Published in final edited form as:

Sci Stud Read. 2016 ; 20(1): 1–5. doi:10.1080/10888438.2015.1118480.

Introduction to the Special Issue: Advancing the State-of-the-Science in Reading Research through Modeling

Jason D. Zevin and
University of Southern California

Brett Miller
Eunice Kennedy Shriver National Institutes of Health and Human Development, U.S. National Institutes of Health

Abstract

Reading research is increasingly a multi-disciplinary endeavor involving more complex, team-based science approaches. These approaches offer the potential of capturing the complexity of reading development, the emergence of individual differences in reading performance over time, how these differences relate to the development of reading difficulties and disability, and more fully understanding the nature of skilled reading in adults. This special issue focuses on the potential opportunities and insights that early and richly integrated advanced statistical and computational modeling approaches can provide to our foundational (and translational) understanding of reading. The issue explores how computational and statistical modeling, using both observed and simulated data, can serve as a contact point among research domains and topics, complement other data sources and critically provide analytic advantages over current approaches.

Keywords

statistical modeling; computational modeling; reading development; reading disability; individual differences

Understanding the factors that contribute to successful acquisition of literacy is, necessarily, a multi-disciplinary enterprise. Even skilled reading may be characterized by detailed psycholinguistic investigations of multiple, somewhat dissociable, outcomes, such as generalization to novel pseudowords and reading for comprehension. Thus it is not surprising that research on the development of reading generally takes a wide variety of cognitive and linguistic skills into account, in an attempt to understand how these factors interact, and how they may put children at risk for reading difficulty. As a result, reading research is often a locus of innovation in statistical analysis and simulation modeling techniques, and reading researchers have tended to be "early adopters" as new techniques

Correspondence: Jason D. Zevin, Dept. of Psychology, University of Southern California - SGM 501, 3620 South McClintock Ave., Los Angeles, CA 90089. zevin@usc.edu.
Jason D. Zevin, PhD, Associate Professor of Psychology and Linguistics, University of Southern California;
Brett Miller, PhD. Program Director, Child Development and Behavior Branch, *Eunice Kennedy Shriver* National Institutes of Health and Human Development, U.S. National Institutes of Health;

become available. This Special Issue surveys recent advances in computational approaches to understanding the development of reading, with a particular emphasis on predicting and understanding reading difficulties.

We think this issue is timely, as larger and more conceptually complex data sets raise new challenges. For instance, how, as a field, do we provide meaningful conceptual bridges to further integrate behavioral and other data sources (i.e., genetic and neurobiological) to provide coherent, integrated conceptual and theoretical models and insights? How can we meaningfully constrain likely approaches to effectively identify and service children with reading difficulties based upon existing data or simulated data sets with known structure? What approaches are necessary to advantage the reading research community as the broader research community increasingly moves to “Big Data”, bottom-up approaches? To support integration across multiple levels of analysis (e.g., neurobiological and behavioral), we need to more effectively utilize approaches that can provide analytic and conceptual bridges for interpreting and integrating data across methodological approaches within individual projects; we argue this need can be naturally filled by increased integration of advanced statistical and computational modeling approaches.

Successfully adapting new statistical or computational approaches is admittedly complicated. This is in part due to the difficulty of learning to use unfamiliar methods and the difficulty of understanding whether and how a new approach can be applied to particular questions of interest. With team-based science increasingly becoming the norm, it is important for researchers with complementary expertise to work together and recognize and appreciate the contribution and insights from researchers trained in other domains. This shift necessitates that the next generation of scholars have the appropriate training to evaluate and adopt novel, sophisticated analysis approaches as they are developed. Our goal in bringing together the papers selected for this special issue is to both highlight work that exemplifies this ethos and introduces new or novel computational approaches to research on reading development and reading difficulties and disability.

The papers composing this issue illustrate how novel modeling efforts can form an important contact point among research domains and topics, and provide conceptual bridges for converging data and platforms for experimentation to inform structure and design of both lab and field studies. The first three contributions focus on examples of different approaches to statistical modeling of reading data. These papers tackle classic issues in the reading research field, and data from commonly used instruments and methodological tools, with novel and distinct modeling assumptions. They demonstrate the potential of sophisticated analytic tools and utilization of simulated data to provide unique insights and analytic advantages over existing approaches.

In the first paper, Sideridis and colleagues (Sideridis, Simos, Mouzaki, & Stamovlasis, 2016) utilize the cusp-catastrophe model to examine the moderating role of Rapid Automatized Naming (RAN; Denckla & Rudel, 1976) on reading achievement in children (2nd–4th grade). This approach comes out of a Nonlinear Dynamic Systems framework and permits us to study nonlinear relationships between two variables. The basic idea is that past a certain threshold in RAN performance (i.e., poor performance on the RAN number task) the relationship between variables (e.g., decoding and reading fluency) changes and becomes nonlinear. To gauge potential advantages of this approach, the authors compared the fit of the cusp-catastrophe model to alternative linear and logistic regression models and found that the cusp model consistently provided a better characterization of the relationship among decoding, reading fluency and RAN performance. More specifically, the insight provided by the cusp-catastrophe approach is that there appear to be qualitatively different functional patterns between fluency and decoding as a function of performance on RAN, and that this pattern holds longitudinally for this age range and cohort.

In the second paper, Matsuki and colleague (Matsuki, Kuperman, & Van Dyke, 2016) take on the challenge of working with datasets that have large number of predictors (p) and modest number of behavioral observations (n) (i.e., small n large p problem) and dealing explicitly with issues relating to overfitting data and concomitant problems. The authors examine the potential strengths of random forests models to handle these analytic problems and provide an alternative to the use of composite scores, which may not be possible given the specific predictors in a data set, or less desirable for interpretive or conceptual reasons. Part of the intent of their analyses is to investigate the value of data-driven analytic approaches that can provide insight into later hypothesis and theory development and subsequent or concomitant confirmatory testing. This approach has particular promise for recent efforts promoting more efficient and extensive utilization of extant data sets[1]. Here Matsuki et al. use the random forests approach to gauge the relative importance of variables, as compared to Dominance Analysis and Multimodel Inference approaches for data sets with and without problems related to small n and large p. Their data concern eye-movements during reading, as these are related to a collection of individual difference measures that share substantial variance associated with verbal processing. As background, the random forests approach is a non-parametric classification and regression method that has the benefit of providing the rank importance of variables compared to others within the dataset, while avoiding overfitting due to issues of collinearity. The analyses show that the random forests handled the issues of collinearity very well when utilizing all predictors in the small n and large p dataset without overfitting, and, importantly, performed better than the two alternative analyses. This pattern held for the larger dataset unencumbered by the small n and large p problem, suggesting the potential of this approach for bottom-up, data-driven analytic efforts.

The final paper in this section, by Schatschneider and colleagues (Schatschneider, Wagner, Hart, & Tighe, 2016) utilizes simulated data to examine the stability of various reading disability schemes over the course of a year. Currently, the lack of a clear consensus on how to most appropriately identify and classify individuals with a reading disability (RD), the relatively poor classification stability within an approach (e.g., Barth, et al., 2008; Brown Waesche, Schatschneider, Maner, Ahmed, & Wagner, 2011), and low or moderate agreement between classification approaches (Spencer, et al., 2014) leads to clearly undesirable consequences in our ability to identify children with RD. Schatschneider et al.

[1] For example, at the U.S. National Institutes of Health, there is a recent initiative focused on data science with the explicit goal of “harnessing the potential of the computational and quantitative sciences to elevate the impact and efficiency of biomedical research”. https://datascience.nih.gov/

(2016) conducted simulation studies to examine the likelihood that 1) measurement error, 2) regression to the mean and 3) true inter-individual differences in change of disability categorization underlie some of this longitudinal instability. Simulation data is particularly well suited to examine these issues – that is, one can manipulate simulated datasets to systematically investigate each of these three scenarios with a known "ground truth". Their results suggest that inter-individual differences in growth have a relatively small impact on the stability of RD classification, that the basic version of the constellation model produced the most stable classifications of RD compared to a range of competitors. Finally, in the case of their constellation model, operationally applying the rubric that any two of the other five possible classification methods need be positive for an RD classification generated greater stability rates than other approaches, and interestingly, incidence rates were consistent with the applied cut-points. This is broadly suggestive that models incorporating multiple indicators can provide greater longitudinal stability for RD classification. The modeling efforts in this paper provide a nice illustration of the strengths and insights that data simulations can provide to our foundational understanding of reading disability that would be difficult or impossible to obtain by only utilizing observed data sets.

The computationally focused papers included in the second half of this issue reflect a distinction between models that help identify and explain patterns in data from children as they learn to read on the one hand, and models that simulate the process of learning to read on the other. Simulations of learning to read can provide important insights about reading difficulties by generating hypotheses about how learning mechanisms operate over the orthographic, phonological and semantic representations that are relevant to learning, and how learning might be impacted by strengths and weaknesses of the learner's other abilities. For example, in a connectionist model that explicitly simulated the relationship between phonological processing and acquisition of decoding abilities, Harm and Seidenberg (1999) demonstrated that manipulations that reduced the efficiency and stability of phonological representation impaired the acquisition of reading skill in a way that was consistent with what is often observed in phonological dyslexia.

Simulations of how decoding abilities are acquired have been a focus of computational modeling in this vein for some time, but connectionist models do not have a monopoly on simulations of learning to decode. For example, Pritchard et al. (Pritchard, Coltheart, Marinus, & Castles, 2016) present a method for learning spelling-to-sound correspondences from a corpus of English words. The resulting rules are appropriate for use in a revision of the Dual Route Cascaded (DRC) model of reading (Coltheart, Curtis, Atkins, & Haller, 1993; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), a computational model largely inspired by double-dissociations from the cognitive neuropsychology of reading and simulates a wide array of results from behavioral studies of healthy adults and patients with acquired dyslexias. In contrast to connectionist models, the DRC assumes a sharp distinction between the processes engaged by lexical processing of words — simulated as an interactive activation model (McClelland & Rumelhart, 1981) — and a rule-based grapheme-to-phoneme conversion (GPC) process that is responsible for nonword reading and also contributes to decoding for regular words. Incorporating a procedure for learning these GPC rules is an important step in making this model more relevant to research on the development of reading and related reading difficulties.

Increasingly, simulations of learning to read have turned to cross-linguistic issues. Much of this work has the goal of assessing the degree to which assumptions that have motivated models that successfully capture the performance of English readers generalize to other alphabetic (Perry, Ziegler., & Zorzi, 2014a; 2014b; Lerner, Armstrong, & Frost, 2014) or morphosyllabic (Yang, McCandliss, Shu, & Zevin, 2009; Yang, Shu, McCandliss, & Zevin, 2013) orthographies. Chang et al. (Chang, Plaut, & Perfetti, 2016) push this theme in a new direction by simulating the acquisition of orthographic representations across an unprecedentedly large number of orthographies (more than 100). Orthographic complexity is an under-appreciated feature of variability across writing systems (cf. Nag & Snowling, 2012), but it may contribute to the diverse presentation of reading difficulties cross-linguistically. For example, the differential contribution of orthographic processing skills to reading outcomes in Chinese as compared to English is likely related to very obvious differences in the complexity of orthographic processing in the two writing systems (e.g., Ho, 2014). By providing a quantitative measure of orthographic complexity based on learning principles, the work by Chang et al. (Chang, Plaut, & Perfetti, 2016) represents a step in this direction. Direct comparisons of the model to human performance in an orthographic processing task further suggest some important considerations for the continued development of such a measure.

The culminating paper of this sub-section proposes a theoretical view of the potential of computational modeling and argues for a strong emphasis on the role of learning and plasticity and its impact in the development and end state of the learning system. Rueckl (2016) lays out a broad theoretical view of the potential directions simulation models of learning to read might take in the future, with a strong emphasis on the role of learning and plasticity in determining both the eventual organization of the reading system, and the developmental trajectory traversed in arriving at that organization. This approach holds out the prospect of bringing simulation modeling of reading acquisition into contact with computational approaches such as those introduced in the empirical contributions presented here, and invites us to imagine ways in which a mechanistic theory of the representational and learning abilities that support the acquisition of reading skill can be used to understand individual variability in developmental trajectories as they unfold over time.

Early, meaningful integration of advanced statistical and computational modeling, as illustrated in this issue, should speed advances in our understanding of individual variation inherent in acquisition of reading, embedded within a broader developmental framework. We have seen demonstrations of novel analysis approaches that can reveal complex linear and nonlinear relationships among predictors of reading ability (Sideridis et al., 2016), and change over time (Schatschneider et al., 2016), as well as an application of a more well-known approach that is under-utilized in reading research in dealing with issues of collinearity and designs with large numbers of predictors common to reading research (Matsuki et al., 2016). It may be helpful, in the context of Rueckl's (2016) contribution, to think of these approaches as refinements to our ability to identify and isolate the "control parameters" that influence individuals' developmental trajectories, or to trace these trajectories out in more detail.

In contrast, the computational models presented here focus on how specific processes and representations that support skilled reading performance, such as spelling-to-sound correspondences (Pritchard et al., 2016) or basic orthographic structure (Chang et al., 2016) might be learned by developing readers. This reflects a complementary approach to understanding the development of reading, but also suggests a way forward that combines simulations with sophisticated statistical modeling. For example, Chang et al. (2016) present the time taken to train an autoencoder on a particular orthography as a dependent measure. In their case, the comparisons of interest are across orthographies, but it is not too difficult to imagine that simulation models will soon be more generally evaluated for their fit to a fuller characterization of changes in children's reading ability over time than to ultimate outcomes, as is common now.

In short, expanding our toolkit and applying it strategically (and appropriately to the scientific question and goal) as needed could have real benefits to building our understanding of individual differences in a developmental context, and provide insight into the control parameters that impact learning, and shape the observed reading behavior we measure. We leave, and hope you will too, optimistic about building a further enhanced understanding of reading development and its variations – modeling should be the foundation for building these conceptual bridges. And, to return to Rueckl (2016), we hope you will accept his invitation to imagine a broader vision of reading and reading development.

Acknowledgments

The opinions and assertions presented in this article are those of the authors and do not purport to represent those of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, the U.S. National Institutes of Health, or the U.S. Department of Health and Human Services. Support for Jason Zevin is provided in part by P01 HD001994 and P01 HD070837 from the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

References

Barth A, Stuebing K, Anthony J, Denton C, Mathes P, Fletcher J, Francis D. Agreement among response to intervention criteria for identifying responder status. Learning and Individual Differences. 2008; 18:296–307. [PubMed: 19081758]

Brown Waesche J, Schatschneider C, Maner J, Ahmed Y, Wagner R. Examining agreement and longitudinal stability among traditional and RTI-based definitions of reading disability using the affected-status agreement statistic. Journal of Learning Disabilities. 2011; 44(3):296–307. [PubMed: 21252372]

Chang L-Y, Plaut D, Perfetti C. Visual complexity in orthographic learning: Modeling learning across writing system variation. Scientific Studies of Reading. 2016

Coltheart M, Curtis B, Atkins P, Haller M. Models of reading aloud: Dual-route and parallel-distributed processing approaches. Psychological Review. 1993; 100:589–608.

Coltheart M, Rastle K, Perry C, Langdon R, Ziegler J. DRC: A dual route cascaded model of visual word recognition and reading aloud. Psychological Review. 2001; 108:204–256. [PubMed: 11212628]

Denckla M, Rudel R. Rapid automatized naming (R.A.N.): Dyslexia differentiated from other learning disabilities. Neuropsychologia. 1976; 14:471–479. [PubMed: 995240]

Harm M, Seidenberg M. Reading acquisition, phonology, and dyslexia: Insights from a connectionist model. Psychological Review. 1999; 106:491–528. [PubMed: 10467896]

Ho C. Preschool predictors of dyslexia status in Chinese first graders with high or low familial risk. Reading and Writing: An Interdisciplinary Journal. 2014; 27(9):1673–1701.

Lerner I, Armstrong B, Frost R. What can we learn from learning models about sensitivity to letter-order in visual word recognition? Journal of Memory and Learning. 2014; 77:40–58.

Matsuki K, Kuperman V, Van Dyke J. The Random Forests statistical technique: An examination of its value for the study of reading. Scientific Studies of Reading. 2016

McClelland JL, Rumelhart DE. An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychological Review. 1981; 88(5):375.

Nag S, Snowling MJ. Reading in an alphasyllabary: implications for a language universal theory of learning to read. Scientific Studies of Reading. 2012; 16(5):404–423.

Perry C, Ziegler J, Zorzi M. CDP++. Italian: Modelling sublexical and supralexical inconsistency in a shallow orthography. Plos One. 2014a; 9(4):e94291. [PubMed: 24740261]

Perry C, Ziegler J, Zorzi M. When silent letters say more than a thousand words: An implementation and evaluation of CDP++ in French. Journal of Memory and Language. 2014b; 72:98–115.

Pritchard S, Coltheart M, Marinus E, Castles A. A computational model of the implicit acquisition of phonological decoding via training on whole-word spellings and pronunciations. Scientific Studies of Reading. 2016

Rueckl J. Towards a theory of variation in the organization of the word reading system. Scientific Studies of Reading. 2016

Schatschneider C, Wagner R, Hart S, Tighe E. Using simulations to investigate the longitudinal stability of alternative schemes for classifying and identifying children with reading disabilities. Scientific Studies of Reading. 2016

Sideridis G, Simos P, Mouzaki A, Stamovlasis D. Efficient word reading: Automaticity of print-related skills indexed by Rapid Automatized Naming (RAN) through cusp-catastrophe modeling. Scientific Studies of Reading. 2016

Spencer M, Wagner R, Schatschneider C, Quinn J, Lopez D, Petscher Y. Incorporating RTI in a hybrid model of reading disability. Learning Disability Quarterly. 2014; 37:161–171. [PubMed: 25422531]

Yang J, McCandliss B, Shu H, Zevin J. Simulating language-specific and language-general effects in a statistical learning model of Chinese reading. Journal of Memory and Language. 2009; 61:238–257. [PubMed: 20161189]

Yang J, Shu H, McCandliss B, Zevin J. Orthographic influences on division of labor in learning to read Chinese and English: Insights from computational modeling. Bilingualism: Language and Cognition. 2013; 16:354–366.