Psychological Methods
2012, Vol. 17, No. 4, 582–599

1082-989X/12/$12.00 DOI: 10.1037/a0029253

Individual Influence on Model Selection

Sonya K. Sterba
Vanderbilt University

Jolynn Pek
York University

Researchers in psychology are increasingly using model selection strategies to decide among competing models, rather than evaluating the fit of a given model in isolation. However, such interest in model selection outpaces an awareness that one or a few cases can have disproportionate impact on the model ranking. Though case influence on the fit of a single model in isolation has been often studied, case influence on model selection results is greatly underappreciated in psychology. This article introduces the issue of case influence on model selection and proposes 3 influence diagnostics for commonly used selection indices: the chi-square difference test, Bayesian information criterion, and Akaike's information criterion. These 3 diagnostics can be obtained simply from the byproducts of full information maximum likelihood estimation without heavy computational burden. We provide practical information on the interpretation and behavior of these diagnostics for applied researchers and provide software code to facilitate their use. Simulated and empirical examples involving different kinds of model comparison scenarios encountered in cross-sectional, longitudinal, and multilevel research as well as involving different kinds of outcome distributions illustrate the generality of the proposed diagnostics. An awareness of how cases influence model selection results is shown to aid researchers in understanding how representative their sample level results are at the case level.

Keywords: model selection, case influence, sensitivity analysis

Supplemental materials: http://dx.doi.org/10.1037/a0029253.supp

Psychologists are increasingly using model selection strategies to decide among competing population generating models, rather than simply evaluating the adequacy or fit of a single model in isolation (Hamaker, van Hattum, Kuiper, & Hoijtink, 2011; MacCallum, 2003; Maxwell & Delaney, 2004; Myung, Forster, & Browne, 2000; Rodgers, 2010). As no one model is true, and all models are approximations to a more complex reality (Box, 1979), the logic of a model selection strategy is to find a working model that provides a better approximation than competing alternatives (e.g., MacCallum, 2003). Although a model selection strategy has historically been more common for some statistical frameworks (e.g., structural equation modeling [SEM]), this strategy is now being recommended and applied more broadly (e.g., single-level regression, multilevel regression models [MLM], item response theory [IRT]; A. S. Cohen & Cho, in press; Hamaker et al., 2011; Kang & Cohen, 2007; Rodgers, 2010).

Popular indices used in model selection, such as chi-square difference tests or information criteria (e.g., Akaike's information criterion [AIC; Akaike, 1974] or the Bayesian information criterion [BIC; Schwarz, 1978]) determine which of the competing models is preferable at the sample level—that is, aggregating across all cases. (Here, a *case* will often be a person, but in general it is the highest level unit in an analysis.) However, all cases' data may not be best fit by the model that is selected at the sample level; such a generalization would constitute an ecological fallacy (Robinson, 1950). One model might provide a relatively better fit to one case, but the other might provide a relatively better fit to another case. Furthermore, an underappreciated issue in psychology is that the results of model selection, at the sample level, could be influenced by one or a few cases' data. For instance, selection of one model at the sample level could be driven by one or a few cases' data that strongly support that model, despite most cases' data modestly supporting the alternative model.

Traditionally, *case influence*—how a given case may impact conclusions drawn about study results (Cook, 1977, 1986)—has been evaluated for a single model at a time in psychology (e.g., Cadigan, 1995; Lee & Wang, 1996; Lee & Xu, 2003b; Zu & Yuan, 2010). Case influence is typically evaluated via a *sensitivity analysis*—a quantification of the uncertainty in statistical results due to the introduction of small, controlled changes or perturbations to data or modeling conditions (e.g., Pek & MacCallum, 2011; Tanaka, Watadani, & Moon, 1991). A popular and well-studied perturbation is a case deletion scheme.[1] When removing a case results in different conclusions from the original statistical test or

This article was published Online First July 30, 2012.

Sonya K. Sterba, Department of Psychology and Human Development, Vanderbilt University; Jolynn Pek, Department of Psychology, York University, Toronto, Ontario, Canada.

We would like to thank Robert C. MacCallum for helpful comments on a previous draft of this article.

Correspondence concerning this article should be addressed to Sonya K. Sterba, Department of Psychology and Human Development, Vanderbilt University, Peabody #552, 230 Appleton Place, Nashville, TN 37203. E-mail: Sonya.Sterba@Vanderbilt.edu

[1] Sensitivity analyses involving case deletion are often called a *global* influence approach, dating to Cook (1977), to be discriminated from *local* influence. Local influence, dating to Cook (1986), introduces an infinitesimal perturbation to the model or data set and then uses differential geometry techniques to assess the behavior of what Cook (1986) termed the *likelihood displacement function* (or other quantities) obtained from the perturbation. Local influence analyses have been performed for a variety of models with a variety of outcome distributions (e.g., Lee & Xu, 2003b; Cadigan, 1995; Poon & Poon, 2002; Zhu & Lee, 2003; Zu & Yuan, 2010). However, for an applied researcher such procedures have not yet been

index, the case is deemed *influential,* and the results are said to be sensitive to the presence of that particular case in the sample. When such influential cases are identified, data integrity may be verified or such cases may serve as interesting case studies. In the situation where no influential cases are found, confidence is gained in interpreting results. However, sensitivity analyses involving case deletion via iteratively refitting the model *N* times, omitting one case per iteration (e.g., Bruce & Martin, 1989; Cadigan, 1994; Pek & MacCallum, 2011; Rensvold & Cheung, 1999; Tanaka et al., 1991, pp. 3811–3814) can be time consuming. Hence, influence diagnostics that serve to approximate case deletion statistics without requiring iterative refitting have been developed (e.g., Cook, 1977; Lee & Lu, 2003; Pregibon, 1981; Reise & Widaman, 1999; Tanaka et al., 1991; Xu, Lee, & Poon, 2006). In particular, Reise and Widaman (1999); Tanaka et al. (1991), and Pregibon (1981) created diagnostics that approximate individual contributions to the chi-square statistic; some have been used to assess case influence on the fit of a single model in isolation.

It is important to clarify at this juncture that diagnosing case influence, the topic of this article, differs subtly in objective and approach from *outlier detection* (e.g., Draper & John, 1981; Mullen, Milne, & Doney, 1995; Stevens, 1984) as well as from what is often referred to in the IRT literature as *person- fit assessment* (e.g., Karabatsos, 2003; Meijer, 2003; Meijer & Sitjsma, 2001). *Outliers* with respect to sample characteristics can have extreme or unusual scores on predictor or outcome variables and may be identified without fitting a model (i.e., can be model-free). In contrast, case influence is evaluated with respect to fitted model(s) (i.e., always model-based). Outliers may or may not be influential with respect to, say, model fit, and influential cases may or may not be outliers. Further, employing outlier detection is typically not sufficient to determine the possibility of influential cases (e.g., Chatterjee & Hadi, 1986; Pek & MacCallum, 2011). *Person-fit assessment* (the model-based or parametric variety) often portrays individual contributions to the fit of a model and identifies cases that fit a model relatively better or worse—with a typical goal of classifying misfitting cases as aberrant test responders (e.g., cheaters, those "faking good," those unfamiliar with computer test equipment). But there is little focus on determining whether the presence/absence of designated aberrant cases would change sample level conclusions (i.e., whether they are influential), which is our concern here. Sometimes when person-fit statistics are applied, the most extreme 1% or 5% of misfitting cases are a priori designated as aberrant (e.g., Drasgow, Levine, & Zickar, 1996). Using such arbitrary cutoffs to identify cases is not consistent with our influence diagnosis goals here because there is no guarantee that any cases among the 1% or 5% will be influential with respect to sample level conclusions (e.g., overall model fit). In many analyses no cases may show influence (examples given later). Finally, different person-fit statistics have historically been used depending on the modeling framework and outcome distributions (e.g., IRT with categorical outcomes [Karabatsos, 2003; Meijer, 2003] vs. SEM with continuous outcomes [Coffman & Millsap, 2006; Reise & Widaman, 1999]), a practice that differs from the approach taken here.

Although little discussed in psychology, the potential for individual cases to influence the *ranking* of competing models has been mentioned several times in the statistics literature (e.g., Cook & Wang, 1983; Greenland, 1989; Hoeting, Raftery, & Madigan, 1996; McCann, 2006; Ronchetti, 1997; Ronchetti, Field, & Blanchard, 1997). Prior research on AIC and BIC has found them sensitive to influential cases (e.g., Atkinson & Riani, 2008; Chik, 2002; Laud & Ibrahim, 1995; Le, Raftery, & Martin, 1996), as has limited prior research on the chi-square difference test in the context of competing models (Sadray, Jonsson, & Karlsson, 1999). Still, Atkinson and Riani (2008) lamented that the sensitivity of model selection indices, such as AIC, to influential cases is an often overlooked issue:

> Professor Akaike's 1974 paper on model selection (Akaike, 1974) is one of the most highly cited papers in statistics . . . Akaike's elegant solution penalizes the maximized log-likelihood by twice the number of parameters in the model. However, the loglikelihood is an aggregate statistic, a function of all the observations. AIC provides no evidence of whether or how individual observations or unidentified structure are affecting the model choice. (p. 3)

Due to this potential for case influence on model selection, researchers are in need of user-friendly diagnostic tools capable of portraying the sensitivity of existing model selection results to influential cases. Existing diagnostics for case influence on the fit of a single model in isolation, on the parameter estimates for a single model, or on the predicted values for a single model are in no way guaranteed to identify cases that have a disproportionate impact on model ranking. Moreover, since researchers typically consider not one but multiple model selection indices when comparing models, researchers need multiple corresponding options for model selection influence diagnostics. Finally, given that psychologists currently use model selection across a wide variety of modeling frameworks (Rodgers, 2010), researchers need model selection influence diagnostics that are generally applicable to many modeling frameworks (e.g., SEM, IRT, single- or multilevel regression) and outcome distributions (e.g., binary, normal, count). Such generality would be convenient—a researcher need only master one kind of diagnostic regardless of what models are to be compared on what kind of data. Further, such generality is essential if, for example, data are to be fit with competing models assuming alternative outcome distributions (e.g., Poisson vs. negative binomial).

However, no diagnostics for case influence on model selection have been disseminated to a psychology audience. Furthermore, no diagnostics for case influence on information criteria have ever been proposed, to our knowledge. Whereas a diagnostic for case influence on the chi-square difference test has been previously mentioned for one model comparison context in the pharmacology literature (Sadray et al., 1999), its generality has not been recognized.

Consequently, the goals of this article are to develop and describe the interpretation of several influence diagnostics for popular model selection indices, which are widely applicable across modeling frameworks, and also to demonstrate their implementa-

implemented in commercial software for a flexible class of models (although specific routines have been made available for particular models; e.g., a three variable mediation model in Zu & Yuan, 2010). The global influence approach can be motivated as a special case of the local (e.g., Lee & Wang, 1996). Rather than focusing on the local influence approach, this article focuses on approximations involving the global influence approach.

tion. The remainder of this article proceeds as follows. First, we briefly review full information maximum likelihood estimation and the case-wise decomposition of the likelihood as background to later developments. Second, we briefly review three popular model selection indices: the chi-square difference test, BIC, and AIC. Third, we describe exact case deletion influence diagnostics for these three model selection indices, and we describe approximations to these exact case deletion diagnostics that do not require iterative model refitting. Fourth, we provide two application examples (one empirical and one simulated) that incorporate extensions not discussed in detail earlier: having >2 models to compare and having a case correspond to a cluster (e.g., a school). Fifth, we describe and illustrate some potential causes of case influence on model selection and some conditions under which it may be more likely. Our illustrations and examples exemplify the generality of the proposed diagnostics in that they span alternative modeling frameworks (e.g., SEM, MLM, IRT) and outcomes (categorical and continuous). Also, illustrations concern often-used model comparisons in order to link concretely to practice (e.g., one- vs. two-factor confirmatory factor analysis [CFA]; one-parameter logistic [1PL] vs. 2PL IRT; MLM with vs. without a cross level interaction; longitudinal factor analysis with different levels of across-time invariance). We conclude by providing software code to calculate the developed diagnostics and by providing recommendations regarding their application in practice.

Full-Information Maximum Likelihood Estimation

Maximum likelihood (ML) is a widely used technique for obtaining parameter estimates. The diagnostics we develop to examine case influence on model selection are predicated on ML estimation. ML requires making distributional assumptions about the observed outcome variables, as well as any latent variables (called factors in SEM; random effects in MLM; traits in IRT), if present.[2] The probability density function (pdf) for the conditional distribution of the outcome is denoted $f(.)$, and the density function for the latent variable(s), if present, is denoted $h(.)$. In a sample of cases $i = 1 \ldots N$—where, as stated earlier, a *case* is a highest level unit in the analysis—the marginal likelihood L_i for case i can be written

$$L_i(\boldsymbol{\theta}|\mathbf{Y}_i) = \int f(\mathbf{Y}_i|\mathbf{u}_i,\boldsymbol{\theta})h(\mathbf{u}_i|\boldsymbol{\theta})d\mathbf{u}_i, \quad (1)$$

where $\boldsymbol{\theta}$ is a $k \times 1$ vector of model parameters, $\mathbf{Y}_i$ is a $p \times 1$ vector of outcomes for case i, and $\mathbf{u}_i$ is their vector of random effects (also called latent variables, factors, or traits). Typically $h(.)$ is assumed to be multivariate normal. If $f(.)$ is also multivariate normal, the integral within the likelihood resolves analytically, and the marginal likelihood for $\mathbf{Y}_i$ is the multivariate normal density function. (Otherwise averaging or integrating over all possible values of the random effects is necessary to obtain the marginal L_i, because these random effects are unobserved.) Under the assumption that the N cases are independent and identically distributed, the sample likelihood L is the product of the casewise likelihoods. However, to improve computational stability its log is typically taken, making the sample loglikelihood, $\ln L$, the sum of the casewise loglikelihoods.

$$\ln L(\boldsymbol{\theta}|\mathbf{Y}) = \sum_{i=1}^{N} \ln L_i(\boldsymbol{\theta}|\mathbf{Y}_i). \quad (2)$$

Historically, ML estimation was conducted by optimizing monotonic transformations of the sample loglikelihood based on sufficient statistics. The use of sufficient statistics—which summarize all relevant information contained in the data about the parameters $\boldsymbol{\theta}$—in place of each case's raw data was required due to computational limitations. Increased computing power now allows ML estimation to be conducted on raw data (termed full-information maximum likelihood, FIML, direct ML, or raw ML; e.g., Neale, Boker, Xie, & Maes, 2003). Advantages of FIML over traditional ML algorithms include the ability to account for missing data (Arbuckle, 1996) and the opportunity to compute casewise contributions to the loglikelihood (e.g., Lange, Westlake, & Spence, 1976; McArdle, 1997; Neale, 2000). We illustrate these properties for the situation where $f(.)$ and $h(.)$ are normal, and in this example, $\ln L_i$ resolves to the multivariate normal pdf:

$$\ln L_i = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i(\boldsymbol{\theta})| - \frac{p_i}{2}\ln(2\pi) - \frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}))'\boldsymbol{\Sigma}_i(\boldsymbol{\theta})^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})). \quad (3)$$

Here p_i denotes the number of outcome variables present for case i, and $\mathbf{Y}_i$ has dimension $p_i \times 1$. $\boldsymbol{\mu}_i(\boldsymbol{\theta})$ is a $p_i \times 1$ model implied mean vector and $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$ is a $p_i \times p_i$ model implied covariance matrix. $\boldsymbol{\mu}_i(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$ can be thought of as resulting from deleting elements of $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where case i has missing outcomes. The FIML sample $\ln L$ is the sum of these casewise $\ln L_i$.

Three Model Selection Indices

After competing models are estimated using FIML, focus typically turns to comparing the fit of these models and selecting a final model. We next briefly review three indices often used for this purpose: the chi-square difference test, BIC, and AIC. Whereas BIC and AIC are rooted in a model selection tradition (Burnham & Anderson, 2002), the chi-square difference test can be considered as coming from a distinct hypothesis testing tradition. However, we are using all three criteria for the same goal in this article: to rank models. Thus, in line with Maxwell and Delaney (2004) and Rodgers (2010), we deemphasize this traditional distinction and consider all three united in their current purpose as indices for model selection.

Chi-Square Difference Test

Recall that a chi-square test can be used to evaluate fit for a single hypothesized model, as reviewed first. More generally, pertaining to the topic of this article, a chi-square difference test can be used to compare fit of alternative models, as reviewed second. In the context of a single hypothesized model, the null

[2] Distributional assumptions about observed predictors may also be made, and in some cases distributional assumptions of latent variables can be partially relaxed.

hypothesis of the chi-square test is that this model fits the data perfectly, or that model-implied moments exactly reproduce sample moments. This test is also known as a likelihood ratio test (LRT) as it has the form

$$\chi^2 = -2\ln(L/L^S) = -2[\ln L - \ln L^S], \quad (4)$$

where $\ln L$ is the estimated loglikelihood based on the hypothesized model and $\ln L^S$ is the loglikelihood based on a saturated model—one that perfectly reproduces the moments of the sample data.[3] For instance, in our multivariate normal example, $\ln L^S$ would be obtained by substituting observed means, denoted $\bar{\mathbf{Y}}_i$, and observed (co)variances, denoted $\mathbf{S}_i$, in place of their model-implied counterparts, $\boldsymbol{\mu}_i(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$, in Equation 3. The χ^2 statistic will equal 0 when the ratio (L/L^S) is 1 (i.e., when the hypothesized model fits as well as the saturated model) and will be asymptotically central chi-square distributed under the null hypothesis with degrees of freedom (*df*) equal to the difference in the number of estimated parameters between the hypothesized and saturated models. For the critical value (χ^2_{crit}) determined by *df* and the desired α, the null hypothesis can be rejected when the obtained $\chi^2 > \chi^2_{crit}$.

Although saturated models can provide useful information in the search for a more parsimonious model, other competing models may be used in place of the saturated model. Hence, we now consider the more general situation where we wish to compare two competing models (A and B); these two models may be the only ones under consideration or may be part of a larger set of competing models (discussed later). Suppose Model A is *nested* in Model B such that, for example, imposing equality constraints or fixing some free parameters in Model B yields Model A. Since Model B is less restricted than A, its loglikelihood, $\ln L^B$, is necessarily $\geq$ the loglikelihood of Model A, $\ln L^A$. The test of perfect fit may therefore be extended to a test comparing the fit of competing Models A and B.

$$\Delta\chi^2 = -2\ln(L^A/L^B) = -2[\ln L^A - \ln L^B] = \chi^2_A - \chi^2_B \quad (5)$$

The null hypothesis is that there is no difference in fit between Models A and B in the population, a circumstance corresponding with a L^A/L^B ratio of 1 and a $\Delta\chi^2$ of 0. Under this null hypothesis, the $\Delta\chi^2$ statistic is asymptotically distributed as a central chi-square with *df* defined as $df^A - df^B = \Delta df$. For the χ^2_{crit} determined by Δdf and α, the null hypothesis is rejected when $\Delta\chi^2 > \chi^2_{crit}$. Rejecting the null implies that Model B fits the data significantly better than A, so B is selected. Failing to reject the null implies A fits as well as B, so A may be retained as it is more parsimonious.

Penalized Model Selection Criteria

When selecting among competing models, researchers may consider criteria other than $\Delta\chi^2$ for several reasons. First, $\Delta\chi^2$ requires competing models to be nested. Second at large *N*, $\Delta\chi^2$ becomes increasingly sensitive to small deviations from the null hypothesis, favoring more complex models. Penalized model selection criteria such as BIC and AIC overcome these limitations (Kuha, 2004). Suppose competing models A and B *may or may not* be nested.[4] For this situation, penalized model selection criteria have the form

$$-2[\ln L^A - \ln L^B] + q(k^A - k^B). \quad (6)$$

Since the number of free parameters can be regarded as one indicator of complexity, the difference in free parameters between the more and less restricted models (here, $k^A - k^B$) portrays the relative complexity of Model A versus B. Hence, the second term may be regarded as a penalty afforded to more complex models. Here, q is a known multiplicative factor by which we want to weight the contribution of complexity (AIC and BIC correspond with different values of q, as discussed shortly). When the penalized model selection criterion in Equation 6 is negative, Model A is selected over B, and vice versa when the criterion is positive. Notice that when nested models are involved, the first term is the $\Delta\chi^2$, and the second term is a constant. Whether competing models are nested, the difference of the loglikelihoods reflects the relative fit of the models to the data. Similar to the $\Delta\chi^2$ LRT, the first term in Equation 6 tends to favor more complex models.

BIC. The BIC is motivated to select the true model from a set of competing models (see, e.g., Wagenmakers & Farrell, 2004, for other properties). For a single model, $\text{BIC} = -2\ln L + k \ln N$ (lower is better). In the context of two competing models, the penalized model selection criterion for BIC is obtained when $q = \ln N$:

$$\Delta\text{BIC} = -2[\ln L^A - \ln L^B] + \ln N(k^A - k^B), \quad (7)$$

which is the between model difference in BIC. Negative ΔBIC indicates Model A is more likely to be the true model (or, closer to the true model) compared with B; the opposite conclusion is drawn for positive ΔBIC. There are different approaches to interpreting the magnitude of ΔBIC (Burnham & Anderson, 2002); a common one is based on the fact that ΔBIC is an approximation to a *Bayes factor*—a measure of the evidence provided by the data in favor of one model against the other (Raftery, 1995). For instance, the degree of evidence for Model B over A could be labeled "weak" if $0 < \Delta\text{BIC} \leq 2$, "positive" if $2 < \Delta\text{BIC} \leq 6$, "strong" if $6 < \Delta\text{BIC} \leq 10$, and "very strong" if $\Delta\text{BIC} > 10$. The negative equivalent would be used to describe the degree of evidence for Model A over B, "weak" if $-2 \leq \Delta\text{BIC} < 0$, "positive" if $-6 \leq \Delta\text{BIC} < -2$, etc.

AIC. An aim of model selection using AIC is to choose the most generalizable model. AIC is closely related to other measures of predictive validity (Stone, 1977), such as the expected cross-validation index (ECVI; Browne & Cudeck, 1992). For a single model, $\text{AIC} = -2\ln L + 2k$ (lower is better). When two competing models are assessed using the AIC, the penalized model selection criterion for AIC is obtained when $q = 2$.

$$\Delta\text{AIC} = -2[\ln L^A - \ln L^B] + 2(k^A - k^B) \quad (8)$$

Negative values of ΔAIC indicate that Model A has more predictive validity than (or cross-validates better than) Model B; vice versa for positive values of ΔAIC. Additionally, it can be shown that ΔECVI (the difference in ECVI for Models A and B) would only differ by a constant $(1/N)$ from ΔAIC. There are also different approaches to interpreting the magnitude of ΔAIC (Wagenmakers & Farrell, 2004). For instance, Burnham and Anderson (2002) suggested guidelines, based on simulation results, that involve determining the model with the best AIC and then subtracting its

[3] A saturated model is not always available, as is discussed later.

[4] Later in this article, when we compute penalized selection criteria for nested models, Model A corresponds with the more restricted model.

AIC from the AIC of poorer fitting model(s). Differences between 0 and 2 suggest that the poorer fitting model still retains substantial support; differences >10 suggest the poorer fitting model has essentially no support. However, these guidelines are subject to a variety of qualifications involving *N,* nestedness, and other issues (Burnham & Anderson, 2002, pp. 131, 170) and are not considered further here.

Exact Case Deletion Influence Diagnostics for Model Selection

One approach for assessing case influence on model selection would be to iteratively remove each case, one at a time, and calculate the change in the sample level selection index associated with deleting a case. That is,

$$\Delta\chi_i^2 = \Delta\chi^2 - \Delta\chi^2_{(-i)}, \tag{9}$$

$$\Delta\text{BIC}_i = \Delta\text{BIC} - \Delta\text{BIC}_{(-i)}, \tag{10}$$

$$\Delta\text{AIC}_i = \Delta\text{AIC} - \Delta\text{AIC}_{(-i)}, \tag{11}$$

where the subscript $(-i)$ denotes "calculated after empirically removing case i from the sample." Using this approach, for the diagnostic $\Delta\chi_i^2$ in Equation 9 we define an influential case with respect to the $\Delta\chi^2$ test as one whose presence/absence could alter the sample level decision about rejecting the null hypothesis. Specifically, denote $d = (\Delta\chi^2 - \chi^2_{crit})$. When $d > 0$, a researcher would diagnose case i as influential for the $\Delta\chi^2$ test if $\Delta\chi_i^2 > d$. When $d \leq 0$, case i would be influential if $\Delta\chi_i^2 < d$.

For the diagnostic ΔBIC_i in Equation 10, we define case influence to mean that the presence/absence of the case could alter the *sign* of ΔBIC (and thus model ranking) at the sample level, and/or could alter the *magnitude* of ΔBIC enough to change the designated degree of evidence for a given model (in terms of Bayes factors; Raftery, 1995) at the sample level. If ΔBIC is positive, influence on its sign would require: $\Delta\text{BIC}_i > \Delta\text{BIC}$. If ΔBIC is negative, influence on its sign would require: $\Delta\text{BIC}_i < \Delta\text{BIC}$. Consider an example in which ΔBIC is –2, meaning Model A is "weakly" preferable to B at the sample level. Case i with $\Delta\text{BIC}_i = -2.5$ would have influence on the sign of ΔBIC, since it would alter the sample level model ranking (making B "weakly" preferable to A) if removed. An example of case influence on magnitude but not sign would be if ΔBIC was –2 and ΔBIC_i was +1. Case i would simply alter the degree of evidence for Model A (from "weak" to "positive") if removed.

Finally, it will be useful for later developments to note how ΔBIC_i is calculated in terms of $\Delta\chi_i^2$:

$$\begin{aligned}\Delta\text{BIC}_i &= (-2[\ln L^A - \ln L^B] + (\ln N)(k^A - k^B)) - (-2[\ln L^A_{(-\text{i})} \\ &\quad - \ln L^B_{(-i)}] + (\ln (N-1))(k^A - k^B)) \\ &= (-2[\ln L^A - \ln L^B] - (-2[\ln L^A_{(-i)} - \ln L^B_{(-i)}]) \\ &\quad + ((\ln N)(k^A - k^B) - (\ln (N-1))(k^A - k^B)) \\ &= (\Delta\chi_i^2) + ((\ln N)(k^A - k^B) - (\ln (N-1))(k^A - k^B)) \\ &= (\Delta\chi_i^2) + (k^A - k^B)\ln (N/(N-1)). \end{aligned} \tag{12}$$

For the diagnostic ΔAIC_i in Equation 11, we define case influence on sign (ranking) as occurring when the presence/absence of the case alters the sign of ΔAIC at the sample level. For positive ΔAIC, influence on its sign would require $\Delta\text{AIC}_i > \Delta\text{AIC}$, whereas for negative ΔAIC, influence on its sign would require $\Delta\text{AIC}_i < \Delta\text{AIC}$. It would be also possible to determine influence on degree of evidence for a given model according to AIC using, for instance, the Burnham and Anderson (2002) guidelines mentioned earlier, but that is not pursued here. Additionally, given the close relationship between ΔAIC and ΔECVI, a researcher could instead choose to use $\Delta\text{ECVI}_i = (1/N)\Delta\text{AIC}_i$. Finally, it can be seen that another way of calculating ΔAIC_i is in terms of $\Delta\chi_i^2$:

$$\begin{aligned}\Delta\text{AIC}_i &= (-2[\ln L^A - \ln L^B] + 2(k^A - k^B)) \\ &\quad - (-2[\ln L^A_{(-i)} - \ln L^B_{(-i)}] + 2(k^A - k^B)) \\ &= (-2[\ln L^A - \ln L^B]) - (-2[\ln L^A_{(-i)} - \ln L^B_{(-i)}]) \\ &= \Delta\chi_i^2. \end{aligned} \tag{13}$$

Since $\Delta\text{AIC}_i = \Delta\chi_i^2$ in Equation 13, these two diagnostics do not differ in their calculation; however, they do differ in their implementation and interpretation. If researchers want to assess influence with respect to $\Delta\chi^2$—or goodness of fit between models—they need to compare case i's diagnostic value with d. On the other hand, if researchers want to assess influence with respect to ΔAIC—or cross-validity between models—they need to compare case i's diagnostic value with ΔAIC.

Whereas it would be possible to compute the case deletion influence diagnostics in Equations 9–11 exactly, this would require N (jackknife) iterative refittings of Model A (plus its full-sample solution) and N (jackknife) iterative refittings of Model B (plus its full-sample solution) for each model comparison. For complex models and/or models without closed-form likelihood expressions (e.g., CFAs with categorical outcomes), $2(N + 1)$ fittings (i.e., one time per model for the full sample; N times per model for the delete-one samples) would potentially be prohibitively time consuming. In fact, in the context of evaluating a single model in isolation, Lee and Xu (2003a) argued that standard case deletion diagnostics would be intractable for CFAs with categorical outcomes. In contrast, as illustrated in a later empirical example, our approximation diagnostics allow us to perform a sensitivity analysis for selecting between alternative three-timepoint longitudinal factor analysis models with categorical outcomes.

Approximate Case Influence Diagnostics for Model Selection

It would be useful to have diagnostics for case influence on model selection that do not require time consuming, potentially computationally intractable iterative model refitting. We develop such non-computationally-intensive selection diagnostics as extensions of an existing diagnostic, termed $\text{ind}_{\text{CHI}_i}$, which applies to a single model in isolation. The $\text{ind}_{\text{CHI}_i}$ is reviewed first.

In the context of a single model in isolation, Reise and Widaman (1999) and Coffman and Millsap (2006) proposed decomposing the χ^2 test of perfect fit for a given model in isolation to obtain case-specific contributions to the χ^2, which they termed:

$$\text{ind}_{\text{CHI}_i} = -2\ln (L_i/L_i^S). \tag{14}$$

(Here, we use ind to stand for index.) $\text{ind}_{\text{CHI}_i}$ sum to the sample chi-square statistic:

$$\chi^2 = \sum_{i=1}^{N} \text{ind}_{\text{CHI}_i}. \quad (15)$$

Although χ^2 is bounded below by 0, a case's contribution is not. Reise and Widaman (1999) and Coffman and Millsap (2006) used $\text{ind}_{\text{CHI}_i}$ only as a person-fit statistic, not in the context of influence detection. If case i has a positive value of $\text{ind}_{\text{CHI}_i}$, this means that its presence in the sample worsens overall model fit. Conversely, if case i has a negative $\text{ind}_{\text{CHI}_i}$, its presence in the sample improves overall model fit.

Although this has not been previously noted, for a single sample in isolation $\text{ind}_{\text{CHI}_i}$ is an approximation to its case deletion counterpart $\chi^2_i = \chi^2 - \chi^2_{(-i)}$ at the point where $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{full}$—that is, where all model parameters are fixed at their values in the full-sample analysis (more on this later). Hence, $\text{ind}_{\text{CHI}_i}$ can be used as an approximate case deletion influence diagnostic for a single model in isolation, without the need for iterative refitting. Specifically, we can define an influential case with respect to the χ^2 test for a single model in isolation as one whose presence/absence could alter the decision about whether to reject the null hypothesis at the sample level. Denote $d = (\chi^2 - \chi^2_{crit})$. When $d > 0$, case i would be flagged as influential for the χ^2 test if $\text{ind}_{\text{CHI}_i} > d$. Conversely, when $d \leq 0$, case i would be flagged as influential if $\text{ind}_{\text{CHI}_i} < d$. For instance, if the sample $\chi^2(7) = 15.96$ and $\chi^2_{crit}(7) = 14.07$ for $\alpha = .05$, a researcher would conclude that the hypothesis of perfect fit can be rejected ($p < .05$). Since $d = 1.89$, a case with $\text{ind}_{\text{CHI}_i} > 1.89$ would be flagged—this case's absence could potentially lead to the opposite conclusion, that perfect fit cannot be rejected. Confirmation of a flagged case's influential status via case deletion is recommended because $\text{ind}_{\text{CHI}_i}$ serves only as an approximation to χ^2_i. In other words, the approximation $\text{ind}_{\text{CHI}_i}$ can be a useful screener, allowing calculation of χ^2_i only for flagged cases.

$\Delta\text{ind}_{\text{CHI}_i}$ as an Approximate Model Selection Influence Diagnostic

Now we extend this approximate diagnostic for a single model, $\text{ind}_{\text{CHI}_i}$, to the context of model selection. We can determine case i's relative contribution to $\Delta\chi^2$, denoted $\text{ind}_{\text{CHI}_i}$, by replacing its contribution to the saturated likelihood in Equation 14 with its contribution to a competing model likelihood:

$$\Delta\text{ind}_{\text{CHI}_i} = -2\ln(L_i^A/L_i^B) = \text{ind}^A_{\text{CHI}_i} - \text{ind}^B_{\text{CHI}_i}. \quad (16)$$

Just as the $\text{ind}_{\text{CHI}_i}$ summed to the sample χ^2 for testing perfect fit of the hypothesized model, the $\Delta\text{ind}_{\text{CHI}_i}$ sum to the sample $\Delta\chi^2$ between competing models

$$\Delta\chi^2 = \sum_{i=1}^{N} \Delta\text{ind}_{\text{CHI}_i}. \quad (17)$$

Unlike the sample $\Delta\chi^2$, $\Delta\text{ind}_{\text{CHI}_i}$ can be positive or negative. Recall that Model A is nested in B. Descriptively, positive $\Delta\text{ind}_{\text{CHI}_i}$ indicates that case i is relatively better fit by the less restricted Model B. Negative $\Delta\text{ind}_{\text{CHI}_i}$ indicates that case i is relatively better fit by the more restricted Model A. Near-zero $\Delta\text{ind}_{\text{CHI}_i}$ indicates that case i is about equally consistent with both competing models.

$\Delta\text{ind}_{\text{CHI}_i}$ serves as an approximation of its exact case deletion counterpart, $\Delta\chi^2_i$, at the point where $\boldsymbol{\theta}_{\text{A}} = \hat{\boldsymbol{\theta}}_{A\,full}$ and $\boldsymbol{\theta}_{\text{B}} = \hat{\boldsymbol{\theta}}_{B\,full}$—that is, where all model parameters are fixed to their values in the full sample analyses for Models A and B. In other words, $\Delta\text{ind}_{\text{CHI}_i}$ affords us an approximation of how much the sample level selection index would change if we removed case *i,* without us actually having to remove case *i.* To illustrate the relationship between $\Delta\text{ind}_{\text{CHI}_i}$ and $\Delta\chi^2_i$, in Figure 1 the approximation diagnostic was plotted against its exact case deletion counterpart (computed iteratively for all cases) for each of the three simulated model comparisons considered in detail later. These three comparisons are a main-effect-only versus interactive MLM example, a 1PL versus 2PL IRT illustration (for one sample), and a one- versus two-factor normal-theory CFA example.

In all three examples in Figure 1, the correlation between $\Delta\text{ind}_{\text{CHI}_i}$ and $\Delta\chi^2_i$ was $\geq .99$. Similarly high correlations were found between $\Delta\chi^2_i$ and $\Delta\text{ind}_{\text{CHI}_i}$ in Sadray et al. (1999) for a nonlinear mixed model. Although $\Delta\text{ind}_{\text{CHI}_i}$ and $\Delta\chi^2_i$ are strongly linearly associated, Figure 1 shows that $\Delta\text{ind}_{\text{CHI}_i}$ still does not exactly equal $\Delta\chi^2_i$ due to approximation error (as also described in later examples).[5] Nevertheless, their near-1.0 correlation implies that cases' rank order on $\Delta\text{ind}_{\text{CHI}_i}$ will very closely correspond with cases' rank order on $\Delta\chi^2_i$; this close correspondence in ranking was confirmed by calculating Kendall's Tau-b for $\Delta\text{ind}_{\text{CHI}_i}$ and $\Delta\chi^2_i$ in each model comparison. Kendall's Tau-b = .99, .94, and .97 for Figures 1A, 1B, and 1C, respectively. Closeness of rank order is important because it means that when the sample level $\Delta\chi^2$ is significant [nonsignificant], the case with the most positive [negative] $\Delta\text{ind}_{\text{CHI}_i}$ is likely the case with greatest potential for influence according to $\Delta\chi^2_i$. Indeed, if that case did show influence, the sensitivity of the sample level model ranking to an individual case contribution would already be apparent.

Relatedly, our recommendation is for researchers to use $\Delta\text{ind}_{\text{CHI}_i}$ as a screening tool to flag cases that could be influential for the sample level decision about rejecting the null hypothesis. For instance, as stated earlier for $d = (\Delta\chi^2 - \chi^2_{crit})$, when $d > 0$, a researcher would flag case i as influential for the $\Delta\chi^2$ test if

[5] This approximation error stems from the fact that when computing $\Delta\chi^2_i$ parameters are allowed to change from their estimates in the full-N analysis (which produced $\Delta\chi^2$) to their reestimates in the $N - 1$ analysis (which produced $\Delta\chi^2_{(-i)}$ in Equation (9)); whereas, when computing $\Delta\text{ind}_{\text{CHI}_i}$ parameters are held at their estimates from the full-N analysis. A full-scale simulation portraying $\Delta\text{ind}_{\text{CHI}_i}$'s approximation quality under diverse conditions is outside the scope of this article; however, one degenerate special circumstance can be noted in which the approximation does not result in the strong positive linear association in Figure 1. This degenerate circumstance is unlikely to be seen in practice. When the likelihood for Model A *exactly* equals the likelihood for Model B *in the sample,* all $\Delta\text{ind}_{\text{CHI}_i} = 0$ (i.e., cases equally support both models in the full-N analysis), however $\Delta\chi^2_i$ can take on a variety of values other than 0 (because sample likelihoods for A and B may diverge in an $N - 1$ analysis, thus allowing cases to support one model over another). This situation is unlikely to be seen in practice because even if the null is true in the population it is unlikely that $\Delta\chi^2 = 0$ in the sample (indeed, its expectation is Δdf).

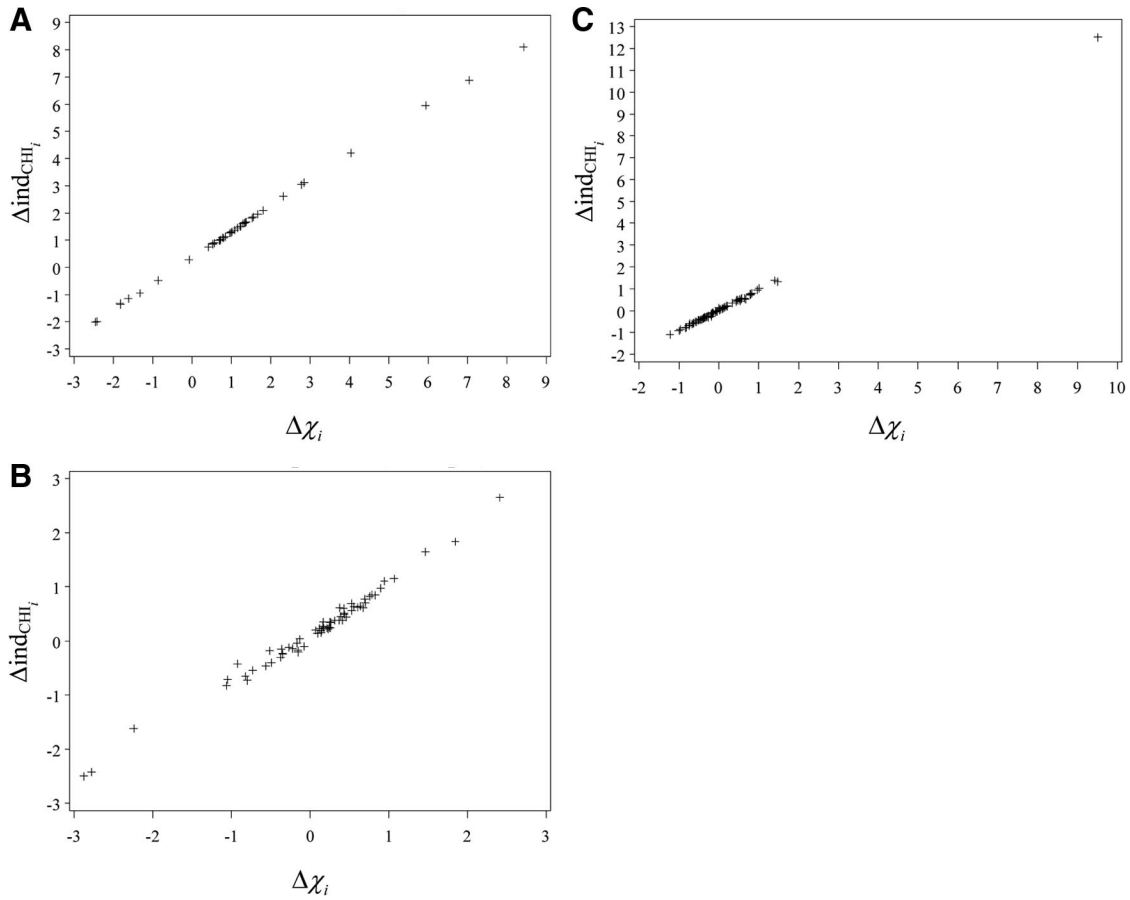

Figure 1. Approximate $\Delta\text{ind}_{\text{CHI}_i}$ versus case deletion $\Delta\chi_i^2$ for three simulated model comparisons discussed later. A. Multilevel regression model (MLM) model comparison. B. Item response theory (IRT) model comparison (one sample). C. Confirmatory factor analysis (CFA) model comparison 1.

$\Delta\text{ind}_{\text{CHI}_i} > d$. When $d \leq 0$, case i would be flagged as influential if $\Delta\text{ind}_{\text{CHI}_i} < d$. Then, for a flagged case only, exact case diagnostics in Equation 9 can be computed to confirm (or disconfirm) suspected influence. Accordingly, for our later simulated and empirical examples, we report our approximation influence diagnostics for all cases (via index plots or the equivalent) and then also provide exact counterparts just for a potential influential case.

It is worth noting that case influence on model selection can occur via different patterns of $\Delta\text{ind}_{\text{CHI}_i}$. For instance, a statistically significant sample $\Delta\chi^2$ might arise because the less restricted model fit one or a few cases' data much better, despite the more restricted model fitting many cases' data modestly better. Or, a nonsignificant sample $\Delta\chi^2$ might arise because one or a few cases' data are much better fit by the more restricted model, although the majority of cases' data are modestly better fit by less restricted model. Another possibility is that one or a few cases' data are much better fit by the more restricted model, and one or a few cases' data are much better fit by the less restricted model, but for the vast majority of cases' data either model is suitable. In this scenario the two sets of influential cases might effectively counterbalance each other, leading to a nonsignificant sample $\Delta\chi^2$.

$\Delta\text{ind}_{\text{BIC}_i}$ and $\Delta\text{ind}_{\text{AIC}_i}$ as Approximate Model Selection Influence Diagnostics

Researchers using alternatives to $\Delta\chi^2$ for model selection with nested *or nonnested* models need a non-computationally-intensive approach for determining how cases influence overall model ranking. To fulfill this need, we approximate the exact case deletion influence diagnostics from Equations 12 and 13 by replacing $\Delta\chi_i^2$ with $\Delta\text{ind}_{\text{CHI}_i}$ in each formula. This yields

$$\Delta\text{ind}_{\text{BIC}_i} = (\Delta\text{ind}_{\text{CHI}_i}) + (k^A - k^B)\ln(N/(N-1)), \quad (18)$$

$$\Delta\text{ind}_{\text{AIC}_i} = (\Delta\text{ind}_{\text{CHI}_i}). \quad (19)$$

Descriptively, for a particular case $\Delta\text{ind}_{\text{BIC}_i}$ and $\Delta\text{ind}_{\text{AIC}_i}$ can be negative (favoring Model A) or positive (favoring Model B).[6] Also, given the aforementioned relation between ΔECVI_i and ΔAIC_i, we can approximate ΔECVI_i as $\Delta\text{ind}_{\text{ECVI}_i} = (1/N)\Delta\text{ind}_{\text{AIC}_i}$. The approximation adequacy of $\Delta\text{ind}_{\text{CHI}_i}$ for $\Delta\chi_i^2$

[6] Note that $\Delta\text{ind}_{\text{BIC}_i}$ or $\Delta\text{ind}_{\text{AIC}_i}$ will not sum across i to their sample level statistic ΔBIC or ΔAIC, unlike $\Delta\text{ind}_{\text{CHI}_i}$, which sums to $\Delta\chi^2$.

(discussed earlier) will be the same as the approximation adequacy of Δind_{AIC_i} for ΔAIC_i or the approximation adequacy of Δind_{BIC_i} for ΔBIC_i, because these formulas differ by at most a constant. Although proposed diagnostics (Δind_{CHI_i}, Δind_{BIC_i} and Δind_{AIC_i}) entail some approximation error, and although exact (jackknife) case deletion statistics are theoretically available (Equations 9–11), the latter will be computationally impractical in many settings for applied researchers. Hence, we consider the proposed diagnostics as easy-to-use screeners, whose accuracy can be verified by focused application of case deletion statistics to flagged cases.

Specifically, definitions of case influence on model ranking, or on degree of evidence for a model, that were provided earlier for exact case deletion statistics ΔBIC_i and ΔAIC_i can be applied to flag potential influential cases using Δind_{AIC_i} and Δind_{BIC_i}. For instance, cases can be flagged as potentially influential on model ranking for positive ΔBIC when $\Delta ind_{BIC_i} > \Delta BIC$; for positive ΔAIC when $\Delta ind_{AIC_i} > \Delta AIC$; for negative ΔBIC when $\Delta ind_{BIC_i} < \Delta BIC$; and for negative ΔAIC when $\Delta ind_{AIC_i} < \Delta AIC$. Influence of flagged cases can be confirmed by computing exact case deletion statistics. Like $\Delta\chi^2_i$ and ΔAIC_i, their approximation counterparts Δind_{CHI_i} and Δind_{AIC_i} differ not in their calculation but in their implementation and interpretation. Case i's value of Equation 19 is compared with d to assess influence on the models' relative goodness of fit but compared with ΔAIC to assess influence on the models' relative predictive validity (illustrations given later).

Comparing ind_{CHI_i}, Δind_{CHI_i}, Δind_{AIC_i}, and Δind_{BIC_i}

It is worth noting that Δind_{CHI_i} is more generally applicable than ind_{CHI_i}, in several respects. For example, χ^2 tests (and thus ind_{CHI_i}) cannot be implemented when there is no saturated model (e.g., for MLMs in the presence of unbalanced data or SEMs in the presence of considerable missing data that may result in zero covariance coverage for some cells in $\mathbf{S}_i$; Bollen & Curran, 2006). Also, χ^2 (and thus ind_{CHI_i}) is not recommended for IRT models where the number of outcome variables is modest or large, as in this situation χ^2 does not have the appropriate null distribution (e.g., Jöreskog & Moustaki, 2001; Reise & Widaman, 1999). However, likelihood ratio $\Delta\chi^2$ (and thus Δind_{CHI_i})[7] is applicable under these circumstances. Indeed, likelihood ratio $\Delta\chi^2$ for comparing competing models has often been recommended when there are many categorical outcomes (e.g., when comparing 1PL vs. 2PL IRT models, or unidimensional vs. bifactor IRT models; Embretson & Reise, 2000; Reise, Widaman, & Pugh, 1993; Swaminathan, Hambleton, & Rogers, 2007; Thissen, Steinberg, & Gerrard, 1986), as it does not suffer from as severe limitations regarding number of variables. To our knowledge Δind_{CHI_i} has been applied only once, to a nonlinear mixed model (Sadray et al., 1999). Δind_{CHI_i}'s generality in comparing models from different frameworks has not been recognized. Δind_{CHI_i}'s generality is useful because, in the past, assessing influence or person-fit for competing models from different frameworks involved employing separate diagnostics to each model (Reise & Widaman, 1999). This process would make it difficult to tell if different persons are identified as influential across models due to (a) the differential suitability of the alternative models for particular persons or (b) the differential performance of alternative diagnostics themselves.

Further insight is obtained by comparing the behavior of Δind_{CHI_i}, Δind_{AIC_i}, and Δind_{BIC_i} in the situation where competing models are nested. Comparing Equations 16, 18, and 19, Δind_{BIC_i} will be more different from Δind_{CHI_i} or Δind_{AIC_i}, when Δdf is larger and N is smaller. Researchers can note if particular cases are influential according to some but not all diagnostics. Such inconsistency can be viewed in light of the different purposes of the indices; for instance, a given case might be influential with respect to "generalizability" of the selected model (i.e., how well it cross-validates: AIC results), but not with respect to selection of the "true" model (e.g., BIC results). In general, *index plots*—scatterplots of case ID number against case diagnostic value (either Δind_{CHI_i}, Δind_{AIC_i}, or Δind_{BIC_i})—can aid in visualizing potential influential cases. When employing all three diagnostics, it would not be necessary to make three separate index plots. For instance, an index plot of Δind_{CHI_i} (Δind_{AIC_i}) could be reported, together with the parsimony correction factor for Δind_{BIC_i}, which indicates the constant by which all points would be negatively shifted to yield a Δind_{BIC_i} index plot. Beyond using Δind_{CHI_i}, Δind_{AIC_i}, and Δind_{BIC_i} for detecting influence on sample level results, it may be of interest in some clinical or educational applications to descriptively report the model ranking for a particular case, and/or report the percentage of cases whose Δind_{AIC_i}, Δind_{BIC_i}, or Δind_{CHI_i} favor a particular model.

Example Applications and Extensions

Next we demonstrate the application and interpretation of the proposed diagnostics in two examples (one empirical and one simulated). These examples' model comparison settings concern practically useful extensions of those considered thus far: (a) influence diagnosis for >2 models under comparison and (b) influence diagnosis when a case is a cluster (e.g., school). In each example, we consider whether sample level conclusions are dependent on one or a few cases.

Example 1: More Than Two Models to Be Compared

Until this point, we have been concerned with the comparison of only two models, labeled A and B. Often, a researcher will need to compare several models (A, B, C, etc.). These alternative models can be arranged into a sequence of pairwise model comparisons; we already do this explicitly when performing chi-square difference tests but usually only do this implicitly when comparing AIC and BIC values (Kuha, 2004). Sometimes all possible pairwise comparisons are of interest (e.g., A vs. B, A vs. C, B vs. C). Other

[7] Note that all diagnostics described here can be applied in the context of missing data. Missing data have been known to interfere with identification of aberrant cases (e.g., cheaters) for a single model in isolation in the person-fit literature (Neale, 2000). However, our goals here are different; our purpose is assessing each case's potential for influence on sample level conclusions *given* that case's available data. We are not, for instance, concerned with what a case's potential for influence would have been if the data had been complete. So, even if a particular case's potential for influence on model selection would have been different in a complete data set, for the data set at hand, application of our diagnostics provides a researcher with an accurate assessment of whether any cases are influential *for their chosen model comparison*, *conditional on their available data*.

times a subset of possible comparisons are of interest. The case influence diagnostics proposed here can be applied to each pair of models under consideration. It thus would be possible to find that a given case is influential for all, none, or some of the pairwise model comparisons considered. We next empirically illustrate the application and interpretation of the developed model selection diagnostics when more than two models are to be compared.

This example concerns the assessment of longitudinal measurement invariance (MI) of ordinal items on a unidimensional scale using categorical longitudinal factor analysis, where several models are to be compared. Longitudinal MI testing addresses the question, "Are we measuring the same construct across time?" Such testing generally involves comparing nested models that impose increasingly restrictive stages of invariance (Millsap, 2010). In the context of categorical longitudinal factor analysis, these nested models can include (Millsap & Yun-Tein, 2004):[8] noninvariant loadings and thresholds over time (Model B), invariant loadings but noninvariant thresholds over time (Model A), and invariant loadings and thresholds over time (Model C). Model C is nested in A, and A is nested in B. Typically all possible pairwise comparisons are not made in the MI context; rather, adjacent nested models are compared, moving from lesser to greater invariance (B vs. A, A vs. C). The sequence of model comparisons is stopped before completion when invariance is rejected. The conclusion traditionally drawn at that point is that at least some item parameters do differ across time, in that sample. However, another possibility raised in the IRT person-fit literature in related contexts (e.g., Johanson & Alsmadi, 2002; Meade, Ellington, & Craig, 2004) is that there are one or a few persons whose item parameters are noninvariant (perhaps due to data/coding errors, chance, population heterogeneity, etc., as discussed later) but for most persons, item parameters are invariant over time. This possibility motivates a sensitivity analysis for case influence on model selection in invariance testing.

For this empirical example, our data set contains $N = 599$ girls from the National Institute of Child Health & Human Development Study of Early Child Care, whose internalizing behavior was evaluated with the Child Behavior Checklist (CBCL; Achenbach, 1991, 1992) at three repeated measurements: ages 24, 36, and 54 months. Eight internalizing symptoms from the CBCL[9] served as indicators of a unifactorial internalizing construct at each of the three timepoints. Each CBCL item has three ordered categories: 0 = not true; 1 = sometimes true; 2 = often true. Due to the use of these categorical outcomes, $f(.)$ in Equation 1 is a multinomial probability mass function (for more details, see Bauer & Hussong, 2009; Moustaki, Jöreskog, & Mavridis, 2004), whereas it is assumed that $h(.)$ is a normal pdf; to obtain the marginal likelihood, numerical integration is required. Δdf for Model A versus B is 14; Δdf for Model A versus C is 30. In Model B, 21 loadings (seven per factor, with one per factor fixed for identification), 46 thresholds (two per ordinal item, with two fixed due to sparseness), and six factor (co)variances were estimated; in Model A, seven loadings, 46 thresholds, and six factor (co)variances were estimated; in Model C, seven loadings, 16 thresholds, and six factor (co)variances were estimated.

For the Model A versus B comparison, at the level of the sample, $\Delta\chi^2(14) = 24.79$, $p < .05$, where $\chi^2_{crit}(14) = 23.69$. ΔBIC = −64.75, and ΔAIC = −3.22, meaning that chi-square selects the more unrestricted model of girls' internalizing behavior over time (Model B) with noninvariant thresholds and loadings. But when we take parsimony and predictive validity into account the other indices select the more restricted Model A, with invariant loadings. The upper panel of Figure 2 provides an index plot of $\Delta\text{ind}_{\text{CHI}_i}$ (or $\Delta\text{ind}_{\text{AIC}_i}$). Additionally, the parsimony corrective term for $\Delta\text{ind}_{\text{BIC}_i}$ was −.02. Descriptively, Model A is a better fit for 43% of persons, according to $\Delta\text{ind}_{\text{CHI}_i}$ or $\Delta\text{ind}_{\text{AIC}_i}$, or 48% of persons, according to $\Delta\text{ind}_{\text{BIC}_i}$. These differences can be visualized by comparing the zero point for $\Delta\text{ind}_{\text{CHI}_i}$ or $\Delta\text{ind}_{\text{AIC}_i}$ in the upper panel of Figure 2 (the zero-point on y-axis) to the zero-point of $\Delta\text{ind}_{\text{BIC}_i}$ (dotted reference line).

No cases have large enough negative $\Delta\text{ind}_{\text{BIC}_i}$ or $\Delta\text{ind}_{\text{AIC}_i}$ to potentially reverse ΔBIC's or ΔAIC's selection of Model A at the sample level. However, since $d = 1.1$, several cases with $\Delta\text{ind}_{\text{CHI}_i} > d$, would be independently flagged as having the potential to influence $\Delta\chi^2$'s selection of Model B at the sample level. To demonstrate case influence on the $\Delta\chi^2$ results, we need only confirm via case deletion that the presence/absence of one of these cases can reverse the $\Delta\chi^2$ results. We use case ID 195 as an example; it had $\Delta\text{ind}_{\text{CHI}_i} = 3.06$. Without ID 195's contribution, the sample $\Delta\chi^2$ became nonsignificant: $\Delta\chi^2(14) = 20.62$, $p > .05$, now indicating support for Model A. ΔBIC (−68.89) and ΔAIC (−7.38) continued to select A. Although the approximate $\Delta\text{ind}_{\text{CHI}_i}$ and its exact deletion counterpart, $\Delta\chi^2_i(4.08)$, both led to the same decision that case 195 was influential, they were not numerically identical. Recall this discrepancy occurs because $\Delta\text{ind}_{\text{CHI}_i}$ corresponds with case 195's contribution when parameter estimates are at the values from the full-N sample analysis. Refitting models with $N - 1$ to get $\Delta\chi^2_i$ allows parameter estimates to change.

If we were to rely solely on the chi-square results at the sample level, we might stop our invariance testing here, retain Model B, and not perform the A versus C comparison. But the support of BIC and AIC for Model A over B, together with the influential $\Delta\text{ind}_{\text{CHI}_i}$ in favor of A, suggest a rationale for continuing. In the A versus C comparison, at the level of the sample, $\Delta\chi^2(30) = 1689.52$, $p < .05$, where $\chi^2_{crit}(30) = 43.77$; ΔBIC = 1497.66, and ΔAIC = 1629.52, meaning that all indices select the invariant loading/noninvariant threshold model (A) over the invariant loading and threshold model (C). The lower panel of Figure 2 provides an index plot of $\Delta\text{ind}_{\text{CHI}_i}$ (or $\Delta\text{ind}_{\text{AIC}_i}$) for the A versus C comparison. Additionally, the parsimony correction term for $\Delta\text{ind}_{\text{BIC}_i}$ was −.05; the dotted reference line in Figure 2 denotes the zero-point of $\Delta\text{ind}_{\text{BIC}_i}$. There are no potentially influential cases ($d = 1645.75$ and no $\Delta\text{ind}_{\text{CHI}_i} > d$; no $\Delta\text{ind}_{\text{BIC}_i} > 1497.66$; no $\Delta\text{ind}_{\text{AIC}_i} > 1629.52$). Taken together, these results suggest that there is considerable support for selecting A over B and A over C. When reporting these results, a researcher could explain that a sensitivity

[8] In categorical factor models, some authors have allowed invariance of loadings and thresholds to be evaluated in separate steps (e.g., Millsap & Yun-Tein, 2004) as we do here, and others have tested their invariance together in one step (L. K. Muthén & Muthén, 1998–2011).

[9] Eight items appear as internalizing domain items on both the Child Behavior Checklist (CBCL) 2/3 form and CBCL 4–18 form (Achenbach, 1991, 1992): "underactive, slow-moving, or lacks energy"; "unhappy, sad or depressed"; "withdrawn/doesn't get involved with others"; "overtired"; "nervous, high-strung, tense"; "too fearful or anxious"; "shy or timid"; "self-conscious or easily embarrassed."

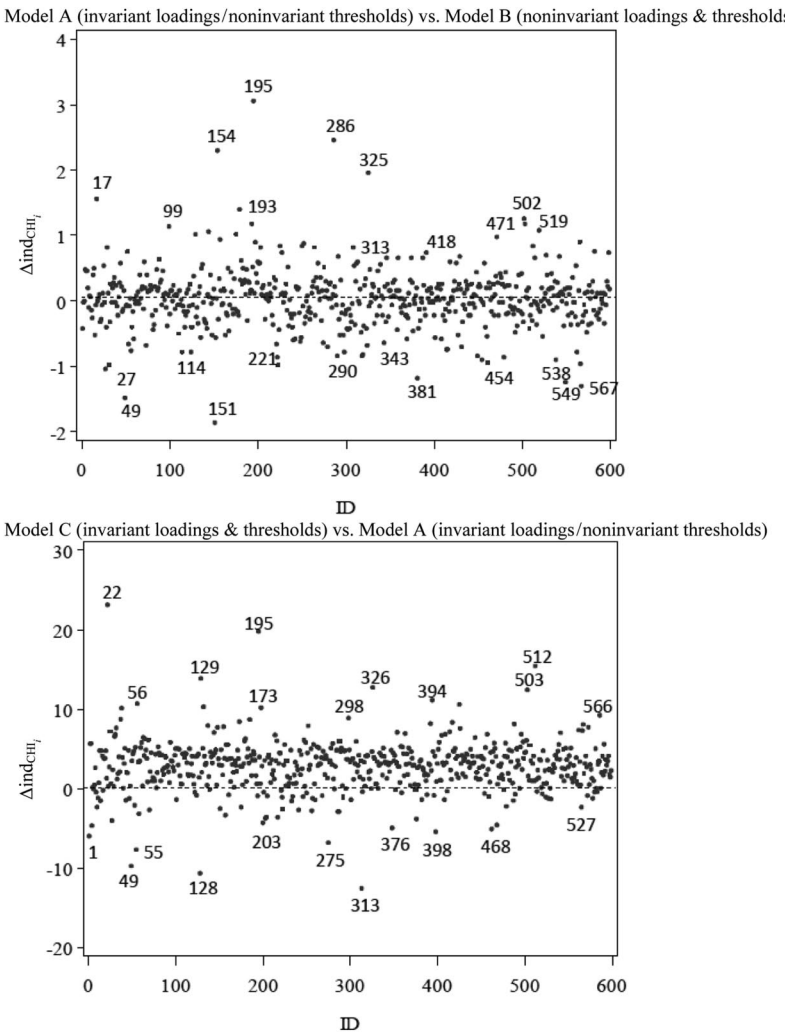

Figure 2. Example 1—$\Delta\text{ind}_{\text{CHI}_i}$ index plots for comparing longitudinal factor models. Dotted horizontal reference line denotes 0 for $\Delta\text{ind}_{\text{BIC}_i}$.

analysis had been done that indentified influence with respect to the chi-square for the first but not the second model comparison. If a researcher was, for instance, most interested in predictive validity, the insensitivity of ΔAIC to influential cases could be emphasized, alongside selecting Model A at the sample level. Another option would be to follow-up on potential cases that dominated $\Delta\chi^2$ results to investigate why they were influential on the A–B comparison; some ideas for such follow-up investigations are discussed in a later section, entitled "Why could a case be influential on model selection?" In sum, this empirical example highlights that, using these diagnostics, richer insights can be obtained about MI at the individual versus sample level across multiple model comparisons, which can, in turn, improve confidence in the final sample level model ranking.

Example 2: Cases as Clusters

Until this point, in illustrations a case has corresponded to a person in a single-level analysis. It was stated earlier that, more generally, the proposed diagnostics consider a case to be the highest level unit in an analysis. Hence, in a multilevel model with mice nested within litters, a case is a litter, but in a daily diary longitudinal model with day nested within week, and week nested within person, a case is a person. Once researchers recognize which unit corresponds with a case, no further complexities arise in applying the diagnostics to a hierarchical or multilevel modeling context. Here we consider the application of our approximation diagnostics $\Delta\text{ind}_{\text{CHI}_i}$, $\Delta\text{ind}_{\text{AIC}_i}$, $\Delta\text{ind}_{\text{BIC}_i}$ to a common model selection problem in multilevel modeling for which influence diagnosis has been previously recommended (Snijders & Bosker, 1999, pp. 137–138). Specifically, we consider a comparison between a multilevel model that includes a cross-level interaction (Model B) and one that does not (i.e., a main effects only model; Model A) using simulated data. Simulated data were *all* generated from Model B, and so we expect support for the interaction model at the sample level, although not necessarily at the case level, and we generally would not expect influential cases.

The generating multilevel model (MLM) has normally distributed outcomes and random effects; hence, $f(.)$ and $h(.)$ are normal for Equation 1. This MLM has $j = 1 \ldots 10$ Level 1 units (student) nested within each of $i = 1 \ldots 40$ Level 2 units or clusters (schools). Since a case is a highest level unit, here a case refers to a cluster (school), rather than a student. Therefore $N = 40$. Outcomes Y_{ji} are an additive linear combination of an intercept, a Level 1 predictor X_{ji}, a Level 2 predictor W_i and their (cross-level) interaction W_iX_{ji}; their respective population coefficients are 2, 1, 1, and 1.[10] Intercepts vary randomly across schools (variance = .5), as do slopes of X_{ji} (variance = .3); the residual (Level 1) variance is 1.0. To this data set we fit two multilevel models: Model B (the true generating model) and Model A (omitting the cross-level interaction term); A is nested in B.

At the sample level, all model selection indices support Model B: $\Delta\chi^2(1) = 58.34$, $p < .05$, (where $\chi^2_{crit}(1) = 3.85$), ΔBIC = 52.64 ("very strong" evidence for Model B against Model A), and ΔAIC = 56.63, consistent with the fact that B is the generating model. An index plot of Δind_{CHI_i} (or Δind_{AIC_i}) is given in Figure 3, and the parsimony-corrective factor for Δind_{BIC_i} was = –.05. As expected, no Δind_{CHI_i}, Δind_{BIC_i}, or Δind_{AIC_i} are influential in the sense that their values, if excluded, could potentially alter the model ranking or the degree of support for Model B ($d = 54.49$ and no $\Delta\text{ind}_{CHI_i} > d$; no $\Delta\text{ind}_{BIC_i} > 52.64$; no $\Delta\text{ind}_{AIC_i} > 56.63$). The largest $\Delta\text{ind}_{CHI_i} = \Delta\text{ind}_{AIC_i} = 8.1$, and $\Delta\text{ind}_{BIC_i} = 7.8$.

Additionally, despite the fact that all cases were generated from B, descriptively, all cases do not support B at the case level. A few cases' data are relatively much better fit by B; most are modestly better fit by B; and some cases' data show slight support for A (18% of cases according to Δind_{CHI_i}, Δind_{BIC_i}, or Δind_{AIC_i}). To understand why, consider the implications of the two models at the case level. A cross-level interaction term implies that the slope of Y_{ji} on X_{ji} depends on case (cluster) i's value of W_i. Moreover, a positively signed cross-level interaction term implies that clusters for which this slope and W_i were larger and positive, or clusters for which this slope and W_i were larger and negative would see much more improved fit from the inclusion of the interaction term than would other clusters (e.g., clusters in which the slope was larger and positive but W_i was larger and negative). This pattern is borne out in Figure 4. In Figure 4 we see that the cases whose Δind_{CHI_i} indicated that their data were much better fit by Model B (e.g., ID#s 22, 28, 35) indeed had a high X_{ji} slope and high W_i value (or low on both). Conversely, the cases whose data were better fit by A (e.g., ID# 33) could have relatively large and opposite-signed coordinates. Only if we were to increase the population effect size difference between the models enough (here, by increasing the coefficient of W_iX_{ji} fourfold), holding constant N and Δdf, would all cases' data eventually be better fit by the generating Model B, than Model A.

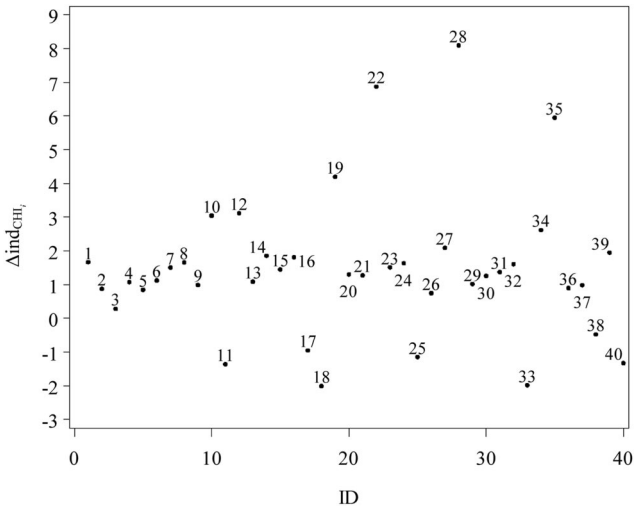

Figure 3. Example 2—Δind_{CHI_i} index plots for comparing a main-effects-only multilevel regression model (MLM; Model A) versus an MLM with an interaction (Model B): $N = 40$ cases generated from Model B. Model A is an MLM including main effects of a Level 1 predictor X_{ji} and a Level 2 predictor, W_i. Model B is an MLM that also includes a cross-level interaction of the Level 1 and 2 predictors.

In sum, we showed that the proposed diagnostics can be straightforwardly applied to detect influence of highest-level units in a multilevel analysis. In this example, all cases were generated from the same fitted model; no influential cases were expected and none were found. This example also served to illustrate why a true generating model need not have the support of all cases at the case level, despite having support at the sample level. For this reason, generalizing sample level conclusions about model ranking to a given case constitutes an ecological fallacy (Robinson, 1950).

Why Could a Case Be Influential on Model Selection?

Thus far we have defined *case influence with respect to model selection* as occurring when a single case's presence/absence alters model ranking (or alters the degree of evidence for a given model according to, say, Bayes factors). We suggested that a case flagged as influential on selection by our approximate influence diagnostics be confirmed as such using exact case deletion. It is also crucial to recognize that model selection influence diagnostics tell us *whether,* but not *why*, a case is influential. This point is not unique to influence diagnostics for model selection; influence diagnostics in general (e.g., with respect to fit of a single model in isolation, or with respect to parameter estimates) do not tell us why a case is influential. Once a diagnostic indicates that a case is influential, it is up to the researcher to investigate and weigh alternative potential causes for influence and decide how to use this information. It would not be advisable to automatically permanently omit an influential case from an analysis based on the influence diagnostic value alone (more cautions in this regard are given in the Discussion). Understanding *why* a case is influential is a separate task requiring additional qualitative or quantitative information beyond the diagnostic.

In general, gathering information on why a case is influential on model selection may involve checking data collection instruments for malfunctions, checking codebooks for data contamination or checking interview records for signs of fatigue or inattention (e.g., Rensvold & Cheung, 1999). Similar suggestions have been made in the IRT literature with respect to investigating potential causes

[10] X_{ji} and W_i were both standard normally distributed in the population.

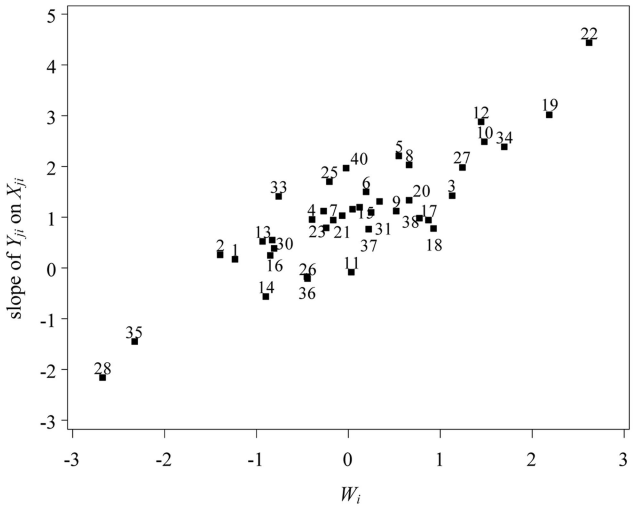

Figure 4. Example 2—Plot of each case's slope of Y_{ji} on X_{ji} versus that case's value of W_i. See Figure 3 notes.

of person misfit (Reise, 2000; Reise & Waller, 2009). Patterns of covariate values may be inspected, potentially leading to consideration of additional models. Sometimes, a researcher may not fully resolve why a case is influential. Case influence on model selection can occur for a variety of reasons, including data entry/coding errors or unobserved population heterogeneity (≥1 case from a different population; discussed shortly). Further, certain data/model conditions increase the chance of finding an influential case, even in the absence of data coding errors or unobserved population heterogeneity. Whereas the possibility for influence stemming from data coding errors seems self-explanatory, the possibility for influence to arise from unobserved population heterogeneity and the possibility for other data/model conditions to increase the chance of finding an influential case are less clear cut. These latter two topics are discussed and illustrated in the next two subsections.

Unobserved Population Heterogeneity

Unobserved population heterogeneity arises when at least one case in the sample is generated from a different population than the rest of the sample, unbeknownst to the researcher (e.g., B. O. Muthén, 1989). Whereas case influence on model selection can be caused by such population heterogeneity, unobserved population heterogeneity is neither necessary nor sufficient to guarantee the presence of influential cases. Hence, diagnostics for case influence on model selection should not be treated as a de facto test for a mixture. Whether population heterogeneity gives rise to influential case(s) depends on the particular characteristics of the heterogeneous case(s) in relation to the models under consideration. To explicate this point, a brief simulated example is used involving selecting between competing numbers of factors in CFA in the presence of alternative heterogeneous cases.[11]

Suppose that we have two nested generating models: a one-factor congeneric CFA model (Model A) versus a two-correlated-factor congeneric CFA model (Model B). Both have 20 normally distributed items and normally distributed factor(s); hence, $f(.)$ and $h(.)$ are normal in Equation 1. In both generating Models A and B, factor loadings = .7, factor variances = 1, and residual variances = .51; also, in Model B, the factor correlation = .5, and 10 items load on each factor. We are concerned with the consequences for model selection between A and B if a researcher's sample contains all 75 persons generated from A plus one person generated from B ($N = 76$). In this context, given that A is unidimensional and B is two-dimensional, if our one heterogeneous case generated from B had factor scores very different from each other, it would be worse fit by A and likely influential on model selection. On the other hand, if this heterogeneous case's scores on the two factors were similar, it would be reasonably consistent with A and, thus, not be influential on model selection. These two circumstances are depicted in the $\Delta\text{ind}_{\text{CHI}_i}$ (or $\Delta\text{ind}_{\text{AIC}_i}$) index plots in Figure 5 for a Model A versus B comparison ($\Delta\text{ind}_{\text{BIC}_i}$'s parsimony-corrective term from Equation 18 is –.01).

First consider the left panel of Figure 5. Here, the analysis data set contained the 75 cases generated from A plus a heterogeneous case from B (ID#B1) that had a relatively large factor score difference (–2.19; which can be thought of as more than a two standard deviation difference between z-scores). Here, at the sample level, Model B would be selected: $\Delta\chi^2(1) = 9.76$, $p < .05$, (for $\chi^2_{crit}(1) = -3.85$), $\Delta\text{BIC} = 5.43$ ("positive" evidence for Model B), and $\Delta\text{AIC} = 7.76$. The heterogeneous case ID#B1 has $\Delta\text{ind}_{\text{CHI}_i} = \Delta\text{ind}_{\text{AIC}_i} = 12.52$ and $\Delta\text{ind}_{\text{BIC}_i} = 12.51$. These diagnostics suggest that case ID#B1 is influential with respect to the model ranking for all selection indices, since $12.52 > d$ (recall $d = \Delta\chi^2 - \chi^2_{crit}$), $12.51 > \Delta\text{BIC}$, and $12.52 > \Delta\text{AIC}$. The influence of case ID#B1 is confirmed via exact case deletion. Excluding ID#B1 reverses the model ranking (in favor of A): $\Delta\chi^2(1) = 0.25$, $p > .05$, $\Delta\text{BIC} = -4.07$ ("positive" evidence for Model A), and $\Delta\text{AIC} = -1.75$, implying that $\Delta\chi^2_i = \Delta\text{AIC}_i = 9.51$ and $\Delta\text{BIC}_i = 9.50$.

Now consider the right panel of Figure 5. Here, the analysis data set contained the 75 cases generated from A plus a different heterogeneous case from B (ID#B2) with a smaller factor score difference (–.70). Here, at the sample level, Model A would be selected: $\Delta\chi^2(1) = .32$, $p > .05$, $\Delta\text{BIC} = -4.01$ ("positive" evidence for Model A), and $\Delta\text{AIC} = -1.68$. The heterogeneous case ID#B2 has $\Delta\text{ind}_{\text{CHI}_i} = \Delta\text{ind}_{\text{AIC}_i} = .39$, $\Delta\text{ind}_{\text{BIC}_i} = .37$ and is not influential with respect to model ranking.

More generally, there would be a whole range of possibilities for case influence on model selection depending on which one or few cases generated from B happened to be mixed with the sample generated from A prior to model fitting. To put this in perspective, Figure 6 depicts 75 cases generated from A (in black) as well as $N = 75$ cases generated from B (in gray)–including ID#B1 and ID#B2. Gray cases show increasing $\Delta\text{ind}_{\text{CHI}_i}$ (favoring B) as their factor score differences increase.

[11] This example was chosen as substantively relevant because chi-square difference tests are still commonly used for this purpose (e.g., Bollen, 1989; Fitzmaurice, Laird, & Ware, 2004; Hedeker & Gibbons, 2006; Mulaik, 2009) with the understanding that they will be slightly conservative when setting a factor/random effect variance to 0 or a factor/random effect correlation to 1 in the restricted model. Whether to adjust chi-square tests in this context and how to do so in general settings is currently unresolved (e.g., Savalei & Kolenikov, 2008; Stoel, Garre, Dolan, & van den Wittenboer, 2006). Other model comparisons in other modeling frameworks could have been used to explicate the same point.

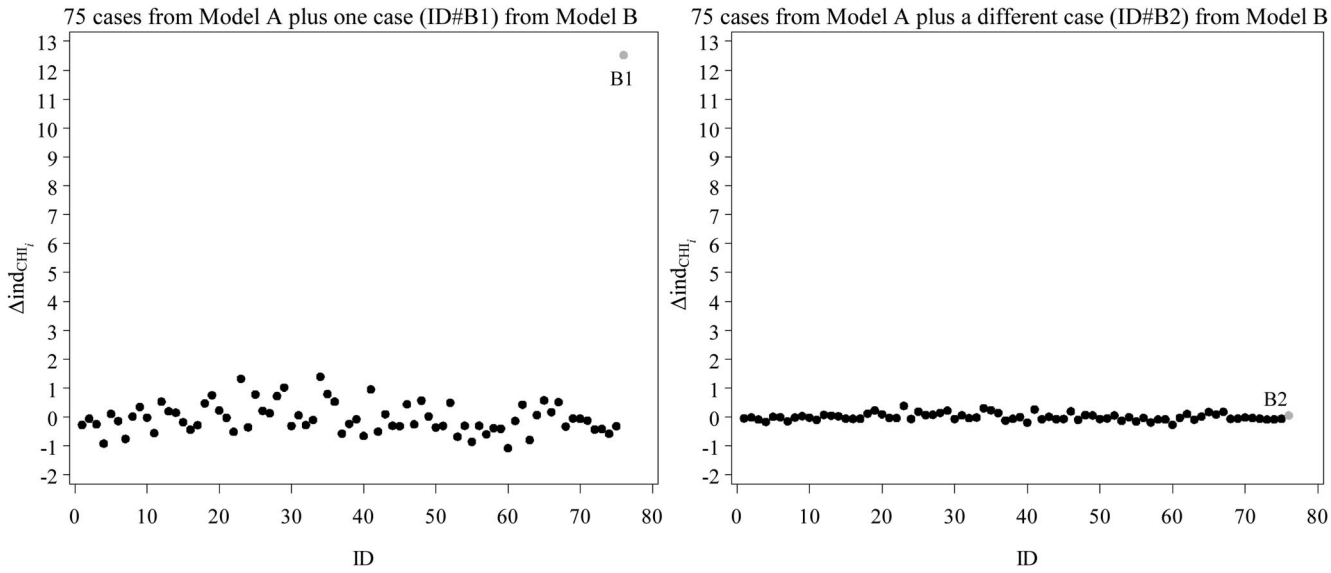

Figure 5. Population heterogeneity does not necessarily imply case influence on model selection: Δind_{CHI_i} index plots for a one-factor confirmatory factor analysis (CFA; Model A) versus a two-factor CFA (Model B) comparison, where all but one case are generated from Model A.

In sum, population heterogeneity is one hypothesis to consider for explaining case influence on model selection. But this illustration showed that a case's potential for influence is not just determined by whether it was literally generated from a second population; it is also determined by that case's characteristics with respect to the models under consideration.

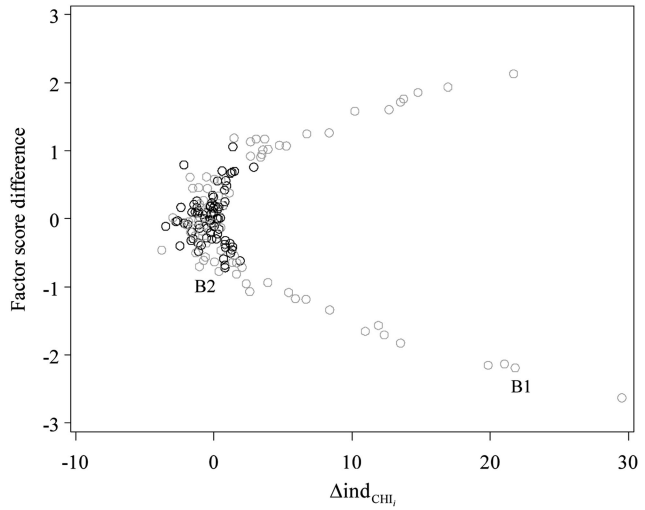

Figure 6. Cases' estimated factor score differences versus cases' Δind_{CHI_i} scores, for 75 (black) cases generated from a one-factor confirmatory factor analysis (CFA; Model A) and 75 (gray) cases generated from a two-factor CFA (Model B). The factor score difference for case *i* was calculated as case *i*'s estimated score on Factor 1 minus case *i*'s estimated score on Factor 2, when Model B was fitted. B1 and B2 refer to the cases described in Figure 5. Models A and B refer to the models described in Figure 5.

Other Data/Model Conditions That Can Increase the Chance of Case Influence on Model Selection

We have thus far considered a few possible reasons for case influence on model selection, including unobserved population heterogeneity. It is also useful to recognize that certain data/model conditions pose a greater chance of finding case influence on model selection, all else equal. These are conditions that result in differences in fit between models close to the selection index's decision threshold, where for the $\Delta\chi^2$ statistic, the decision threshold would be χ^2_{crit} and for ΔBIC and ΔAIC it would be 0. For instance, consider $\Delta\chi^2$, and recall $d = \Delta\chi^2 - \chi^2_{crit}$. When $|d|$ is smaller, there is generally a greater chance of finding case influence (i.e., $\Delta ind_{CHI_i} >$ positive d or $<$ negative d). Given that a researcher's competing models are not exactly the same, N has a positive monotonic relationship with d. Also, the *effect size* difference between models (e.g., size of parameters constrained to 0 in Model A but not B or size of differences between parameters constrained to equality in Model A but not B)—has a nonmonotonic relationship with d. Specifically, all else equal, there is a greater chance of $\Delta\chi^2$ influence for low N combined with small effect size. Holding N constant, a very tiny effect size or a large effect size makes influence unlikely (the former makes d larger and negative; the latter makes d large and positive).

Since influence on model selection can occur by chance alone (i.e., due to sampling variability), we illustrate the nature of relations among effect size, N, and case influence on $\Delta\chi^2$ by comparing the proportion of samples having an influential case by chance alone under conditions that differ in effect size and N—and thus differ in d. Although any model comparison context could be chosen for this illustration, we chose a frequently used nested IRT model comparison (see Kang & Cohen, 2007): a one-parameter logistic (1PL or Rasch, with equal item discriminations) vs. a

two-parameter logistic (2PL with unequal item discriminations); $\Delta df = 9$. Five hundred samples of 10 binary items were generated from a 2PL model with item difficulty parameters chosen from Embretson and Reise's (2000, p. 69) abstract reasoning test results, Items 11–20. Twelve data conditions were defined by four *N*s (100, 500, 750, 1000) crossed with three effect size differences between models. Effect size differences between models were manipulated by reducing or increasing the range of item discriminations in the generating 2PL: *large* discriminations (range = .5–2.5 by .222);[12] *small* discriminations (range = 1–1.5 by .055); *very small/trivial* discriminations (range = 1–1.05, by .0056). 1PL and 2PL models were fitted to the 500 samples per cell. Incidences of a case flagged as influential by $\Delta\text{ind}_{\text{CHI}_i}$ were recorded. Here $f(.)$ in Equation 1 is a binomial probability mass function, $h(.)$ is a normal pdf, and integration is required to obtain the marginal likelihood.

Figure 7 summarizes the results obtained from the simulation. Figure 7 shows that there is little chance of influence for large effect size (dotted line), regardless of *N.* And there is little chance of influence for large *N,* regardless of effect size. There is also little chance of influence for very small/trivial effect size (dashed line), at most *N.* As anticipated, the combination of small *N* and small effect size posed the highest chance of influence: 17% of samples had at least one influential case flagged by $\Delta\text{ind}_{\text{CHI}_i}$.

This brief illustration considered influence on model selection occurring due to sampling variability only. Yet, more generally, a case whose contribution is relatively more extreme than other cases for *any other reason* (data coding error, population heterogeneity, etc.) would *also* be more likely to have influence on sample level conclusions when differences in fit between models are close to a selection index's decision threshold. Further, though this illustration concerned only IRT models, the phenomenon demonstrated is not limited to this modeling context; at issue is the *closeness* of the models' fit difference to a selection index's decision threshold, not *which* models (e.g., SEMs, IRT models, MLMs, single-level regressions, etc.) gave rise to a given fit difference.[13] Finally, this illustration considered only $\Delta\chi^2$; closeness of ΔAIC and ΔBIC to their decision thresholds of 0 depends on Δdf over and above *N* and effect size (see Equations 7 and 8). In sum, researchers should be aware that there is a higher chance of case influence on model selection under certain conditions, all else equal.

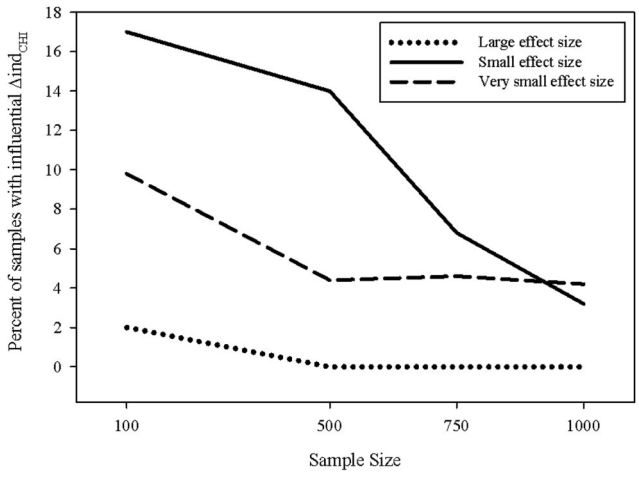

Figure 7. Percentage of samples with at least one influential $\Delta\text{ind}_{\text{CHI}_i}$ across alternative simulated data conditions for model comparisons that differ in *N* and effect size.

To summarize, we revisit this section's question: "Why could a case be influential on model selection?" We discussed that there are many causes of case influence on selection and that there are certain data conditions under which we are more likely to find influence. Researchers can investigate alternate potential causes of case influence and try to adjudicate between them and can also report results of influence diagnostics in the context of a sensitivity analysis (described in the Discussion).

Software Implementation of $\Delta\text{ind}_{\text{CHI}_i}$, $\Delta\text{ind}_{\text{BIC}_i}$, and $\Delta\text{ind}_{\text{AIC}_i}$

Two steps are needed to produce the approximate casewise influence diagnostics discussed here. First, a researcher needs to fit each of their competing models using one of several software packages that provide FIML *and* also allow the saving and exporting of casewise loglikelihood values (e.g., Mplus, Version 6.1 [L. K. Muthén & Muthén, 1998–2011], see Save = Loglikelihood option; Mx [Neale, Boker, Xie, & Maes, 2003], see MX%P = < filename> option). Second, the researcher needs to use these casewise loglikelihood values to compute model selection influence diagnostics for each pair of models. In our online Appendix, we provide example Mplus code for exporting casewise loglikelihood values. We also provide SAS code for calculating $\Delta\text{ind}_{\text{CHI}_i}$, $\Delta\text{ind}_{\text{BIC}_i}$, and $\Delta\text{ind}_{\text{AIC}_i}$ from these exported casewise loglikelihood values and obtaining index plots for each diagnostic. Given the generality of the proposed individual contributions to model selection indices, this two-step approach can be used with any model type or outcome type, so long as FIML estimation was used.

Discussion

Researchers' increasing use of model selection in psychology far outpaces an awareness or examination of individual-specific influences on model selection. This article began by highlighting the underappreciated issue that one or a few cases can have a disproportionate impact on the selection of a model at the sample level. We then developed several approximate influence diagnostics for commonly used model selection indices—$\Delta\chi^2$, ΔBIC, and ΔAIC—that are obtained simply from byproducts of FIML estimation using available software, without computationally heavy iterative model refitting. We described and provided simulated demonstrations of the diagnostics' interpretation and behavior, along with code to facilitate their use in practice.

Here we summarize key take-home points. Case influence on model selection refers to whether a single case can impact the sample level conclusion, reflecting the approach of the case influ-

[12] Reise and Waller (2009, p. 30) reported that discriminations often exceed 2.5 in the clinical but not cognitive psychology literature.

[13] For instance, selected cells of the simulation were conducted with nested CFA models with normal outcomes, rather than IRT models, and similar findings were obtained. More information is available from the first author upon request.

ence literature more generally. Unlike outlier diagnostics and person-fit diagnostics, our focus is not on comparing cases' contributions against one other but ultimately to sample level decision thresholds for model ranking. It is therefore possible and common for *no* cases to be influential on model selection (e.g., the MLM example); this differs from some person-fit diagnostics that may automatically siphon off the top 1% or 5% of cases (Drasgow et al., 1996). Like other influence diagnostics, our diagnostics for case influence on model selection convey *whether* rather than *why* a case is influential. Earlier we discussed several potential causes of case influence that a researcher could investigate and some data/model conditions under which influence may be more common.

In addition to using the developed diagnostics for influence detection (e.g., by comparing a case's Δind_{CHI_i}, Δind_{BIC_i}, and/or Δind_{AIC_i} with sample level *d,* ΔBIC, and ΔAIC, respectively), we noted that these diagnostics also provide descriptive, ideographic information about a given case's own model ranking (by comparing a case's Δind_{CHI_i}, Δind_{BIC_i}, and/or Δind_{AIC_i} with 0). Our empirical and simulated examples highlighted the applicability of the Δind_{CHI_i}, $\Delta ind_{\mathbf{BIC}_i}$, and Δind_{AIC_i} influence diagnostics to diverse modeling frameworks and outcome types (e.g., categorical or continuous). Although all of our simulated and empirical model comparisons happened to be nested, this was merely so that all three selection diagnostics could be applied and compared for pedagogical purposes. In general, however, $\Delta ind_{\mathbf{BIC}_i}$, and Δind_{AIC_i} can be readily applied to nonnested comparisons, such as between an autoregressive model and a growth curve model. Furthermore, although our empirical example pertained to clinical psychology, potential application areas are diverse. For instance, these diagnostics are particularly relevant for mathematical modeling of cognitive psychology processes (e.g., memory; decision making), an area that has historically employed a model selection approach (Myung, Pitt, & Kim, 2005). Next we provide some practical recommendations on reporting results of sensitivity analyses using such diagnostics and conclude with a discussion of limitations and potential future extensions of these techniques.

Recommendations for Reporting Sensitivity Analyses for Case Influence on Selection

We suggest using these diagnostics in the context of a sensitivity analysis. A sensitivity analysis framework is increasingly used for gauging the practical impact of particular data conditions or assumption violations (Verbeke & Molenberghs, 2000). Researchers can conduct a sensitivity analysis for case influence on model selection by using Δind_{CHI_i}, Δind_{BIC_i}, and/or Δind_{AIC_i} as screeners to flag potential influential cases, and then by confirming the influence of flagged cases by exactly computing $\Delta\chi^2_i$, ΔAIC_i, and/or ΔBIC_i. If no influential cases are found, researchers can report that their model selection conclusions are robust to influential cases. But, for descriptive purposes, researchers may still want to report the model ranking for individual case(s) of particular substantive interest. If an influential case (or cases) are found, researchers may consider follow-up investigations of those cases.

At the least, information about case influence is something for researchers to be aware of when articulating and framing their substantive conclusions. How researchers use this information will depend on their analytic goals. For instance, suppose a researcher was comparing five models (A, B, C, D, E) and had found two models (A, B) strongly preferable to the other three models (C, D, E) at the sample level using BIC. But suppose the best fitting two models A and B were only "weakly" differentiable from each other at the sample level, and the researcher had decided to postpone judgment between them. If the researcher found that there was an influential case that could reverse the ranking from weak support of B over A to weak support for A over B, such sensitivity may have little consequence for decision making in this context. Given different analysis goals, or given a case that exerted stronger influence (e.g., reversing the degree of evidence for A over B to "very strong" rather than "weak" levels) this influential case might be considered more substantively consequential. In other situations, an influential case could lead us to contextualize or qualify what would have otherwise been an all-or-nothing decision about model ranking. For instance, if the difference in fit between competing models is very close to the chosen index's decision threshold (e.g., the small effect size condition from our IRT model comparison simulation) and if *N* is very small, a researcher finding influence for unexplained reasons might report the following caveat alongside their sample level results. Under their data/model conditions, their chosen sample level model ranking could be materially altered by a single case. In future studies the researcher could consider alternative models that were more distinct and/or larger *N*s to try to avoid this situation.

We caution that unless separate investigations into the cause of an influential case can confirm a particular cause (e.g., a data coding error), it may not be advisable to omit this case from the sample permanently. Case omission is certainly not a necessary step and should only be considered in light of multiple kinds of additional information beyond the diagnostic value, such as the nature of the researchers' measurement, in conjunction with their sample characteristics and theory—to avoid the premature dismissal of a substantively important pattern. Also, researchers should avoid deciding whether to delete a case solely based on whether the model ranking would change for or against hypotheses (this could be framed as an ethical issue; Panter & Sterba, 2011).

Taking a step back, we recognize that, in practice, the results of model selection—even when exclusively considering sample level model ranking—are often equivocal. Indices' rankings may not agree; no model may have a strong degree of evidence over others. Here we have further complicated the picture by adding ideographic-level information to the nomothetic. An individual could be influential on one model comparison but not another. Further, influence diagnostics may not agree on what case is influential, just as they do not always agree at the sample level. Yet we consider the added complexity informative and worthwhile. These indices fall in line with a progression of methodological recommendations away from evaluating one nomothetic model in isolation, to considering nomothetic and ideographic influences on one model (e.g., Reise & Widaman, 1999), to considering multiple competing models (Rodgers, 2010), to considering nomothetic and ideographic influences on model comparison (as done here). We emphasize that the primary motivation for a sensitivity analysis of case influence in the selection context is to avoid unknowingly

obtaining and interpreting a model ranking that is materially driven by one (or very few) cases.

Limitations and Extensions

In some applications researchers may be most interested in (a) whether model selection can be influenced by *at least one* case; (b) whether *any one of several* flagged cases could individually influence selection results (e.g., ID#s 195, 286, 154, 325 in the Figure 2 empirical example); or (c) whether *jointly two or three* cases, for instance, could influence selection. Objective (a) has been our concern in earlier examples. Objective (b) could be assessed by individually confirming the influence of each flagged case of interest, one at a time. Regarding objective (c), one-at-a-time case deletion statistics (whether approximate or exact, of which Δind_{CHI_i}, Δind_{BIC_i}, Δind_{AIC_i}, and $\Delta\chi^2_i$, ΔAIC_i, ΔBIC_i would be no exception) are not designed to simultaneously identify an influential "clump" (J. Cohen, Cohen, West, & Aiken, 2003) of cases (e.g., Poon & Poon, 2002; Xu et al., 2006). A clump could dominate or drive model selection results to the extent that such diagnostics would not flag a given case within the clump as influential, although the clump as a whole is influential; this phenomenon is generally called *masking* (e.g., Atkinson & Riani, 2008; Bendre & Kale, 1987). Potentially, approximation influence diagnostics could be calculated iteratively after one case from the clump of interest is deleted at a time; at each iteration the original versus current model selection results at the sample level could be compared.

More general approaches for addressing masking have been proposed in the context of evaluating single models in isolation; future research could consider explicitly extending these to the context of model selection. One approach is leave-*m*-cases out case deletion (e.g., Bruce & Martin, 1989), wherein the model is iteratively refit leaving out, say, each possible pair ($m = 2$), and/or triplet ($m = 3$) of cases, which escalates computational demand (Rensvold & Cheung, 1999). A second approach is a forward search procedure (e.g., Mavridis & Moustaki, 2008; Yang, Tanaka, & Nakaya, 2006), which aims to begin with an initial subset of noninfluential cases and then monitor changes in fit as cases are added in order of their consistency with the fitted model. A third approach is to employ local influence diagnostics (see footnote 1), which are considered somewhat less vulnerable to masking than case-deletion-type (i.e., global) diagnostics (Poon & Poon, 2002). On balance, since detecting individual influence on model selection currently receives no attention in applied psychology research, we believe the proposed Δind_{CHI_i}, Δind_{BIC_i}, and Δind_{AIC_i} diagnostics provide a worthwhile and simple-to-implement first step to allow detection of certain kinds of influence, even if they do not detect all potential kinds of influence well.

Another important direction for future work involves an expanded evaluation of the adequacy of the proposed diagnostics' approximation of their case deletion counterparts, perhaps involving simulation studies across a broad variety of data and model conditions. A third interesting direction for expansion would be to define case influence on magnitude of ΔBIC and ΔAIC in terms of Schwarz weights and Akaike weights, respectively (e.g., Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004).

Finally, a fourth direction for additional research involves a complementary but philosophically different avenue for handling case influence on model selection besides the diagnostic approach considered here. Specifically, robust model comparison methods may be used to downweight the impact of influential cases without necessarily bringing such cases to the attention of the researcher. Several such methods have been developed, for instance: a robust version of AIC for time series models using least squares estimation (Chik, 2002), a robust version of AIC implemented for autoregressive models (Ronchetti, 1997), and a robust version of Mallows' C_p for regression models using least squares estimation (Atkinson & Riani, 2008). Future research could expand such methods to other model types, other selection indices, and other estimation algorithms to increase the generality of this approach. These methods differ philosophically from our diagnostics in the sense that they may consider influential cases a nuisance to be controlled more so than a potentially theoretically meaningful occurrence to explore.

Conclusions

Model selection is a useful and increasingly popular endeavor in psychology that should be encouraged. We can often learn more from model comparisons than we can from the evaluation of a single model in isolation. A summary message from our demonstrations and empirical example is that, under at least some conditions, researchers may not recognize how often their model selection results are contingent on one or a few cases in a sample. Awareness of how individuals influence model selection results can help researchers understand how representative sample level results are at the individual level. We hope that available user-friendly software tools will facilitate researchers' greater exploration of case influence on model selection.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4–18 and 1991 profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1992). *Manual for the Child Behavior Checklist/2–3 and 1992 profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723. doi:10.1109/TAC.1974.1100705

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–276). Mahwah, NJ: Erlbaum.

Atkinson, A. C., & Riani, M. (2008). A robust and diagnostic information criterion for selecting regression models. *Journal of the Japanese Statistical Society, 38,* 3–14.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14,* 101–125. doi:10.1037/a0015583

Bendre, S. M., & Kale, B. K. (1987). Masking effect on tests for outliers in normal samples. *Biometrika, 74,* 891–896. doi:10.1093/biomet/74.4.891

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach.* Hoboken, NJ: Wiley.

Box, G. E. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association, 74,* 1–4.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21,* 230–258. doi:10.1177/0049124192021002005

Bruce, A. G., & Martin, R. D. (1989). Leave-*k*-out diagnostics for time series. *Journal of the Royal Statistical Society, Series B, 51,* 363–424.

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

Cadigan, N. G. (1994). *Computing case diagnostics for structural equation models.* Retrieved from http://www.sascommunity.org/sugi/SUGI94/Sugi-94–188%20Cadigan.pdf

Cadigan, N. G. (1995). Local influence in structural equation models. *Structural Equation Modeling, 2,* 13–30. doi:10.1080/10705519509539992

Chatterjee, S., & Hadi, A. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science, 1,* 379–393. doi:10.1214/ss/1177013622

Chik, Z. (2002). The effect of outliers on the performance of order selection criteria for short time series. *Pakistan Journal of Applied Sciences, 2,* 912–915.

Coffman, D. L., & Millsap, R. E. (2006). Evaluating latent growth curve models using individual fit statistics. *Structural Equation Modeling, 13,* 1–27. doi:10.1207/s15328007sem1301_1

Cohen, A. S., & Cho, S.-J. (in press). Information criteria. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory: Models, statistical tools, and applications.*

Cohen, J., Cohen, P., West, S., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Mahwah, NJ: Erlbaum.

Cook, R. (1977). Detection of influential observations in linear regression. *Technometrics, 19,* 15–18. doi:10.2307/1268249

Cook, R. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B, 48,* 133–169.

Cook, R. D., & Wang, P. C. (1983). Transformations and influential cases in regression, *Technometrics, 25,* 337–343.

Draper, N. R., & John, J. A. (1981). Influential observations and outliers in regression. *Technometrics, 23,* 21–26.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9,* 47–64.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fitzmaurice, G., Laird, N., & Ware, J. (2004). *Applied longitudinal analysis.* Hoboken, NJ: Wiley.

Greenland, S. (1989). Modeling and variable selection in epidemiological analysis. *American Journal of Public Health, 79,* 340–349. doi:10.2105/AJPH.79.3.340

Hamaker, E. L., van Hattum, P., Kuiper, R. M., & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 231–255). New York, NY: Taylor & Francis.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis.* New York, NY: Wiley.

Hoeting, J., Raftery, A. E., & Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics and Data Analysis, 22,* 251–270. doi:10.1016/0167-9473(95)00053-4

Johanson, G., & Alsmadi, A. (2002). Differential person functioning. *Educational and Psychological Measurement, 62,* 435–443.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36,* 347–387. doi:10.1207/S15327906347-387

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31,* 331–358. doi:10.1177/0146621606292213

Kankaras, M., Moors, G., & Vermunt, J. K. (2009). Testing for measurement invariance with latent class analysis. In E. Davidov, P. Schmidt, & J. B. Billiet (Eds.), *Cross-cultural analysis: Methods and applications.* Routledge.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16,* 277–298. doi:10.1207/S15324818AME1604_2

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33,* 188–229. doi:10.1177/0049124103262065

Lange, K., Westlake, J., & Spence, M. (1976). Extensions to pedigree analysis: III. Variance components by the scoring method. *Annals of Human Genetics, 39,* 485–491. doi:10.1111/j.1469-1809.1976.tb00156.x

Laud, P. W., & Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Series B, 57,* 247–262.

Le, N. D., Raftery, A. E., & Martin, R. D. (1996). Robust Bayesian model selection for autoregressive processes with additive outliers. *Journal of the American Statistical Association, 91,* 123–131.

Lee, S. Y., & Lu, B. (2003). Case-deletion diagnostics for nonlinear structural equation models, *Multivariate Behavioral Research, 38,* 375–400. doi:10.1207/S15327906MBR3803_05

Lee, S. Y., & Wang, S. J. (1996). Sensitivity analysis of structural equation models. *Psychometrika, 61,* 93–108. doi:10.1007/BF02296960

Lee, S. Y., & Xu, L. (2003a). Case-deletion diagnostic for factor analysis model with continuous and ordinal categorical data. *Sociological Methods & Research, 31,* 389–419. doi:10.1177/0049124102239081

Lee, S. Y., & Xu, L. (2003b). On local influence analysis of full information item factor models, *Psychometrika, 68,* 339–360. doi:10.1007/BF02294731

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38,* 113–139. doi:10.1207/S15327906MBR3801_5

Mavridis, D., & Moustaki, I. (2008). Detecting outliers in factor analysis using the forward search algorithm. *Multivariate Behavioral Research, 43,* 453–475. doi:10.1080/00273170802285909

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

McArdle, J. J. (1997). Modeling longitudinal data by latent growth curve models. In G. Marcoulides (Ed.), *New statistical models with business and economic applications* (pp. 359–406). Mahwah, NJ: Erlbaum.

McCann, L. (2006). *Robust model selection and outlier detection in linear regression.* Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Meade, A., Ellington, J. K., & Craig, S. (2004). *Exploratory measurement invariance: A new method based on item response theory.* Symposium presented at the 19th annual Conference for Industrial and Organizational Psychology, Chicago, IL.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8,* 72–87. doi:10.1037/1082-989X.8.1.72

Meijer, R. R., & Sitjsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107–135. doi:10.1177/01466210122031957

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4,* 5–9. doi:10.1111/j.1750-8606.2009.00109.x

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479–515. doi:10.1207/S15327906MBR3903_4

Moustaki, I., Jöreskog, K. G., & Mavridis, D. (2004). Factor models for ordinal variables with covariate effects on the manifest and latent variables: A comparison of LISREL and IRT approaches. *Structural Equation Modeling, 11,* 487–513. doi:10.1207/s15328007sem1104_1

Mulaik, S. A. (2009). *Linear causal modeling with structural equations.* Boca Raton, FL: Chapman & Hall.

Mullen, M. R., Milne, G. R., & Doney, P. M. (1995). An international marketing application of outlier analysis for structural equations: A methodological note. *Journal of International Marketing, 3,* 45–62.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585. doi:10.1007/BF02296397

Muthén, L. K., & Muthén, B. O. (1998–2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Myung, I. J., Forster, M., & Browne, M. W. (2000). Special issue on model selection. *Journal of Mathematical Psychology, 44,* 1–2. doi:10.1006/jmps.1999.1273

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. In K. Lambert & R. Goldstone (Eds.), *The handbook of cognition,* pp. 422–436. London, England: Sage.

Neale, M. (2000). Individual fit, heterogeneity, and missing data in multigroup structural equation modeling. In T. Little, K. Schnabel, & J. Baumert (Ed.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 249–267). Mahwah, NJ: Erlbaum.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical modeling* (6th ed.). Richmond, VA: Virginia Commonwealth University Department of Psychiatry.

Panter, A. T., & Sterba, S. K. (2011). *Handbook of ethics in quantitative methodology.* New York, NY: Routledge.

Pek, J., & MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: Cases and their influence. *Multivariate Behavioral Research, 46,* 202–228. doi:10.1080/00273171.2011.561068

Poon, W.-Y., & Poon, Y. S. (2002). Influential observations in the estimation of mean vector and covariance matrix. *British Journal of Mathematical and Statistical Psychology, 55,* 177–192. doi:10.1348/000711002159644

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics, 9,* 705–724. doi:10.1214/aos/1176345513

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163. doi:10.2307/271063

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35,* 543–568. doi:10.1207/S15327906MBR3504_06

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5,* 27–48. doi:10.1146/annurev.clinpsy.032408.153553

Reise, S. P., & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods, 4,* 3–21. doi:10.1037/1082-989X.4.1.3

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566. doi:10.1037/0033-2909.114.3.552

Rensvold, R. B., & Cheung, G. W. (1999). Identification of influential cases in structural equation modeling using the jackknife method. *Organizational Research Methods, 2,* 293–308. doi:10.1177/109442819923005

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357. doi:10.2307/2087176

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

Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica, 7,* 327–338.

Ronchetti, E., Field, C., & Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association, 92,* 1017–1023.

Sadray, S., Jonsson, E. N., & Karlsson, M. O. (1999). Likelihood-based diagnostics for influential individuals in non-linear mixed effects model selection. *Pharmaceutical Research, 16,* 1260–1265. doi:10.1023/A:1014857832337

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods, 13,* 150–170. doi:10.1037/1082-989X.13.2.150

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London, England: Sage.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95,* 334–344. doi:10.1037/0033-2909.95.2.334

Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11,* 439–455. doi:10.1037/1082-989X.11.4.439

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B, 39,* 44–47.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 683–718). Amsterdam, the Netherlands: Elsevier.

Tanaka, Y., Watadani, S., & Moon, S. H. (1991). Influence in covariance structure analysis: With an application to confirmatory factor analysis. *Communications in Statistics: Theory and Method, 20,* 3805–3821. doi:10.1080/03610929108830742

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128. doi:10.1037/0033-2909.99.1.118

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York, NY: Springer.

Wagenmakers, E., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192–196. doi:10.3758/BF03206482

Xu, L., Lee, S. Y., & Poon, W. Y. (2006). Deletion measures for generalized linear mixed effects models. *Computational Statistics & Data Analysis, 51,* 1131–1146. doi:10.1016/j.csda.2005.11.009

Yang, W., Tanaka, Y., & Nakaya, J. (2006). Forward search algorithm for robust influence analysis in maximum likelihood factor analysis. *International Journal of Computer Science and Network Security, 6,* 43–49.

Zhu, H. T., & Lee, S. Y. (2003). Local influence for generalized linear mixed models, *Canadian Journal of Statistics, 31,* 293–309. doi:10.2307/3316088

Zu, J., & Yuan, K.-H. (2010). Local influence and robust procedures for mediation analysis. *Multivariate Behavioral Research, 45,* 1–44. doi:10.1080/00273170903504695

Received August 26, 2011
Revision received April 20, 2012
Accepted April 28, 2012 ■