3 Statistical Modeling in Literacy Research

Christopher Schatschneider and Yaacov Petscher
Florida State University

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.

(Box & Draper, 1987, p. 74)

Statistical modeling is ubiquitous in scientific research. Across the vast differences in the subject matter studied by biologists, psychologists, anthropologists, physicists, and educational researchers, one of the constants is the widespread use of statistical modeling. This is not to say that there are not other forms of research that do not use statistical models. In fact, if one were to look at the *Handbook of Reading Research: Volume III* (Kamil, Mosenthal, Pearson, & Barr, 2000), one might get the impression that few literacy researchers engage in statistical modeling. Of the 10 chapters on methods of literacy research, only one chapter (Shanahan, 2000) explicitly talks about using statistical modeling as a part of the methodological approach. With the recent broadening of our “methodological tool bag” in literacy research, it would be useful to examine the benefits and limitations of statistical modeling. To that end, this chapter provides a brief conceptual overview of statistical models, including a discussion of what statistical modeling entails as well an exploration into the philosophical underpinnings of statistical modeling. We also provide a brief overview of statistical modeling in literacy research and end with a discussion of limitations inherent in these models.

**WHAT ARE STATISTICAL MODELS?**

A formal definition of a statistical model is, oddly enough, beyond the scope of this chapter. A useful starting place would be to define a statistical model as an equation or set of equations that attempts to describe a phenomena being studied. Statistical models accomplish this by using mathematical equations that generate predictions about the observed phenomena. These predictions can be compared to the actual observations to assess how well the model performs in replicating the observed phenomena. By using probability distributions, statistical models also provide inferences about the likelihood that a relation being modeled from a set of observations is due to chance. Finally, in the construction of most statistical models there is an attempt to balance two competing goals—fit and parsimony. A simple illustration may be useful, both in terms of understanding the purpose and usefulness of statistical models, and also revealing some of the basic assumptions that often go unexamined. Let’s say that a reading researcher has collected 100 measurements of student reading ability. These observations may show that some students are quite skilled at reading, others are not as skilled, and still others struggle with reading. From these observations, we can attempt to fit some simple statistical models. Models can be useful in describing these reading scores by distilling a large amount of quantitative information into a
smaller set of numbers that conveys some potentially useful information about the observations. With these 100 observations, we could compute a mean and variance, for example. Computing a mean and variance is one of the most basic of statistical models. This simple act contains much of the same benefits and limitations that apply to almost all statistical modeling procedures. In computing a mean, we are using a mathematical equation in an attempt to describe a particular phenomenon. In computing the variance, we are using the mean as a predictor of each score, in an attempt to see how well the mean adequately describes the set of observations. The variance is a direct measure of how well the model fits the data. A relatively large variance estimate implies that the mean does not provide as good of a description of those observations while a smaller variance estimate implies that the mean is a better representation of those observations.

This simple example can also shed light on the balancing act of adequate fit to the data versus parsimony of explanation. All statistical models strive to adequately describe the observations to which they are fit. But many researchers also value parsimony as a general principle of science (although others view parsimony as an epistemological preference and not a general principle; see Courtney & Courtney, 2008). Parsimony in the scientific sense refers to the idea that “the simplest explanation is usually the best.” Statistical parsimony typically refers to the number of components or parameters in a model (in our example of computing variance, the statistical model had one parameter—the mean) in such a way that all else being equal, fewer parameters are regarded as better or more parsimonious. In our previous example, we used the mean of the observations as our parameter in a statistical model to try and predict each observation. This is one of the simplest models that one can fit to a set of observations, but it is rarely adequate in describing a set of observations. That is, this model is parsimonious, but probably does not represent the data well.

The real strength of statistical modeling comes when we develop two or more models and compare them in terms of their parsimony and model fit (Maxwell & Delaney, 2004). Almost any statistical model that one can think of either implicitly or explicitly entails the comparison of two or more models based on the increase in model fit in relation to the increase in model complexity. Statistical models have a structured way in which they can be compared in order to judge whether a more complicated model sufficiently improves the description of the observations to justify the relative loss of parsimony. Models accomplish this with the help of probability distributions. A probability distribution describes the likelihood that a particular value or estimate will occur under a particular set of conditions. Some probability distributions give the likelihood of observing a particular value or estimate when nothing is operating except differences due to random sampling. These are called central distributions. In common statistical analyses, the t distribution for tests of differences between means and the F distribution for analyses of variance represent just such “known” central distributions. If the value of a statistic derived from a model (such as an F-ratio or t-ratio) is found to be large enough that values that large or larger only occur 5% of the time in a distribution where only random sampling is operating, then most researchers will conclude that it is not likely that this particular F – or t – ratio occurred by chance, and this model provides a significant improvement in explaining the data. The use of known probability distributions allows researchers to assign a probability value to the difference between two models that provides a basis by which we can select the more complex model over the simpler model.

PHILOSOPHICAL UNDERPINNINGS IN THE USE OF STATISTICAL MODELS

Statistical models serve as useful tools for scientists who are grounded in the philosophy of empiricism and analytic reductionism. Empiricism is the idea that knowledge arises from experience and observation. Empiricism often favors observations derived from experimentation
over passive observations, and rejects evidence based on intuition or reasoning alone. Analytic reductionism is the belief that we can gain understanding of a complex system by understanding the parts of the system and how they interact.

A foundational assumption of empiricism is that nature is lawful (Underwood, 1957; Maxwell & Delaney, 2004). The assumption of lawfulness implies that observations made about nature and used by statistical models are not random or chaotic, and that an inspection of observations can reveal general principles. In this sense, statistical models could be used as mathematical formalizations of our ideas about the general principle being observed (but do not necessarily need to). The assumption of lawfulness has a number of corollaries. One is that nature is also consistent or uniform. The assumption of uniformity implies that the regularity seen in a limited amount of observations should generalize universally. For social science research, this assumption is questionable. But in terms of statistical models, it implies that the model developed on a particular set of observations should be applicable to a new set of observations if the observations come from the same population of observations.

Analytic reductionism rests on the idea of finite causation (Maxwell & Delaney, 2004), or that a phenomena being studied have a finite number of causes. The ability to read, for example, is a complex behavior that has many causes and correlates. The reasons why any one person can comprehend texts are numerous (ranging from cognitive, genetic, neuroanatomical to instruction, home environments, and cultural contexts). Some would argue that the potential causes of reading we investigate are not causes at all, but are actually INUS conditions (Shadish, Cook, & Campbell, 2002). INUS stands for an Insufficient but Nonredundant part of an Unnecessary but Sufficient condition. Let’s take reading instruction in the classroom as an example of an INUS condition for a student learning to read. Reading instruction in the classroom is insufficient for the student learning how to read because other factors must be present (the child must be able to attend to the instruction, must not have any neurological damage that would prevent reading, etc.), and it is nonredundant in that it can contribute something unique to all the other reasons a child might learn how to read (either on their own or with help from their parents), it is unnecessary in that there are other mechanisms by which the student could learn how to read (perhaps at home), but it is part of sufficient condition in that in combination with other factors, can also lead to students learning to read. But in terms of finite causation, this implies that there are only a limited number of causes (or INUS conditions) that will explain the ability to read. If there are an unlimited number of causes that can produce reading behavior, then generalizations of findings would be impossible.

Most statistical models are a mathematical outgrowth of empiricism and analytic reductionism. Many scientists believe that nature is lawful and consistent, and that knowledge about a phenomenon can be obtained by making observations. Many scientists also believe that it is fruitful to study only parts of a phenomenon as a way of understanding the whole. Statistical models facilitate this endeavor by using observations as a means for inferring lawful relationships. Statistical models are mathematical representations of the scientist’s ideas about the lawful relations that exist in nature. Given a set of observations, scientists develop mathematical models that they believe will explain this relationship. Statistical models take into account the fallible nature of observations and that they contain errors. They provide a mathematical basis for judging the adequacy with which the proposed lawful relation explains the observations. But it is important to remember that the use of these models typically implies an adherence to a philosophical tradition to which not all literacy researchers subscribe.

One final comment about statistical models in general. As we stated previously, the inputs into a statistical model are the observations made by the researcher. These observations can come from a variety of environments (research designs), and it is these environments that determine what can be concluded from the results of statistical modeling. As Lord (1953, p. 750) stated, the “numbers don’t know where they came from.” Statistical models don’t know where
the numbers come from either. The researcher is responsible to knowing how the observations were obtained, and what conclusions can be drawn from the results of the statistical model.

**STATISTICAL MODELING IN LITERACY RESEARCH**

Statistical modeling has had an enormous impact on literacy research. The uses of statistical modeling in advancing literacy research can be grouped into a few broad areas that we describe next.

**Observation and Measurement**

The foundation of an empirical approach to science is observation. However, these observations must take on some kind of symbolic form in order to be useful. Whether these observations are turned into hand-written notes, transcribed verbal recordings, or numeric representations, some kind of symbol-system must be used to record observations. Additionally, it is rare that a researcher in the social sciences is studying something that can be continuously observed. Instead, observations must be sampled from the phenomena being studied, with the assumption that a sample of observations will be a good representation of the phenomena under investigation. Finally, it is often the case that observations of certain behaviors are easier to obtain if they are elicited by the researcher. Literacy researchers will often set up an artificial environment where reading behaviors can be recorded, for example, or they may ask a teacher to deliver a particular activity or lesson that can provide the researcher with just such behaviors.

Recorded observations take on many forms and serve many purposes. But all observations have one thing in common: To a greater or lesser degree, all observations are prone to error. The idea that observations and measurements contain errors has been around for a long time, and the systematic study of errors in observations since the early 19th century (Tabak, 2005). Every measurement has the possibility of being influenced by error. A researcher can miss something that is occurring in the classroom, students may not answer a question correctly because they are not feeling well, or perhaps they didn’t understand what was being asked, or conversely, a student may get an answer correct by luck in guessing. A researcher could also simply miscode a particular response or observation. These are just a couple of examples of potential errors that occur in literacy research. Errors in measurement are unavoidable when conducting science. However, statistical models can aid the researcher in a number of ways. First, the effectiveness of statistical models in helping researchers minimize the impact of error of measurement stems from the notion that most errors in measurement are random rather than systematic; that is, there is no systematic way in which errors are influencing the observations. If the influence of errors on observation is random, then, over the long run, they will cancel each other out. Statistical models can estimate the amount of random error that is influencing a set of observations.

Classical Test Theory (CTT: Nunnally & Bernstein, 1994) employs a number of statistical models for which the primary purpose is to estimate the errors in measurements. CTT assumes that anything that can be measured has a “true score” component and an “error score” component, and the models developed in CTT are designed to estimate the percentage of each.

These models are useful to the literacy researcher in a number of ways. First, they help to refine measurements and observations. Statistical models can estimate the amount of error in an observation and can point to ways in which these errors can be minimized. Item-total correlations can be estimated to inspect an items utility in representing the domain being sampled, estimates of inter-observer agreements can be calculated to see where two or more people observing the same phenomena agree or disagree on what is being observed and can give an indication as to which behaviors are most troublesome to find agreement. Statistical models
produce estimates of error that can be expected for published observation instruments which aids the literacy researcher in selecting a pre-developed observation instrument if the researcher chooses that route.

More recently, advances in Item Response Theory (IRT; Hambleton, Swaminathan, & Rogers, 1991) have proved to be useful to literacy researchers and have been seen as an improvement over classical test theory. IRT rests on a simple idea—the probability that someone will respond in a particular way is directly related to how much of a particular attribute a person possesses. For example, the probability that a student will be able to read a particular word is dependent upon how much “word reading ability” the student has. The statistical models that support IRT are fairly complex. However, these models also provide a number of useful products. IRT provides for a straightforward way to develop alternate assessment forms so students can be assessed over time while minimizing practice effects. IRT also provides a very powerful means to detect item bias (termed differential item function in IRT) so that items that behave differently for different populations can be removed or minimized. IRT also aids other statistical models in that the ability estimates obtained via IRT are on a true interval scale, which is often one of the assumptions of other statistical models.

Many of the norm-referenced and standardized tests of reading and reading related ability used by researchers have been developed using IRT. Tests such as the Wide Range Achievement Test – 4 (Wilkinson & Robertson, 2006), Woodcock Reading Mastery Test – Revised (Woodcock, 1998), and the Peabody Picture Vocabulary Test – Third Edition (Dunn & Dunn, 1997) are examples of assessments that use IRT as the basis for test development and score reporting. Using scores from an IRT based assessment provides the researcher with a number of benefits including an assurance that the scores obtained from the instrument are on the same scale over the entire age range that is covered in the assessment, which is critical for longitudinal research.

**Construct Validation**

All areas of science are concerned with establishing the viability of potentially important variables and constructs and statistical modeling can be useful in this endeavor. Construct validity is a problem of generalizing the observations made on a given phenomena to the higher order constructs these observations represent (Shadish et al., 2002). Constructs such as “decoding ability” and “vocabulary knowledge,” for example, are unobservable constructs, assumed to exist by researchers; however, the existence of these constructs must be inferred by researchers from samples of behaviors. It would scarcely be possible to conduct science without operationally defining our constructs of interest. Nunnally and Bernstein (1994) argue that there are three major aspects to construct validation: (a) Specify the domain of observables thought to be related to the construct, (b) determine whether those observables are measuring the same thing or different things, and (c) conduct research to see which of those constructs (or really the measures created to represent them) behave in ways are consistent with hypotheses made about those constructs. Statistical modeling is the primary tool for the second aspect of construct validation—determining whether the relationships among our observations are being produced by the same construct or different constructs.

One statistical modeling technique that is ideally suited to this task is structural equation modeling (SEM: Bollen, 1989). SEM provides a means by which a researcher fit different statistical models to a set of observations to see if the covariation seen among the observations is best captured by one or more constructs. One influential example of the use of statistical modeling to assess construct validity is the study conducted by Wagner, Torgesen, Laughon, Simmons, and Rashotte (1993). In this particular study, Wagner and colleagues were attempting to operationally define the construct of phonological processing that was originally posed.
in Wagner and Torgesen (1987). They constructed a number of assessments that were thought to measure different aspects of phonological processing ability. In this study, they first used SEM to investigate the individual factors that they believed comprised phonological processing. The statistical models that were fit to these data were able to provide these researchers with evidence the items that they had developed were either all tapping the same factor or were tapping different factors. Once they were satisfied with their measurement models, they used SEM to test alternate models of phonological processing that would best represent the construct of phonological processing. As stated in the beginning of this chapter, one of the biggest strengths of statistical modeling is to compare different models against each other in terms of fit and parsimony. These researchers started with the simplest model—that all of the observations developed to assess phonological processing measure just one construct. They then proposed more complex models that may also best represent the covariation seen among the observations. The results of these model comparisons revealed two phonological awareness constructs (phonological analysis and synthesis), a phonological memory construct, and two phonological code retrieval efficiency constructs. Further research on phonological awareness found that the two phonological awareness constructs are probably better characterized as one unitary construct (Schatzschneider, Francis, Foorman, Fletcher, & Mehta, 1999; Anthony et al., 2002). This study is an excellent example of the use of statistical modeling in literacy research to inform construct validation. Further research has also validated the usefulness of phonological processing as a useful construct when studying how students learn to read.

**Identifying Relationships**

Studies of individual differences have fruitfully used statistical modeling to identify important relationships between constructs as they relate to reading (Bowey, 2005). A variety of statistical models have been employed to analyze data from these individual differences studies. Simple correlations, multiple regression and SEM are the most common statistical models in an individual difference study. In most individual difference studies of literacy, researchers attempt to find important correlates of word reading and/or comprehension skills. Many of the individual differences studies of cognitive correlates of reading ability are searching for the cognitive sub-components that are thought to be necessary for efficient word reading or reading comprehension skills. Bowey (2005) reviewed a large body of individual difference studies and identified six constructs that have consistently been found to reliably relate to reading ability in the early grade: (a) verbal ability (Bowey, 1995; Cronin & Carver, 1998), (b) phonological memory (Badian, 2000; Wagner, Torgesen, & Rashotte, 1994); (c) speech perception and production (Scarborough, 1990); (d) phonological awareness (Wagner et al., 1994; Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004); (e) letter name knowledge (de Jong & van der Leij, 1999; Schatschneider et al., 2004); and (f) rapid automatized naming (Wagner et al., 1994; Wolf, Bally, & Morris, 1986; Schatschneider et al., 2004).

Individual differences studies in reading are not restricted to the search for cognitive correlates of reading skill. Many literacy researchers have also looked for environmental and social factors that correlate with reading ability. Broad environmental factors such as socioeconomic status (Arnold & Doctoroff, 2003) and mother’s education (Riciuti, 1999) as well as parental expectations of achievement (Hill, 2001), and home literacy environments (Burgess, Hecht, & Lonigan, 2002) have all been shown to be consistently correlated with children’s reading ability.

Statistical models such as multiple regression and SEM have helped identify these constructs as important correlates of reading ability. These models have done so by providing the researcher the ability to identify which constructs have a unique relationship with reading, above and beyond other constructs. This is an important tool for researchers to have because these models provide evidence that a particular construct provides non-redundant information in its relation
to reading ability. If we view reading ability as something with multiple causes, the use of statistical models in individual differences studies can provide the literacy researcher with clues as to which construct may be an INUS condition for reading ability.

Of course, it is well known that correlations do not prove causation. Many constructs may be correlated with reading because of a shared relationship with another construct. However, correlations do provide a minimally necessary condition for a causal relationship between two constructs. If there is no correlation, there can be no causal relationship. Individual differences studies provide crucial information to literacy researchers in that they point to constructs that, if interventions can be developed to improve, may help children become better readers. Not all of the constructs identified through correlational research will be helpful, and some of them may only be helpful for certain students in certain contexts. However, the statistical models used in individual difference studies can provide us with good leads in the search for identifying possible causes in the individual differences seen in students reading ability. But it remains for experimental research to determine whether interesting correlations represent real causes.

Experiments and Causal Conclusions

Empiricism has its basis in observation, and especially in observations of experiments (Maxwell & Delaney, 2004). The goal of experimentation is to discover the effects of potential causes (Shadish et al., 2002). How can we know if something is a “cause” of something else? Philosopher John Stuart Mill (1843/1987) proposed three conditions for inferring cause: (a) The potential cause must precede the effect, (b) the potential cause must be related to the effect, and (c) there is no other plausible alternative explanation for the effect. Shadish and colleagues (2002) suggest that experiments are well suited to studying cause and effect relationships because they (a) ensure that a presumed cause is deliberately manipulated and, thereby, precedes the observed effect, (b) incorporate procedures that help determine whether the cause is related to the effect, and (c) incorporate procedures to minimize and/or assess the influence of extraneous factors that could produce the effect presumed to be attributed to the cause.

To be able to infer that a potential cause has had an effect in an experiment, it is crucial that we have some knowledge of what would have happened if the potential cause is absent. Inferring effects by comparing them to what would have happened if the potential cause had been absent is called counterfactual inferencing (Shadish et al., 2002). The essence of counterfactual inferencing is that we can only know if event C caused event E if it were the case that if C had not occurred, E would not have occurred.

In literacy research, it is not possible to know what would have happened if a particular intervention, for example, had not been delivered. This is why hypothetical counterfactuals are developed through the use of random assignment. The control group in a randomized experiment represents our best guess as to what would have happened to the treatment group had they not been given treatment.

Statistical models play a large role in the analysis of observations from experiments in literacy research. Because hypothetical counterfactuals have to be employed based on other groups of students, it is certainly possible that the control group may differ from the treatment group based solely on chance. Statistical models are ideally suited to help the researcher discern chance differences between groups from real effects. Statistical models do so by assigning a probability value to the possibility that the observations made in an experiment are solely due to chance factors. This is accomplished by comparing two models—one where the only actor operating in the experiment is chance differences due to sampling variation and another where both chance differences and treatment group differences explain the results. If the statistical model that includes the treatment group in its equation is a better model than the alternative model that does not include the experimental grouping, then the researchers have evidence to conclude that
the experimental manipulation is the cause of the difference. A “better model” in this context is defined as one that provides a significantly better prediction of the subjects observed scores by knowing whether or not they received the treatment, in relation to the loss of parsimony that comes from the addition of this grouping variable into the equation. Whether the model that contains the group membership variable is a significantly better model is determined by comparing the improvement in prediction obtained using this model to a known probability distribution (at distribution, for example), and if the probability that this improvement in prediction occurred by chance is sufficiently low (most commonly set at less than a 5% chance), then we would conclude that this model is a better model. The number of experimental studies in literacy has grown rapidly over the past decade. In the United States, the Institute for Education Science (IES) was created in 2002 as a part of the U.S. Department of Education. Its mission is “to provide rigorous evidence on which to ground education practice and policy. By identifying what works, what doesn’t, and why, we intend to improve the outcomes of education for all students, particularly those at risk of failure” (http://ies.ed.gov/director/). To advance that mission, IES provides millions of dollars for randomized controlled research trials to determine whether particular programs and practices have a causal relationship to improved reading outcomes.

However, it bears repeating that in literacy research, we are not typically studying cause and effect relationships, but INUS conditions. In a true causal relationship, when a cause occurs, an effect will occur every time. INUS conditions, however, are a sufficient condition in that only in combination with other factors, do they contribute to reading ability. This has profound implications for drawing conclusions from randomized controlled experiments in literacy research. What this means is that an effect will most likely not occur every time for every student. Because the ability to read has multiple causes, and the experiment being conducted could not possibly intervene on all of them, some students will not respond to a particular intervention, or some may respond more than others. In this sense, our models become “probabilistic prediction models” (Stanovich, 2003) in that an identified “effective” intervention only increases the probability that a student will benefit from the intervention. It is of critical importance that literacy researchers explain what they mean when they say they are running studies that draw “causal conclusions,” and discuss the limitations of their findings. Additionally, consumers of science need to be made aware that these studies do not guarantee that an intervention will work for every student, and also that if an intervention does not work for one student, it does not negate the possibility that it would work for another.

USE OF STATISTICAL MODELING IN LITERACY RESEARCH: AN EXEMPLAR

Statistical models serve as valuable tools for conducting empirical research. From assessing the quality of our measurements to determining the effectiveness of our interventions, statistical models provide an objective means by which we can evaluate our research questions. In this chapter, we have touched upon a number of broad uses of statistical models. At this point, we thought it would be illuminating to describe the use of statistical models in the context of a single study. Out of the hundreds of potential studies to describe in more detail, we selected Wagner et al. (1993). We chose Wagner et al. as an exemplar because their work has been influential in our study of early reading skills and they used some relatively sophisticated statistical models to perform construct validation and to identify important relationships.

Wagner et al. (1993) administered a number of assessments that were hypothesized to be related to early reading development to a group of kindergarten and second grade students. Many of the assessments used in the study were developed by Wagner and colleagues and the choice of which assessments to give to the students was grounded in the theoretical work done by Wagner and Torgesen (1987). In their 1987 paper, Wagner and Torgesen hypothesized the existence of
three correlated but distinct constructs that they believed to be causally related to early reading development: Phonological awareness, phonological recoding in lexical access, and phonetic recoding in working memory. These three conceptual constructs comprise the superordinate construct of phonological processing. In order to provide evidence for the existence of these constructs, Wagner and colleagues employed or created measures that they believed would tap the skills that were thought to encompass these three constructs and administered them to kindergarten and second grade children in order to obtain estimates of how much these measures covary with each other. These covariances provide the basis by which researchers hypothesize what causes some measures to covary. Measures that covary highly are thought to do so because performance on those measures is determined by the same underlying latent (unobserved) ability. Performances on assessments that do not covary strongly are thought to be driven by different cognitive constructs.

Wagner et al. (1993) then fit a series of structural equation models to the covariance matrix of the measures of phonological processing. Structural equation models create predicted covariance matrices that can be compared to the observed covariance matrix obtained from the data. The comparison of the predicted covariance matrices to the observed covariance matrix determines the model fit, or how well the model explains the original variances and covariances of the observed variables. The difference between the predicted and observed covariance matrix from any structural equation model can be expressed as a chi-square value (which represents the sum of the squared difference between the elements in the predicted and observed covariance matrices weighted by sample size). This chi-square value can then be compared to a chi-square distribution table to determine statistical significance. Because the chi-square is based on the squared differences between observed and predicted covariances, models that fit better will have smaller chi-square values, and nonsignificant probability values associated with a model will typically indicate a good fit.

A stronger use of structural equation models comes when we can compare two predicted models against each other in terms how well each explains the observed covariance matrix. Wagner et al. (1993) employed this model comparison approach when they proposed to test a series of five alternate models of phonological processing abilities. They argued for the potential viability of each of these models including models where all of the covariance observed can be explained by one common latent factor (the most parsimonious solution) and another model where each of the abilities thought to measure phonological processing was its own latent construct (least parsimonious but most likely will provide the best fit to the data).

But as is true with many proposed hypotheses, the models did not directly conform to their pre-conceived ideas. As stated before, the researchers proposed that three abilities were thought to comprise phonological processing: Phonological awareness, phonological recoding in lexical access, and phonetic recoding in working memory. However, when testing their models, they found that their measures of phonological awareness were better modeled as having two latent factors instead of only one. They named the two factors phonological analysis and phonological synthesis. Additionally, the measures thought to tap phonological recoding in lexical access were also best represented by two latent constructs that they subsequently named isolated naming and serial naming. This unexpected finding raised the number of models they wanted to test from five to seventeen. After fitting a series of models, assessing their fit, and comparing them to each other, Wagner et al. (1993) arrived at a four factor solution for kindergarten students and a five factor solution in second grade. These models were found to be the best balance between model fit and parsimony. Once they arrived at these solutions, they then used the best fitting model from second grade and used those factors to predict word recognition skills. After controlling for general cognitive ability, they found that the phonological processing factors accounted for an additional 20% of the variance observed in word recognition skills above and beyond general cognitive ability.
In Wagner et al. (1993) we see many of the advantages of using statistical modeling. First, by formally testing their measurement models, they uncovered that in their data phonological awareness and phonological recoding in lexical access should be represented by two factors each instead of only one. Without examining this first, the researcher can only argue on conceptual grounds that measures tap the same latent ability. In the case of Wagner et al., measures that tap these constructs may have been aggregated together inappropriately. Second, their use of the model comparison approach gave them an empirical justification for deciding which constructs comprise the domain of phonological processing. Finally, the usefulness of assessing phonological processing was supported by the statistical models that demonstrated the explanatory power of these constructs in predicting word reading skills above and beyond general cognitive processing.

LIMITATIONS OF STATISTICAL MODELING

Statistical modeling is an incredibly powerful tool for literacy researchers. Statistical models aid in the construction of observational tools, the development and identification of constructs, and in understanding the relations of constructs to one another and to reading. They can be used for prediction, and they also assist in the identification of “causal” relationships. It is our belief that every literacy researcher should have statistical modeling in their methodological toolbag as one of the useful means by which questions in literacy can be answered.

To that end, it is also helpful to discuss the limitations of statistical models. There is a tool for every job, and it’s important to know the strengths and weaknesses of each tool. First, there is the obvious limitation that all statistical models rest upon certain mathematical assumptions. The consequences of violating these assumptions range from minimal to severe. Much has been written about the assumptions that underlie our models and the consequences of violating those assumptions. Less obvious are the conceptual limitations of statistical modeling. First, statistical models are limited to observations that can be turned into numbers (or at least categories). This limits models to those areas of literacy research where our observations of literacy can be quantified. Statistical models are not designed to directly analyze or synthesize qualitative information. This can often limit the types of questions that can get asked if a literacy researcher only uses statistical models. Just as we recommended that all literacy researchers should be versed in the use of statistical models, we also advocate that literacy researchers need to be able to do more than just employ statistical modeling. As the saying goes, if the only tool one has is a hammer, then one tends to see every problem as a nail.

Another limitation of statistical models is that they are only as good as the observations made by the researcher. The researcher interjects his bias in regards to which observations about a phenomena are made and which are not collected. In studying literacy, it would be almost impossible for a researcher to collect observations on all the biological, cognitive, social, emotional, and environmental influence on reading behavior, even though most literacy researchers would acknowledge the importance of all these areas. Statistical models can only provide information in regards to the observations that are collected, not the information that is ignored.

Another issue arises when two or more statistical models cannot be differentiated from each other. That is, is oftentimes the case that two or more models can provide a reasonable explanation of a set of observations (Breiman, 2001). These models may provide different estimates as to the importance of particular constructs, or perhaps may not even include the same constructs. Statistical models also have a difficult time distinguishing between nonlinear relationships and interactions (Lubinski & Humphreys, 1990), which also has implications for inferring which constructs are more closely related to reading ability than others.
A final limitation is related to the quote that opens this chapter. All models are wrong. But what does this statement mean? It means that all models are imperfect representations of what is occurring in nature. Models by their very nature are reductionistic and will not be able to fully explain the complex relationships we are attempting to understand. Statistical models will never tell us with certainty whether something is an exact cause of something else. Probability is inherent in every statistical model, and probabilities imply the possibility that a model is incorrect. But simply because this is true, it does not imply that models are not useful. Statistical models provide us with an incrementally better understanding of reading ability and development. They aid us in advancing our understanding of the components and correlates of reading, they give us ideas about which correlates may be fruitful to intervene with in order to enable more students to read, and they provide us with the ability to predict if our interventions will work on future students.

REFERENCES


