Fitting Problems: Evaluating Model Fit in Behavior Genetic Models

S. Mason Garrison (garrissm@wfu.edu)
Wake Forest University https://orcid.org/0000-0002-4804-6003
Joseph Lee Rodgers
Vanderbilt University

Research Article

Keywords: Methods, Structural Equation Modeling, Best Practices, Model Selection, Model Fit

Posted Date: May 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-496590/v1

Abstract

In behavior genetics, like many fields, researchers must decide whether their models adequately explain their data – whether their models "fit" at some satisfactory level. Well-fitting models are compelling, whereas poorly-fitting models are not (Rodgers & Rowe, 2002). Oftentimes, researchers evaluate model fit by employing "universal" rules of thumb (e.g., Hu and Bentler, 1999). However, these rules are not universal, and should be treated as model specific (Kang et al., 2016). Accordingly, we focused on developing fit criteria emulating Hu and Bentler (1999) for classic univariate models (ACE; CE; AE) by fitting simulated twin data with correctly- and incorrectly-specified models. Ideal criteria should consistently accept correct models and reject incorrect models. Classic ACE models were indistinguishable and virtually all fit indices were non-informative because (or especially when) they were obtained in saturated models. For non-ACE models, criteria were informative. Nevertheless, every fit metric employed, except TLI, differed markedly across models and/or conditions. Universal solutions remain elusive, but promising and valid approaches include nested model comparisons, increasing degrees of freedom, and ruthless skepticism.

Introduction

Behavior genetics is not immune to the "replication crisis" (c.f. Lee & McGue, 2016; Turkheimer, 2016). Although general behavior genetics findings replicate (Plomin et al., 2016), many specific and once-classic findings have failed to do so, such as candidate genes for intelligence (Chabris et al., 2012; Sniekers et al., 2017) and stress-by-gene interactions (Risch et al., 2009). The specific reasons for those failed replications are beyond the scope of this paper, but their general causes are not. Many of the failed replications addressed methodological weaknesses from the original paper, often by following best practices. Accordingly, we should continue to implement methodological best practices in all behavior genetics studies, rather than just a select and anomalous few.

In behavior genetics, like other areas of psychology, researchers decide whether their models adequately explain their data – whether their models "fit" at some satisfactory level. Well-fitting models are compelling, poorly-fitting models are not (Rodgers & Rowe, 2002; cf. Roberts & Pashler, 2000). Often, researchers evaluate model fit by employing "universal" rules of thumb, which provide fit-index-specific criteria to declare a model well-fitting. Unsurprisingly, papers believed to provide such criteria are among the most highly-cited academic articles. For example, Hu and Bentler (1999) has been cited over 56,000 times. Although popular and listed in many guidelines (e.g., Appelbaum et al., 2018; Cumming, 2014; Hancock & Mueller, 2010; Mueller & Hancock, 2010), these rules of thumb should not be treated as universal.

Rather, these model-fit criteria should be viewed as model-specific, and influenced by the measurement model (Hancock & Mueller, 2011; Kang et al., 2016). Cutoff criteria in Hu and Bentler (1999) cannot be generalized beyond the single-factor model. For example, if these criteria were applied to five-factor data, the best fitting model would have four or fewer factors (Garrison, 2017). Moreover, many works, including

Savalei (2012), Millsap (2012), and Marsh, Hau, and Wen (2004) highlight that incorrect models can appear to be well-fitting when universal criteria are applied. Accordingly, cutoff criteria need to be generated for the specific model of interest. Otherwise, researchers risk rejecting well-fitting models or supporting poor-fitting models, and committing errors of the first and second kind, respectively.

This paper aims to address two major questions:

1) How are behavior geneticists evaluating model fit, and are those methods conforming to best practices?

2) What criteria are effective to evaluate model fit for behavior genetic models?

In the next section, we provide an overview of behavior genetic models, focusing on

matrix specification and model identification. Then, we review evaluating model fit for SEM, where we attend to best practices and evaluation of just-identified models. Following, we present a study of current practices in behavior genetic model fitting, followed by a simulation study. We conclude with summary recommendations.

Behavior Genetic Modeling

The aim of twin research and, more broadly, behavior genetics research is to “distinguish[…] between the effects of tendencies received at birth, and of those that were imposed by the circumstances of their after lives; in other words, between the effects of nature and of nurture” (Galton, 1876, p. 392). Such designs have been used to understand the genetic and environmental influences on virtually every construct on which people vary (Polderman et al., 2015). These individual constructs can range from the benign, such as height (Holzinger, 1929; Rodgers et al., 2019; Silventoinen et al., 2003) to potentially controversial, including intelligence, educational attainment, and economic success (Burks, 1938; Burt, 1966; Hernstein & Murray, 1994; Jensen, 1969; Murray, 1998; Thorndike, 1905). Researchers have used a variety of kinship groups in their biometrical comparisons, ranging from twins (Galton, 1876), different types of siblings (Thorndike, 1905), cousins (Fisher, 1919), and adoptees (Snygg, 1938); this review will focus on twin studies because they are the most popular within this discipline. Indeed, the “classic” behavior genetic study is a twin design (Liew et al., 2005; Rende et al., 1990), comparing correlations (or covariances) among monozygotic twins to correlations (or covariances) among dizygotic twins.

The “classic” way to estimate heritability from twin designs has developed nearly in lockstep with methodological advancements (Jinks & Fulker, 1970; Rende et al., 1990). Before the development of the correlation (Galton, 1896; Pearson, 1909; Stigler, 1989) and the wider acceptance of twin types (Fisher, 1919; Newman, 1917; Siemens, 1924; Smith, 1856), the method was qualitative (Galton, 1876), with nebulous categories of nature and nurture. After those developments, the method became algebraic and model-based (Falconer, 1952; Holzinger, 1929), with important underlying assumptions and less nebulous categories of genetic and environmental influences. Other developments led to incremental modeling improvements (see Jinks & Fulker, 1970 for added treatment on this topic).

Incremental algebraic gains were upended with the insight that confirmatory factor analytic models could be repurposed for twin designs (Eaves & Gale, 1974; Loehlin & Vandenberg, 1968; Martin & Eaves, 1977). This insight led behavior geneticists to embrace factor analytic models and more general structural equation models (SEM). Structural equation modeling has facilitated great advancements in behavior genetic research, ones that were impossible under the assumptions of older designs. Indeed, some assumptions, such as no gene-by-environment correlations, have been transformed into their own research areas.

Nevertheless, classic twin designs have pushed the limits of structural equation modeling, in terms of model identification. In general, a covariance-based model is identified when there exists a unique set of parameters that define the maximum of the fit function (Bekker & Wansbeek, 2003). In less formal terms, a model is identified when free parameters are unique, and cannot trade off with one another to produce the same covariance. We direct readers to Hunter et al. (2021) for analytic solutions and further discussion of this issue. The classic ACE twin design produces a just-identified model, whereas the fuller ACDE model is under-identified (Hunter et al., 2021; Jinks & Fulker, 1970; Neale & Maes, 2004). In the fuller ACDE model, the covariance structure for monozygotic twins, reared-together is

$$Cov_{MZt} = \begin{bmatrix} a^2 + d^2 + c^2 + e^2 & a^2 + d^2 + c^2 \\ a^2 + d^2 + c^2 & a^2 + d^2 + c^2 + e^2 \end{bmatrix}$$

where a^2 is additive genetic variance; d^2 is dominant genetic variance; c^2 is shared environmental variance; and e^2 is non-shared environmental variance. For dizygotic twins, reared-together, the covariance structure is

$$Cov_{DZt} = \begin{bmatrix} a^2 + d^2 + c^2 + e^2 & \frac{1}{2}a^2 + \frac{1}{4}d^2 + c^2 \\ \frac{1}{2}a^2 + \frac{1}{4}d^2 + c^2 & a^2 + d^2 + c^2 + e^2 \end{bmatrix}$$

In this model, there are more unknowns (4; a, d, c, and e) than knowns (3; Cov MZ, Cov DZ, and phenotypic variance Vp), resulting in an under-identified model. Without more information (i.e., data point), these unknowns cannot be uniquely calculated. To address this problem, researchers must either add information or reduce the number of parameters. Additional information can be achieved by adding another group (such as full siblings, or twins reared-apart) or using individual-level data. In the case in the classic ACE model, the number of parameters is reduced by removing dominance from the model. Then, the covariance structure for monozygotic twins is

$$Cov_{MZt} = \begin{bmatrix} a^2 + c^2 + e^2 & a^2 + c^2 \\ a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix},$$

whereas the covariance structure for dizygotic twins is

$$Cov_{DZt} = \begin{bmatrix} a^2 + c^2 + e^2 & \frac{1}{2}a^2 + c^2 \\ \frac{1}{2}a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix}.$$

Because we have assumed a lack of dominance in this model, the model now contains three unknowns and three knowns., and we have a just-identified model (Hunter et al., 2020). Although just-identified models can be used to estimate parameters, they cannot be used to evaluate model fit in the traditional sense. Traditional model fit relies on comparing the model of interest to a null model, in the same manner that one would evaluate any null hypothesis. The default null model is often a saturated model, where as many parameters as possible can be defined from the data, and then reproduce perfectly the means and covariances (Widaman & Thompson, 2003). In other words, the default model is a just-identified (saturated) model with a χ^2 value of zero and zero residual degrees of freedom. Accordingly, if one were to report a χ^2 test comparing the ACE model to a saturated model, the test would indicate that the model fit perfectly. However, this perfect fit is a consequence of the classic ACE model being just-identified. Regardless, nested model comparisons are possible for just-identified models, including classic ACE models. For example, one could compare statistically a nested AE or CE model (both of which are over-identified) to the full ACE model.

The classic covariance-based model (or other similar “two-kinship-category” models) is just-identified, but other models, such as twins-raised-apart-and-together or children-of-twins, are over-identified. In those cases, classic methods of evaluating fit ought to apply. Moreover, one can include additional information, such as group means or individual-level data, in addition to covariances. That additional information increases total degrees of freedom and results in an over-identified model, with positive residual degrees of freedom. In these non-classic cases, traditional methods of evaluating model fit can and should be used. Many behavior genetic models are specific applications of structural equation models, and can be treated as such.

Model Fit for SEM

Best Practices for Evaluating Model Fit

Since the advent of the replication crisis (Open Science Collaboration, 2015; but see, Shrout & Rodgers, 2017; or Ioannidis, 2005), much of the behavioral and social sciences has become interested in improving research practices. Beyond an increased interest in replication, the replication crisis has encouraged social and behavioral scientists to follow best practices for their statistical analyses. These guidelines tend to be broad and flexible, aiming to give scholars enough information to evaluate the model being tested without excessively burdening the authors. Further, this flexibility recognizes that methodological advances, and guidelines for evaluating fit may change in accordance with those advances.

Within the context of structural equation modeling, the guidelines for best practice are broad (Appelbaum et al., 2018; Lance et al., 2016; Mueller & Hancock, 2010). For evaluating model fit, best practice recommends that multiple indices across classes be used in conjunction with the χ^2 statistic. The inclusion of the χ^2 statistic allows the reader to calculate most other fit indices – or at least those that are log-likelihood- or χ^2-based.

However, these guidelines can be criticized. For example, Marsh, Hau, and Wen (2004) noted that best practices and recommendations of "golden rules" tended to deemphasize limitations of each metric and omit caveats from the original papers. Every fit index that we discuss in the next section has limitations. Some of those are obvious (such as χ^2 being sensitive to sample size; Gerbing & Anderson, 1985), whereas others are less obvious (such as RMSEA being insensitive to omitted cross-loadings; Savalei, 2012).

Specific Methods for Evaluating Model Fit

Because numerous methods exist to evaluate model fit for SEM, we will focus on those recommended in current best practices (Appelbaum et al., 2018; Lance et al., 2016; Mueller & Hancock, 2010). For a classic (and more thorough) overview, see Bollen and Long (1993).

The two primary ways to evaluate model fit are through model comparison and goodness-of-fit indices. The classic model comparison is the χ^2 test (a.k.a. the likelihood ratio test), where the proposed model is compared to the saturated model. Significant values indicate less than perfect fit. Some methodologists consider the χ^2 test is "overly strict" (Mueller & Hancock, 2010, p. 395) , "almost always statistically significant" (Kenny, n.d.), and overpowered (Bentler, 1990; Hu et al., 1992). Recent work by McNeish (2018) has indeed found that the χ^2 test's Type I error rate is inflated at smaller samples.[1] Nested models can also be compared using the χ^2 test. However, nested comparisons using the likelihood ratio test suffer inflated error rates for certain applications, such as classically-specified ACE models (Carey, 2005; Verhulst et al., 2019).[2] Although best practice remains to report the χ^2 test, many goodness-of-fit indices were developed to address the limitations of the χ^2 test (Mulaik et al., 1989).

These fit indices can be broadly classified into three classes: absolute, parsimonious, and incremental. The three broad classes prioritize different mixtures of fit with parsimony. Absolute fit indices describe the overall discrepancy between observed and model implied covariance values. The addition of more parameters improves fit without penalty. Examples include: SRMR (Standardized Root Mean Square Residual), where values below .08 are considered good fit (Hu & Bentler, 1999).

Although parsimonious fit indices also describe the overall discrepancy between observed and model implied covariance values, these indices penalize added model complexity. Model fit should improve with the addition of parameters only if the benefits outweigh additional model complexity. Examples include RMSEA (Root Mean Square Error of Approximation), where the lower bound of the 90% confidence interval falls below .05 (Hu & Bentler, 1999). Incremental (sometimes called relative) indices compare fits to a baseline model, typically the null model. Examples include TLI (Tucker-Lewis Index), and CFI

(Comparative Fit Index), where values above .95 are considered good fit (Hu & Bentler, 1999; Marsh, 1995).

Evaluating Just-identified Models

Many classic methods for evaluating model fit cannot be used on just-identified models, because the null comparison model is saturated and just-identified (Widaman & Thompson, 2003). However, some methods can be applied. We have already discussed nested model comparisons comparing the saturated to the proposed model. However, that approach can be used for selecting between nested models, including the classic ACE model and models with fewer components, such as the AE or CE model. In addition to nested model comparisons, information criteria can be used, including AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and aBIC (Adjusted BIC). These criteria are typically used to compare non-nested models, but they can be used in nested models, if they are fit using the same data.

Like the fit indices discussed earlier, these criteria aim to balance parsimony and fit, using the likelihood function rather than the χ^2. The idea is to minimize information lost between the data generation process and the model used to approximate it. Better fitting models lose less information. For example, AIC is a function of the number of parameters and maximum value of the likelihood function; models with smaller AIC indicate relatively better fit as they have lost less information using the same data. Rules of thumb have been developed for evaluating whether these model differences are meaningful, such as differences larger than 10 (Burnham and Anderson, 2002). However, such rules are problematic within SEM because of the large sampling variability for fit measures, including BIC (Preacher & Merkle, 2012). Nevertheless, non-nested models can be compared, even within the context of SEM (Merkle et al., 2016).

Study 1: Current State of the Field

How are behavior geneticists evaluating model fit?

We examined how behavior geneticists are evaluating their model fits. Work by Jackson (2009) on general reporting practices for CFA revealed that most papers followed contemporary guidelines. However, a priori, we suspected that this practice would not extend to the field of behavior genetics. In a recent review of the field's flagship journal, *Behavior Genetics*, the first author found that many articles were using out-of-date software (Garrison, 2018). In 2016 and 2017, 22% of SEM-based articles in *Behavior Genetics* used Mx. Mx was last released in 2009, (Version 1.7.03; Neale, 2009); it is no longer compatible with most operating systems; and its developers have moved onto openMx (Boker et al., 2011; Neale et al., 2016). Garrison observed that some articles omitted standard fit statistics (e.g., χ^2), even in cases when the model was over-identified. In addition, many of those papers focused on AIC and a nested-model comparison. This method mirrors the last release of a Mx-companion behavior genetics ebook (Neale & Maes, 2004). That mirroring is suggestive that these papers used that book as a guide, without considering whether their own models were just- or over-identified.[3] Accordingly, this first study aims to quantify those observations.

[1] Inflated Type I errors for ACE models might be the result of non-normal data, rather than an over-powerful test.

[2] The cause of these inflated rates again is likely more from mis-specified sampling distributions. However, the result is the same – an incorrectly performing test.

[3] Neale and Maes (2004) provides an excellent discussion on model identification for classic twin designs.

Methods

As described in Garrison (2018), the first author reviewed two years of publications from the flagship journal *Behavior Genetics* (2016 and 2017). That paper focused on determining which SEM programs were being employed. The number of articles identified is summarized in Table 1. There were identified 473 "publication units" (all publications) during those two years, using Publish or Perish (Harzing, 2018). Published abstracts from the annual conference were excluded (n=310), as well as articles indexed across multiple databases (n=46). Further, the remaining 117 articles were refined to human participants (n=88), by excluding editorials (n=4; e.g., Ayorech et al., 2016), simulations (e.g., Verhulst, 2017) and non-human subjects (n=20; e.g., whales, Whitehead et al., 2017). Lastly, non-SEM analyses (n=48) were eliminated, such as co-twin control studies (e.g., Mosing et al., 2016) and molecular genetic studies (e.g., van den Berg et al., 2016). The remaining 40 articles employed structural equation modeling for quantitative genetic modeling on humans.

Results

Garrison (2018) focused on SEM software. In the current investigation, we used the same articles to identify which methods were being used to evaluate model fit in the flagship journal *Behavior Genetics*. The variability in practice of how much attention was paid to model fit was considerable, ranging from dedicated discussion subsections (e.g. Smolkina, Morley, Rijsdijk, et al, 2017) to the total absence of their usage (n=8; 20%). We have provided summary statistics in Table 2. Given the variety of methods used and the broad guideline for best practice, we have classified each paper on their model-fitting practice. If a paper used more than one method to evaluate fit, we classified that paper as *acceptable*. If the paper used multiple classes of methods to evaluate fit, such as fit indices (e.g., absolute, parsimonious, incremental), we classified that paper as meeting the standard for *better* practice. Lastly, if the paper used multiple classes in addition to reporting either χ^2 statistic or log likelihood, we classified that paper as meeting the standard for *best* practice. Accordingly, 5% met the standard for best practice (n=2), 50% (n=20[1]) for better practice, and 72.5% (n=29) for acceptable practice. Further, if nested model comparisons are included as a "class", these percentages increase to 42.5% (n = 17) for best, 72.5% (n=29) for better; and 75% (n=30) for acceptable.

For context, 80% (n=32) of articles reported a formal evaluation of model fitting. In those remaining 20% (n=8) of articles, models were selected for a variety of reasons, including “parsimony”, visible inspection of covariances, significance of A or C parameters, and findings from previous literature. In general, if an article evaluated fit, it typically seemed to draw inspiration from Neale and Maes (2004). Further, if an article evaluated fit, it used the criteria correctly. Therefore, rather than further focusing on evaluating current practice, Study 2 will focus on developing clear criteria.

Table 2 Number of Papers per Fit Method

Method of Fit	Number of Papers
AIC/BIC	19
Nested Model Comparison	28
RMSEA/SRMR	7
CFI/TLI	2
Loglikelihood/χ	26
Any	32
Total	40

[1] This count includes the two articles that met the best practice standard, as well as 18 addition articles that met the better practice standard.

Discussion

To summarize, we evaluated whether 2016-2017 *Behavior Genetic* papers met the standard for best practice (5%), better practice (50%), and acceptable practice (72.5%). Very few papers met the best standard (Appelbaum et al., 2018; Lance et al., 2016; Mueller & Hancock, 2010), which recommended that multiple indices across classes (parsimonious, relative, and absolute) be used in conjunction with the χ^2 statistic. We believe that these recommendations are both generous and reasonable across all SEM models. Unfortunately, most papers published in *Behavior Genetics* did not meet this standard. They often did not report a second index from another class, such as an absolute measure, RMSEA, or a relative measure, TLI. Instead papers were likely to include a nested model comparison instead of a second index. Accordingly, the “better practice” standard we developed allowed the nested model comparison to count as a class for evaluating fit. Even then, only half of papers met this standard. Frankly, the greatest concern is that 20% of papers did not evaluate fit at all. In those cases, it would be impossible to create a metric to include them as meeting some or any standard of practice.

Nevertheless, and a positive finding, if an article evaluated fit, that article typically used the criteria correctly. Our interpretation of these findings is that most researchers recognize that evaluating model fit

is part of the research process. Instead, rather than criticizing current practice, or focusing on the current weaknesses in the field, Study 2 will focus on developing clear criteria that can be used to successfully evaluate model fit. Ideally, those criteria will be simple, universal, and meet the standards for best practice.

Study 2: Evaluating Model Fit

Study 2 builds upon Study 1, by determining whether clear criteria for evaluating model fit can be developed, and whether those differ from "classic" SEM criteria. Specifically, we focused on developing custom fit criteria, in the style of Hu and Bentler (1999) for classic univariate ACE models, by fitting models to data generated from both correctly and incorrectly specified models, using the classic twin design comparing MZ and DZ twin correlations/covariances.

Evaluation Plan

Methods of evaluating model fit included (when possible), RMSEA, SRMR, AIC, BIC, aBIC, χ^2, nested model comparison, CFI, TLI, as well as an additional method identified in Study 1 – the model was kept if it had significant A or C parameters. Our evaluation plan emulated Hu and Bentler's (1999) approach by examining the Type I and Type II error rates for correctly and incorrectly specified models. Unacceptable rates were those that deviated substantially from 5% Type I error rates and 20% Type II error rates. The ideal method will be the one that can achieve both values, across all factors within the simulation designs. It may be unlikely that any measure/combination of measures will meet all these ideal values.

Methods

All simulations were conducted in R version 3.6.1 (R Core Team, 2020) and Mplus version 7.3 (Muthén & Muthén, 2017), using the R packages: MASS 7.3 (Venables and Ripley 2002), MplusAutomation (Hallquist & Wiley, 2018), and discord (Garrison et al., 2020).

Data Generation

We examined whether the classic cutoff values for model fit criteria are effective for classifying models correctly for univariate models. In other words, are misspecified models rejected, and are correctly specified models accepted? In the spirit of Hu and Bentler (1999), we fit a correctly specified model as well as many incorrectly specified models. We have produced a summary table, Table 3, of the conditions outlined below; we use 5,000 replications.

Table 3 Simulation Conditions

	Condition	Values
Genetic Variance	Additive	0, 1, 2
Environmental Variance	Shared	0, 1, 2
	Non-Shared	1, 2
Design	Sample Size (total pairs)	200, 500, 1000, 2000
	Data	Raw, Covariance

Variance Components. We have included one genetic component (additive) and two environmental components (shared- and non-shared). Each component condition, except for non-shared, can be 0, 1, or 2. We excluded the 0 non-shared variance condition because practically every measure has non-shared environmental variance (Polderman et al., 2015; Turkheimer & Waldron, 2000). These variance components can be converted into classic heritability estimates, by taking the proportion of variance components. For example, if the variance components were 1 additive genetic and 1 non-shared environmental, then that would result in $a^2 = .5$, $c^2 = 0$, and $e^2 = .5$. Each variance component is generated separately and then combined to create a total score, using the discordsim function from the discord package (Garrison et al., 2020) in R. This function generates biometrically-informed data for up to two measures for each member of a kin pair, using multivariate normal functions for each variance component. These kin pairs can vary in their genetic and environmental similarity. Further, the two measures can have overlapping biometrical influences, that are either bidirectional (e.g., correlated factors model, direct symmetric model) or unidirectional (Cholesky model).

Sample Size. We have varied sample size ranging from 200 total pairs to 2000 total pairs (N=200, 500, 1000, 2000). These sample sizes are based on typical sample sizes from Polderman's et al. (2015)'s meta-analysis. Pairs will be balanced to 50% Monozygotic Twins and 50% Dizygotic twins, as is typical for classic studies (Polderman et al., 2015).

Data Input Method.[1] The generated data were used directly (i.e., at the individual level) or indirectly, with estimated covariance matrices from those same generated data.Each dataset was fit using the following models: ACE, AE, and CE, using the raw data and the covariance matrix. In each case, the data were fit to two incorrect models, and the correctly specified model. For example, if the correctly specified model is an AE model, then the incorrectly specified models fit would be ACE and CE. This simulation is a 3 x 3 x 2 x 4 x 2 (Genetic variance x Shared environmental variance x Non-shared environmental variance x sample size x covariance versus raw data; see Table 3) design, with 144 total cells. Data were generated using R. Models were estimated with Mplus. Select conditions were used to validate the data generation process by fitting ACE models to data generated from an E model[2].

Rejection Criteria

Rejection criteria were initially planned to be guided wholly by observed methods. However, given the small sample of those who used any given fit index, we selected criteria based on a mixture of those observed practices in the survey along with common practice and popular articles, books, and websites (Hooper et al., 2008; Kenny, n.d.; Kline, 2015). Those criteria are documented in Table 4. Please note that aBIC, AIC, and BIC as well as the nested model comparison are nested, and in that regard are compared to the saturated model (i.e., ACE) fit to those same data.

Table 4 Model Rejection Criteria

Metric	Criteria	Nested
aBIC, AIC, BIC	Minimum value	Yes
ChiSqM Test	≤ .05	No
CFI, TLI	<.95	No
RMSEA Estimate	>.06	No
RMSEA 90% CI Lower Bound	>0	No
RMSEA 90% CI Upper Bound	>.06	No
p(RMSEA)<.05	<1	No
SRMR	>.08	No
Non-significant parameter	>.05	No
Nested Model Comparison	≤.05	Yes

[1] An additional series of models were fit using correlation matrices, instead of the covariance or raw data. However, Mplus does not allow for multi-group SEM models to be estimated with correlations. Preliminary attempts to work around this restriction by fitting the correlations as covariances with means of 0 led to inconclusive results.

[2] The E model describes any model where only non-shared environmental variance was generated and all other variance components (A & C) were 0.

Results

Given the number of conditions and methods of comparisons, we have focused on general trends and the three most promising methods for evaluating fit: TLI, χ^2 Test, and Nested Model Comparison. We direct readers to detailed tables in the Appendix and to the accompanying data files for complete results. Results did not differ meaningfully across data input condition (raw data or covariance-based summary statistics), which typically did not differ until the third decimal place. Accordingly, focus on the raw-data based method.

First, we conducted basic validity checks to determine whether the data generation and the model fitting process were performing correctly. Specifically, we examined whether correct models performed “better” than incorrect models in the expected direction for each fit statistic. We also examined how data generated from an E model – that is, a model with only non-shared environmental variance – would fit an ACE model and if the descriptive statistics from those models performed as expected.

Second, we present descriptive statistics for models by whether the correct or incorrect model was fit, and also by whether the fitted model had three variance components (i.e., an ACE model) versus a model with only two variance components (i.e., AE or CE). Third, we present Type I error rates and power for model selection, using conventional benchmarks. After each subsubsection, we provide brief discussions.

Validity Checks

In this subsubsection, we verify that the simulation worked as intended with the discord package. We examined for each fit statistic whether correct models performed “better” than incorrect models in the expected direction; and whether general summary statistics for ACE models fit to data generated from E models performed as expected. Tables 5- 12 contain descriptive statistics for each condition, and can be found in the appendix.

As expected, the following measures had lower median levels when the correct model was fit relative to incorrect models: aBIC, AIC, BIC, $p(\chi^2)$, χ^2, CFI, TLI, and p(RMSEA<.05). In contrast, the following measures had higher median levels when the correct model was fit relative to incorrect models: estimated RMSEA, RMSEA’s upper bound of the 90% CI, and SRMR. All indices performed as expected, indicating that correct models are at least better fitting than those that are not correctly specified.

Next, we examined descriptive statistics for data generated from an E-only model, but fit to an ACE model. As expected across conditions, median standardized (and unstandardized) parameter estimates for A and C were 0 across all conditions. Further, median standardized parameter estimates for E ranged from 0.995 to 0.999 across conditions. These p-values are more extreme than one would hope; however, they are consistent with recent work by Verhulst and colleagues (2019) and reflect the fact that estimated values for C and A are at the boundary of 0 and are restricted to be positive. In addition, distributions of $p(\chi^2)$ performed as expected; specifically, median values ranged from 0.331 to 0.358. Their 5th percentile ranged from 0.02 to 0.026 across conditions.

Summary Statistics

In this subsubsection, we present descriptive statistics for models by whether the correct or incorrect model was fit, for TLI, the significance level of χ^2 test of model fit, and significance level for nested model comparisons, comparing the ACE model to the selected model of interest. Table 13 in the appendix provides median values for standardized measures of fit with 1%, 2.5%, 5%, 10%, 90%, 95%, 97.5%, and 99% quantiles for correctly-specified models. These values compare favorably to the majority of classic benchmarks. For example, approximately 95% of all correct models reported $p(\chi^2)$ greater than .05.

Unfortunately, however, the distributions for these measures (as well as non-standardized methods of fit) overlapped considerably with the distributions from incorrectly specified models (see Table 14 Appendix).

Select Findings. Figures 1, 2, and 3 illustrate this general trend in overlapping distributions with box and whisker plots. For each figure, the whiskers span the 5% and 95% quantiles, and the Interquartile Range spans the 25% and 75% quantiles. As illustrated in Figure 1, median levels of TLI differed between correct models (1.001) and incorrect models (0.988). However, the 5% and 95% quantiles overlapped; for correct models, the quantiles ranged from 0.98 to 1.018, and for incorrect models from 0.913 to 1.015. Similarly, distributions overlapped for $p(\chi^2)$ where median levels were 0.495 for correct models and for incorrect models were 0.066 as illustrated in Figure 2. The quantiles ranged from 0.049 to 0.949 for correct models and from 0 to 0.845 for incorrect models. For nested model comparisons, the trend again was similar with differing median levels, but overlapping quantiles, for correct (1, 90% QI [0.2, 1]) and incorrect models (0.021 , 90% QI [0, 1]), as illustrated in Figure 3. It is worth noting that although the nested model comparison does overlap, a close inspection of figure three reveals that the large bulk of models are being distinguished correctly.

Next, we distinguished by the type of model fit. Specifically, we compared models with all three variance components (i.e., ACE) to models with only two of the three variance components (AE and CE models). Median levels generally differed, but correct and incorrect models continued to have overlapping distributions. As expected, incorrect ACE models tended to be indistinguishable from correct models. Indeed, practically all measures showed this consistent pattern where median levels were indistinguishable for ACE models (which are saturated models) but differed for AE and CE models (which are not saturated). For example, TLI levels were indistinguishable for ACE models (1, 90% QI [0.987, 1.011] for correct vs 1, 90% QI [0.806, 1.028] for incorrect; see Figure 4). However, median TLI values were lower when incorrect two component models (0.983) were fit compared to correct two component models which also had median TLI values of (1.001). A similar trend was observed for $p(\chi^2)$, where median levels of $p(\chi^2)$ for incorrect ACE models (0.417) were indistinguishable from correct ACE models (0.494). Similarly, median $p(\chi^2)$ values were lower when incorrect two-component (0.008) models were fit compared to correct two-component models which also had median $p(\chi^2)$ values of 0.495 (see Figure 5). Lastly, for nested comparisons, the trend again was similar with indistinguishable median values for ACE models (incorrect 1; correct 1), but lower values when two-component models were evaluated (incorrect 0; correct 1; see Figure 6). Nevertheless, even when distinguished by type, all measures overlapped in their distributions..

When distinguished by both type of model fit (ACE vs AE or CE) and sample size, distributions tended to narrow with larger sample sizes. The spread reduction was most drastic between 200 and 500, as one would expect. The same trends occurred as described previously. When fit to incorrect models, ACE models still resembled correct models, often with nearly indistinguishable distributions. As was the case without accounting for sample size, incorrect AE and CE models had distinguishable median levels when compared to incorrect ACE or correct models in general.

We also examined whether these summary statistics differed across levels of total variance[1]. Total variance ranged from 1 to 6. In general, more variance did not impact model fits when correct models were fit. However, a linear relationship was observed for most statistics when fit to incorrect models, either in terms of decreasing variability (e.g., TLI; Figure 7) or in declining median levels. For example, without accounting for type of model fit, significance levels for the χ^2 test of overall fit did not vary as a function of total variance when the correct model was fit (e.g., median levels of significance were 0.496 when total variance was 2; 0.499 when total variance was 6, etc; see Figure 8). However, they did vary for incorrect models. Increases in total variance were associated with decreases in p-value. Lastly, nested model comparisons behaved similarly (see Figure 9); however, this behavior seems to be more the product of total variance being partially confounded with type of model fit.

The same relationships with total variance seemed to hold even when accounting for type of model fit (ACE vs AE or CE), with one expected exception. The ACE models fit to incorrect models still tended to be indistinguishable from correct ACE models. we have included Figures 10, 11, and 12 to illustrate the trends for TLI, significance levels for the χ^2 test of overall fit, and nested model comparisons. For example, significance levels for the χ^2 test of overall fit did not vary as a function of total variance when the correct model was fit (regardless of kind) or when the incorrect ACE model was fit. Those values did level off when fit to incorrect AE or CE models, median significance leveled off quickly when fit to AE or CE, see Figure 11 (e.g., 0.444 when total variance was 1; 0.106 when total variance was 2; 0.001 when total variance was 3, etc).

Brief Discussion. The main finding from the descriptive statistics revealed that although indices for correctly-fit models tended to compare favorably to classic benchmarks, there was considerable overlap in distributions between correctly and incorrectly specified models. These distribution overlaps were present even when compared across type of model fit (ACE, CE, AE), sample size, and total variance. These overlaps were most problematic when comparing incorrect ACE models to correct models, where the distributions were practically indistinguishable. However, those indistinguishable distributions are likely the product of the ACE model being saturated.

Power and Type I Error

Next, we examined model rejection rates, using classic benchmarks and nested model comparisons. Type I errors were defined as rejecting a correctly specified model. Power was defined as rejecting an incorrectly specified model. Type II error rates were defined as β (1-power), or incorrectly specified models that were not rejected. These rates were calculated by condition. For the sake of brevity, we will again focus on the three most promising metrics: TLI, $p(\chi^2)$, and the nested model comparison. We direct readers to Table 15 in the appendix for these rejections as well as all the other tested metrics by condition.

Select Findings. For correctly-specified models, the median rejection rate across conditions for correct models (i.e., Type I errors) ranged from 0 (significance of model parameters, TLI) to 0.992 for p(RMSEA)

<1. The χ^2 test had median rejection rates near 0.05 (0.051; min 0.043, max 0.06). In contrast, TLI and nested model comparisons had rejection rates well below .05. For TLI, the median rejection rate was 0 (min 0, max 0.222), and for nested χ^2 model comparisons it was 0.009 (min 0, max 0.03). Figures 13 - 15 illustrate variability and differences in rejection rates with box and whisker plots. In each figure, power and Type I error rates are benchmarked with horizontal lines at 0.80 and 0.05.

For incorrectly specified models, the median rejection rate across conditions for incorrect models (i.e., power) ranged from 0 (significant model parameters) to 1 for p(RMSEA)<1. All three of the metrics on which we focused had median rejection rates well below .80: TLI (0.028; min 0, max 0.578; see Figure 13), $p(\chi^2)<.05$ (0.333; min 0.057, max 1; see Figure 14); and nested χ^2 model comparisons (0.631; min 0, max 1; see Figure 15).

Given that distributions of fit statistics seemed to differ as a function of sample size as well as total variance, we examined whether rejection rates were associated with these values. All rates for correctly specified models were negatively associated with sample size. Virtually all correlations were large (>.4). These correlations are provided in Table 16 and ranged from -0.972 for RMSEA <.06 to -0.024 for χ^2 nested model comparisons. Many rates for incorrectly specified models were positively associated with sample size, but this trend was not universal. $p(\chi^2)$ and nested model comparisons were positively associated with sample size, while TLI was negatively associated.

In addition, many rates were associated with total variance for both correctly and incorrectly specified models, but not as universally as sample size. $p(\chi^2)$ was positively associated with total variance for both model specifications. TLI was negatively associated for both. However, for nested model comparisons the association with total variance was positive for incorrectly specified models and negative for correctly specified models.

Like the descriptive statistics, rejection rates varied across type of model fit (ACE, CE, and AE) – primarily ACE compared to either CE or AE. Figures 16, 17 and 18 illustrate these differences using box and whisker plots. These plots show the distribution of rejection rates by condition. An ideal metric would have low variability across conditions and exceed the minimum standards of power>.80.

For correctly specified models, median rejection rates ranged across type of model fit. For correct ACE models, the median rejections rate ranged from 0 (TLI, nested χ^2 model comparisons) to 0.992 for p(RMSEA)<1. For our three metrics of focus: $p(\chi^2)<.05$ had median rejection rates around .05, specifically 0.051 (min 0.043, max 0.06). Our other two metrics had rejection rates well below .05: TLI (0; min 0, max 0.058), and nested χ^2 model comparisons (0; min 0, max 0). For correctly specified two-parameter (AE or CE) models, the median rejection rate ranged from 0 (significant model parameters) to 0.99 for p(RMSEA) <1. $p(\chi^2)<.05$ median rejection rates around .05: (0.051; min 0.046, max 0.06), while TLI's median rejection rate was well above .05: TLI (0; min 0, max 0.222). Lastly, nested χ^2 model comparisons had rejection rates well below .05 (0.025; min 0.018, max 0.03).

For incorrectly specified models, the median rejection rate ranged across conditions (i.e., power). Incorrect ACE models ranged from 0 (significant model parameters) to 1 for p(RMSEA)<1. All three metrics had median rejection rates below .80: TLI (0.005; min 0, max 0.286), nested χ^2 model comparisons (0; min 0, max 0), and p(χ^2)<.05 (0.07; min 0.059, max 0.112). For incorrectly specified AE or CE models, the rejection rates ranged from 0 (significant model parameters) to 1 for p(RMSEA>.05). For our three metrics, nested χ^2 model comparisons had median rejection rates of at least .80: (0.943; min 0, max 1). However, TLI and p(χ^2)<.05 rejection rates were below .80: TLI (0.045; min 0, max 0.578); and p(χ^2)<.05 (0.719; min 0.057, max 1) and had considerable variability.

[1] Total variance was not randomized and was not balanced across correct versus incorrect models as well as across the type of model fit (ACE, AE, CE). For example, no correct model had a total variance of 1.

Discussion

Brief Discussion. This stage examined power and Type I error rates for distinguishing between correctly specified and incorrectly specified models using classic criterion. These examinations revealed that many classic criteria had Type I error rates that deviated from .05, drastically (and usually conservatively). Moreover, within each metric, there was considerable variability in Type I error rates across conditions, even when grouped by type of model fit, sample size, and total variance. All metrics had inflated Type I error rates in at least one condition, except for nested χ^2 comparisons and RMSEA's Lower Bound > 0.

Similarly, the power to reject an incorrect model using classic criteria deviated considerably across conditions ranging from 0 to 1, even when grouped by type of model fit (ACE versus CE or AE), sample size, and total variance. The metrics had decreased power rates (<.8) in at least one condition, with two exceptions. Only p(RMSEA>.05) < 1 and Lower Bound of RMSEA > 0 were consistently and sufficiently powered.

Broad Discussion

Study 1 asked how behavior geneticists evaluate model fit, by examining two years of recent publications in the flagship journal *Behavior Genetics*. We evaluated whether these papers met the standard for best practice (5%), better practice (50%), and acceptable practice (72.5%). Very few papers met the best standard (Appelbaum et al., 2018; Lance et al., 2016; Mueller & Hancock, 2010), whereas 20% of papers did not evaluate fit at all.

Study 2 asked, what are effective criteria to evaluate model fit for behavior genetic models? It addressed this question by conducting a simulation study. Data were generated from a 3 x 3 x 2 x 4 x 2 design,

where variance components (A, C, and E), sample size, and estimation method varied. These data were then fit to both incorrectly and correctly specified ACE, AE, and CE models in a standard twin design. The analyses were broken into three stages. Stage 1 examined descriptive statistics for correct and incorrect models. Stage 2 examined model rejection rates, using classic benchmarks as well as nested model comparisons. Stage 3 identified alternative thresholds, by using three common methods, with the intent of developing a universal metric.

Stage-specific results are discussed at the end of each sub-subsection in the Results section. Broadly, the answer is no, there are not universal criteria that can consistently distinguish between data fit to correct versus incorrect models with anything close to good sensitivity and specificity. This status will be discussed separately for the ACE models, and the AE and CE models.

The fundamental problem seems to be that for data fit to ACE models, it is borderline impossible to distinguish correct models from incorrect models. In the twin design used here, the ACE model is saturated, leaving no degrees of freedom to test model fit, which creates the challenge identified in this study. Further, power and Type I error rates revealed that incorrect ACE models were difficult to distinguish from correct models of any type, using classic benchmarks. The ACE models were harder to correctly identify and thus, disproportionately contributed to the reduced power and inflated Type I error rates. Accordingly, these two stages indicated that criteria would likely have greatest difficulty with detecting model misspecification in saturated models, i.e., ACE models.

From the study's onset, it was clear that just-identified models like ACE models were going to be challenging for fit statistics to provide useful diagnostic information. Indeed, the fact that Neale and Maes (2004) dedicated an entire section to model identification and fit suggests that this challenge has been long apparent – and most likely previously discovered and just accepted as an inherent headache for modeling in the field of behavior genetics. Indeed, those metrics most used by Neale and Maes (2004) performed relatively better than those metrics not included in their text. This pattern supports that they may have recognized the problems identified within the current study. However, there is, apparently no discussion in the behavior genetic literature to support this possibility.

It is possible that some combination of indices that included both a highly-specific metric and a highly-sensitive metric might be created to jointly produce model selection rules that achieve both power greater than .8 and Type I error rates at or below .05. Such a combination of indices would be consistent with best practices for evaluating fit for SEM models and would have the chance of providing universal selection criteria – a series of golden rules rather than a single golden rule. Specifying such a combination was not a goal of the current study but is a useful suggestion for fruitful future research.

Nevertheless, if we had to select a single metric, it would have to be TLI. TLI was the closest metric that could be universally applied, and yet TLI still had considerable variability in sensitivity and specificity. Furthermore, many of the "optimized" thresholds were right at the boundary of what was possible: TLI $\geq$ 1. Accordingly, although this metric could be used universally, we still would not recommend it, as it is

impractical, might encourage HARKing (hypothesizing after the results are known; Kerr, 1998; Rubin, 2017), and does not require that the researcher think critically about the model they are fitting.

Fit statistics thresholds at the boundary are impractical because they set an impossible expectation. Perfectly-fitting models using real data can only be achieved through overfitting or selectively reporting fit indices. Researchers could be incentivized to tinker with their models to meet those thresholds. The resulting exploratory model might be presented and/or interpreted as a confirmatory one. It is just an impossible expectation that does more harm than good. And frankly, any model that can perfectly reproduce its covariance matrix is not useful. A model is designed to be a simplification of reality, not a reproduction (Rodgers, 2010).

Concluding Thoughts

We were tempted to provide criteria that differed depending on whether a saturated model (ACE) model was being used, or if a fully identified model (AE or CE) was being used. However, after considering the long-reaching implications and the confidence that declaring a well-fitting model brings, we decided against giving specific criteria. Readers who insist on custom criteria can find them in Tables 17-19 in the appendix. However, we caution against eagerly accepting these benchmarks, especially for the ACE model and any saturated models. Even so, if you are evaluating a saturated model, in this paper's case the ACE model, we encourage you to be ruthlessly skeptical when evaluating your fit statistics. Further, we recommend that you be skeptical of any model whose fit statistics noticeably deviate from perfection. Even if your model "fits," we encourage you to fit additional models to their data that are over=identified, and not merely just-identified. Researchers can either use a design that has more than two kinship categories, add covariates, or at least fit a restricted model like an AE or CE model. If your model does indeed fit well, it will survive this skepticism, and indicate that it fits better relative the other models.

More broadly, we think that combining ruthless skepticism with the general best practices of using multiple metrics from across classes of indices, should improve our ability to select the better fitting model – ideally the best identifiable model. Although the optimized metrics are using individual indices, future work should examine whether a combination of these metrics could achieve sufficient power and acceptable alpha levels. Indeed, if only one lesson is to be learned from these findings, it is the following: behavior genetic researchers should evaluate models beyond a "well-fitting" ACE model. A "well-fitting" ACE model does not indicate that that model is actually "well-fitting," even that the model is likely to be correct. In general, behavior genetics researchers should be ruthlessly skeptical of model fitting results.

Declarations

Funding: The authors were supported by NICHD grant R01-HD087395.

Conflicts of interest: The authors declare that they have no conflicts of interest.

Ethics approval: Not Applicable

Consent to participate: Not Applicable

Consent for publication: Not Applicable

Availability of data and materials: Due to the large file size (>6 GB), the raw simulated data are available upon request. Model-level results are available here: https://github.com/R-Computing-Lab/Dissertation_Data Data took approximately two weeks to generate and model.

Code availability: Code are available in Appendix A and the discord R package.

References

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Ayorech, Z., Selzam, S., Smith-Woolley, E., Knopik, V. S., Neiderhiser, J. M., DeFries, J. C., & Plomin, R. (2016). Publication Trends Over 55 Years of Behavioral Genetic Research. *Behavior Genetics*, *46*(5), 603–607. https://doi.org/10.1007/s10519-016-9786-2

Bekker, P., & Wansbeek, T. (2003). Identification in Parametric Models. In B. H. Baltagi (Ed.), *A Companion to Theoretical Econometrics* (pp. 144–161). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470996249.ch8

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10/dbj

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T. C., Mehta, P., & Fox, J. (2011). OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika*, *76*(2), 306–317. https://doi.org/10/fvqbph

Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models*. Sage Publications, Inc.

Burks, B. S. (1938). On the Relative Contributions of Nature and Nurture to Average Group Differences in Intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *24*(7), 276–282. https://doi.org/10/c6nx97

Burt, C. (1966). The genetic determination of differences in intelligence: A study of monozygotic twins reared together and apart. *British Journal of Psychology*, *57*(1-2), 137–153. https://doi.org/10/bfcmfp

Carey, G. (2005). Cholesky problems. *Behavior Genetics*, *35*(5), 653–665. https://doi.org/10.1007/s10519-005-5355-9

Chabris, C. F., Hebert, B. M., Benjamin, D. J., Beauchamp, J., Cesarini, D., van der Loos, M., Johannesson, M., Magnusson, P. K. E., Lichtenstein, P., Atwood, C. S., Freese, J., Hauser, T. S., Hauser, R. M., Christakis, N., & Laibson, D. (2012). Most Reported Genetic Associations With General Intelligence Are Probably False Positives. *Psychological Science*, *23*(11), 1314–1323. https://doi.org/10/f25m66

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Eaves, L. J., & Gale, J. S. (1974). A method for analyzing the genetic basis of covariation. *Behavior Genetics*, *4*(3), 253–267. https://doi.org/10.1007/BF01074158

Falconer, D. S. (1952). The Problem of Environment and Selection. *The American Naturalist*, *86*(830), 293–298. https://doi.org/10/d3dtjx

Fisher, R. A. (1919a). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(2), 399–433. https://doi.org/10.1017/S0080456800012163

Fisher, R. A. (1919b). The Genesis of Twins. *Genetics*, *4*(5), 489–499.

Friedrich, J., Brand, B., Graunke, K. L., Langbein, J., Schwerin, M., & Ponsuksili, S. (2017). Adrenocortical Expression Profiling of Cattle with Distinct Juvenile Temperament Types. *Behavior Genetics*, *47*(1), 102–113. https://doi.org/10.1007/s10519-016-9816-0

Galton, F. (1876). The History of Twins, as a Criterion of the Relative Powers of Nature and Nurture. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *5*, 391–406. JSTOR. https://doi.org/10/d8g7tp

Galton, F. (1896). Note to the Memoir by Professor Karl Pearson, F.R.S., on Spurious Correlation. *Proceedings of the Royal Society of London*, *60*, 498–502. JSTOR.

Garlapow, M. E., Everett, L. J., Zhou, S., Gearhart, A. W., Fay, K. A., Huang, W., Morozova, T. V., Arya, G. H., Turlapati, L., St. Armour, G., Hussain, Y. N., McAdams, S. E., Fochler, S., & Mackay, T. F. C. (2016). Genetic and Genomic Response to Selection for Food Consumption in Drosophila melanogaster. *Behavior Genetics*, 1–17. https://doi.org/10.1007/s10519-016-9819-x

Garrison, S. M. (2018). Popular Structural Equation Modeling Programs for Behavior Genetics. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(6), 972–977. https://doi.org/10.1080/10705511.2018.1493385

Garrison, S. M., Trattner, J., & Ream, C. (2020). *Discord R Package*. https://cran.r-project.org/web/packages/discord/index.html

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behavioral Research*, *20*(3), 255–271.

Gerlai, R., Poshusta, T. L., Rampersad, M., Fernandes, Y., Greenwood, T. M., Cousin, M. A., Klee, E. W., & Clark, K. J. (2017). Forward Genetic Screening Using Behavioral Tests in Zebrafish: A Proof of Concept Analysis of Mutants. *Behavior Genetics*, *47*(1), 125–139. https://doi.org/10.1007/s10519-016-9818-y

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling*, 1–18. https://doi.org/10.1080/10705511.2017.1402334

Hancock, G. R., & Mueller, R. O. (2010). *The Reviewer's Guide to Quantitative Methods*.

Hancock, G. R., & Mueller, R. O. (2011). The Reliability Paradox in Assessing Structural Relations Within Covariance Structure Models. *Educational and Psychological Measurement*, *71*(2), 306–324. https://doi.org/10.1177/0013164410384856

Harzing, A. W. (2018). *Publish or perish*.

Hernstein, R. J., & Murray, C. (1994). *The bell curve: The reshaping of American life by differences in intelligence* (1st Free Press pbk. ed). Free Press.

Holzinger, K. J. (1929). The relative effect of nature and nurture influences on twin differences. *Journal of Educational Psychology*, *20*(4), 241–248. https://doi.org/10.1037/h0072484

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60. https://doi.org/10.1037/1082-989X.12.1.58

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hu, L.-T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351–362. https://doi.org/10.1037/0033-2909.112.2.351

Hunter, M. D., Garrison, S. M., Burt, S. A., & Rodgers, J. L. (2020). *A note on the analytic identification of variance component models common to behavior genetics*.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), 0696–0701. https://doi.org/10.1371/JOURNAL.PMED.0020124

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting Practices in Confirmatory Factor Analysis: An Overview and Some Recommendations. In *Psychological Methods* (Vol. 14, Issue 1, pp. 6–23). AMER PSYCHOLOGICAL ASSOC. https://doi.org/10/cbd6zd

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement. *Harvard Educational Review*, *39*(1), 1–123. https://doi.org/10/gdkxt4

Jinks, J. L., & Fulker, D. W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin*, *73*(5), 311–349. https://doi.org/10/bmfrvb

Kang, Y., McNeish, D. M., & Hancock, G. R. (2016). The Role of Measurement Quality on Practical Guidelines for Assessing Measurement and Structural Invariance. *Educational and Psychological Measurement*, *76*(4), 533–561. https://doi.org/10/ggfnx3

Kenny, D. (n.d.). *SEM: Fit*. Retrieved April 11, 2019, from http://davidakenny.net/cm/fit.htm

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10/dnqm8w

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Publications.

Lance, C. E., Beck, S. S., Fan, Y., & Carter, N. T. (2016). A taxonomy of path-related goodness-of-fit indices and recommended criterion values. *Psychological Methods*, *21*(3), 388–404. https://doi.org/10/gcz6zx

Lee, J. J., & McGue, M. (2016). Why Behavioral Genetics Matters. *Perspectives on Psychological Science*, *11*(1), 29–30. https://doi.org/10/gf8cw4

Liew, S. H. M., Elsner, H., Spector, T. D., & Hammond, C. J. (2005). The first “classical” twin study? Analysis of refractive error using monozygotic and dizygotic twins published in 1922. *Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies*, *8*(3), 198–200. https://doi.org/10.1375/1832427054253158

Loehlin, J. C., & Vandenberg, S. G. (1968). Genetic and environmental components in the covariation of cognitive abilities: An additive model. In S. G. Vandenberg (Ed.), *Progress in Human Behavior Genetics*. Johns Hopkins University Press.

Marsh, H., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler’s (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Martin, N. G., & Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity*, *38*(1), 79–95. https://doi.org/10.1038/hdy.1977.9

McNeish, D. (2018). Should We Use *F* -Tests for Model Fit Instead of Chi-Square in Overidentified Structural Equation Models? *Organizational Research Methods*, 109442811880949. https://doi.org/10.1177/1094428118809495

Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychological Methods*, *21*(2), 151–163. https://doi.org/10/f8thxp

Millsap, R. E. (2012). A simulation paradigm for evaluating approximate fit. In M. C. Edwards & R. C. MacCallum (Eds.), *Current topics in the theory and application of latent variable models* (pp. 165–182). Routledge.

Mosing, M. A., Cnattingius, S., Gatz, M., Neiderhiser, J. M., & Pedersen, N. L. (2016). Associations Between Fetal Growth and Self-Perceived Health Throughout Adulthood: A Co-twin Control Study. *Behavior Genetics*, *46*(3), 457–466. https://doi.org/10.1007/s10519-015-9776-9

Mueller, R. O., & Hancock, G. R. (2010). Structural Equation Modeling. In *The Reviewer's Guide to Quantitative Methods*.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*(3), 430–445. https://doi.org/10.1037/0033-2909.105.3.430

Murray, C. (1998). *Income inequality and IQ.* American Enterprise Inst. for Public Policy Research.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus: Statistical analyses with latent variables. User's guide version 8* (Vol. 3).

Neale, M. C. (2009). *Mx Graphical User Interface*. https://vipbg.vcu.edu/resources/statistical-software/mxgui/#RASmfIwc

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H., & Boker, S. M. (2016). OpenMx 2.0: Extended Structural Equation and Statistical Modeling. *Psychometrika*, *81*(2), 535–549. https://doi.org/10/f8rfrg

Neale, M. C., & Maes, H. H. M. (2004). *Methodology for Genetic Studies of Twins and Families* (Vol. 48). Kluwer Academic Publishers B.V. https://doi.org/10.1136/jmg.30.9.800-a

Newman, H. H. (1917). *The biology of twins (mammals)*. University of Chicago Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/SCIENCE.AAC4716

Pearson, K. (1909). Determination of the coefficient of correlation. *Science*, *30*(757), 23–25. JSTOR. https://doi.org/10.1126/SCIENCE.30.757.23

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2016). Top 10 Replicated Findings From Behavioral Genetics. *Perspectives on Psychological Science*, *11*(1), 3–23. https://doi.org/10.1177/1745691615617439

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, *47*(7), 702–709. https://doi.org/10/f3nbfg

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, *17*(1), 1–14. https://doi.org/10/fzbzph

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rende, R. D., Plomin, R., & Vandenberg, S. G. (1990). Who discovered the twin method? *Behavior Genetics*, *20*(2), 277–285. https://doi.org/10.1007/BF01067795

Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., Griem, A., Kovacs, M., Ott, J., & Merikangas, K. R. (2009). Interaction Between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression: A Meta-analysis. *JAMA*, *301*(23), 2462–2471. https://doi.org/10.1001/jama.2009.878

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*(1), 1–12. https://doi.org/10/cc8fhq

Rodgers, J. L. (2019). Degrees of freedom at the start of the second 100 years: A pedagogical

treatise. Advances in Methods and Practices in Psychological Science, 2(4), 396-405.

https://doi.org/10.1177/2515245919882050

Rodgers, J. L., Garrison, S. M., O'Keefe, P., Bard, D. E., Hunter, M. D., Beasley, W. H., & Oord, E. J. C. G. van den. (2019). Responding to a 100-Year-Old Challenge from Fisher: A Biometrical Analysis of Adult Height in the NLSY Data Using Only Cousin Pairs. *Behavior Genetics*. https://doi.org/10.1007/s10519-019-09967-6

Rodgers, J. L., & Rowe, D. C. (2002). Theory development should begin (but not end) with good empirical fits: A comment on Roberts and Pashler (2000). *Psychological Review*, *109*(3), 599–604; discussion 605-607. https://doi.org/10.1037/0033-295x.109.3.599

Rubin, M. (2017). When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress: *Review of General Psychology*. https://doi.org/10.1037/gpr0000128

Savalei, V. (2012). The Relationship Between Root Mean Square Error of Approximation and Model Misspecification in Confirmatory Factor Analysis Models. *Educational and Psychological Measurement*, *72*(6), 910–932. https://doi.org/10.1177/0013164412452564

Shrout, P. E., & Rodgers, J. L. (2017). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*. https://doi.org/10.1146/ANNUREV-PSYCH-122216-011845

Siemens, H. W. (1924). *Die zwillingspathologie; ihre bedeutung, ihre methodik, ihre bisherigen ergebnisse*. Springer.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., de Lange, M., Harris, J. R., Hjelmborg, J. V. B. B., Luciano, M., Martin, N. G., Mortensen, J., Nisticò, L., Pedersen, N. L., Skytthe, A., Spector, T. D., Stazi, M. A., Willemsen, G., & Kaprio, J. (2003). Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries. *Twin Research*, *6*(05), 399–408. https://doi.org/10.1375/TWIN.6.5.399

Smith, W. T. (1856). A Course of Lectures ON THE THEORY AND PRACTICE OF OBSTETRICS. *The Lancet*, *67*(1703), 419–423. https://doi.org/10.1016/S0140-6736(02)55452-0

Sniekers, S., Stringer, S., Watanabe, K., Jansen, P. R., Coleman, J. R. I., Krapohl, E., Taskesen, E., Hammerschlag, A. R., Okbay, A., Zabaneh, D., Amin, N., Breen, G., Cesarini, D., Chabris, C. F., Iacono, W. G., Ikram, M. A., Johannesson, M., Koellinger, P., Lee, J. J., … Posthuma, D. (2017). Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature Genetics*, *49*(7), 1107–1112. https://doi.org/10/b7gm

Snygg, D. (1938). The Relation Between the Intelligence of Mothers and of Their Children Living in Foster Homes. *The Pedagogical Seminary and Journal of Genetic Psychology*, *52*(2), 401–406. https://doi.org/10.1080/08856559.1938.10534325

Stigler, S. M. (1989). Francis Galton's Account of the Invention of Correlation. *Statistical Science*, *4*(2), 73–79. https://doi.org/10.1214/ss/1177012580

Thorndike, E. L. (1905). Measurement of Twins. *The Journal of Philosophy, Psychology and Scientific Methods*, *2*(20), 547–553. JSTOR. https://doi.org/10/c8c7xk

Turkheimer, E. (2016). Weak Genetic Explanation 20 Years Later: Reply to Plomin et al. (2016). *Perspectives on Psychological Science*, *11*(1), 24–28. https://doi.org/10.1177/1745691615617442

Turkheimer, E., & Waldron, M. (2000). Nonshared environment: A theoretical, methodological, and quantitative review. *Psychological Bulletin*, *126*(1), 78–108. https://doi.org/10/c6ndk8

van den Berg, S. M., de Moor, M. H. M., Verweij, K. J. H. H., Krueger, R. F., Luciano, M., Arias-Vásquez, A., Matteson, L. K., Derringer, J., Esko, T., Amin, N., Gordon, S. D., Hansell, N. K., Hart, A. B., Seppälä, I., Huffman, J. E., Konte, B., Lahti, J., Lee, M., Miller, M., … ... (2016). Meta-analysis of Genome-Wide Association Studies for Extraversion: Findings from the Genetics of Personality Consortium. *Behavior Genetics*, *46*(2), 170–182. https://doi.org/10.1007/s10519-015-9735-5

Verhulst, B. (2017). A Power Calculator for the Classical Twin Design. *Behavior Genetics*, *47*(2), 255–261. https://doi.org/10.1007/s10519-016-9828-9

Verhulst, B., Prom-Wormley, E., Keller, M., Medland, S., & Neale, M. C. (2019). Type I Error Rates and Parameter Bias in Multivariate Behavioral Genetic Models. *Behavior Genetics*, *49*(1), 99–111. https://doi.org/10.1007/S10519-018-9942-Y

Whitehead, H., Vachon, F., & Frasier, T. R. (2017). Cultural Hitchhiking in the Matrilineal Whales. *Behavior Genetics*, *47*(3), 324–334. https://doi.org/10.1007/s10519-017-9840-8

Widaman, K. F., & Thompson, J. S. (2003). On Specifying the Null Model for Incremental Fit Indices in Structural Equation Modeling. *Psychological Methods*, *8*(1), 16–37. https://doi.org/10/dpps6n

Figures

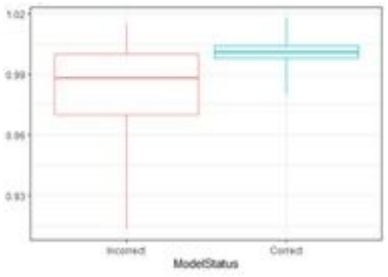

Figure 1

Box and whisker plot of TLI by Model Status

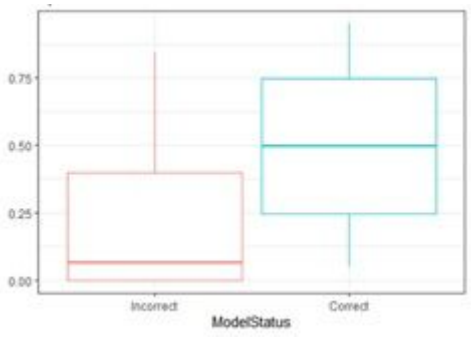

Figure 2

Box and whisker plot of p(χ^2) by Model Status

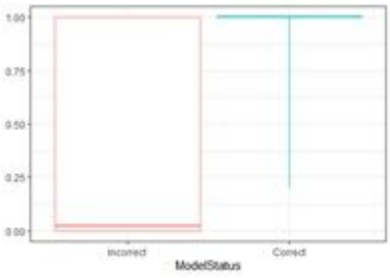

Figure 3

Box and whisker plot of p(Nested Model Comparison) by Model Status

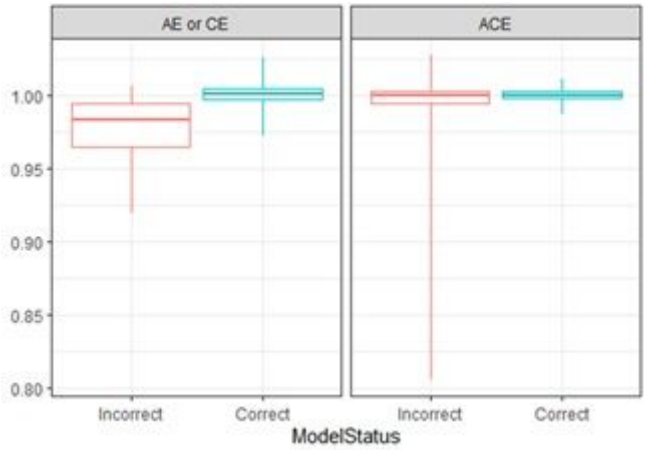

Figure 4

Box and whisker plot of TLI by Model Status and Type

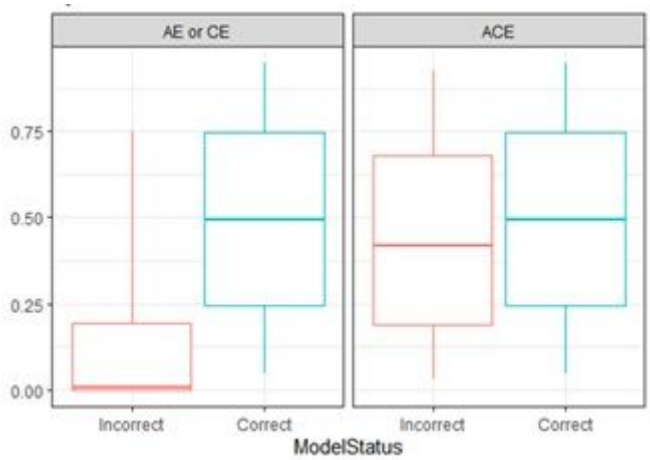

Figure 5

Box and whisker plot of p(χ^2) by Model Status and Type

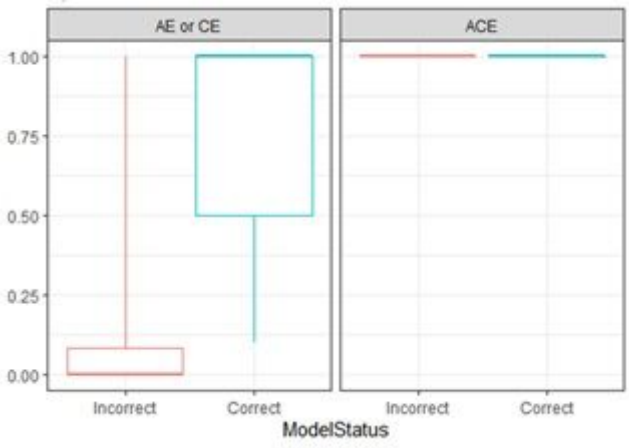

Figure 6

Box and whisker plot of p(Nested Model Comparison) by Model Status and Type

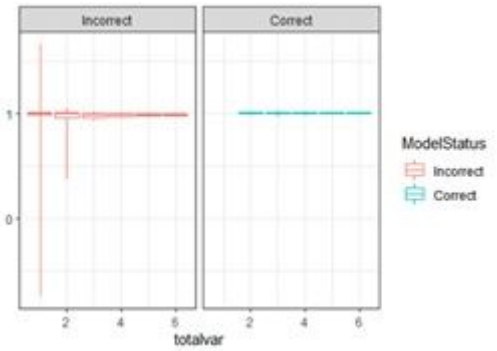

Figure 7

Box and whisker plots of TLI by Model Status and Total Variance

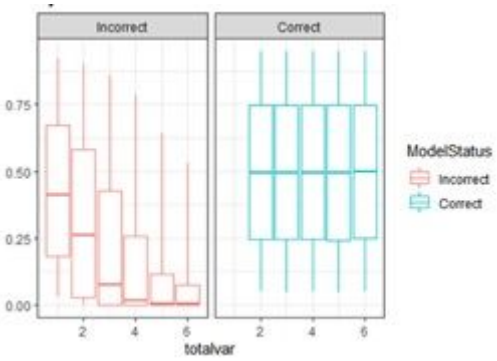

Figure 8

Box and whisker plot of p(χ^2) by Model Status and Total Variance

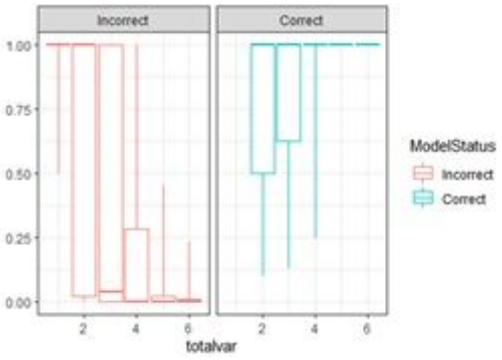

Figure 9

Box and whisker plot of p(Nested Model Comparison) by Model Status and Total Variance

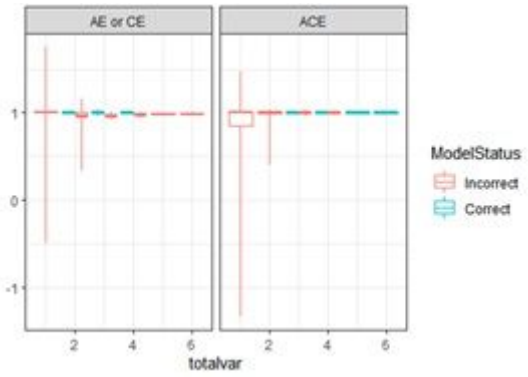

Figure 10

Box and whisker plots of TLI by Model Status, Model Type, and Total Variance

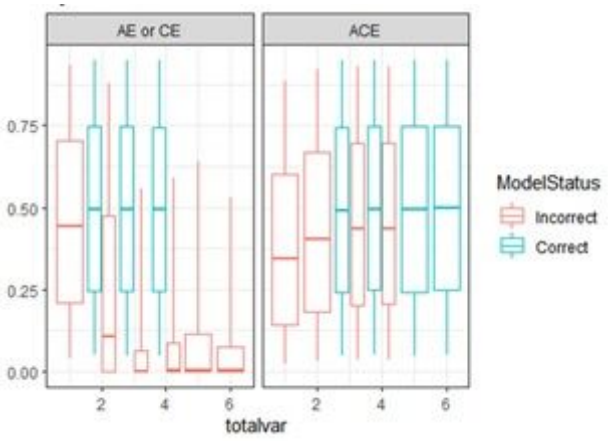

Figure 11

Box and whisker plots of p(χ^2) by Model Status, Model Type, and Total Variance

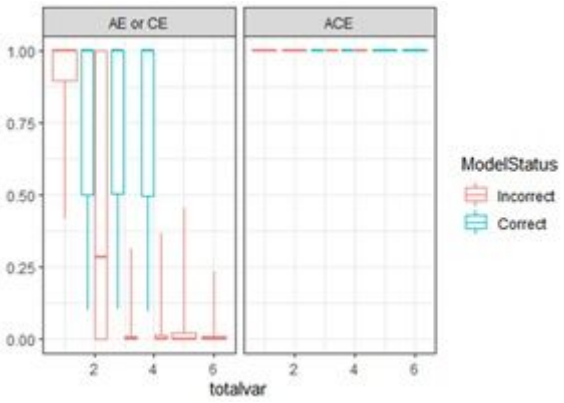

Figure 12

Box and whisker plots of p(Nested Model Comparison) by Model Status, Type, & Total Variance

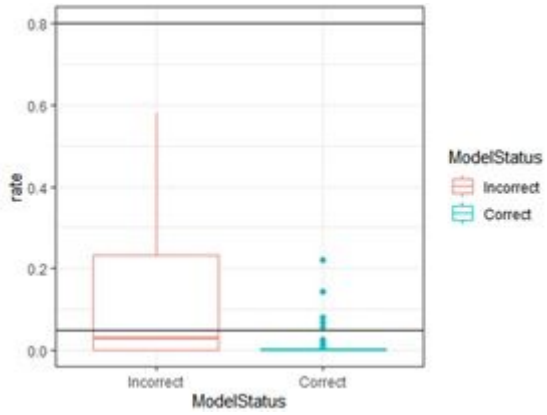

Figure 13

Distribution of model rejection rates using TLI across conditions

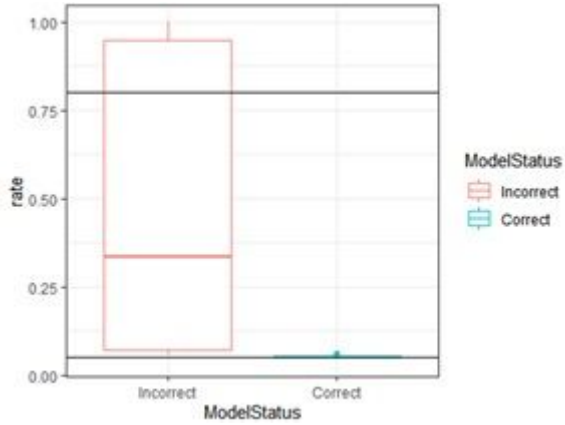

Figure 14

Distribution of model rejection rates using p(χ^2) across conditions

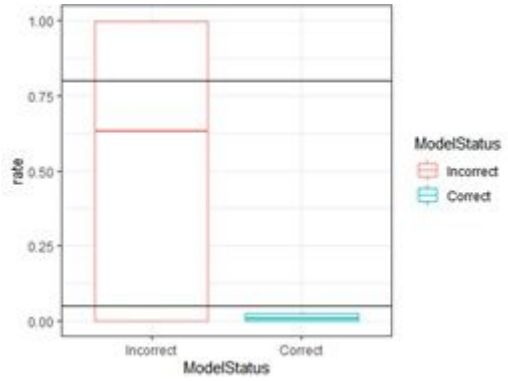

Figure 15

Distribution of model rejection rates using nested model comparisons across conditions

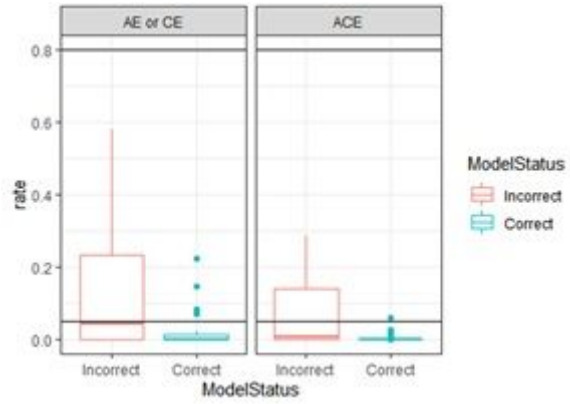

Figure 16

Distribution of model rejection rates using TLI across conditions by model type

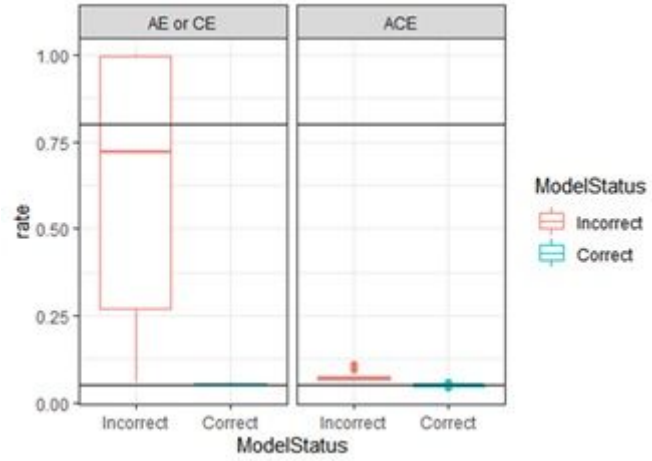

Figure 17

Distribution of model rejection rates using p(χ^2) across conditions by model type

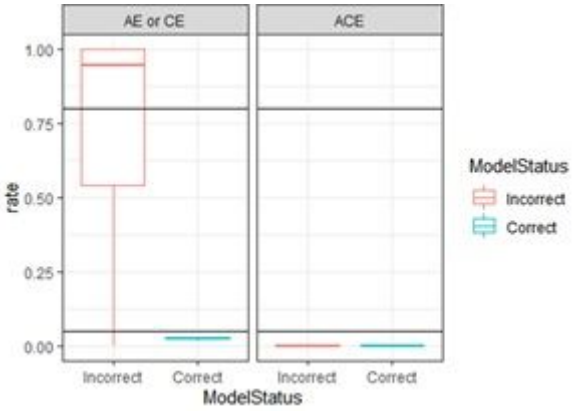

Figure 18

Distribution of model rejection rates using nested model comparisons across conditions by model type

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 5AppendixCode.docx
- AppendixTables.docx