ignoring indirect effects. They are accounted for in the zero-order correlations.

Use of fixed versus random effects in meta-analysis

In addition to the broad criticisms of the field of early literacy research, Paris and Luo raised methodological questions that were specific to the NELP report. These had to do with our choice of fixed-effect versus random-effect models for the meta-analyses. There has been much debate in the literature about choosing fixed-effect versus random-effect models in meta-analysis. The Cochrane Collaborative, the international organization dedicated to producing systematic reviews of health care research, frames the debate as follows (The Cochrane Collaboration Open Learning Material [n.d.]):

> The debate is not about whether the underlying assumption of a fixed effect is likely (clearly it isn't) but more about which is the better trade off, stable robust techniques with an unlikely underlying assumption (fixed effect) or less stable, sometimes unpredictable techniques based on a somewhat more likely assumption (random effects).

The issue is not whether one model is correct or not but which produces interpretable findings. What is important to note is that the choice between fixed-effects models and random-effects models does not produce any *systematic* differences in effect size estimates. They produce highly similar results. We chose fixed-effects for all of correlational models because those analyses occurred first, and, at that time, the fixed-effects models were more popular and better understood. When we analyzed the intervention studies, we took a different strategy and used the fixed-effects models unless there was enough heterogeneity in the effect sizes to warrant applying a random-effects model.

Conclusions

None of these statistical, conceptual, or methodological arguments pose a serious challenge to the evidence summarized and the conclusions made in the NELP Report. In some cases, the concerns raised by the authors of the critiques are understandable mistakes and, in others, the concerns are based on an idiosyncratic perspective on statistics and an unsupported conceptual argument about the importance of different variables. CST offers no basis for rejecting any of the NELP findings or of the individual studies of early literacy

summarized in the Report. The foundational statistical claims of CST are incorrect, and it is not clear that the empirical expectations of CST regarding distributions of scores on measures are realized in actual data. Whereas it is the case that variation among individuals for some domains may be present early but not late in development, such a pattern does not render the early variation uninformative regarding later development. For the early literacy skills labeled as “constrained” by CST, the variation among young children on those variables consistently relates to later variation among them on conventional literacy skills. .

References

Byrne B, Fielding-Barnsley R. Evaluation of a program to teach phonemic awareness to young children. Journal of Educational Psychology. 1991; 83:451–455.

Byrne B, Fielding-Barnsley RF. Evaluation of a program to teach phonemic awareness to young children: A one year follow-up. Journal of Educational Psychology. 1993; 85:104–111.

Byrne B, Fielding-Barnsley R. Evaluation of a program to teach phonemic awareness to young children: A 2- and 3-year follow-up and a new preschool trial. Journal of Educational Psychology. 1995; 87:488–503.

Guilford, J. Psychometric methods. New York: McGraw Hill; 1936.

Havlicek LL, Peterson N. Effect of the violation of the assumptions upon significance levels of the Pearson *r*. Psychological Bulletin. 1977; 84:373–377.

Hayes, WL. Statistics for psychologists. New York: Holt, Reinhart, and Winston; 1963.

Lonigan CJ. Reading to preschoolers exposed: Is the emperor really naked? Developmental Review. 1994; 14:303–323.

Lonigan, CJ. Conceptualizing phonological processing skills in prereaders. In: Dickinsen, DK.; Neuman, SB., editors. Handbook of Early Literacy Research: Second Edition. New York: Guilford Press; 2006. p. 77-89.

Lonigan, CJ. Vocabulary development and the development of phonological awareness skills in preschool children. In: Wagner, RK.; Muse, A.; Tannenbaum, K., editors. Reading Related Vocabulary Development. New York, NY: The Guilford Press; 2007. p. 15-31.

Lonigan, CJ.; Wagner, RK.; Torgesen, JK.; Rashotte, C. Test of Preschool Early Literacy. Austin, TX: ProEd; 2007.

Mueller, RO. Basic principles of structural equation modeling: An introduction to LISREL. New York: Springer-Verlag; 1996.

Paris SG. Reinterpreting the development of reading skills. Reading Research Quarterly. 2005; 40:184–202.

Paris, SG.; Carpenter, RD.; Paris, AH.; Hamilton, EE. Spurious and genuine correlates of children's reading comprehension. In: Paris, SG.; Stahl, SA., editors. Children’s reading comprehension and assessment. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2005. p. 131-160.

Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and quasi-experimental designs for generalized causal inference. New York: Houghton Mifflin Company; 2002.

Shih WJ, Huang W. Evaluating correlation with proper bounds. Biometrics. 1992; 48:1207–1213.

The Cochrane Collaboration Open Learning Material (n.d.). Fixed and random effects meta-analysis. Retrieved January 27: 2010, from http://www.cochrane-net.org/openlearning/HTML/mod13-4.htm