CHAPTER 11

Model Selection in Structural Equation Modeling

Kristopher J. Preacher
Haley E. Yaremych

A concept of central importance in the application of SEM is model fit. Fit indices such as the *comparative fit index* (CFI; Bentler, 1990), the *Tucker–Lewis index* (TLI; Tucker & Lewis, 1973), *standardized root mean squared residual* (SRMR; Bentler, 1995), and *root mean square error of approximation* (RMSEA; Steiger & Lind, 1980) are useful mainly for gauging the quality of a single model in isolation. These indices reflect different facets of the discrepancy between observed data and what a model implies the data should look like. A great deal of attention has been devoted to this topic. (For excellent overviews of model fit in SEM, see Hu & Bentler, 1998; Marsh, Hau, & Grayson, 2005; see also West, Wu, McNeish, & Savord, Chapter 10, this volume.)

Here we focus instead on a closely related topic, *model selection* in SEM. Model selection is the process of specifying and fitting a set of rival, theoretically motivated models to a given data set, comparing their performance on some key criteria, and using the results to select at least one of them for continued attention. We begin by discussing why it is not sufficient to simply gauge the fit of a model in isolation, and why model selection is usually the preferred strategy. We present the philosophical background that justifies this focus. Next, we discuss the characteristics, advantages, and disadvantages of specific model selection criteria. We conclude with recommendations for practice. We limit the scope of our discussion to traditional applications of SEM, including path analysis, factor analysis, and most models involving latent variables with multiple indicators. For the sake of brevity, we do not consider model selection in the context of extensions of SEM, such as Bayesian SEM, dynamic SEM, multilevel SEM, or mixture models. We deliberately cite many sources because we hope this chapter can serve as an entrée to the literature for applied users and methodologists alike.

THE CASE FOR MODEL SELECTION

Problems with Model Fit

Because no model can perfectly reflect the truth, we instead regard models as approximations to reality (Leamer, 1974; MacCallum, 2003). Thus, numerous fit indices have been suggested to quantify various facets of *verisimilitude* (a model's proximity to the objective truth). However, these indices have many limitations. For example, if a model is saturated (i.e., just-identified with $df = 0$), perfect fit is guaranteed, regardless of how factually correct or incorrect the model's theoretical implications might be. Overidentified models will have imperfect fit, but model misfit will be determined by a variety of factors, only one of which is the match

between the model and observed data. Many extraneous factors will also influence indices of absolute fit, including sample size (*N*), model complexity, sampling variability, violation of distributional assumptions, measurement error, degree of nonnormality in the data (Curran, West, & Finch, 1996), degrees of freedom (Kenny, Kaniskan, & McCoach, 2014), near-zero error variances (Browne, MacCallum, Kim, Anderson, & Glaser, 2002), what kinds of parameter constraints are responsible for the misfit (e.g., constrained factor loadings, error covariances, intercepts, or path coefficients; Savalei, 2012), and other factors; for more detail, see West et al., chapter 10, this volume. Additionally, *researcher degrees of freedom*—a catchall term for the collection of subjective, often questionable decisions made by a researcher when specifying and fitting a model—also plays a role (Simmons, Nelson, & Simonsohn, 2011). Decisions such as what variables to include or exclude, how to deal with missing data, what software and estimation method to use, what fit indices to report, what benchmarks to use, and whether or how to modify the model in response to poor fit can cause model fit to fluctuate wildly. At best, good fit conveys that the model is but one plausible representation of the actual underlying structure (Tomarken & Waller, 2003).

Additionally, SEM is not immune to the serious philosophical problems associated with point null hypothesis testing. Typically, null hypotheses are precise and known a priori to be false, whereas theories are imagined as vague alternatives. Therefore, theories are rarely confronted with severe tests or falsified. Meehl (1978, 1986) criticized psychology for its overreliance on these flawed, uninformative tests of isolated null hypotheses. SEM, on the other hand, flips the traditional roles of null and alternative hypotheses. The researcher's scientific hypothesis is cast as a null hypothesis, which would seemingly make the enterprise more scientific because the theory is subject to direct tests via examination of model fit. However, this reversal of null and alternative hypotheses incentivizes bad research practices that reduce power to reject the null—reliance on small samples, the use of unreliable instruments, and the abuse of researcher degrees of freedom among them.

The well-meaning social scientist would seem to be at an impasse: One choice is to attempt to support a theory by demonstrating the falsity of a straw-man null hypothesis known a priori to be false, with repeated replications of this process yielding illusory support for one's theory (Meehl, 1986). The other choice is to attempt to support a theory by casting it as the null hypothesis (again, known to be false) and failing to reject it, under conditions in which such failure can be attributed to many reasons apart from the theory's verisimilitude. Neither choice is acceptable. For the social sciences to be truly cumulative, theories need opportunities to be compared to each other and accrue track records of successes and failures over time (Lakatos, 1970; Meehl, 1986, 1990). The comparison of multiple competing theories via statistical comparison of models is a strong step in that direction.

A Better Alternative: Model Selection

Given that no model can be literally true, and there are many factors other than verisimilitude that can influence fit indices, what characteristics should we value in a model? Many qualities make up a good model (e.g., interpretability, parsimony, generalizability), but perhaps most importantly, a good model *should perform better on some key dimensions than rival models* (Myung & Pitt, 1998). According to popular metatheories of science (e.g., Lakatos, 1970), theories are refuted not by experiment but rather by the emergence of superior theories that better explain existing facts and anticipate new facts. Scientific goals are advanced best when two or more theories are compared against each other, and the ones that least agree with the data are discarded or modified (Lakatos, 1970; MacCallum, 2003).

In model selection in the SEM context, two or more rival models vying to explain the same data are ranked by model selection criteria (these criteria reward better fit and penalize complexity). Subjective interpretability serves as a necessary qualitative aid in this process (Browne, 2000; Marsh & Balla, 1994). For example, if one model implies a negative slope where all competing theories expect a positive slope, this fact may serve as a tie-breaker between two models that may otherwise be difficult to rank. Whatever model emerges as victorious is still not the truth, of course, any more than a given model in isolation can be true. Instead, we hope that the selected model can serve as a useful "best working hypothesis" until something better comes along. This strategy mitigates some of the problems associated with model fit and is consistent with strategies advocated by philosophers of science (Meehl, 1990). Indeed, the performance of a model relative to that of other models provides a more reasonable benchmark of success than pitting a model against an unattainable

standard of perfect fit, or against a vague standard of close fit that varies as a function of many irrelevant factors. In model selection, because the same researcher fits a set of models to the same data set using the same software and the same sample size, researcher degrees of freedom are held relatively constant.

Clearly, there is much to be gained by adopting a model selection strategy. Regardless of which theory or model "wins" in a given application, a better theory or model will inevitably emerge to supplant it—one with some mix of better fit, superior cross-validation, and a closer match to theory. This lack of emphasis on absolute fit mitigates a major incentive to misuse researcher degrees of freedom, yielding a more honest approach to science. A focus on model selection also incentivizes using larger samples, more parsimonious models, and more reliable scales, as these factors further increase support for the model should it outperform its rivals (Roberts & Pashler, 2000). In addition, because one model must always emerge as superior, the researcher always has a story to tell; contrast this to the more common scenario in which only one model is tested; the researcher is highly invested in demonstrating good fit and is thus tempted to abuse researcher degrees of freedom. Finally, the model selection approach offers greater flexibility, as we need not select just one model (Burnham & Anderson, 2002; Kadane & Lazar, 2004; Weakliem, 2016); in some situations, we might discard models that are obviously poor and retain a subset of well-performing models for further study.

Model Complexity versus Generalizability

When evaluating or comparing models, it is essential to consider *model complexity*[1]—a model's intrinsic ability to fit diverse data patterns, regardless of theoretical predictions (Dunn, 2000; Myung, 2000; Myung & Pitt, 2002; Pitt & Myung, 2002). Researchers who observe good fit should interpret that fit in light of how well the model could be expected to explain *any* plausible, empirically obtainable data. Models that show good fit to data, when the a priori probability of obtaining good fit is high, pass tests that are not particularly risky (Bamber & van Santen, 2000; Meehl, 1978, 1990, 1992). These models do not accrue much credibility, even when they demonstrate good fit to data, because they have a high probability of fitting any observed data simply by chance (Hogarty, 2000; Meehl, 1978, 1990).

Complexity can also be understood as the complement to a model's parsimony or falsifiability. Complex models have a tendency to *overfit*; that is, the parameter estimates of a complex model will fit a given data set very well, but these same parameter estimates may perform poorly when applied to new data (Forster & Sober, 1994; Myung & Pitt, 2002; Pitt, Myung, & Zhang, 2002). This happens because a complex model capitalizes on meaningful variation as well as random error in a particular data set. Complexity is often operationalized as the number of free parameters (q) in a model (Mulaik et al., 1989). As q increases, the model becomes more likely to fit well to *any* data set (Forster & Sober, 1994; Myung & Pitt, 1998). The danger with adding free parameters—or equivalently, removing constraints—is that it could lead to selecting a model not because it has greater verisimilitude but, rather, because it is simply more complex than its competitors (Mulaik et al., 1989; Myung, 2000; Roberts & Pashler, 2000). Yet the number of parameters is not the only determinant of model complexity. Another major factor influencing complexity is a model's functional form or *structural complexity*, the unique way in which a model's parameters interact and conspire to fit observed data (Myung, 2000; Myung & Pitt, 1998, 2002; Preacher, 2006). All else being equal, models that are more complex make fewer narrow and less risky predictions, and thus are less likely to be falsified when fit to data.

Dunn (2000) asserted that models must be placed on equal footing with respect to their propensities to fit any kind of data before being compared in terms of their abilities to fit observed data; that is, we must consider the *data space* to which a given set of rival models might be applied—the universe of possible data that might be encountered in practice[2] in a given domain (Myung, 2000; Roberts & Pashler, 2000). A given model will fit some regions of the data space well and others poorly. In that sense, models may be said to "occupy" different—and differently sized—regions of the data space. When a model inherently fits well to a large region of the data space, it is more complex, and its good fit to the current data is less compelling (Myung, 2000; Myung & Pitt, 1998). Holding absolute fit constant, we should prefer the model that is less complex.

A related facet of a model's quality is its *generalizability*, which Myung and Pitt (2002, p. 447) define as "a model's ability to fit not only the observed data in hand but also future, unseen data sets generated from the same underlying process." Generalizability can be understood as a model's replicability or ability to *cross-validate* well. If a complex model fits well to a given

data set, it likely will fit well to virtually any data set drawn from the same population, and little is learned. If a model, *with its parameters constrained equal to values estimated in one sample*, still fits well when applied to a new sample, it is said to have cross-validated well (a cross-validation strategy known as *tight replication*; Bentler, 1980). Because more complex models capitalize on meaningful population-level variation *as well as* random fluctuations in a particular sample, they are less likely to cross-validate well and therefore are less generalizable. We thus argue that generalizability is a better criterion for model selection than is goodness of fit per se (Myung, 2000; Pitt & Myung, 2002; Pitt et al., 2002; Preacher, Cai, & MacCallum, 2007).

Summary

Thus far, the primary take-home points are (1) comparison of multiple competing models is an approach to science that is far preferable to evaluating individual models in isolation and (2) models should be compared in terms of their generalizability, not merely fit. More complex models have less generalizability because their fit is enhanced by factors unrelated to verisimilitude. Thus, the better a model selection criterion considers complexity, the better it will be at quantifying generalizability. In the next section we elaborate on several model selection criteria for SEM. Finally, the literature on model selection and related issues is vast. For more detail on the topics broached here, we refer to excellent general texts by Burnham and Anderson (2002), Claeskens and Hjort (2008), and Weakliem (2016).

MODEL SELECTION IN SEM

In the following sections, we assume that the models to be compared are all posited a priori. Otherwise, if models are obtained by proceeding from one model and modifying it to produce a "final model," the process can depend too much on subjective decisions about the sequence; the degree to which this process is data-driven undercuts the role of theory. We also assume that the models to be compared are based on real psychological theories; that is, none of the rival models is created to fail in order to make a preferred model look superior by comparison; all have an a priori nontrivial chance of being selected. Finally, we assume that the rival models are fit to the same data (the same endogenous variables). Models that, say, differ by the inclusion or exclusion of an observed mediator cannot be compared, even informally, because they are models for different sets of observed variables.

The Discrepancy Framework

A *selection criterion* is a statistic used to help decide among competing models. To scaffold our discussion of model selection criteria used in SEM, we use a framework described by Cudeck and Henly (1991), adapted from Linhart and Zucchini (1986), which is commonly used to depict how sources of error contribute to lack of fit in the SEM context (e.g., Browne, 2000). In Figure 11.1, each circle represents a covariance matrix, and each line represents a quantifiable measure of distance or *discrepancy* separating a pair of covariance matrices. *Sample discrepancy* represents the discrepancy between observed data (covariance matrix $\mathbf{S}$) and a model's predictions in the sample ($\hat{\boldsymbol{\Sigma}}$). Sample discrepancy is lack of fit observed in the sample, and its opposite is goodness of fit in the sample with no penalty for model complexity. *Discrepancy due to approximation*, or approximation discrepancy, represents model error, or the discrepancy between the population covariance matrix ($\boldsymbol{\Sigma}_0$) and the model's predictions in the population ($\tilde{\boldsymbol{\Sigma}}_0$). Approximation discrepancy is the complement to verisimilitude. *Discrepancy due to estimation*, or estimation discrepancy, represents sampling variability, which is closely tied to sample size. Finally, *overall discrepancy* represents the difference between the population covariance matrix $\boldsymbol{\Sigma}_0$ and $\hat{\boldsymbol{\Sigma}}$. The complement to overall discrepancy is a model's generalizability, or its potential to cross-validate well.

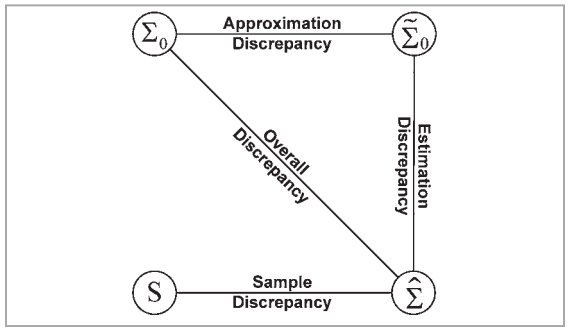

FIGURE 11.1. Cudeck and Henly's (1991) framework for representing relationship between model and real world.

We present Cudeck and Henly's (1991) framework here because it can help us understand what various model selection criteria are intended to quantify, and hence understand the basis of model comparisons used for model selection. Model selection criteria in SEM are usually based on approximation discrepancy (reflecting verisimilitude) or on overall discrepancy (reflecting generalizability). The diagram can be used heuristically to anticipate what happens as sample size changes and/or as absolute model fit improves. In general, holding sample size constant and removing constraints on a model (freeing parameters) will decrease approximation discrepancy but increase estimation discrepancy, and thus can either increase or decrease overall discrepancy (Browne & Cudeck, 1992). As sample size increases, estimation discrepancy decreases, until $\mathbf{S} = \boldsymbol{\Sigma}_0$ and $\hat{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\Sigma}}_0$ in the limit, implying that verisimilitude and generalizability are asymptotically equivalent (Browne, 2000). However, because samples are finite and we are never able to specify a model that is completely true, overall discrepancy > approximation discrepancy > 0. Which model is selected can depend on how the chosen model selection criterion balances verisimilitude and generalizability.

As a model's complexity increases, approximation discrepancy decreases but estimation discrepancy increases. This phenomenon is known as the bias–variance trade-off (Burnham & Anderson, 2002). Greater complexity generally leads to less biased parameter estimates and better model fit, but this better fit is often "bought" at the expense of that model's ability to fit future data (generalizability). Various selection criteria balance the competing goals of maximizing fit and minimizing complexity. Thus, many model selection criteria impose a penalty for model complexity.

Fit Indices that Reflect Sample Discrepancy

The most straightforward model selection criteria are simple model fit indices, such as SRMR (Bentler, 1995) or the chi-square statistic (χ^2). These indices are measures of sample discrepancy because they are based directly on differences between elements of $\mathbf{S}$ and $\hat{\boldsymbol{\Sigma}}$, reflecting raw fit in the sample with no penalty for complexity. Even though these indices are widely used for evaluating the fit of models in isolation, their utility as selection criteria is limited. Selecting a model based only on minimizing sample discrepancy does not acknowledge the possibility that the model could fit virtually any data just as well or better.

Another common approach is to formally compare models in terms of χ^2 (the likelihood ratio test, or LRT; Steiger, Shapiro, & Browne, 1985); however, this test is limited to models that are parametrically nested—those pairs for which the free parameters of one model are a subset of those of the other model (but see Levy & Hancock, 2011; Merkle, You, & Preacher, 2016). The LRT is highly sensitive to sample size, such that even with moderate *N*, virtually any LRT will favor the more complex model even if differences in fit are trivial (Bentler & Bonett, 1980). Furthermore, the LRT involves testing the hypothesis that two nested models fit equally well—a hypothesis known a priori to be false (Hu & Bentler, 1998; Tucker & Lewis, 1973).

Fit Indices that Reflect Overall Discrepancy

Many fit indices reflect overall discrepancy in various ways because they include a penalty for model complexity, although the penalty may not be obvious (Marsh & Balla, 1994). For example, the RMSEA, or ε (Browne & Cudeck, 1992; Steiger & Lind, 1980) is a measure of model discrepancy per degree of freedom (*df*):

$$\text{RMSEA} = \sqrt{\max\left(\frac{\hat{F}_m}{df_m} - \frac{1}{N-1}, 0\right)}$$

where $\hat{F}_m$ is the maximum likelihood (ML) discrepancy function value at convergence for the target model, df_m is its *df*, and *N* is sample size. Holding discrepancy constant and decreasing df_m (i.e., increasing complexity) distributes that discrepancy over fewer *df*, yielding a larger RMSEA, effectively penalizing the inclusion of superfluous parameters. A second example is Bentler's (1990) CFI:

$$\text{CFI} = \frac{\max\left(\chi^2_b - df_b, 0\right) - \max\left(\chi^2_m - df_m, 0\right)}{\max\left(\chi^2_b - df_b, 0\right)}$$

in which the numerator is the difference between the noncentrality (a measure of badness of fit) of the baseline (null) model[3] and that of the model of interest, and the denominator is the baseline model's noncentrality. CFI adjusts for complexity by subtracting df_m in the numerator; adding constraints yields a higher CFI, all else being equal.

Partly in response to the limitations of LRTs, it has become popular to compare models using fit statistics that recognize model complexity in some way, such as CFI, RMSEA, and others (see, e.g., Bentler & Bonett,

1980; Mulaik et al., 1989; see also West et al., Chapter 10, this volume). For example, Cheung and Rensvold (2002) recommended using differences in CFI (i.e., ΔCFI) and other fit indices to investigate measurement invariance. RMSEA has been suggested as an informal selection criterion in factor analysis (a special case of SEM; Huang, 2017; Preacher, Zhang, Kim, & Mels, 2013). However, we recommend exercising caution in using fit indices for model selection in SEM. This practice was suggested primarily in the context of comparing nested models, and specifically for testing measurement invariance or selecting among factor models. It is unclear whether or how this method should extend to SEM more generally. Moreover, the method is ad hoc in that the fit indices used to create differences were not designed for this purpose, so such comparisons are of only heuristic value. In contrast, other criteria, such as those described below, were designed for the explicit purpose of model selection. Using differences in fit indices may not be as useful as its popularity might imply.

Cross-Validation

In classical cross-validation, a sample is split into two smaller random samples. One is treated as a *calibration sample* (**C**) and the other as a *validation sample* (**V**). Model parameter estimates are obtained by fitting the model to **C**, then the model, with parameters fixed to the values estimated in **C**, is fit to **V**. Measures of fit to **V** reflect the degree of predictive success and therefore quantify generalizability (Myung & Pitt, 2002). In the SEM context, the cross-validation index (CVI) is obtained by:

$$\mathrm{CVI} = \hat{F}_m\left(\mathbf{S}_V, \hat{\mathbf{\Sigma}}_C\right)$$

where $\mathbf{S}_V$ is the sample covariance matrix for **V** and $\hat{\mathbf{\Sigma}}_C$ is the model-implied covariance matrix for **C**. In words, CVI quantifies the difference between the observed covariance matrix **V** and what the model (fitted to **C**) expects **V** to look like. When used as a selection criterion, smaller values of CVI are preferred. Cross-validation automatically corrects for overfitting by penalizing a model's capitalization on random error in **C** (Myung & Pitt, 1998). A complex model fits well in **C** partly because it conforms more closely to the idiosyncrasies of the data; however, **V** will not have the same idiosyncrasies. Cross-validation was recommended for model evaluation in SEM by Cudeck and Browne (1983) and MacCallum, Roznowski, Mar, and Reith (1994).

Classic cross-validation is limited in several ways. First, the degree of cross-validation is determined in part by how one splits the sample; different allocations of cases to **C** and **V**, as well as different sample sizes for **C** and **V**, yield different results. Second, information is wasted because the sample is split into two parts. Therefore, Browne and Cudeck (1989, 1992) proposed an *expected cross-validation index* (ECVI):

$$\mathrm{ECVI} = \mathrm{E}\left(\mathrm{CVI}\right) = \hat{F}_m\left(\mathbf{S}, \hat{\mathbf{\Sigma}}\right) + \frac{df_m}{N}$$

where **S** is the sample covariance matrix using the full N, $\hat{\mathbf{\Sigma}}$ is the model-implied covariance matrix, and the expectation is taken over all **C** and **V**. Thus, with ECVI it is not necessary to split one's sample. ECVI is a linear rescaling of the Akaike information criterion (AIC), and thus yields the same model ranking as AIC, to be discussed shortly. For an in-depth discussion of cross-validation in SEM, see Browne (2000).

Information Criteria

Many *information criteria* were deliberately created to select models in terms of generalizability. Different information criteria balance fit and parsimony in different ways. Unlike with the LRT, models compared using information criteria need not be parametrically nested. Most information criteria are based on log-likelihoods with a penalty for complexity, and typically have the form:

$$D + f(q, N) \tag{11.1}$$

where D is a function of the model's log-likelihood (i.e., fit) at convergence, whereas f is a function of the number of free parameters (q) and sample size (N). We focus on the AIC and the Bayesian information criterion (BIC) because most other information criteria are asymptotically equivalent to either AIC or BIC (Burnham & Anderson, 2004; Weakliem, 2016). To use any information criterion of the form in Equation 11.1, one computes the criterion for each model in the candidate set, then selects the model with the lowest value of the criterion. Later we make recommendations about choosing, reporting, and interpreting such criteria in SEM, and empirical illustrations (including R and Mplus code) are shown in the online supplement at the companion website.

AIC

One of the most popular criteria is the AIC (Akaike, 1973):

$$\text{AIC} = -2\ln L\left(x \mid \hat{\theta}_x\right) + 2q$$

where $\ln L\left(x \mid \hat{\theta}_x\right)$ is the model's log-likelihood under ML estimation at convergence and q is the model's number of free parameters. AIC has two components: $-2\ln L\left(x \mid \hat{\theta}_x\right)$ represents lack of fit and $2q$ is a penalty term for complexity. As model fit improves, $-2\ln L\left(x \mid \hat{\theta}_x\right)$ will decrease. As more free parameters are included, $2q$ will increase.

AIC is not easily interpretable at face value.[4] It can be understood as a measure of *entropy* or the *Kullback–Leibler divergence* (or *K-L distance*), a measure of the difference between a model's fit to sample data and reality (Burnham & Anderson, 2004; Ichikawa & Konishi, 1999)—compare this to the definition of overall discrepancy earlier. AIC can also be understood as –2 times a bias-corrected measure of the expected $\ln L\left(x \mid \hat{\theta}_x\right)$ for a model. Though esoteric, these interpretations permit comparisons among rival models for the same data: AIC is computed for each model, and the model with the lowest AIC is selected. In theory, the selected model is the one with the greatest potential to generalize or cross-validate well[5] at a given N (Forster & Sober, 1994; Weakliem, 2016). Numerous extensions of AIC have been proposed for various settings. For discussions of AIC and its extensions, see Bozdogan (1987), Burnham and Anderson (2002, 2004), Claeskens and Hjort (2008), and Ichikawa and Konishi (1999). A closely related criterion is the *deviance information criterion* (DIC) often used in Bayesian modeling, including Bayesian SEM (Cain & Zhang, 2019). When noninformative priors are used, DIC and AIC are expected to yield equivalent model rankings.

AIC has been criticized for being overly sensitive to sample size (e.g., Marsh & Balla, 1994). However, because it reflects overall discrepancy, and overall discrepancy depends on the sample size through estimation discrepancy, it is expected that AIC will select more complex models as N increases (Cudeck & Henly, 1991). In relatively small samples, the penalty for complexity has a greater influence and simpler models are selected. AIC is often accused of overfitting in large samples; however, this criticism has been shown to be exaggerated (Bozdogan, 2000; Claeskens & Hjort, 2008; Haughton, Oud, & Jansen, 1997), as choosing the *correct* model is not the purpose of AIC. AIC is known not to be asymptotically consistent in that it is expected to pick different models at different N's (Bozdogan, 1987; Haughton, 1988; Haughton et al., 1997; Kashyap, 1982; Kuha, 2004; Rissanen, 1978; Schwarz, 1978; Sclove, 1987).

Burnham and Anderson (2002) provide simple but informal guidelines for judging the magnitude of AIC differences ($\Delta_i = \text{AIC}_i - \text{AIC}_{\min}$, simply the difference between model i's AIC and the lowest AIC among the candidate models). AIC differences that are fairly small (say, 0–2) imply that the data are not sufficient to meaningfully distinguish between a pair of models. We caution that these guidelines are somewhat arbitrary, and will depend on the type of model, the type of misspecifications, and data quality.

Bayesian Information Criteria

Bayesian model selection (BMS) invokes the Bayesian framework, which weighs relative probabilities of competing hypotheses. Given a model M_i with unknown parameters Θ_i, the likelihood of the data is given by

$$L_i = \int_{\Theta} p\left(x \mid \Theta_i\right) p\left(\Theta_i, M_i\right) d\Theta_i \tag{11.2}$$

where $\int_{\Theta}$ indicates a multiple integral taken over the entire parameter space. Comparing the likelihood of competing models allows one to weigh the evidence in favor of each model relative to that of the others.

However, the integral in Equation 11.2 is usually intractable, making it difficult to compute the likelihood (Lanterman, 2001). Therefore, many researchers (e.g., Bollen, Harden, Ray, & Zavisca, 2014; Bollen, Ray, Zavisca, & Harden, 2012; Burnham & Anderson, 2002; Claeskens & Hjort, 2008; Haughton, 1988; Haughton et al., 1997; Kashyap, 1982; Kuha, 2004; Raftery, 1995; Sclove, 1987; Weakliem, 1999; Xie, 1999) have presented a Laplace approximation, or second-order Taylor series expansion, of the log of this integral, here multiplied by –2 to maintain comparability to other criteria:

$$\begin{aligned}\text{BMS} = &-2\ln L\left(x \mid \hat{\theta}_x\right) - 2\ln p\left(\theta \mid M\right) \\ &-q\ln\left(2\pi\right) + q\ln N + \ln\left|I\left(\theta\right)\right|\end{aligned}$$

where $p(\theta \mid M)$ represents prior information[6] and $|I(\theta)|$ is the determinant of the expected Fisher information matrix[7] for a single individual. We omit a constant term

whose contribution diminishes sharply with N, and the prior term $-2\ln p(\theta|M)$ is usually omitted in practice.

The BMS approximation is very good under mild regularity conditions that are commonly met in applications of SEM (Lanterman, 2001). Several information criteria can be obtained as special cases of BMS that omit some terms. For example, Bollen and colleagues' (2012, 2014) *information matrix-based BIC* (IBIC) omits only the prior information term $-2\ln p(\theta|M)$:

$$\text{IBIC} = -2\ln L\left(x|\hat{\theta}_x\right) - q\ln(2\pi) + q\ln N + \ln\left|I(\theta)\right|$$

Draper (1995) derived a simplified version of the Laplace approximation that omits both the $-2\ln p(\theta|M)$ and $\ln|I(\theta)|$ terms, which are often negligible asymptotically across the set of models to be compared, but retains $-q\ln(2\pi)$. We refer to it as Haughton's BIC (HBIC), following Bollen and colleagues (2012, 2014):

$$\text{HBIC} = -2\ln L\left(x|\hat{\theta}_x\right) - q\ln(2\pi) + q\ln N$$

HBIC performed well in a simulation study by Haughton and colleagues (1997) using SEM. HBIC has also been discussed by Markon and Krueger (2004) and by Myung, Navarro, and Pitt (2006).

Further omitting $-q\ln(2\pi)$ yields the celebrated BIC, originally proposed by Schwarz in 1978 (Raftery, 1995; Weakliem, 1999; Xie, 1999):

$$\text{BIC} = -2\ln L\left(x|\hat{\theta}_x\right) + q\ln N$$

The penalty term in BIC includes q but weighted by the natural log of the sample size. BIC is superficially similar to AIC, but with a stronger penalty for complexity in most situations, meaning that it tends to select more parsimonious models than AIC. Relative to AIC, BIC is less subject to choosing more complex models as N increases because the penalty term increases with N (Schwarz, 1978).

BIC is appealing for a number of reasons. First, owing to its Bayesian roots and derivation, BIC is easily relatable to *Bayes factors*. In the Bayesian framework, the Bayes factor is the ratio of likelihoods for two models, say, M_1 and M_0, vying to explain the same data. For example, if the Bayes factor is 4, then the data are four times more likely to have arisen from model M_1 than from M_0. The Bayes factor for M_1 with respect to M_0 can be approximated by $e^{\frac{1}{2}(BIC_0 - BIC_1)}$ implying that the observed data are $e^{\frac{1}{2}(BIC_0 - BIC_1)}$ times more likely to have been generated under M_1 than under M_0 (Kass & Raftery, 1995; Weakliem, 2016). Second, BIC is *consistent* in that it tends to select the true data-generating process (DGP) if it is among the candidate models (Haughton, 1988; Haughton et al., 1997; Wasserman, 2000). Even when the DGP is not among the candidate models (which we maintain is always the case), theoretically, BIC will select the model closest to it. If more than one model is "true" in that one of them contains superfluous parameters, BIC will select the simplest "true" model. Third, like AIC, BIC can be used to compare non-nested models.

BIC is not without criticism. First, BIC relies on N as a summary of the information in a data set, but N is a poor measure of information (Weakliem, 1999). Second, its use assumes that the researcher has chosen a particular type of prior for each model parameter, although different researchers may use different priors when operating in the Bayesian framework. However, Raftery (1999) argues that the assumptions about priors that underlie the derivation of BIC are strongly supported in practice, rendering this assumption inconsequential. Finally, Weakliem (1999) also argued that BIC suffers by the omission of the $\ln|I(\theta)|$ term. Raftery (1999) countered by noting that in cases in which this term is included, the resulting model ranking is often very similar to the ranking identified by BIC. However, the $\ln|I(\theta)|$ term (which represents structural complexity) becomes more important as N decreases, and for that reason it is included in more complete information criteria. BIC's best justification might be that it is an easy approximation to minimum description length (MDL; to be discussed next), but the approximation may be poor at sample sizes commonly used in SEM. MDL may be superior because it uses more information.

The MDL Principle and Related Methods

We end this section with a discussion of a class of selection criteria with great potential for development and application in SEM. In a remarkable example of convergent scientific thought, essentially the same idea was developed independently at least three times.

First, Meehl (1990, 1992) suggested a *corroboration index* (C_i) for quantifying the long-run performance of a theory. C_i is essentially a measure of goodness of fit penalized by model complexity (thus, higher values are better). Besides its general similarity to the family of information criteria, C_i quantifies complexity as the proportion of the data space to which a model could fit

well. Hogarty (2000) has suggested variants of C_i for application to SEM.

Second, an idea similar to Meehl's had already been explored by Botha, Shapiro, and Steiger (1988) in the context of factor analysis. They proposed that when comparing competing factor models, the fit of each model should be compared to a distribution of fit statistics (say, SRMR or χ^2) obtained by fitting the model to a set of randomly generated correlation matrices. Their measure, the *uniform index-of-fit* (*i**), is simply the proportion of these random matrices to which the model fits *worse* than the observed matrix. Competing models can be compared on *i** (again, higher values are preferred). Given two models with equal fit, the simpler one will have a higher *i**. Preacher (2006) applied this method to SEM.

Essentially the same idea, but more formally couched in information theory, forms the basis of the MDL principle (Grünwald, 2000; Hansen & Yu, 2001; Myung et al., 2006; Rissanen, 1978, 1989, 1996, 2001; Stine, 2004). Under the MDL principle, rival models are analogous to compression codes written in a fixed programming language, and data are analogous to strings in need of compression. The optimal model (compression code) is one that "compresses" the data the best, capitalizing on regularities in the data (Rissanen, 1978), without itself being too long (complex). A complex code can compress nearly any given string quite compactly, at the expense of being long, whereas a more parsimonious code can compress a more limited array of strings (data) and is not itself too long. The MDL principle has been used in many model comparison settings and shows great promise for application to SEM.

MDL is a principle, not a statistic, but its logic informs at least three selection criteria that are relevant to SEM. A key advantage of these criteria is that they incorporate model complexity that arises not only from the number of free parameters (*q*), but also from structural complexity. One is *normalized maximum likelihood* (NML; Barron, Rissanen, & Yu, 1998; Rissanen, 1996, 2001):

$$\mathrm{NML} = \frac{L\left(x \mid \hat{\theta}_x\right)}{\int_S L\left(z \mid \hat{\theta}_z\right) dz}$$

where $\hat{\theta}_x$ is the ML estimate for the observed data x, $\hat{\theta}_z$ is the ML estimate for hypothetical data z, and the integral (the penalty term) is taken over all possible data patterns z in a bounded data space S. Model selection using NML is asymptotically equivalent to model selection by Bayes factors.[8] Computing NML directly is not possible in any but the simplest cases (where all possible data sets can be generated). Otherwise, the penalty term can be approximated by simulating random data and computing an average likelihood in the denominator, with simulation terminating once a sufficiently precise average has been obtained. As with C_i and *i**, higher values are preferred.

NML can be logged and negated to yield *stochastic complexity* (Rissanen, 1996, 2001):

$$\mathrm{SC} = -\ln L\left(x \mid \hat{\theta}_x\right) + \ln \int_S L\left(z \mid \hat{\theta}_z\right) dz$$

The first term represents badness of fit, and the second is a penalty for complexity equal to the natural log of the integral of MLs of a model when fit to all data sets that could be observed in a given domain. As with AIC and BIC, lower values are preferred. Rissanen (2001) offers additional interpretations.

NML can be interpreted as the fit of a model to one's observed data relative to the fit of the model to a large number of randomly drawn data sets that uniformly span the data space. Approximating NML using simulated random data requires the researcher to define (1) what constitutes the data space and (2) the density function defining the likelihood of drawing a given data set from within the bounds defined in (1). Reasonable researchers may disagree on these definitions. In addition to the indeterminacy in defining the data space, it can be difficult to generate random data of the sort required to approximate NML, and perhaps time-consuming to fit the model to a quantity of data sufficient to obtain a precise estimate of the average likelihood.

A more easily quantifiable expression of the MDL principle is the *Fisher information approximation* (FIA; Barron et al., 1998; Grünwald, 2000; Myung et al., 2006; Myung & Pitt, 2002; Pitt et al., 2002; Rissanen, 1996, 2001; Su, Myung, & Pitt, 2005):

$$\mathrm{FIA} = -\ln L\left(x \mid \hat{\theta}_x\right) + \tfrac{1}{2} q \ln\left(N / 2\pi\right) + \ln \int_\Theta \sqrt{\left|I(\theta)\right|} d\theta$$

where the last two terms constitute *geometric complexity*, which is related to the number of distinguishable probability distributions that a model can explain; the higher this number, the more complex the model (Myung, Balasubramanian, & Pitt, 2000). FIA approximates the negative logarithm of NML. It explicitly quantifies model complexity in terms of the number of free parameters (*q*) as well as structural (functional

form) complexity via $I(\theta)$ (Grünwald, 2000). Although the integral in FIA is far more tractable than that in NML, computation is still challenging in most examples of SEM. When multiplied by 2, FIA closely resembles BIC but includes separate penalty terms for the number of free parameters and structural complexity. Thus, BIC can be seen as an approximation to FIA, which in turn is an approximation to NML.

An approximation to FIA that is easier to compute is *stochastic information complexity* (SIC; Hansen & Yu, 2001; Markon & Krueger, 2004; Preacher, 2006; Rissanen, 1989):

$$\begin{aligned}\text{SIC} &= -\ln L\left(x \mid \hat{\theta}_x\right) + \tfrac{1}{2}\ln\left|\hat{I}(\theta)\right| \\ &= -\ln L\left(x \mid \hat{\theta}_x\right) + \tfrac{1}{2}\ln\left|NI(\theta)\right| \\ &= -\ln L\left(x \mid \hat{\theta}_x\right) + \tfrac{1}{2}q\ln N + \tfrac{1}{2}\ln\left|I(\theta)\right|\end{aligned}$$

where $\hat{I}(\theta) = N \cdot I(\theta)$ is the observed Fisher information matrix for the sample. Although its origins differ, SIC can be seen as an approximate simple special case of BMS after some terms have been dropped. In the limit, under mild conditions usually observed in SEM, $\left|I(\theta)\right| \to 1$ and SIC further reduces to BIC (Lanterman, 2001). In fact, an early MDL criterion was identical to BIC (Rissanen, 1978, 1989). SIC is considered an easy approximation to FIA because it involves no integration and uses the observed sample Fisher information matrix, which can be obtained in practice by simply inverting the asymptotic covariance matrix of parameter estimates,[9] $\hat{\boldsymbol{\Sigma}}(\theta)$, provided by any SEM package[10] (Bonifay & Cai, 2017; Haughton et al., 1997; Markon & Krueger, 2004). Markon and Krueger (2004) conducted two simulation studies of SEM applied to behavior genetic models, finding that SIC performed well relative to other information criteria but was often subject to difficulty in computing $\left|I(\theta)\right|$ as a function of $\hat{\boldsymbol{\Sigma}}(\theta)$.

There is a large literature on MDL, spanning multiple disciplines. Brief but lucid discussions of the MDL principle can be found in Grünwald (2000), Hansen and Yu (2001), Lanterman (2001), Markon and Krueger (2004), Pitt and colleagues (2002), and Stine (2004). Book-length treatments include Grünwald (2007) and the edited volume by Grünwald, Myung, and Pitt (2005). Vandekerckhove, Matzke, and Wagenmakers (2015) contrast MDL (FIA) to AIC and BIC, with examples. Table 11.1 summarizes the criteria we have discussed, multiplied by 2 as needed to place them on comparable scales.

Model Averaging: If You Can't Beat 'Em, Join 'Em

When the model selection process ends in choosing one model, that model is usually treated as the best working model. Its parameters are estimated, interpreted, and reported, along with standard errors or confidence intervals. But there is always uncertainty about whether the "right" model was chosen given sampling error and the various subjective decisions involved in selecting one model from many. *Model selection uncertainty* is the sampling variability in decisions based on selection criteria (Burnham & Anderson, 2002; Lubke & Campbell, 2016; Preacher & Merkle, 2012). It is important to remember that selection criteria are also statistics with sampling variance, implying that model rankings may vary across samples even when best practices are closely followed. Further uncertainty is introduced when one considers that the objectively best model was probably not included in the pool of candidate models in the first place.

Model averaging is an attempt to account for this uncertainty. First, models that clearly perform poorly in the initial comparison may be eliminated upfront. Then, for each of the remaining models, a set of weights is computed based on (say) its AIC or BIC values. Wasserman (2000) argues that BIC is better than AIC for this purpose. For AIC, the weights are

$$w_i = \frac{\exp\left(-\tfrac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R}\exp\left(-\tfrac{1}{2}\Delta_r\right)}$$

for every model i in a set of R models. Model weights can also be obtained through the Bayesian framework (for more detail, see Depaoli, Kaplan, & Winter, Chapter 38, this volume).

Such weights may be used to obtain weighted average parameter estimates, obtain predicted values that give each model a say in determining those values, or to correct standard errors and confidence intervals for model selection uncertainty (Burnham & Anderson, 2002; Wasserman, 2000). We caution that the first use seems questionable in applications of SEM given that the estimates of a parameter taken from different candidate models will have sometimes starkly different interpretations. For example, we would not want to average estimates of the effect of a predictor on a latent factor when one of the models contains a moderator of that effect. For more discussion on model averag-

TABLE 11.1. Selected Information Criteria

Criterion	Abbreviation	Expression
An / Akaike's Information Criterion	AIC	$-2\ln L\left(x \mid \hat{\theta}_x\right) + 2q$
Bayesian Model Selection	BMS	$-2\ln L\left(x \mid \hat{\theta}_x\right) - 2\ln p\left(\theta \mid M\right) + q\ln N - q\ln\left(2\pi\right) + \ln\left\lvert I\left(\theta\right)\right\rvert$
Bayesian Information Criterion	BIC	$-2\ln L\left(x \mid \hat{\theta}_x\right) + q\ln N$
Haughton's BIC	HBIC	$-2\ln L\left(x \mid \hat{\theta}_x\right) + q\ln N - q\ln\left(2\pi\right)$
Information Matrix-Based BIC	IBIC	$-2\ln L\left(x \mid \hat{\theta}_x\right) + q\ln N - q\ln\left(2\pi\right) + \ln\left\lvert I\left(\theta\right)\right\rvert$ $= -2\ln L\left(x \mid \hat{\theta}_x\right) - q\ln\left(2\pi\right) - \ln\left\lvert\hat{\boldsymbol{\Sigma}}\left(\theta\right)\right\rvert$ (computational)
Stochastic Complexity	SC	$-2\ln NML$ $= -2\ln L\left(x \mid \hat{\theta}_x\right) + 2\ln\int_S L\left(z \mid \hat{\theta}_z\right) dz$
Fisher Information Approximation	FIA	$-2\ln L\left(x \mid \hat{\theta}_x\right) + q\ln N - q\ln\left(2\pi\right) + 2\ln\int_\Theta \sqrt{\left\lvert I\left(\theta\right)\right\rvert} d\theta$
Stochastic Information Complexity	SIC	$-2\ln L\left(x \mid \hat{\theta}_x\right) + \ln\left\lvert\hat{I}\left(\theta\right)\right\rvert$ $= -2\ln L\left(x \mid \hat{\theta}_x\right) + q\ln N + \ln\left\lvert I\left(\theta\right)\right\rvert$ $= -2\ln L\left(x \mid \hat{\theta}_x\right) - \ln\left\lvert\hat{\boldsymbol{\Sigma}}\left(\theta\right)\right\rvert$ (computational)

Note. All criteria have been placed on a $-2\ln L$ scale. Minor terms that converge to zero in probability have been omitted for simplicity.

ing, see Burnham and Anderson (2002), Claeskens and Hjort (2008), Draper (1995), Hoeting, Madigan, Raftery, and Volinsky (1999), Kadane and Lazar (2004), Kaplan and Lee (2016), Raftery, Madigan, and Hoeting (1997), and Weakliem (2016).

SUMMARY AND CONCLUSION

Important Considerations in the Use of Model Selection Criteria

In practice, most SEM software will report AIC and/or BIC and one or two other criteria (see supplement at companion website), but most others are computable based on information provided in output. When comparing two or more competing structural equation models, what is the best course of action? How should researchers choose from among the many model selection criteria—including AIC, BIC, and MDL (and their many extensions)—now available?

Depending on the researcher's goal, it may be sensible to choose criteria that emphasize *consistency* or *efficiency* (Shao, 1997; Vrieze, 2012). We may want to identify the candidate model that minimizes squared prediction error, or that is expected to cross-validate the best when applied to new samples. Here, we should use a criterion that emphasizes *efficiency*, such as AIC. On the other hand, BIC emphasizes *consistency*: If a true model exists and is in the set to be compared, BIC selects it almost surely as N increases. Otherwise, BIC will select the model in the set that lies closest to the true DGP. MDL criteria are also consistent (Barron et al., 1998; Grünwald, 2007; Myung et al., 2000; Su et al., 2005). These qualities generally hold for the variants of AIC, BIC, and MDL; variants differ mainly in how complexity is penalized. Consistency and efficiency cannot be maximized simultaneously (Claeskens & Hjort, 2008). Thus, it is usually sensible to report and interpret at least one information criterion that emphasizes consistency (from the AIC family) and at least one that emphasizes efficiency (from the BIC or MDL families). Triangulation using multiple selection criteria in tandem is often the best policy (Kuha, 2004; Xie, 1999), just as with indices of absolute and relative fit.

Practicality is another important consideration. AIC and BIC are easy to compute and are often provided

by SEM software. The best MDL formulations (NML and FIA) share the consistency quality of BIC and use more precise information but are difficult to compute and are not yet incorporated in SEM software. Our position is that MDL has theoretical advantages over BIC and should be used when possible, especially when comparing non-nested models (Myung & Pitt, 2002) and when N is not large, as structural complexity has a stronger influence in smaller samples. However, the most complete expressions of MDL are the most difficult to compute, and those that are easier to compute manually tend to be incomplete in some way that limits their advantages over BIC. We recommend using AIC as well as the most complete versions of BIC and/or MDL that are feasible to compute—perhaps HBIC, IBIC, or SIC,[11] unless one is willing to go the extra mile and use simulation to compute NML or FIA.

A third consideration is to remember the role of researcher expertise. The use of any model selection criterion is pointless in the absence of theoretically informed subjective judgment. Substantive theory must be used to define a small set of defensible models prior to analysis. Theory and judgment must also be used during model selection to determine which models to reject or retain on the grounds of interpretability in addition to fit. And, ultimately, theory is used to scaffold the interpretation of results using the selected model(s). Model selection criteria are tools used to support decision making; they are not the final arbiters (Navarro, 2019).

Finally, an important caveat is that there can be considerable sampling variability in the ranking of models according to indices such as AIC and BIC (Lubke & Campbell, 2016; Preacher & Merkle, 2012). However, selection criteria are usually reported for a set of models fit to data from only a single sample. Although it is possible to obtain confidence intervals for selection criteria, these intervals do not fully capture the variability in model rankings that would be observed across repeated sampling from the same population. It is good to remain mindful of this sampling variability that is rarely, if ever, considered.

Frontiers for Future Research

First, although NML appears to be the best selection criterion, defining the data space S in an uncontroversial way is the primary impediment to its use. Defining S requires establishing boundaries defining the range and interior of the space (the possible data patterns for all of the models under scrutiny), as well as the relative plausibility of specific data patterns inside S. Both decisions are subjective and value-laden, and therefore subject to dispute among researchers. Many open questions remain as to how S should be defined (Bonifay & Cai, 2017; Dunn, 2000; Falk & Muthukrishna, 2021; Lai, Green, & Levy, 2017; Li, Lewandowsky, & DeBrunner, 1996; Preacher, 2006). This reliance on the definition of S is avoided by criteria that use functions of the information matrix to quantify complexity, like FIA and SIC (Li et al., 1996).

The importance of structural complexity diminishes as N increases, but at typical sample sizes for social science research, it is still an important factor (Myung & Pitt, 1998). However, those criteria that capture it well—NML and FIA—do so by using multidimensional integration, which quickly becomes intractable. To avoid integration, Preacher (2006), Preacher and colleagues (2007), and Roos (2008) used MCMC, and Falk and Muthukrishna (2021) provide R code (*ockhamSEM*) to generate random correlation matrices using the onion method (Lewandowski, Kurowicka, & Joe, 2009). Su and colleagues (2005) showed simpler ways to obtain the Fisher information matrix in models for multinomial or normal outcomes; however, the simplified computations are still daunting. Until improved computations are incorporated into SEM software, it is much simpler, if less precise, to use SIC instead of NML or FIA (Markon & Krueger, 2004).

Computation of traditional information criteria assumes complete data. Yet missing data is the rule rather than the exception in practice. Model selection with missing data is an area of active research. Claeskens and Hjort (2008) describe some partial solutions to this problem. Challenges and promising approaches have been described by El Matouat, Moussa, and Hamzaoui (2016), Ibrahim, Zhu, and Tang (2008), Jiang, Nguyen, and Rao (2015), Lai (2021), and Li, Yu, and Zeng (2019).

Finally, although we have argued that model selection is theoretically superior to the evaluation of a model's absolute fit in isolation, it is not a perfect approach. Selection criteria inherit some of the problematic aspects of absolute fit indices, such as lack of robustness to distributional violations, strange behavior at the boundaries of the parameter space, and differential sensitivity to different kinds of misspecification. For example, consistent with the reliability paradox (see West et al., Chapter 10, this volume), as indicator reliabilities become higher, model selection criteria are better able

to detect minute differences in fit. The model selection process is itself subject to abuse, for example, by strategically choosing what models to compare, HARKing (hypothesizing after the results are known; Kerr, 1998), omission of inconvenient cases, and model modification through specification searches (MacCallum, Roznowski, & Necowitz, 1992). In addition, researchers frequently ignore the existence of statistically equivalent models at their peril (Lee & Hershberger, 1990; Williams, 2012); often there are thousands of models equivalent to each of those being compared, all with distinct theoretical implications, yet selection criteria are not able to distinguish among them. Future research may suggest ways to mitigate some of these problems.

There is much ground we have not covered. One related issue is the comparison of nested and non-nested models using LRTs rather than model selection criteria (Levy & Hancock, 2011; Merkle et al., 2016). Another is the comparison of models that differ in their exogenous predictors (information criteria are commonly used for variable selection in regression). We have restricted our attention to the comparison of models that have been specified a priori, yet model selection can also be an important tool in theory-building exploration. We also have not discussed model selection in SEM extensions such as Bayesian SEM, dynamic SEM, multilevel SEM, or mixture models, but we emphasize that these are areas of active and important research.

Take-Home Points

- Selecting one or more models from a pool of rival models is a better approach to science than evaluating single models in isolation.
- Generalizability is a better mark of model quality than is pure model fit.
- There is no such thing as a correct model. There are models that cross-validate better than others, and there are those that better reflect the true DGP.
- It is wise to use more than one (and more than one kind of) criterion. They reflect different aspects of relative model quality. AIC-based criteria are useful when prediction or cross-validation is important; criteria based on BIC or MDL are useful for identifying a model that best approximates the truth. These often agree, but sometimes they will not, and that is OK.
- Theory and content expertise are necessary ingredients for good model selection.

ACKNOWLEDGMENTS

We thank Carl Falk and Terrence Jorgensen for valuable discussions regarding modern information criteria.

NOTES

1. Model complexity is identical or similar to the concepts of *fit/fitting propensity* (Falk & Muthukrishna, 2021; Preacher, 2006), *flexibility* (Cutting, 2000; Roberts & Pashler, 2000), *tolerance* (Meehl, 1990, 1992), *prediction range* (Bamber & van Santen, 2000), and *scope* (Cutting, 2000; Cutting, Bruno, Brady, & Moore, 1992; Li et al., 1996).

2. A reasonable data space for factor analysis models might be, for example, the universe of correlation matrices with all-positive correlations. A reasonable data space for SEM might be the universe of all covariance matrices such that no means or variances exceed the maxima established by the scales of the instruments employed.

3. This is usually, but not always, the independence model that estimates means and variances but constrains covariances to zero (Widaman & Thompson, 2003). For more detail on baseline models, see West et al., Chapter 10, this volume.

4. Clear discussions of AIC's derivation and meaning may be found in Bozdogan (1987), Burnham and Anderson (2002), Forster and Sober (1994), Ichikawa and Konishi (1999), and Weakliem (2016).

5. As alluded to earlier, the ECVI proposed by Browne and Cudeck (1989, 1992) is related to AIC by a factor of $N - 1$ and ranks models in the same order when ML estimation is used (Browne, 2000; Haughton et al., 1997).

6. Prior information in this context is the probability, based on past research and subjective judgment, attached to the collection of model parameters.

7. The Fisher information matrix is a matrix-valued function of the information contained in the data about the parameters of a given model. It is directly related to the precision with which those parameters can be estimated.

8. This is true under the assumption that noninformative Jeffreys priors are used for each compared model.

9. Exploiting some simple identities,

$$\ln\left|\hat{I}(\theta)\right| = \ln\left|\hat{\boldsymbol{\Sigma}}(\theta)^{-1}\right| = \ln\left|\hat{\boldsymbol{\Sigma}}(\theta)\right|^{-1} = -\ln\left|\hat{\boldsymbol{\Sigma}}(\theta)\right|$$

A simpler formula for SIC is thus

$$SIC = -\ln L\left(x \mid \hat{\theta}_x\right) - \tfrac{1}{2}\ln\left|\hat{\boldsymbol{\Sigma}}(\theta)\right|$$

10. Mplus uses the observed or expected information matrix (or a combination) to obtain $\hat{\boldsymbol{\Sigma}}(\theta)$. In models with continuous outcomes, no missing data, and no numerical integration, the expected Fisher information matrix may be used. Otherwise, the observed Fisher information matrix is used. It matters little which one is used to compute information crite-

ria that employ Fisher information, as long as the same type is used consistently across models.

11. Most criteria discussed here—including AIC, BIC, HBIC, IBIC, and SIC, and others—can be obtained using the R package *semTools* (current version 0.5-7; Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2022) in conjunction with *lavaan* (current version 0.6-12; Rosseel, 2012).

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akadémiai Kiadó.

Bamber, D., & van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology, 44*, 20–40.

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory, 44*, 2743–2760.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31*, 419–456.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual*. Temple City, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. *Structural Equation Modeling, 21*, 1–19.

Bollen, K. A., Ray, S., Zavisca, J., & Harden, J. J. (2012). A comparison of Bayes factor approximation methods including two new methods. *Sociological Methods and Research, 41*, 294–324.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research, 52*, 465–484.

Botha, J. D., Shapiro, A., & Steiger, J. H. (1988). Uniform indices-of-fit for factor analysis models. *Multivariate Behavioral Research, 23*, 443–450.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology, 44*, 62–91.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*, 108–132.

Browne, M. W., & Cudeck, R. (1989). Single-sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24*, 445–455.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*, 230–258.

Browne, M. W., MacCallum, R. C., Kim, C.-T., Anderson, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods, 7*, 403–421.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33*, 261–304.

Cain, M. K., & Zhang, Z. (2019). Fit for a Bayesian: An evaluation of PPP and DIC for structural equation modeling. *Structural Equation Modeling, 26*, 39–50.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

Claeskins, G., & Hjort, N. L. (2008). *Model selection and model averaging.* Cambridge, UK: Cambridge University Press.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18*, 147–167.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109*, 512–519.

Curran, P. J., West, S. G., & Finch, J. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16–29.

Cutting, J. E. (2000). Accuracy, scope, and flexibility of models. *Journal of Mathematical Psychology, 44*, 3–19.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General, 121*, 364–381.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B, 57*, 45–70.

Dunn, J. C. (2000). Model complexity: The fit to random data reconsidered. *Psychological Research, 63*, 174–182.

El Matouat, A., Moussa, F. D., & Hamzaoui, H. (2016). Consistency of information criteria for model selection with missing data. *Communications in Statistics—Theory and Methods, 45*, 6900–6914.

Falk, C. F., & Muthukrishna, M. (2021). Parsimony in model selection: Tools for assessing fit propensity. *Psychological Methods*. [Epub ahead of print]

Forster, M. R., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45*, 1–35.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology, 44*, 133–152.

Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

Grünwald, P. D., Myung, I. J., & Pitt, M. A. (Eds.). (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association, 96*, 746–774.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Annals of Statistics, 16*, 342–355.

Haughton, D. M. A., Oud, J. H. L., & Jansen, A. R. G. (1997). Information and other criteria in structural equation model selection. *Communications in Statistics: Simulation and Computation, 26*, 1477–1516.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14*, 382–401.

Hogarty, K. Y. (2000). Risky predictions and damn strange coincidences: An initial consideration of Meehl's index of corroboration. *Florida Journal of Educational Research, 40*, 76–99.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424–453.

Huang, P. H. (2017). Asymptotics of AIC, BIC, and RMSEA for model selection in structural equation modeling. *Psychometrika, 82*, 407–426.

Ibrahim, J. G., Zhu, H.-T., & Tang, N.-S. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association, 103*, 1648–1658.

Ichikawa, M., & Konishi, S. (1999). Model evaluation and information criteria in covariance structure analysis. *British Journal of Mathematical & Statistical Psychology, 52*, 285–302.

Jiang, J., Nguyen, T., & Rao, J. S. (2015). The E-MS algorithm: Model selection with incomplete data. *Journal of the American Statistical Association, 110*, 1136–1147.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling* (R package version 0.5-7). https://cran.r-project.org/package=semtools

Kadane, J. B., & Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association, 99*, 279–290.

Kaplan, D., & Lee, C. (2016). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling, 23*, 343–353.

Kashyap, R. L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-4*, 99–104.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2014). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*, 486–507.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33*, 188–229.

Lai, K. (2021). Using information criteria under missing data: Full information maximum likelihood versus two-stage estimation. *Structural Equation Modeling, 28*, 278–291.

Lai, K., Green, S. B., & Levy, R. (2017). Graphical displays for understanding SEM model similarity. *Structural Equation Modeling, 24*, 803–818.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–195). Cambridge, UK: Cambridge University Press.

Lanterman, A. D. (2001). Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection. *International Statistical Review, 69*, 185–212.

Leamer, E. E. (1974). False models and post-data model construction. *Journal of the American Statistical Association, 69*, 122–131.

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

Levy, R., & Hancock, G. R. (2011). An extended model comparison framework for covariance and mean structure models, accommodating multiple groups and latent mixtures. *Sociological Methods & Research, 40*, 256–278.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*, 1989–2001.

Li, S.-C., Lewandowsky, S., & DeBrunner, V. E. (1996). Using parameter sensitivity and interdependence to predict model scope and falsifiability. *Journal of Experimental Psychology: General, 125*, 360–369.

Li, Y., Yu, J., & Zeng, T. (2019). Deviance information criterion for latent variable models and misspecified models. *Journal of Econometrics, 216*, 450–493.

Linhart, H., & Zucchini, W. (1986). *Model selection*. New York: Wiley.

Lubke, G., & Campbell, I. (2016). Inference based on the best-fitting model can contribute to the replication crisis: Assessing model selection uncertainty using a bootstrap approach. *Structural Equation Modeling, 23*, 479–490.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38*, 113–139.

MacCallum, R. C., Roznowski, M., Mar, C. M., & Reith, J. V. (1994). Alternative strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research, 29*, 1–32.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490–504.

Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics, 34*, 593–610.

Marsh, H. W., & Balla, J. (1994). Goodness of fit in confirmatory factor analysis: The effects of sample size and model parsimony. *Quality and Quantity, 28*, 185–217.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 275–340). Mahwah, NJ: Erlbaum.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Meehl, P. E. (1986). What social scientists don't understand. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science: Pluralisms and subjectivities* (pp. 315–338). Chicago: University of Chicago Press.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108–141.

Meehl, P. E. (1992). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. *Psychological Reports, 71*, 339–467.

Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychological Methods, 21*, 151–163.

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*, 430–445.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44*, 190–204.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA, 97*, 11170–11175.

Myung, J. I., Navarro, D. J., Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology, 50*, 167–179.

Myung, I. J., & Pitt, M. A. (1998). Issues in selecting mathematical models of cognition. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 327–355). Mahwah, NJ: Erlbaum.

Myung, I. J., & Pitt, M. A. (2002). Mathematical modeling. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology: Vol IV. Methodology* (3rd ed., pp. 429–460). New York: Wiley.

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain and Behavior, 2*, 28–34.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Science, 6*, 421–425.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472–491.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41*, 227–259.

Preacher, K. J., Cai, L., & MacCallum, R. C. (2007). Alternatives to traditional model comparison strategies for covariance structure models. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 33–62). Mahwah, NJ: Erlbaum.

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

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research, 48*, 28–56.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.

Raftery, A. E. (1999). Bayes factors and BIC. *Sociological Methods & Research, 27*, 411–427.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92*, 179–191.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*, 465–471.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Hackensack, NJ: World Scientific.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42*, 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory, 47*, 1712–1717.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

Roos, T. (2008). Monte Carlo estimation of minimax regret with an application to MDL model selection. *Proceedings of the 2008 IEEE Information Theory Workshop*, pp. 284–288.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36.

Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational & Psychological Measurement, 72*, 910–932.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–343.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistics Sinica, 7*, 221–264.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50*, 253–263.

Stine, R. A. (2004). Model selection using information theory and the MDL principle. *Sociological Methods and Research, 33*, 230–260.

Su, Y., Myung, J. I., & Pitt, M. A. (2005). Minimum description length and cognitive modeling. In P. Grünwald, J. I. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (pp. 411–433). Cambridge, MA: MIT Press.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of Abnormal Psychology, 112*, 578–598.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–317). Oxford, UK: Oxford University Press.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*, 228–243.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44*, 92–107.

Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods and Research, 27*, 359–397.

Weakliem, D. L. (2016). *Hypothesis testing and model selection in the social sciences.* New York: Guilford Press.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8*, 16–37.

Williams, L. J. (2012). Equivalent models: Concepts, problems, alternatives. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 247–260). New York: Guilford Press.

Xie, Y. (1999). The tension between generality and accuracy. *Sociological Methods and Research, 27*, 428–435.

Handbook of STRUCTURAL EQUATION MODELING

second edition

edited by

RICK H. HOYLE

THE GUILFORD PRESS
New York London

Copyright © 2023 The Guilford Press
A Division of Guilford Publications, Inc.
370 Seventh Avenue, Suite 1200, New York, NY 10001
www.guilford.com

All rights reserved

No part of this book may be reproduced, translated, stored in a retrieval system,
or transmitted, in any form or by any means, electronic, mechanical, photocopying,
microfilming, recording, or otherwise, without written permission from the publisher.

Printed in the United States of America

This book is printed on acid-free paper.

Last digit is print number: 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Names: Hoyle, Rick H., editor.
Title: Handbook of structural equation modeling / edited by Rick H. Hoyle.
Description: Second edition. | New York, NY : The Guilford Press, [2023] |
Includes bibliographical references and index. |
Identifiers: LCCN 2022015675 | ISBN 9781462544646 (cloth)
Subjects: LCSH: Structural equation modeling.
Classification: LCC QA278.3 .H36 2023 | DDC 519.5—dc23
LC record available at *https://lccn.loc.gov/2022015675*