Routledge
Taylor & Francis Group

Research in Human Development

ISSN: 1542-7609 (Print) 1542-7617 (Online) Journal homepage: http://www.tandfonline.com/loi/hrhd20

Getting beyond the Null: Statistical Modeling as an Alternative Framework for Inference in Developmental Science

Kyle M. Lang, Shauna J. Sweet & Elizabeth M. Grandfield

To cite this article: Kyle M. Lang, Shauna J. Sweet & Elizabeth M. Grandfield (2017) Getting beyond the Null: Statistical Modeling as an Alternative Framework for Inference in Developmental Science, Research in Human Development, 14:4, 287-304, DOI: 10.1080/15427609.2017.1371567

To link to this article: https://doi.org/10.1080/15427609.2017.1371567

Published with license by Taylor & Francis Group, LLC© 2017 Kyle M. Lang, Shauna J. Sweet and Elizabeth M. Grandfield.

View supplementary material

Published online: 26 Oct 2017.

Submit your article to this journal

Article views: 300

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=hrhd20

Research in Human Development, 14: 287–304, 2017
Published with license by Taylor & Francis Group, LLC
ISSN: 1542-7609 print / 1542-7617 online
DOI: https://doi.org/10.1080/15427609.2017.1371567

Routledge
Taylor & Francis Group

Check for updates

Getting beyond the Null: Statistical Modeling as an Alternative Framework for Inference in Developmental Science

Kyle M. Lang
Department of Methodology and Statistics, Tilburg University

Shauna J. Sweet
Department of Human Development and Quantitative Methods, University of Maryland, College Park

Elizabeth M. Grandfield
Department of Psychology, University of Kansas

We describe statistical modeling as a powerful alternative to null hypothesis significance testing (NHST). Modeling supports statistical inference in a fundamentally different way from NHST which can better serve developmental researchers. Modeling requires researchers to fully articulate their beliefs about the processes under study and to communicate that understanding through the structure of a probabilistic model before testing specific hypotheses. Research hypotheses are assessed through estimated parameters of the model and by conducting model comparisons. We conclude the paper with a series of worked examples that highlight the merits of the statistical modeling approach as a tool for scientific inference.

Many articles have recently appeared in popular press (e.g., Krueger, 2014; Nuzzo, 2014) that criticize researchers' reliance on null hypothesis significance testing (NHST) as the default approach to building and testing psychological theories. The growing skepticism of NHST calls into question the status quo of how research is conducted and demands rethinking how hypotheses are investigated and how results are communicated. Psychology is certainly not unique in this struggle—researchers in nearly every branch of the social, behavioral, and biomedical sciences are being asked to confront and address their over-reliance on NHST, which is deeply embedded in research practices.

Correspondence should be addressed to Kyle M. Lang, E-mail: K.M.Lang@uvt.nl PO Box 90153, 5000 LE Tilburg, the Netherlands.

© 2017 Kyle M. Lang, Shauna J. Sweet and Elizabeth M. Grandfield.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

The inferential shortcomings inherent in the NHST framework have been the subject of methodological criticism for years (see Gelman & Loken, 2014; Levine, Weber, Hullett, Park, & Lindsey, 2008; Nickerson, 2000). Applied researchers are increasingly aware of NHST's oft-cited methodological limitations (e.g., *p* values' sensitivity to sample size, conceptual divide between statistical and practical significance, overinterpretation or misinterpretation of *p* values, inflation of Type I error rates through repeated testing). Even so, many researchers may not recognize how these methodological shortcomings directly translate into issues of social justice. In particular, developmental research often has direct, real-world policy impact on children, who are among the most vulnerable populations (Foster & Kalil, 2005). Findings from developmental research programs inform policy decisions that have direct and often widespread effects on children's lives. As Little (2015) notes, "research findings are the building blocks of policy and practice" (p. 269). Take, for example, the U.S. Department of Health and Human Services' (HHS) Teen Pregnancy Prevention program that enlists a team of methodological experts to regularly review relevant scientific literature (see Lugo-Gil et al., 2016, for results of the most recent evidence review). The recommendations based on this review help guide HHS funding decisions regarding teen pregnancy prevention programming.

The methods that researchers use to investigate competing hypotheses and to justify their conclusions are important. Lawmakers who appeal to scientific findings as the basis for policy decisions do so with the expectation that the conclusions presented are valid, reliable, and free of bias. Ensuring that evidentiary arguments are appropriately constructed and adequately supported is, therefore, fundamentally an issue of social justice. We define *social justice* as the idea that every person should enter life with equal protection from social structures, equal responsibility to support and improve those social structures, equal opportunity to benefit from the fruits of their labor, and equal obligation to suffer the consequences of their mistakes. Policies that are guided by biased science can violate this mandate by systematically disadvantaging (or advantaging) certain groups regardless of the policy makers' original intent. Consequently, we suggest that the shortcomings of NHST extend far beyond a sterile academic debate about the nuances of "statistical significance."

Although a necessary first step, criticism does not immediately translate into the identification and implementation of best practices. The American Statistical Association (ASA) released an official position statement highlighting the limitations of *p* values in January 2016 (Wasserstein & Lazar, 2016), and the editors of *Basic and Applied Social Psychology* (*BASP*) recently banned reporting *p* values in any submission to their journal. Yet the ASA statement and the BASP policy stopped short of advocating for specific alternatives. This article aims to extend the conversation around NHST from criticism into practice by proposing statistical modeling as a powerful alternative, particularly when viewing evidentiary argumentation through the lens of social justice. We begin by providing working definitions of *NHST* and *statistical modeling*. Through these definitions, we contrast statistical modeling with NHST as an approach to social scientific inquiry and hypothesis testing. In the second half of this article we analyze real data to illustrate the strengths of the modeling approach.

NULL HYPOTHESIS SIGNIFICANCE TESTING

When we refer to *NHST* in this article, we are referring specifically to the hybrid NHST approach originally proposed by Lindquist (1940) that has since become standard practice throughout the social scientific community. Hybrid NHST is a specific framework for conducting null

hypothesis tests that combines the hypothesis testing tradition of Jerzy Neyman and Egon Pearson (Neyman, 1935, 1942, 1955; Neyman & Pearson, 1928, 1933) with that of Sir. Ronald Fisher (Fisher, 1932, 1954, 1955, 1960). Hybrid NHST begins with defining a set of research questions relating to processes within, or characteristics of, the population being studied. These research questions are then used to construct falsifiable null and alternative hypotheses. The alternative hypothesis codifies the expected answer to the research question whereas the null hypothesis is the complement of the alternative hypothesis and contradicts the researchers' expectations. Finally, the research questions are evaluated by collecting and analyzing sample data to test the tenability of the null hypothesis.

LIMITATIONS OF NHST

Before proceeding, it is important to clarify which aspects of NHST pose problems for developmental research. Hypothesis testing itself is certainly not the issue. Science cannot exist without hypothesis testing, and we do not advocate a retreat to purely descriptive, exploratory, or qualitative modes of inquiry. Neither do we suggest a "scorched-earth" ban on NHST in all its forms. In some circumstances, NHST remains a useful tool. In certain experimental contexts, for example, NHST can be a perfectly adequate methodology for gathering inferential evidence (Cohen, 2017). Furthermore, NHST can be a useful tool to guide model building. Many model-building decisions are informed by atomic NHSTs of individual model parameters (e.g., testing the null hypothesis that the value of a regression parameter is zero), and nested model comparisons are often tested with NHST (e.g., using the $\Delta\chi^2$ to test the null hypothesis that the change in fit between two nested models is zero). We simply suggest that NHST, on its own, is an inadequate framework within which to build and evaluate modern developmental theory. You probably would not want to build a house without any nails, but neither should you attempt to build a house using nothing but nails. We do not wish to ban nails, but we will humbly recommend purchasing some lumber and a hammer.

Within the framework of NHST, data are gathered as evidence used to test whether the null hypothesis—the complement of the interesting hypothesis—is tenable. Although the null hypothesis can encode any falsifiable claim, it most commonly states that there is no effect present or no relationship; this specific null hypothesis is called the "nil-null," and it is almost never plausible in developmental science (as noted in Little, Widaman, Levy, Rodgers, & Hancock, this issue). NHST poses the question: "If the null is true in the population, what is the probability of encountering the observed data patterns or some more extreme pattern?" Notably absent from this question is any trace of a substantive hypothesis. A low probability of observing the current data, given a true null hypothesis (i.e., a small p value), suggests that the data were sampled from a population wherein the null is unlikely to be true—nothing more. In a standard application of NHST, finding significance amounts to the underwhelming accomplishment of rejecting a claim that was not expected to hold anyway.

STATISTICAL MODELING

Simplicity is a compelling and oft-cited reason for relying on NHST to build evidentiary arguments (Little, 2015). Statistical modeling is a more complicated approach than NHST,

statistically and in terms of presentation. This added complexity, however, affords researchers the opportunity to quantify evidence in support of specific substantive hypotheses relative to competing hypotheses—not simply against a null hypothesis. As developmental scientists, our goal is to elucidate underlying developmental processes that give rise to the population-level phenomena. This objective demands an inferential framework that is grounded in a thoughtful consideration of those processes; statistical modeling is an ideal candidate.

Statistical models are idealized representations of reality (Pearl, 2009) that require researchers to situate inferences in context by fully describing a hypothesized process in a concrete, mathematical representation. In developmental science, statistical models articulate and operationalize researchers' substantive theories about how latent psychological processes give rise to observed phenomena. Statistical models also quantify the potential interplay between these psychological processes and relevant environmental, cultural, and physiological influences. When encoded as a statistical model, the components of a theory and the constituent pieces of the process under study are therefore quantifiable, testable, and falsifiable. Similar to Collins (2006) and Rodgers (2010a; 2010b), we consider statistical modeling as an epistemological enterprise, whereby statistical models are derived from substantive theory and researchers seek to define, to the extent possible, "a close correspondence between theoretical and statistical [models in order to] provide an elegant test of a scientific hypothesis and a penetrating look at [available] data" (Collins, 2006, p. 509).

As conceptualized here, all statistical models share two core features that contribute to their usefulness as tools for statistical inference. First and foremost, a statistical model must be rooted in theory and specified to parsimoniously quantify the hypothesized data generating process. Second, all statistical models are encoded as probability distributions, with the parameters of these distributions defined according to the theorized associations among the variables. To fix ideas, suppose that a researcher believes children's levels of emotional expressivity are linearly dependent upon their ages and attachment styles. This theory is reflected in a multiple linear regression (MLR) model wherein emotional expression score is regressed onto age and attachment style. In terms of probability distributions, this regression model corresponds to a normal distribution of emotional expressivity with a freely estimated variance (i.e., the residual variance in the regression model), and a mean defined by the weighted sum of age and attachment style (i.e., the predicted values from the regression model).

The purpose of statistical modeling is to represent—as accurately and completely as possible—a data generation process, with the goal of understanding and gathering evidence about its structure. The aim is to gather (relative) evidence for substantive hypotheses rather than simply obtain evidence against a hypothesis of no effects. This evidence is not obtained from direct tests of the substantive hypotheses (i.e., modelers do not simply invert NHST logic to directly test the alternative hypothesis). Rather, relative evidence for substantive hypotheses is obtained by comparing models which represent competing data generating processes. Models are compared in terms of their ability to describe the observed data (i.e., model fit) and to predict future data (i.e., replication and cross-validation). The model that best describes the current data and best predicts future data (where these goals must be balanced according to the infamous bias-variance-tradeoff; Hastie, Tibshirani, & Friedman, 2009) is the most likely candidate (among the pool of tested models) for the true data generating process.

Model comparisons provide relative, not absolute, support because most models are just one member of a set of equivalent models (i.e., their equivalence class) that all describe the data

equally well (Lee & Hershberger, 1990; MacCallum, Wegener, Uchino, & Fabrigar, 1993; Stelzl, 1986). For example, a model that regresses emotional expressivity onto age is equivalent to one that regresses age onto emotional expressivity. The ever-present threat of equivalent models further necessitates firmly grounding statistical models in theory. Much of a model's equivalence class can be ruled out through theory. The second of the two equivalent models listed above, for example, can be ignored because children should continue to age regardless of how their emotional expressivity changes. Despite this limitation, statistical modeling does an excellent job of supporting rigorous scrutiny of the structures that reflect competing theories.

How is Modeling Better than NHST?

If, as we claim above, NHST remains a useful—albeit limited—tool, what unique benefits does statistical modeling bring? In two words: *ecological validity*. Human behavior, development, and psychological processes are all examples of complex systems (Eidelson, 1997). A complex system is one that cannot be fully described simply by quantifying its constituent parts (i.e., a whole that is more than the sum of its parts). Statistical models describe the data generating process of an entire system. Moreover, statistical modeling is a framework in which structural variants of a system can be evaluated and compared. Because every statistical model is a probability distribution, simple models may be combined to form a complex model by collapsing simple probability distributions into a distribution with higher dimensionality. For example, the process model shown in the upper panel of Figure 1 (discussed in more detail below) can be partitioned into five MLR models (i.e., two models wherein an outcome variable is the dependent variable [DV] and three models wherein a mediator is the DV). In other words, statistical models are modular, which is one of their greatest strengths in application.

Statistical models are well-suited to quantifying complex human systems and to testing the competing theories that seek to describe them. NHST, on the other hand, is sanitized; any given NHST can only ask one question, can only probe one isolated dimension of a system. NHST can never provide the same holistic representation of a complex system that statistical modeling can produce. NHST might be useful for testing individual parameters in a model, but without the context provided by a statistical model, NHST cannot stand alone as an adequate tool to quantify the complex systems of developmental science.

FLAVORS OF STATISTICAL MODELS

Statistical models do not need to be intrinsically complicated, but they do vary in complexity. Different modeling frameworks introduce additional levels of complexity in return for more nuanced inferential capabilities. In this section, we discuss some of the key differences in popular modeling paradigms.

Observed versus Latent Variable Modeling

Observed variable modeling entails fitting statistical models to sets of measured variables or observed aggregates thereof (e.g., scale scores). Multiple linear regression is probably the most common type of observed variable modeling. Although observed variable modeling brings all

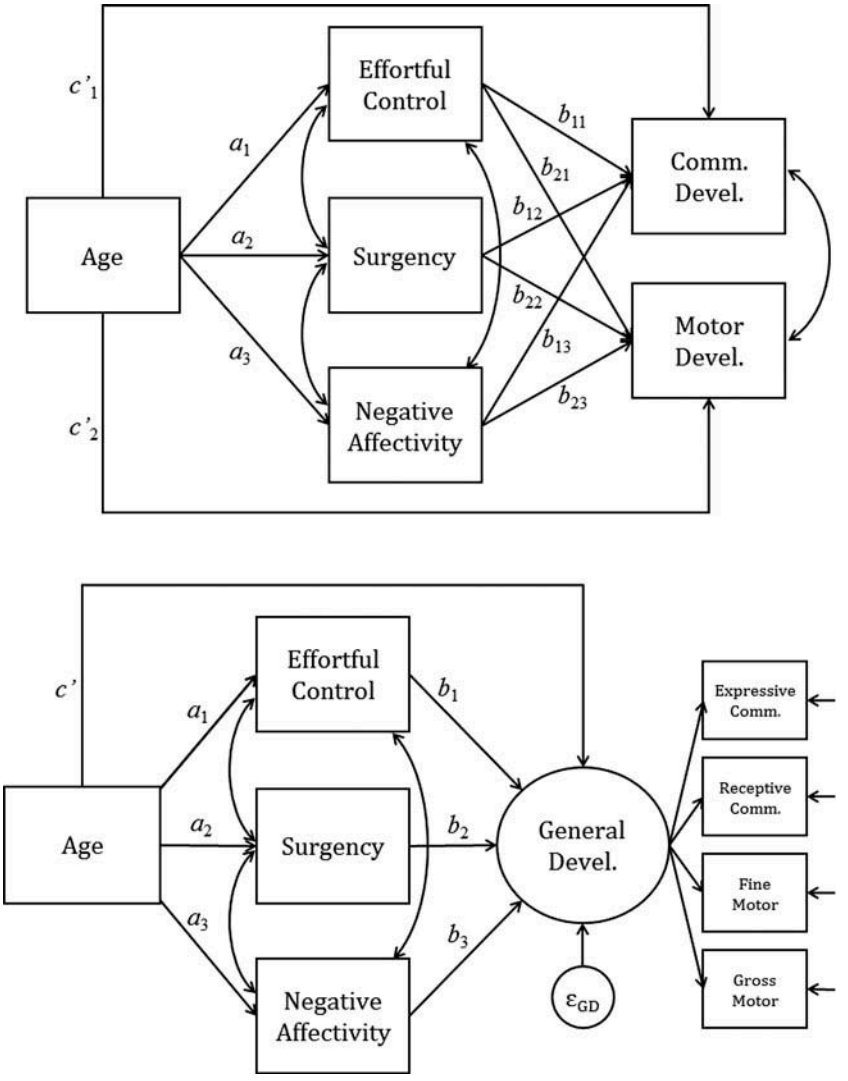

FIGURE 1 Path diagrams of process models.

the basic benefits of the statistical modeling paradigm, it also requires making several strong assumptions about the measurement properties of any aggregate scores used in an analysis. These assumptions may not be warranted or defensible, particularly in the context of human development research. For example, traditional scale scores are assumed to be measured without error and the underlying scales are assumed to demonstrate equivalent psychometric properties across groups in multiple-group modeling.

The latent variable modeling approach admits an explicit submodel for the measurement structure of a scale. In latent variable modeling (which is one of the practices recommended in

Little, Gorrall, Panko, & Curtis, this issue), the multiple indicators of a scale are used to define a latent variable that is free of measurement error (Kline, 2015). Having an explicit model for the measurement structure of these latent variables allows explicit tests for the equivalence of the measurement structure between groups or over time. These so-called measurement invariance tests ensure equivalent psychometrics across groups which, in turn, ensure between-group comparisons of the latent variables are anchored to the same operational definition.

Frequentist versus Bayesian Modeling

The fundamental difference between Bayesian and frequentist models is that parameters are viewed as random variables in the former but are viewed as fixed values in the latter. After estimating the model, therefore, Bayesians can make direct probabilistic inferences about any parameter in the model. Statements about parameters in a frequentist model, on the other hand, must be indirectly supported by probabilistic inferences about the data after conditioning on the estimated parameters. Each parameter in a Bayesian model has an explicitly estimated distribution, whereas Frequentist parameters are assumed to follow asymptotic sampling distributions. So, a Bayesian can ask: "What is the probability that $r > 0.3$?" but a frequentist must ask: "What is the probability of observing these data, or more extreme data, if $r > 0.3$?"

The added inferential flexibility of Bayesian models comes at the cost of somewhat more complex implementation. When data and parameters are viewed as random variables, the data analyst must encode their a priori beliefs about the parameters' distributions to identify the model—a role that is filled by assumptions in Frequentist modeling. By specifying these so-called prior distributions for every parameter in the model, Bayesian modelers must not only consider their hypothesized understanding of the data generating process, but also their degree of a priori knowledge about the system under study.[1] The task of specifying priors is an area of active research that is discussed in much greater detail in Zondervan-Zwijnenburg, Peeters, Depaoli, & van de Schoot (this issue).

As fruit of the exhaustive model specification, a Bayesian parameter estimate is the entire posterior distribution of the model's parameters (McElreath, 2016)—not simply point estimates of the data distribution's parameters. The posterior distribution is a statistical model that represents a weighted mixture of the researcher's prior knowledge about the system (contributed through prior distributions) and the new evidence provided by the data (contributed through the likelihood). In other words, an estimated Bayesian model represents a tangible realization of the abstract statistical model we have been discussing throughout this paper.

LEVERAGING STATISTICAL MODELS

In this section, we utilize real data to demonstrate the strengths of the statistical modeling approach. We do so by demonstrating a simple process analysis (Hayes, 2013) which is a modeling framework that tests hypotheses involving mediation and causal processes.[2] A fully realized process analysis requires a modeling approach because the analyst must explicitly conceptualize the entire process before any hypotheses can be tested. Yet simple NHST-based approximations to a full process analysis (e.g., the causal steps approach; Baron & Kenny, 1986) remain popular and provide a resonant point of comparison for the modeling-based analysis.

Suppose we are interested in exploring how child temperament mediates the relationship between age and developmental progress in communication and motor skills. This question can be addressed via the model shown in the upper panel of Figure 1. Temperament is represented by three subdomains: effortful control (an age-appropriate analogue of self-regulation), surgency (an age-appropriate analogue of extroversion), and negative affectivity. This model is the basis for the three example analyses reported below: a strictly NHST-based implementation of the causal steps approach, a full process analysis using structural equation modeling (SEM), and a Bayesian version of the SEM-based process analysis. All analyses were conducted in R (R Core Team, 2017), and all code used in these analyses is available as online supplementary material. Unfortunately, we are not able to distribute the example data due to restrictions imposed by the consent procedures of the study for which the data were originally collected.

Data

The data analyzed in these examples[3] contain parent reports ($N = 243$) of several developmental measures for a sample of children age 1 to 8 years ($M = 3.87$, $SD = 2.47$). The sample was 51.4% male ($n = 125$), 65.4% white ($n = 159$), 8.5% black ($n = 21$), 8.2% Asian or Pacific Islander ($n = 20$), and 17.7% mixed race ($n = 43$).

Measures

Children's ages, in months, were provided by the parents. The mediators and outcomes were constructed from the following parent-reported scales.

Communication and motor development

The children's overall communication and motor development was assessed with four raw scales (i.e., not the age-standardized versions) from the Vineland Adaptive Behavior Scale: II (VABS-II; Sparrow, Balla, & Cicchetti, 2005): expressive/receptive communication skills and fine/gross motor skills. We did not use the written communications skills item because it was not administered to children younger than age 36 months. Internal consistencies of the communication and motor skills domains were high in the current sample (*Communication*: Cronbach's $\alpha = .95$, 95% confidence interval [CI][4] = [0.93, 0.96]; *Motor*: Cronbach's $\alpha = .94$, 95% CI = [0.92, 0.95]).

Child temperament. The children's levels of effortful control, surgency, and negative affectivity were measured using the appropriate subscales of the Very Short Form of the Early Child Behavior Questionnaire (ECBQ; Putnam, Jacobs, Gartstein, & Rothbart, 2010) for children age 36 months or younger and the Very Short Form of the Child Behavior Questionnaire (CBQ; Putnam & Rothbart, 2006) for children older than age 36 months. Each subscale contained 12 items. In the current sample, internal consistencies for the ECBQ-derived temperament domains were moderate (*Effortful Control*: Cronbach's $\alpha = .69$, 95% CI = [0.59; 0.76]; *Surgency*: Cronbach's $\alpha = 0.73$, 95% CI = [0.64; 0.79]; *Negative Affectivity*: Cronbach's $\alpha = .58$, 95% CI = [0.43, 0.69]). Internal consistencies for the CBQ-derived domains were somewhat higher than the ECBQ versions but also moderate (*Effortful Control*: Cronbach's

α = .69, 95% CI = [0.59, 0.76]; *Surgency*: Cronbach's α = .74, 95% CI = [0.65, 0.80]; *Negative Affectivity*: Cronbach's α = .74, 95% CI = [0.64, 0.81]).

Missing Data

Child age was fully observed, and missing data rates for the other scales were low. Specifically, effortful control was missing 5.1%, surgency was missing 6.6%, negative affectivity was missing 3.4%, and the four VABS indicators were missing between 0.4% and 1.2%. Considering the high response rates, missing values were replaced using a single imputation conducted with the *mice* package (Van Buuren & Groothuis-Oudshoorn, 2011). All data were imputed at the item level (i.e., before constructing any aggregate scores) using 25 iterations of the MICE algorithm. Continuous variables were imputed with Bayesian linear regression, nominal variables were imputed with multinomial logistic regression, and ordinal variables were imputed with proportional odds logistic regression. Age, gender, and presence of siblings were used as predictors in all imputation models. We included these three covariates in all imputation models to retain the possibility of included them in the inferential models (although we did not end up doing so). Additional auxiliary variables were selected using the *quickpred* function provided by *mice* to select variables correlated with the imputation targets or their response indicators at greater than $r = .2$. A full list of potential auxiliary variables subjected to the *quickpred* procedure is available as online supplementary material.

ANALYSIS AND RESULTS

Before analysis, children's ages were converted to years and mean-centered. Mean scores were created from the effortful control, surgency, and negative affectivity items.[5] To facilitate the causal steps analysis, mean scores were also created for communication skills and motor skills. The mean scores and the communication skills and motor skills items were standardized before analysis.

Causal Steps

The causal steps approach (Baron & Kenny, 1986) entails inferring the existence of an indirect effect by satisfying four binary conditions. First, the input must be significantly correlated with the outcome. Second, the input must be significantly correlated with the mediator. Third, the mediator must be significantly correlated with the outcome, controlling for the input. Fourth, the direct effect (i.e., the effect of the input on the outcome, controlling for the mediator) must be significantly smaller than the total effect (i.e., the unconditional effect of the input on the outcome).[6] The most notable limitation of this approach, from the perspective of our current discussion, is the complete absence of the indirect effect itself in any of these four steps. The causal steps approach implies an indirect effect strictly based on a series of naïve significance tests without explicitly considering the estimated indirect effect or the larger process model from which it originates.

Using the causal steps approach to piece together inferential support for the process represented in Figure 1 requires breaking the model into six specific indirect effects (i.e., the indirect effects of age on communication and motor development through effortful control, surgency, and

negative affectivity while controlling for the two temperament domains not acting as mediator). Each of these specific indirect effects must then satisfy the four conditions enumerated above. To check these four conditions without explicitly modeling the underlying process, we can use partial correlations to estimate the various *a* and *b* paths. We wrote an R function (available as online supplemental material) that uses the *ppcor* package (Kim, 2015) to estimate the requisite partial correlations and automatically checks the four causal steps. The Sobel (1982) test was used to assess the significance required in Step 4. The results of this procedure are shown in the first panel of Table 1. Only two of the specific indirect effects failed to achieve statistical significance at the $\alpha = .05$ level (i.e., Age → Negative Affectivity → Communication Skills and Age → Surgency → Motor Skills). Both tests failed because the mediators and outcomes were not significantly correlated after controlling for the input.

Latent Variable Modeling

Immediate improvement on the strictly NHST-based causal steps analysis can be made simply by estimating the *a* and *b* paths within MLR models. Doing so refocuses emphasis on estimating the indirect effect *ab*, as opposed to simply stepping through a series of binary decisions about its significance. To improve on MLR-based process analysis, the entire process model can be simultaneously estimated using path analysis. In the current data, however, we can go one step further because the multiply-indicated outcome measures allow modeling the process as a structural equation model (SEM; which is a type of latent variable model). By modeling the process shown in Figure 1 using SEM, we can simultaneously estimate the entire process model and explicitly model (and test) the measurement structure of the outcome variables.

Measurement model. All latent variable models were estimated with the *lavaan* package (Rosseel, 2012). The first step in the SEM analysis was to assess the marginal measurement model of the outcome variables. To do so, we fit a two-factor confirmatory factor analysis (CFA) model wherein communication development and motor development were modeled as correlated latent variables that were indicated by two observed items each. We set the scales by fixing the latent variances to 1.0. Mean structures were not modeled. Although this model fit the data well ($\chi^2 = 1.134$, $df = 1$, $p = .287$, Comparative Fit Index [CFI] = 1.00, Root Mean Square Error of Approximation [RMSEA] = .023, Standardized Root Mean Square Residual [SRMR] = .003), and produced sensible estimates for the measurement parameters ($\lambda \in [0.918, 0.973]$, $\theta \in [0.049, 0.153]$), the latent factors were perfectly correlated ($\psi_{21} = 1.00$). We, therefore, tested a more parsimonious measurement model with a single latent factor underlying all four indicators. This model also fit the data well ($\chi^2 = 1.138$, $df = 2$, $p = .566$, CFI = 1.00, RMSEA = .000, SRMR = .003) and produced sensible estimates of measurement parameters ($\lambda \in [0.918, 0.973]$, $\theta \in [0.049, 0.153]$). A $\Delta\chi^2$ test between the two-factor and one-factor models showed that the one-factor representation did not fit the data significantly worse than the two-factor version ($\Delta\chi^2 = .004$, $\Delta df = 1$, $p = .952$). We, therefore, chose to employ the more parsimonious measurement representation in the structural models. Note that even though this model comparison was evaluated using an NHST-based $\Delta\chi^2$ test, our conclusion was to support the substantive hypothesis that the data were generated via the more parsimonious of the compared processes.

TABLE 1
Parameter Estimates from Frequentist and Bayesian Structural Equation Models (SEMs)

Causal Steps

Effect	*Z-Statistic*	p-*Value*	*Step Progress*
Age → Con → Com	4.61	<.001	All Passed
Age → Srg → Com	−2.45	.014	All Passed
Age → Neg → Com	—	—	Failed Step 3
Age → Con → Mot	4.25	<.001	All Passed
Age → Srg → Mot	—	—	Failed Step 3
Age → Neg → Mot	3.12	.002	All Passed

Frequentist SEM

Effect	*Point Estimate*	*Standardized Estimate*	*95% CI LB*	*95% CI UB*
$a_{control}$	0.180	0.445	0.136	0.224
$a_{surgency}$	−0.094	−0.232	−0.144	−0.045
$a_{negative}$	0.244	0.604	0.197	0.289
$b_{control}$	0.513	0.189	0.314	0.691
$b_{surgency}$	0.109	0.040	−0.034	0.252
$b_{negative}$	0.310	0.114	0.120	0.503
c’	0.841	0.765	0.713	1.050
$IE_{control}$	0.092	0.084	0.053	0.140
$IE_{surgency}$	−0.010	−0.009	−0.028	0.003
$IE_{negative}$	0.076	0.069	0.033	0.133
TIE	0.158	0.144	0.092	0.233

Bayesian SEM

Effect	*Posterior Mode*	*Standardized Mode*	*95% CI LB*	*95% CI UB*	*80% CI LB*	*80% CI UB*
$a_{control}$	0.180	0.445	0.134	0.227	0.150	0.211
$a_{surgency}$	−0.096	−0.236	−0.145	−0.043	−0.126	−0.060
$a_{negative}$	0.244	0.604	0.204	0.285	0.217	0.270
$b_{control}$	0.501	0.184	0.344	0.673	0.401	0.617
$b_{surgency}$	0.109	0.042	−0.035	0.249	0.017	0.203
$b_{negative}$	0.311	0.115	0.132	0.484	0.192	0.423
c’	0.828	0.763	0.702	0.964	0.742	0.911
$IE_{control}$	0.087	0.081	0.054	0.130	0.065	0.115
$IE_{surgency}$	−0.008	−0.007	−0.026	0.004	−0.019	0.000
$IE_{negative}$	0.074	0.066	0.031	0.120	0.044	0.103
TIE	0.152	0.142	0.095	0.218	0.116	0.195

Note. CI = confidence interval; LB = lower bound; UB = upper bound.

Mediation model. The structural mediation model is represented in the lower panel of Figure 1. As noted in Little et al. (this issue), the ability to “dialogue with data” and update our theory from the structure represented in the upper panel of Figure 1 to that represented in the lower panel is one of the great strengths of statistical modeling. After defining the measurement structure

of the outcome variable, the observed input and mediators were simply introduced as diagrammed, and the indirect effects were defined as the relevant products of the three a and b paths (i.e., a_1b_1, a_2b_2, a_3b_3). The shift to effect estimation, rather than naïve significance testing, admits the possibility of using more robust methods to test the indirect effects' significance. Current best-practice in mediation analysis calls for testing the indirect effects using nonparametric bootstrapping (Hayes, 2013; Jose, 2013). Nonparametric bootstrapping is a technique introduced by Efron (1979) that resamples the data, with replacement, many times (e.g., $B = 1000$). The focal effect is estimated in each of the resamples and the B replicates of the effect produce an empirical sampling distribution from which nonparametric confidence intervals can be computed. We employed $B = 1000$ resamples, and we used bias corrected confidence intervals (Efron, 1981) to control for the potential asymmetry in the indirect effects' sampling distributions.

The structural model did not fit the data especially well ($\chi^2 = 224.21$, $df = 14$, $p < .001$, CFI = .90, RMSEA = .249, SRMR = .031).[7] The modification indices for this model did not suggest any plausible adjustments to the model specification. We re-estimated the model with separate motor development and communication development constructs as outcomes to see if fit would be improved by allowing for differential effects of temperament on motor development and communication development. Doing so significantly improved model fit ($\Delta\chi^2 = 58.45$, $\Delta df = 3$, $p < .001$, Change in Akaike Information Criterion [ΔAIC] = 52.45, Change in Bayesian Information Criterion [ΔBIC] = 41.97), but the pattern of effects was identical for the separate motor development and communication development outcomes and their magnitudes were also very similar. Furthermore, the domain-specific patterns and effect sizes mirrored those of the univariate outcome model. In other words, the more complex model—although better fitting—did not provide any novel insight into the process under study. We, therefore, chose to proceed with inference based on the more parsimonious model. Estimates from the domain-specific model are available as online supplementary material.

The middle panel of Table 1 contains the estimated structural parameters from this model. As demonstrated by the standardized coefficients, age had a moderately large, positive effect on effortful control, a small, negative effect on surgency, and a large, positive effect on negative affectivity. Effortful control and negative affectivity had small, positive effects on overall development, but surgency was only trivially associated with development. Furthermore, after controlling for the three temperament dimensions, age maintained a very strong positive association with overall development. The 95% bias corrected CIs suggest nonzero indirect effects through effortful control and negative affectivity but not through surgency. The fully standardized versions of these indirect effects suggest that, although significantly different from zero, they were small in magnitude.

Bayesian Modeling

To fit the mediated SEM as a Bayesian model, we only needed to specify prior distributions for the model parameters and fit the model using software capable of Bayesian estimation. We fit the Bayesian SEM model using the *RStan* package (Stan Development Team, 2017) which is an R wrapper for the *Stan* C++ library (Stan Development Team, 2016). The likelihood for the Bayesian SEM (i.e., the structural associations among variables as represented in Figure 1) was defined equivalently to the frequentist version described above except that we incorporated mean structures into the Bayesian SEM.[8] The scale of the latent outcome was set by fixing its mean to 0.0 and its variance to 1.0. All intercept parameters were given Normal(0.0, 10.0) priors, factor loadings

were given Half-Normal(0.0, 5.0) priors, unique factor variances were given Half-Cauchy(0.0, 5.0) priors, latent regression coefficients were given Normal(0.0, 5.0) priors, and the residual covariance matrix of the mediators was left with *Stan*'s default prior (i.e., each element was assigned an improper uniform prior over its legal range). We chose the weakly-informative priors to assign minimal prior probability to parameters with large magnitudes since age only ranged from −2.85 to 5.11 and all other variables were standardized. The prior for the factor loadings was centered at 0.5 because loadings in a one-factor SEM with unit latent variance should take values in [0.0, 1.0]. As recommended by Depaoli and Van De Schoot (2015), we visualized the prior distributions before estimating the model to check that they encoded reasonable parameter ranges. These visualizations are available as online supplementary material.

We estimated the model using two parallel Markov chains that each discarded 10,000 burn-in iterations and retained 10,000 post-burn-in samples. Convergence of the Markov chains was assessed by checking that the potential scale reduction factor ($\hat{R}$; Gelman & Rubin, 1992) was less than 1.1 for all stochastic parameters and using the *coda* package (Plummer, Best, Cowles, & Vines, 2006) to generate traceplots of all stochastic parameters. Both criteria suggested convergence. The maximum $\hat{R} \approx 1.005$, and the traceplots demonstrated good mixing of the post-burn-in chains. The traceplots are available as online supplementary material. To check for local convergence, we re-estimated the model with 20,000 burn-in iterations and 20,000 post-burn-in samples retained. The $\hat{R}$ values and traceplots indicated that this longer run also converged—suggesting that our initial samples did not suffer from local convergence.

We also conducted a sensitivity analysis to assess the impact of the informative priors. We reran the model with two alternative prior specifications. One using *Stan*'s default priors for all model parameters (i.e., improper uniform priors over the parameter's declared ranges), and one that doubled the size of the prior *SD*s we initially employed (i.e., $SD = 5 \rightarrow SD = 10$, $SD = 10 \rightarrow SD = 20$). Neither of these changes dramatically altered the size of the focal effects' posterior modes or posterior *SD*s (i.e., the most discrepant mode was 5.36% smaller and the most discrepant *SD* was 5.21% smaller). This sensitivity analysis is fully documented in online supplementary material. Finally, we plotted the histograms of the indirect effects and latent regression coefficients to ensure that the number of posterior samples (i.e., 20,000) was sufficient to smoothly approximate the posterior distribution. These histograms (which are available as online supplementary material) all suggest that 20,000 samples were sufficient.

The third panel in Table 1 shows the posterior modes, as well as the 95% and 80% highest posterior density credible intervals for the focal effects. As noted above, the Bayesian estimate is the full posterior model, so the summary statistics reported in Table 1 are only meant to convey information about the shape of the posterior distribution. Following the recommendations of McElreath (2016) we report two sets of credible intervals in Table 1 to provide a more detailed picture of the posterior. The Bayesian estimates are all very similar to the frequentist estimates because the priors we used were only weakly informative. The interpretation of the Bayesian estimates is much more intuitive than their frequentist counterparts, however. For example, the posterior distribution of the total indirect effect is centered at 0.152 and there is a 95% chance that the total indirect effect is between 0.095 and 0.218 and an 80% chance that it is between 0.116 and 0.195. Therefore, we can be quite certain that the total indirect effect of temperament is non-zero but small. The posterior distributions of the three specific indirect effects, the total indirect effect, the direct effect, and the total effect are visualized in Figure 2. By computing the proportion of posterior

samples that exceed a given threshold, we can make direct inferences about the posterior probability of certain effect sizes. For example, applying this procedure to the standardized direct effect suggests a 96.6% chance that $c'_{std} \geq 0.7$, a 66.4% chance that $c'_{std} \geq 0.75$, and a 16.2% chance that $c'_{std} \geq 0.8$. Therefore, we can be very certain that the residual effect of age on development, after controlling for the total indirect effect of the three temperament domains, is strongly positive.

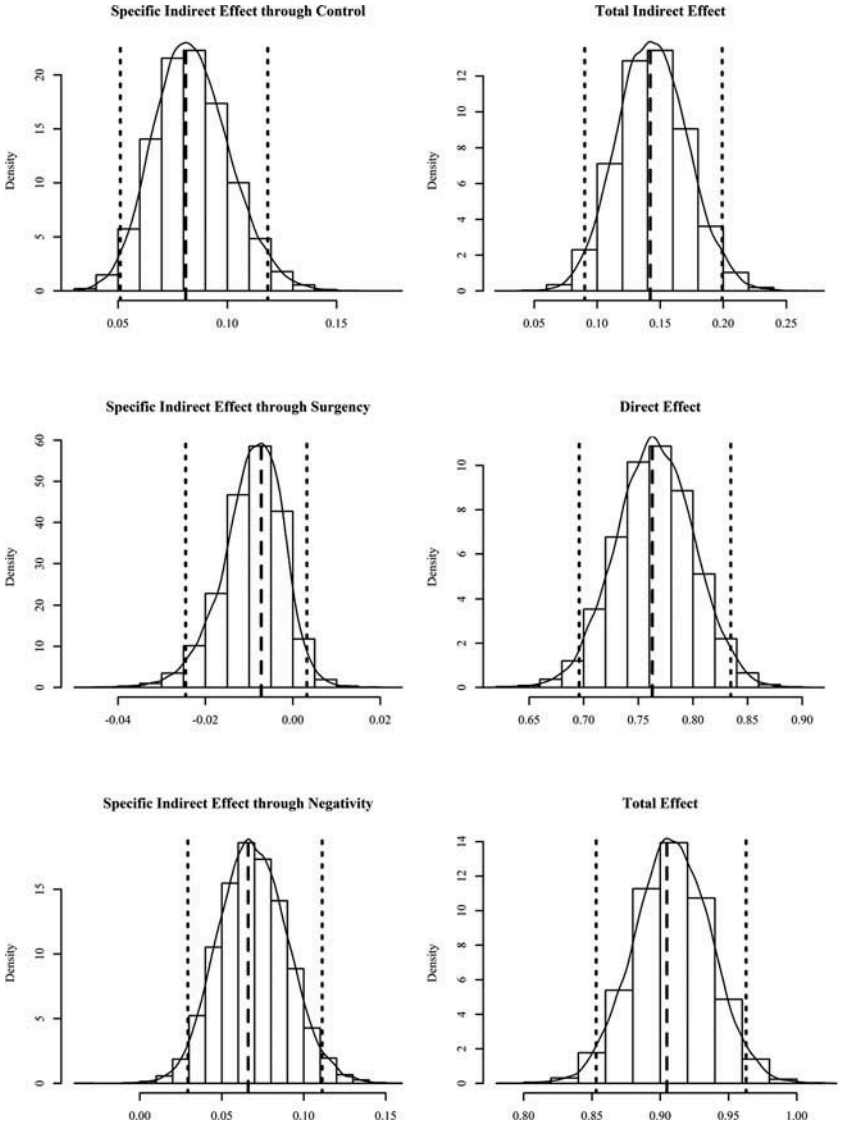

FIGURE 2 Posterior distributions of specific indirect, total indirect, direct, and total effects. Dashed lines represent the posterior modes, dotted lines represent the 95% Highest Posterior Density Credible Intervals

CONCLUSION

With the goal of meaningfully informing practice, we propose statistical modeling as a superior inferential framework that developmental researchers can employ to strengthen the rigor of their science. Statistical models are modular combinations of probability distributions that are sufficiently flexible to allow for easy "mixing and matching" of component submodels to produce complex, theoretically-relevant inferential tools. The example analyses we report above demonstrated these strengths of the statistical modeling paradigm using real data.

In closing, we reiterate that statistical modeling and NHST need not be incompatible. Many of the issues that arise from NHST are not inherent limitations of the NHST framework but are contaminants introduced when researchers do not take the time to carefully consider their hypotheses, their data, and the underlying data generating process. Furthermore, we note that modeling is not the only way to improve on NHST. Williams, Bååth, and Philipp (this issue) for example, describes how Bayes factors can replace NHST-based tests of point hypotheses. Researchers who prefer the familiarity of NHST can rest assured that they will still have the option of using NHST when comparing models in the statistical modeling framework. Yet, in rethinking their approach to statistical inference, applied researchers can be actively working to address the methodological crisis lamented by Nuzzo (2014) and Krueger (2014)

NOTES

1. Even when using so-called uninformative priors (i.e., those that encode no prior knowledge), the Bayesian approach incorporates this explicit acknowledgment of ignorance as the substantive claim that any legal values of the parameters are equally likely—which is a nontrivial statement (Gelman, Carlin, Stern, & Rubin, 2014).
2. Although the term *mediation* implies causation, the example data we analyze are not experimental or longitudinal, so we have no basis for making causal statements. We will, however, use the language of mediation to improve readability, with the understanding that we are not inferring causation.
3. We thank Marc Bornstein, Diane Putnick, and Melissa N. Richards for providing the data used in these examples.
4. 95% bootstrap confidence intervals were computed using the *psych* package (Revelle, 2017).
5. Although the child temperament domains were derived from multiply indicated scales, merging the ECBQ and CBQ was only possible at the domain level. So, we collapsed the temperament indicators into three mean scores when combining the $\leq$ 36 months and > 36 months samples.
6. In a single mediator model this step is equivalent to testing the significance of the indirect effect. In multiple mediator models, this step is equivalent to testing the significance of the total indirect effect (Hayes, 2013).
7. The especially large value for the RMSEA, relative to the other fit indices, is most likely due to the small model (i.e., one with few degrees of freedom between which to divide the χ^2).
8. Including the mean structures improved convergence of *Stan*'s sampling algorithm, for this problem.

ACKNOWLEDGMENTS

The authors wish to thank Drs. Diane Putnick, Marc Bornstein, and Melissa N. Richards for sharing the data used in the example analyses.

SUPPLEMENTAL DATA

Supplemental data for this article can be access on the publisher's website.

ORCID

Kyle M. Lang http://orcid.org/0000-0001-5340-7849
Shauna J. Sweet http://orcid.org/0000-0003-2458-5528
Elizabeth M. Grandfield http://orcid.org/0000-0002-3188-4086

REFERENCES

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Cohen, B. H. (2017). Why the resistance to statistical innovations? A comment on sharpe (2013). *Psychological Methods*, *22*, 204–210. doi:10.1037/met0000058

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, *57*, 505–528. doi:10.1146/annurev.psych.57.102904.190146

Depaoli, S., & Van De Schoot, R. (2015). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*. Advance online publication. doi:10.1037/met0000065

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. Psychological Methods, 22(2),240–261. doi: 10.1037/met0000065

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, *9*(2), 139–158.

Eidelson, R. J. (1997). Complex adaptive systems in the behavioral and social sciences. *Review of General Psychology*, *1*(1), 42–71.

Fisher, R. A. (1932). Inverse probability and the use of likelihood. *Proceedings of the Cambridge Philosophical Society*, *28*, 257–261. doi:10.1017/S0305004100010094

Fisher, R. A. (1954). *Statistical methods for research workers* (*12th* ed.). Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B*, *17*, 69–78.

Foster, E. M., & Kalil, A. (2005). Developmental psychology and public policy: Progress and prospects. Developmental Psychology, 41(6),827–832. doi: 10.1037/0012-1649.41.6.827

Fisher, R. A. (1960). *The design of experiments* (*7th* ed.). Edinburgh, Scotland: Oliver & Boyd.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York: Springer.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Jose, P. E. (2013). *Doing statistical mediation and moderation*. New York, NY: Guilford Press.

Kim, S. (2015). *Ppcor: Partial and semi-partial (part) correlation (R package version 1.1)*. Retrieved from https://CRAN.R-project.org/package=ppcor

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York, NY: Guilford publications.

Krueger, J. I. (2014, February 17). *The quest for replicable results in psychology* [web log post]. Retrieved from https://www.psychologytoday.com/blog/one-among-many/201402/the-quest-replicable-results-in-psychology

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in structural equation modeling. *Multivariate Behavioral Research*, *25*, 313–334.

Levine, T. R., Weber, R., Hullett, C. R., Park, H. S., & Lindsey, L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, *34*, 171–187.

Lindquist, E. F. (1940). *Statistical analysis in educational research*. Boston, MA: Houghton Mifflin.

Little, T. D., Widaman, K. F., Levy, R., Rodgers, J. L., & Hancock, G. R. (in press). Error, Error in my Model, Who's the Fairest of Them All? Research in Human Development.

Little, T. D., Gorrall, B. K., Panko, P., & Curtis, J. D. (in press). Modern Practices to Improve Human Developmental Research. Research in Human Development.

Little, T. D. (2015). Methodological practice as matters of justice, justification, and the pursuit of verisimilitude. *Research in Human Development*, *12*, 268–273.

Lugo-Gil, J., Lee, A., Vohra, D., Adamek, K., Lacoe, J., & Goesling, B., & U.S. Department of Health and Human Services. (2016). *Updated findings from the HHS teen pregnancy prevention evidence review: July 2014 through August 2015* (Research Report). Retrieved from https://tppevidencereview.aspe.hhs.gov/pdfs/Summary_of_findings_2015.pdf

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*, 185–199.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.

Neyman, J. (1935). On the problem of confidence intervals. *Annals of Mathematical Statistics*, *6*, 111–116. doi:10.1214/aoms/1177732585

Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society*, *105*, 292–327. doi:10.2307/2980436

Neyman, J. (1955). The problem of inductive inference. *Communications in Pure & Applied Mathematics*, *3*, 13–46. doi:10.1002/cpa.3160080103

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Particle I Biometrika*, *20A*, 175–240. doi:10.2307/2331945

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, *231*, 289–337. doi:10.1098/rsta.1933.0009

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*, 150–152.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, England: Cambridge University Press.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*, 7–11.

Putnam, S. P., Jacobs, J., Gartstein, M. A., & Rothbart, M. K. (2010, March). *Development and assessment of short and very short forms of the early childhood behavior questionnaire*. Poster presented at International Conference on Infant Studies, Baltimore, MD.

Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the children's behavior questionnaire. *Journal of Personality Assessment*, *87*(1), 103–113.

R Core Team. (2017). *R: A language and environment for statistical computing*., Vienna, Austria: R Foundation for Statistical Computing.Retrieved from https://www.R-project.org

Revelle, W. (2017) *Psych: Procedures for personality and psychological research* (R package version 1.7.5). Retrieved from https://CRAN.R-project.org/package=psych Version = 1.7.5

Rodgers, J. L. (2010a). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*(1), 1–12. doi:10.1037/a0018326

Rodgers, J. L. (2010b). Statistical and mathematical modeling versus NHST? There's no competition! *Journal of Modern Applied Statistical Methods*, *9*(2), 340–347.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology, 13, 290–312. doi: 10.2307/270723

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (2005). *Vineland-II adaptive behavior scales*. Circle Pines, MN: American Guidance Service Publishing.

Stan Development Team. (2016). *The stan C++ library* (version 2.15). Retreived from http://mc-stan.org/

Stan Development Team. (2017). *RStan: The R interface to stan (*R package version 2.15.1). Retrieved from https://CRAN.R-project.org/package=rstan

Stelzl, I. (1986). Changing the causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, *21*, 309–331.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician* 70, 129–133. doi:10.1080/00031305.2016.1154108

Williams, M. N., Bååth, R. A., & Philipp, M. C. (in press). Using Bayes factors to test hypotheses in developmental research. Research in Human Development.

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., & van de Schoot, R. (in press). Where do priors come from? A Bayesian evaluation of a latent growth model with informative priors. Research in Human Development.