IES grant information:
Grant number R305D140046
Funded by National Center for Education Research (NCER)

On the Complexity of Item Response Theory Models

Wes Bonifay

University of Missouri

Li Cai

University of California, Los Angeles

Author Note

A preliminary version of this work was presented at the 2014 meeting of the Society for Multivariate Experimental Psychology in Nashville, TN, and the abstract accompanying this presentation appeared in *Multivariate Behavioral Research, 50*(1), 128.

Abstract

Complexity in item response theory (IRT) has traditionally been quantified by simply counting the number of freely estimated parameters in the model. However, complexity is also contingent upon the functional form of the model. The information-theoretic principle of minimum description length provides a novel method of investigating complexity by considering the inherent propensity of a model to fit well to any possible data. We examine four popular IRT models—exploratory item factor analytic, bifactor, DINA, and DINO—with different functional forms but the same number of free parameters. In comparison, a simpler (unidimensional 3PL) model was specified such that it had 1 more free parameter than the previous models. All five models were then fit to 1,000 data sets that were randomly and uniformly sampled from the complete data space and each model was assessed using global and item-level fit and diagnostic measures. The findings revealed that the factor analytic and bifactor models possess excessive flexibility and therefore a strong tendency to fit any possible data. The unidimensional 3PL model displayed minimal fitting propensity, despite the fact that it included an additional free parameter. The DINA and DINO models did not demonstrate a proclivity to fit any possible data, but they did fit well to meaningfully different data patterns. These findings suggest that applied researchers and psychometricians should consider functional form—and not goodness-of-fit alone—when selecting and applying an IRT model.

Keywords: item response theory, minimum description length, bifactor model, diagnostic classification model, model evaluation

Abstract

Complexity in item response theory (IRT) has traditionally been quantified by simply counting the number of freely estimated parameters in the model. However, complexity is also contingent upon the functional form of the model. We examined four popular IRT models—exploratory factor analytic, bifactor, DINA, and DINO—with different functional forms but the same number of free parameters. In comparison, a simpler (unidimensional 3PL) model was specified such that it had 1 more parameter than the previous models. All models were then evaluated according to the minimum description length principle. Specifically, each model was fit to 1,000 data sets that were randomly and uniformly sampled from the complete data space and then assessed using global and item-level fit and diagnostic measures. The findings revealed that the factor analytic and bifactor models possess a strong tendency to fit any possible data. The unidimensional 3PL model displayed minimal fitting propensity, despite the fact that it included an additional free parameter. The DINA and DINO models did not demonstrate a proclivity to fit any possible data, but they did fit well to distinct data patterns. Applied researchers and psychometricians should therefore consider functional form—and not goodness-of-fit alone—when selecting an IRT model.

Keywords: item response theory, minimum description length, bifactor model, diagnostic classification model, model evaluation

On the Complexity of Item Response Theory Models

In psychological and educational measurement research, model complexity has traditionally been quantified by tallying the number of freely estimated parameters in a given model. Simply put, the general assumption is that the greater the number of free parameters, the more complex the model becomes. However, complexity involves more than just a parameter count. Complexity, in the words of Myung, Pitt, and Kim (2005), is “a model’s inherent flexibility that enables it to fit a wide range of data patterns” (p. 12). While the number of estimated parameters certainly contributes to this flexibility, there is a second but no less important aspect to consider – the functional form of the model, that is, the way in which the parameters and random variables are combined and specified in the model’s equations (Collyer, 1985; Cutting, Bruno, Brady, & Moore, 1992). A counterintuitive corollary follows from considering both sides of model complexity. That is: two models with the same number of parameters but disparate functional forms may differ markedly in complexity. For example, the models y = x + b and y = e^{xb} have the same number of parameters, but they differ in complexity, such that the pliable exponential function will be much better at fitting data than the rigid linear function (Myung, Pitt, & Kim, 2005).

In item response theory (IRT) modeling, complexity has been gauged almost exclusively by counting parameters. Popular likelihood-based model evaluation indices such as the Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1974) include penalties for the number of free parameters in an IRT model, but they do not address the functional form issue. Bayesian measures of model quality, such as the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 1998; Fox, 2010), improve upon AIC and BIC by calculating more sophisticated versions of complexity (e.g., the effective number of parameters (Moody, 1992)). Yet, however elaborate the accounting of effective parameters may be, the functional form of the model receives little direct discussion. Suffice it to say that commonly used model evaluation indices take into consideration the first contributor to complexity (the number of free parameters), but they do not directly address the second (the model's functional form).

In the present study, we deviate from the reigning frequentist and Bayesian paradigms and focus instead on model evaluation from the perspective of information theory, a field that may be less familiar to researchers in psychological and educational measurement. The goal in the information-theoretic approach to modeling is to *compress* the data as much as possible by identifying *regularities* (i.e., patterns or trends) in the data, and thereby to better predict unknown data. Early proponents of this idea include Kolmogorov (1963), who presented the definitive notion of extreme data compression (aptly named Kolmogorov complexity), and Solomonoff (1964), who sought to mathematically formalize Occam's razor by synthesizing techniques from mathematics and computer science (e.g., Shannon, 1948; Huffman, 1956). These and other ideas from early information theorists served as a philosophical and statistical basis for Rissanen (1978, 1983, 1989) to put forward the principle of minimum description length (MDL).

MDL is a method of inductive inference, based on the idea that data can be represented by a set of symbols—or *code*—that is shorter than the literal length of the data set itself. MDL states that the more regularities that exist in the data, the more the data can be compressed. Conversely, the more one is able to compress the data, the more one can learn about the data (i.e., by understanding the regularities in the data). A primary tenet of the philosophy underlying the MDL principle is that the goal of inductive inference should be to "squeeze out as much regularity as possible" from the data (Grünwald, 2005). The main task is to separate structure (i.e., meaningful information) from noise (i.e., accidental information); to correctly model the data, one must identify the structure and minimize the noise. Of course, noise is defined relative to the specific model under consideration. In information-theoretic terms, noise is represented as the residual number of *bits* or *nats*[1] needed to encode the data after the optimal model has been selected. In that sense, noise is not a random variable; it is a function of the selected model and the observed data.

MDL is especially useful when choosing between competing models. If the choice between candidate models is based simply on goodness-of-fit to the observed data, then there is a risk that the better fitting model will overfit the data. Although a model with fewer parameters may do a better job of identifying the underlying trend in the data, a model with more parameters will achieve better fit by capturing more of the random noise. In order to select the best model, a tradeoff is needed between goodness-of-fit and model complexity.

The two-part version of the MDL principle addresses this tradeoff directly, by taking into consideration both the number of freely estimated parameters and the model's functional form. Let H_1, H_2, ..., H_n be a list of candidate models that each represent a different hypothesis about the data. In information-theoretic terms, the best hypothesis H to explain the data D is the one that minimizes the sum of two parts: $L(\mathrm{H}) + L(\mathrm{D}|\mathrm{H})$, where $L(\mathrm{H})$ is the length, in bits or nats, of the description of the hypothesis H, and $L(\mathrm{D}|\mathrm{H})$ is the length of the description of the data once it has been encoded according to the hypothesis. In more common terminology, $L(\mathrm{H})$ represents the model itself and $L(\mathrm{D}|\mathrm{H})$ represents the goodness-of-fit of the model to the data. One can usually find a very complicated model (i.e., a model with large $L(\mathrm{H})$) to explain the data, and it

[1] A *bit* is the base-2 unit of information (i.e., 0 or 1) (see e.g., Shannon & Weaver, 1949); a *nat* is the base-*e* unit of information (Boulton & Wallace, 1970). One nat ≈ 1.443 bits.

may have excellent fit (i.e., small L(D|H)). Alternately, one can find a simplistic model (small L(H)) that has very poor fit (large L(D|H). Under the MDL principle, the sum of these two parts will be minimized to arrive at a hypothesis/model that is relatively (but not overly) simple and has good (but not perfect) fit. In the early articles on MDL, Rissanen (1978, 1983) advocated choosing a minimax code that minimizes the shortest total description length L(H) + L(D|H) over all possible data sequences.

While no one statistical perspective will provide a complete and unequivocal evaluation of a model, MDL offers a unique perspective that complements the established methods. Frequentists often evaluate their models by considering goodness-of-fit to the observed data or applying a tool such as the parametric bootstrap (Efron & Tibshirani, 1993), which involves drawing many resamples from a parametric estimate of the population (i.e., a model that has been fit to the observed data). Bayesian model evaluation is now routinely carried out via prior (Box, 1980) or posterior predictive model checking (Guttman, 1967; Rubin, 1984; Gelman, Meng, & Stern, 1996), wherein data are generated from the prior or posterior predictive distributions and compared in some particular way (using test quantities sensitive to specific kinds of misfit) with the observed data. The information-theoretic approach differs markedly from these techniques in that it does not rely on observed data; instead, a model is evaluated relative to the *complete* data space. In assessing how a particular model performs relative to any possible data, researchers can use MDL to identify important features of the model (i.e., its inherent performance regardless of the observed data) that cannot be detected using more common methods.[2]

[2] An accessible introductory overview of MDL is given by Grünwald (2005) and explicit comparisons with frequentist and Bayesian methods are given by Vitányi and Li (2000), Markon and Kreuger (2004), and Lee and Pope (2006).

The present study was inspired by the work of Preacher (2006), who explored the MDL principle in the context of structural equation modeling (SEM). Specifically, he examined the concept of *fitting propensity* (FP)—a structural model's ability to fit diverse patterns of data, all else being equal. He found that models with the same number of free parameters, but different structures, may exhibit different FP. That is, the arrangement of the latent variables and associated parameters in the model may result in an inbuilt tendency to fit any possible data. While this line of research is quite promising, it has yet to be embraced by SEM scholars, partly because it involves the complicated task of generating and estimating uniformly distributed random covariance/correlation matrices, and perhaps also because of the unfamiliar reasoning underlying the principle of MDL.

Although the philosophical and logical elements of MDL may be alien to many psychometricians, item response theory (IRT) appears to be more accommodating than SEM with regard to various technical aspects of MDL analysis. IRT, unlike SEM, was developed exclusively for modeling categorical item level data, and this greatly simplifies the process for exploring the data space, as discussed below. Further, there are a number of statistics that can be derived from an IRT analysis and evaluated in accordance with the MDL principle. These statistics include item-fit measures, local dependence indices, and other aspects of item-level analysis that are uncommon in SEM research. Thus, while Preacher (2006) invoked the MDL principle to provide valuable insight regarding the global fit (via the standardized root mean square) of competing structural models, the IRT analysis presented herein explores not only global fit, but also several statistics specific to model appraisal in routine item analysis.

Models under investigation

In Preacher's (2006) work on SEM, the structural models were more or less *ad hoc* arrangements of causal paths between a few latent variables. Although we believe that his findings about FP in SEM were profound, a potential reader not familiar with MDL could easily dismiss the differences as too specific to the nature of the models he chose, making it difficult to generalize MDL for use in other research scenarios. Common IRT models, on the other hand, are given labels that identify the construction of their item trace lines (e.g., 1PL, 2PL, 3PL, graded response, etc.) and/or their multidimensional factor/attribute structure (e.g., bifactor, 2nd-order, correlated traits, two-tier, etc.), or both. This enables one to draw important MDL-based conclusions about certain named and widely used models that are not only popular in research settings, but are used throughout the educational and psychological measurement community at large.

It is important to note that this investigation is not a standard model selection analysis, in which competing models of some theory are evaluated in order to determine which best corresponds to the data. Rather, four of the five models below were chosen solely because they have different functional forms, but the same number of parameters. The remaining model allowed us to explore the FP of a relatively simpler model that includes more parameters. Excepting functional form and number of parameters, all other aspects of the models under consideration (e.g., theoretical underpinnings, practical applications, etc.) are only tangential to the goal of our comparisons. Our intention is not to provide guidelines regarding whether a practitioner should choose between, say, a bifactor and unidimensional model if a certain fit index is above or below a certain value. Instead, we take an approach that aims to uncover an intrinsic property of a few popular IRT models (of the many possible), regardless of whether these models would ever appear alongside one another in a typical theory-guided model selection scenario.

Exploratory item factor analytic model

The first IRT model under consideration was an exploratory item factor analysis (EIFA; Bock, Gibbons, & Muraki, 1988; Gorsuch, 1997; Wirth & Edwards, 2007; Cai, 2010) model. In utilizing an exploratory (rather than confirmatory) multidimensional IRT model, the researcher does not fix *a priori* any of the paths between the latent and observed variables; rather, the model is free to “explore” the combination of latent factors that best represent the manifest variables (i.e., with optimal interpretability and parsimony). Figure 1(a) provides a visual representation of the EIFA model under analysis, following the standard graphical practices common to structural equation modeling; in this case, two factors (circles) were selected to represent the seven items (rectangles). As this was an exploratory model, all of the items were free to load on both of the factors, save the path from Factor 2 to Item 1, which was constrained to zero for model identification purposes.

Bifactor model

The second model under investigation was an item bifactor model. The bifactor model (Holzinger & Swineford, 1937) is a factor structure wherein the covariance among a set of items is explained by a single primary dimension (or “general factor”) and multiple specific dimensions (or “group factors”). The primary dimension in a bifactor model represents the overall construct that the test was designed to assess, while the specific dimensions represent additional variation due to narrow subconstructs among non-overlapping groups of items. A psychiatric screening questionnaire, for instance, might measure overall depression (the primary dimension) by including small clusters of questions about mood, sleeping habits, diet, and so forth (the specific dimensions).

The bifactor model has enjoyed a resurgence of late (Reise, 2012), partly because of its strong performance in a number of model comparison studies. Rodriguez, Reise, and Haviland (2015), for example, examined 50 recent psychological research articles in which the bifactor model was selected as the best choice among several competing models. All too frequently, this decision was based primarily on the superior goodness of fit of the bifactor model, with minimal regard for its complexity or ability to generalize to future data. In some studies, goodness-of-fit alone was offered not only as a representation of the sample data, but as support for some posited theory or hypothesis. For example, Longley, Calamari, Wu, and Wade (2010), developed competing models of anxiety symptoms and concluded, "The better fit of the bifactor model indicates congruence with the integrative model and our hypotheses about hypochondriasis and [obsessive-compulsive disorder] and panic attack symptoms" (p. 461).

However, the tendency of the bifactor model to exhibit superior goodness-of-fit may be due to its inherent ability to capture random noise in the data. That is, the functional form of the bifactor model may enable it to overfit the sample data, thereby causing researchers to draw conclusions that do not generalize to other scenarios. As Thomas (2012) cautioned, "Indiscriminate use of the bifactor model without proper regard for theory is highly questionable . . . Simply put, the bifactor model's added benefit may not excuse its complexity" (p. 108). Indeed, one of the motivations for this study was the need to formally evaluate the bifactor model, and by doing so, to lessen its "indiscriminate use."

The particular bifactor structure that was analyzed in the present study is shown in Figure 1(b). This model included a primary dimension, upon which all seven items loaded, and two specific factors. The first specific factor explained variance among items 1 through 5 and the second factor explained variance among items 6 and 7. For model identification purposes, the item factor loadings associated with the second specific factor were constrained to be equal. This particular arrangement was chosen only because it includes the same number of free parameters found in the EIFA model.

Diagnostic classification models

Another type of latent variable model involves classifying individuals with regard to the discrete attributes underlying the items on a test. For example, presence of symptoms (in psychological assessment) or mastery of skills (in educational assessment) can be specified as discrete latent variables that are related to the responses on observed items. Models that include such variables yield attribute profiles (i.e., latent classes)—patterns of presence/absence of psychological symptoms or mastery/non-mastery of skills—that can be used to diagnose psychological disorders or ascertain academic achievement. These sorts of models are referred to as diagnostic classification models, cognitive diagnostic models, cognitive assessment models, or restricted latent class models, among other labels (Rupp, Templin, & Henson, 2010). While this burgeoning area of psychometrics comprises many interesting item response models, the present study focused on two popular diagnostic classification models: the deterministic input noisy and-gate (DINA) model and the deterministic input noisy or-gate (DINO) model.

The DINA model (Haertel, 1989; Junker & Sijstma, 2001) is non-compensatory, or conjunctive, meaning that presence/mastery of one attribute will not compensate for absence/non-mastery on other attributes. The “and-gate” portion of the DINA acronym indicates that all item attributes must be present/mastered in order to endorse an item/produce the correct response. For example, part A of the DSM-5 diagnostic criteria for autism spectrum disorder requires the presence of deficits in social-emotional reciprocity *and* deficits in nonverbal communicative behaviors *and* deficits in developing, maintaining, and understanding relationships (American Psychiatric Association, 2013). The presence of just one or two of these attributes would not suffice for clinical diagnosis.

Of course, not all attributes will be non-compensatory. For instance, it may be that presence/mastery of any one of the assessed attributes associated with an item will compensate for the absence/non-mastery of the other attributes (e.g., due to the existence of multiple solution strategies in a math item). The DINO model (Templin & Henson, 2006) specifies this particular hypothesis. Here, an “or-gate” models the odds of endorsement/success response and is indifferent to which one or more of the attributes that the respondent possesses. For example, consider Part B of the DSM-5 diagnostic criteria for autism spectrum disorder, which requires the presence of at least two of the following: repetitive motor movements *or* inflexible adherence to routines *or* intensely fixated interests *or* hyperreactivity to sensory input (American Psychiatric Association, 2013). The presence of all four symptoms is not necessary for diagnosis; rather, the presence of any pair of these symptoms will be sufficient. Although this example does not strictly lead to a DINO model, our point is that such indifference in classifications might occur quite frequently in practical settings.

An informative property of both the DINA and DINO models is the ability to model the attribute space. That is, the pattern of symptomatology or mastery that makes up an attribute profile can itself be measured with an item factor model utilizing either the logistic or normal-ogive parameterization (de la Torre & Douglas, 2004). In this so-called “structured tetrachoric model” (Rupp, Templin, & Henson, 2010), each of the discrete latent attribute variables loads on one or more continuous higher-order latent factors. High loadings in the attribute space would indicate a strong relationship between the latent factor(s) and the diagnostic or classification criteria represented by the attributes.

The DINA and DINO models that were analyzed in this study are illustrated in Figure

1(c) and (d). Several diagrammatic conventions have been introduced to represent the distinct characteristics of diagnostic classification models. First, the latent attribute variables are divided by a chord, which serves as a visual reminder that these are discrete variables. Second, each diagram includes a pair of cross-loadings, or “interaction effects,” which showcase the key difference between these two models. Consider Item 3 for example. In the DINA model, a correct response to Item 3 would require mastery of both Attributes 1 *and* 2. In the DINO model, a correct response to Item 3 would require mastery of either Attribute 1 *or* 2. The remaining paths (denoted as λs) from the attributes to Items 1, 2, 4, 6, and 7 are termed “main effects,” indicating items that are associated with a single attribute. Finally, the higher-order latent variable represents a continuous dimension and so does not include a horizontal chord.

In typical diagnostic classification modeling applications, it is essential that the Q-matrix, which represents relationships among items and attributes, is specified *a priori* (e.g., Rupp & Templin, 2008b) or freely estimated (e.g., Chen, Liu, Xu, & Ying, 2015). However, in the present investigation, the DINA and DINO models were not evaluated with regard to real data in which proper identification of and alignment with the underlying attributes is necessary. Rather, these models were fit to random data, as discussed below. Q-matrix specification in the context of FP is a topic of future investigation.

Unidimensional 3PL model

The final model under investigation was a unidimensional 3-parameter logistic (3PL) IRT model (Birnbaum, 1968), a psychometric model that is widely used in educational measurement (e.g., in the National Assessment of Educational Progress). As depicted in Figure 1(e), this model included a single latent dimension to account for the covariance among the seven items. As far as model complexity related to dimensionality is concerned, a unidimensional structure exemplifies the simplest possible functional form. As discussed below, however, the complexity of each item within the model may allow a unidimensional structure to be more flexible than certain multidimensional models.

Differences in free parameters

Table 1 enumerates the freely estimated parameters in each of the models. Although the EIFA, bifactor, DINA, and DINO models differed in functional form, each model was specified to include exactly 20 freely estimated parameters (whether factor loadings, slopes, or attribute main or interaction effects). Controlling for the number of free parameters in the multidimensional model structures ensured that observed differences in FP were due to functional form rather than the number of free parameters. The fifth model under investigation was a simple unidimensional structure, but each item was measured using a 3PL function. Specification of the 3PL for all seven items in the unidimensional model resulted in 21 free parameters; relative to the multidimensional models under consideration, the unidimensional model had an extra free parameter. In keeping with the traditional "parameter counting" view of model complexity, the enhanced flexibility of the 21-parameter unidimensional 3PL model should cause it to have a higher FP than each of the 20-parameter multidimensional models.

Hypotheses

Regarding the performance of these models in the context of any possible data, we offer two hypotheses.

Hypothesis 1: The EIFA model will exhibit, on average, the highest fitting propensity.

The EIFA model was included as a baseline of sorts, since the exploratory nature of this model should imbue it with the highest degree of FP. That is, unless the underlying "true" data generating mechanism of the chosen data set just happens to represent at random a bifactor, DINA, DINO, or unidimensional 3PL model, then the EIFA model should always fit best.

Hypothesis 2: The bifactor model will display higher fitting propensity than the DINA and DINO models.

The second hypothesis is that the bifactor item response model, relative to the DINA and DINO models, will fit a greater number of data sets that are randomly sampled from (and uniformly distributed over) the complete data space. As discussed earlier, the bifactor model has become increasingly popular in recent years (e.g., Reise, 2012), due in part to its ability to closely fit the observed data. However, we hypothesize that the functional form of the bifactor model instills in it an undesirable tendency to fit any possible data.

Finally, the unidimensional 3PL model was included in this analysis to better understand the notion of complexity in IRT models. Among the five factor structures included in the present study, the unidimensional model has the simplest functional form; it is the only model in which each item loads on a single latent variable. This economy of functional form may cause the unidimensional 3PL model to have a drastically reduced FP. However, all items in the unidimensional model were fit using 3PL trace lines, which increased the model's complexity, as gauged by traditional metrics (i.e., the number of free parameters). Because of this incongruity between functional form and parametric complexity, we do not offer a clear hypothesis regarding the unidimensional 3PL model; we choose instead to simply observe its performance relative to the competing multidimensional models.

Method

Preacher (2006) noted that mathematical expressions of the MDL principle (see Grünwald, 1997, for many examples) are intractable due to integration over the complete data space. He concluded, "Until a good analytic approximation can be identified, calculation of an

MDL index in the SEM context involves fitting a model to a large number of random data sets" (p. 249). This echoes the earlier recommendation of Cutting, Bruno, Brady, & Moore (1992) that a baseline for model fit can be established by fitting models to random data. As Cutting et al. argued, "Without baseline comparisons of models run on random data, we think any approach that proceeds by comparing models with matched numbers of parameters may be in jeopardy' (p. 380). Herein, we intend to establish FP baselines for the five IRT models by examining their functioning relative to many data sets that were randomly sampled from and uniformly distributed across the complete data space.

Data generation

SEM is sometimes referred to as a "moment structure analysis" (e.g. Bentler & Weeks, 1979) in which the moments to be analyzed are the covariances (or correlations) between the manifest variables rather than the raw data. IRT can also be thought of as a type of moment structure analysis in which the moments to be analyzed are not the raw data, but the probabilities of obtaining each response pattern associated with the full underlying multinomial contingency table formed by the item-by-item cross-classifications (Holland, 1990). Instead of the multinomial, Teugels (1990) used the multivariate Bernoulli distribution, and in this representation, the IRT model may be thought of a true moment structure model for all marginal moments of the multivariate Bernoulli. In either case, to represent the complete data space and to avoid imposing any *a priori* structure within the data generating mechanism, it is necessary (a) that we generate the probability vectors for every possible response pattern (rather than the raw response data itself), and (b) that these probability vectors are uniformly distributed and that the probabilities within each vector sum to 1.0.

Part (a) above is straightforward. On a given test, each examinee or respondent provides a pattern of responses to the *n* items. When test items are scored dichotomously, such as in the case of this research, the data are frequencies for 2^n possible response patterns. Each pattern will be represented by a probability (greater than or equal to 0) that corresponds to either the proportion of examinees with that observed pattern from a sample (the observed data) or the probability of the population of examinees that are expected to have that response pattern (the IRT model), but the summed total probability across all possible patterns or actually observed patterns must equal 1.0.

Part (b) of our data generation strategy exploits a fundamental statistical property of IRT. In item response modeling, the complete data space therefore consists of all points on a multinomial simplex, and the size of the simplex is completely pre-determined by the number of items. This is the data space we must explore. We acknowledge that there may be a number of ways to generate random data that best represent this space, and that different random data generation methods may produce different results (see e.g., Botha, Shapiro, & Steiger, 1988). Here, we adopt the methodology of Smith and Tromble (2004), who established that sampling from a simplex is ideal for situations in which the goal is to obtain random multinomial probability distributions that are uniformly sampled across a range from 0.0 to 1.0.

Our example test included seven dichotomously-scored items and the simulated response proportions of each of the $2^7 = 128$ possible response patterns were sampled from a multinomial simplex. To ensure comprehensive coverage of the data space, this sampling process was replicated such that 1,000 unique random data sets were created. The data generation script, written in the R statistical software program (R Core Team, 2014), is presented in the Appendix.

It is important to note that our data generation process will not result in data that have any known underlying structure. The data were explicitly designed to represent the *complete* data space, which implies that each of the models should fit well to at least some subset of the data sets. To be precise, the complete data space necessarily includes data that truly are unidimensional in nature, data that truly align to a bifactor structure, and so on. Any of the five candidate models that demonstrates a predisposition to fit well to a relatively large number of these data sets may be regarded as remarkably (perhaps overly) flexible. If the bifactor model tends to fit well to many data sets, for example, that is not because the data were necessarily generated from a bifactor structure; rather, such a finding would highlight a property of the bifactor model itself, as an excessively flexible model that bends to fit an exorbitant number of data sets.

Estimation specifications

Once the randomly weighted data were generated, an R script was written to fit each of the five models to the same 1,000 data sets using the flexMIRT software program (Cai, 2013). In all models, all item parameters were estimated using the Bock-Aitkin expectation-maximization (EM) algorithm (Bock & Aitkin, 1981). For the models with continuous dimensions, 49 equally spaced quadrature points between -6.0 and 6.0 were used. The EIFA, bifactor, and unidimensional models used cross-product approximation to compute standard errors, while the two diagnostic classification models used the Richardson extrapolation method (Tian, Cai, Thissen, & Xin, 2013). The diagnostic models also differed from the others by specifying a maximum of 5 M-step iterations, rather than the flexMIRT default of 100. These changes in estimation of the cognitive diagnostic models were advised by the authors of the flexMIRT software (Houts & Cai, 2013). Additionally, to improve the stability of estimation of the unidimensional 3PL model, a Beta(1.0, 4.0) prior was specified for the pseudo-guessing parameter of each item.

In all cases, the potential difficulties in iteratively fitting models to random data where no good known starting values or solution paths exist were partly mitigated by setting the E-step tolerance at .001 (rather than the more stringent flexMIRT default of .00001) and increasing the maximum number of E-step iterations to 20,000 (from the flexMIRT default of 2,000). For the purposes of this study, estimation convergence was defined as the detection of a local maximum according to a 2nd-order test performed by the software (Houts & Cai, 2013). Despite the relaxed tolerance and the increase in estimation iterations, there were still a number of replications that did not settle on stable parameter estimates. Table 2 displays the convergence rates for each of the dichotomous IRT models under investigation. The unidimensional 3PL model had the highest non-convergence rate—when fit to the unidimensional model, 24.3% of the data sets failed to converge on a stable solution. Non-convergence rates were slightly lower for the EIFA (21.0%) and bifactor (18.4%) models. Interestingly, the two diagnostic classification models had far greater success with regard to estimation convergence. The DINA model converged on stable estimates in all but 4.7% of the data sets; the DINO model fared even better, obtaining stable solutions in all but 3.9% of the data sets. We echo the reasoning of Preacher (2006), who argued that in demonstrating FP, estimates computed after 10,000 iterations can be accepted as the final (converged) estimates, notwithstanding their possible instability. By specifying a maximum of 20,000 E-step iterations, our aim was to afford further confidence in the non-converged results.

Measures of fit

To appraise the FP of various structural models, Preacher (2006) selected the root mean squared residual (RMSR; Jöreskog & Sörbom, 1996) as the appropriate metric of model fit. RMSR was chosen specifically because it does *not* adjust for the number of free parameters or the functional form of the model. RMSR is, in a sense, a “pure” measure of fit that is unswayed by the characteristics of the model. Thus, RMSR allows one to measure the FP of competing models simply by comparing differences in their fit to the same (random) data.

However, similar to other common fit measures in SEM, the computation of RMSR requires a correlation matrix based on continuous variables, which makes it unsuitable for IRT investigations. Thus, the present analysis focused on two diagnostic measures that were developed specifically for use in categorical data analysis: the Y2/*N* statistic and the LD X^2 local dependence index.[3]

Y2/N statistic

Perhaps the closest analog to RMSR that currently exists for discrete data is the Y2 statistic (Bartholomew & Leung, 2002; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006). This fit statistic is found by summing all the univariate and bivariate marginal chi-squares derived from the contingency tables of item response probabilities:

$$\mathrm{Y2} = N\left[\sum_{i=1}^{I}\frac{(o_i - e_i)^2}{e_i(1-e_i)} + \sum_{i=1}^{I-1}\sum_{j=i+1}^{I}\frac{(o_{ij} - e_{ij})^2}{e_{ij}(1-e_{ij})}\right], \tag{1}$$

where *N* is the sample size, *I* is the number of items, o_i and e_i are the observed and expected response frequencies for the endorsement/correct response to item *i*, and o_{ij} and e_{ij} are the observed and expected response frequencies for item pair *ij*, wherein both items are endorsed or correct. Y2 denotes the magnitude of the discrepancy between the data and the statistical model; it is a “badness-of-fit” index in that higher values indicate worse fit. In the present study, Y2 was divided by the sample size *N* to produce the Y2/*N* statistic. This slightly modified version of the Y2 statistic is independent of sample size. To date, no benchmark values have been established

[3] We also explored FP by investigating the S-X^2 item fit index (Orlando & Thissen, 2000; 2003), the D^2 latent distribution fit index (Li & Cai, 2012), and the marginal χ^2 values of each of the five IRT models. Due to page limitations, discussion of these indices and their bearing on FP is included in the online supplementary material.

for the Y2/*N* statistic, but it is similar in nature to a population level discrepancy measure of the degree of misfit in first and second order marginal moments implied by the multinomial distribution of response patterns.

LD X^2 local dependence index

IRT models assume that items are only correlated through the underlying latent construct that the item set is designed to measure (Lord & Novick, 1968). If residual correlations exist after accounting for the correlations explained by the latent trait, then the assumption of local independence has been violated. Chen and Thissen (1997) developed the LD X^2 index to address local dependence (LD) violations in IRT models. To compute this index, phi correlations are calculated for the observed and expected bivariate contingency tables. When the observed correlation is higher than the model-implied correlation for an item pair, the result is positive LD; if the model-implied correlation is higher, then negative LD has been detected within that item pair. The absolute value of the LD X^2 statistics can then be tested against some critical value to determine whether the violation is ignorable (Houts & Cai, 2013).

Results

Y2/*N* statistic

For all five models, Table 3 displays the overall means and standard deviations of the Y2/*N* statistic for the total, converged, and non-converged analyses, as well as the difference between the converged and non-converged analyses. This table provides a general comparison between all models, as well as a more detailed comparison of the converged and non-converged analyses within each model. Beginning with the between-model comparisons across all 1,000 data sets, Table 3 reveals that on average, the EIFA and bifactor models produced Y2/*N* values of .05 or lower. That is, on average, the bifactor model was almost as capable as the EIFA model with regard to fitting any possible data. The DINA and DINO models tended to have Y2/*N* values of .10, and the unidimensional 3PL model yielded an average Y2/*N* of .13. Table 3 also facilitates within-model comparisons of the converged and non-converged data sets. Although there were differences in each model's convergence rate (as discussed earlier with regard to Table 2), the Y2/*N* results were fortunately not affected by the lack of convergence. The indistinguishability of the converged and non-converged analyses gives credence to the deeper investigation of Y2/*N* shown below, wherein the results are based on all 1,000 data sets regardless of non-convergence.

Although the descriptive statistics hint at differences between the models, FP is better expressed through visualizations of Y2/*N*. Figure 2 displays the empirical cumulative percentage distribution of the Y2/*N* statistic in each of the five models. The curves in the figure simply display the percentage of data sets that achieved a particular value of Y2/*N* when fit to each model. This type of figure allows for the models to be compared in two ways. The first is by investigating the vertical distance between the curves at some particular value of Y2/*N*. For example, consider the vertical grid line at Y2/*N* = .05. The EIFA curve intersects with this line at *y* = 79%, meaning that 790 of all 1,000 data sets had Y2/*N* values of .05 or lower when fit with the EIFA model. The curve associated with the bifactor model reveals that this popular structure produced a Y2/*N* of .05 or less in 63.5% of all data sets. The diagnostic classification and unidimensional 3PL models were far less likely to yield Y2/*N* values as low as .05. Specifically, Y2/*N* values of .05 or lower were only obtained in 5.0% of data sets fit to the DINA model, 5.2% of the data sets fit to the DINO model, and 2.3% of the data sets fit to the unidimensional 3PL model.

The second way to compare the cumulative percentage distributions is to consider the horizontal discrepancy between the curves in Figure 2. Suppose that a researcher is interested in evaluating each model, not by selecting some referent value of Y2/*N*, but instead against some benchmark percentage. The horizontal grid line at $y =$ 80%, for instance, indicates that 80% of all EIFA fittings achieved Y2/*N* $\leq .05$, 80% of all bifactor fittings resulted in Y2/*N* $\leq .06$, 80% of all DINA and DINO fittings had Y2/*N* $\leq .13$, and 80% of all unidimensional 3PL fittings produced Y2/*N* $\leq .17$. Indeed, an inspection of the Y2/*N* values at every 10th percentile reveals a consistent pattern: the EIFA model always had the lowest Y2/*N* value, the bifactor model followed closely behind, the two diagnostic classification models produced higher values (and performed almost identically), and the unidimensional 3PL model offered the highest Y2/*N* values. A few interesting comparisons can be made. For example, 40% of EIFA model fittings yielded Y2/*N* values of .03 or lower, but not a single DINA or unidimensional model fitting produced Y2/*N* values of that magnitude. An even more drastic disparity is found by comparing the highest and lowest deciles: 90% of EIFA and bifactor model fittings revealed Y2/*N* ≤ .06 and .08, respectively, but only 10% of DINA, DINO, and unidimensional model fittings resulted in similar Y2/*N* statistics.

It is clear from the Y2/*N* results that the EIFA and bifactor models possessed much greater propensities to fit any possible data. These findings, while informative, do not offer any details about the degree of overlap between the models. In the MDL literature, it is not uncommon to see figures showing hypothetical regions of the complete data space that are "occupied" by competing models (see e.g., Pitt, Myung, & Zhang, 2002). It could be, for instance, that even though the DINA and DINO models tend to fit well to approximately the same *percentage* of data sets, the actual data sets that they fit well could be completely different. A series of visualizations were created to better understand how the five models under investigation interacted within the complete data space.

The “amoeba” plots presented in Figures 3 and 4 depict the FP of each model at various levels of Y2/*N*, but they also reveal the overlap (and lack thereof) that characterizes these models. In each of these figures, the square area represents the complete data space. The transparent colored regions represent the percentage of all 1,000 data sets that were fit by the corresponding model at a specific value of Y2/*N*. The regions are drawn roughly to scale; the values that accompany each figure indicate the size of each region as well as the precise degree of overlap between regions.

The top panel of Figure 3 shows one of the simplest scenarios: Y2/*N* ≤ .01. Here, the EIFA (black) model occupied just 1.4% of the complete data space and the bifactor (green) model occupied 0.9%. That is, at this strict Y2/*N* criterion, the EIFA model fit well to 14 of the 1,000 random data sets and the bifactor model fit well to 9 data sets. This figure reveals that the bifactor region was not fully subsumed by the EIFA region; there were some data sets that were fit well by the EIFA model but not the bifactor model, and vice versa. As the figure shows, the overlap between the EIFA and bifactor models (denoted as region A) occupied 0.4% of the data space, meaning that 4 out of 1,000 data sets were fit extremely well (Y2/*N* ≤ .01) by both models.

Regions B and C in this first amoeba plot highlight the unique data sets that were fit well by each model. The EIFA model fit 1.0% (region B), or 10 data sets that were not fit by the bifactor model; the bifactor model fit 0.5% (region C), or 5 data sets that were not fit by the EIFA model. Finally, the values presented alongside these regions show that at Y2/*N* ≤ .01, the DINA, DINO, and unidimensional 3PL models did not occupy any part of the data space, and that 98.1% of the complete space remained unoccupied by any of the candidate models.

The bottom panel of Figure 3 depicts regions of the complete data space that were occupied by each model when Y2/*N* ≤ .03. In this case, the EIFA model fit 36.2% of all data sets and the bifactor model fit 27.0%. These two models overlapped such that 22.7% (region A) of all data sets were fit well by both models. Note, however, that 4.3% (region C) of the data sets were fit by the bifactor model but *not* the EIFA model. The DINA, DINO, and unidimensional 3PL models made an appearance when Y2/*N* ≤ .03, though the regions they occupied were quite small and the overlap between them was extensive. Specifically, the diagnostic classification and unidimensional models each fit the same two data sets (region D), and each also fit one unique data set (regions E, F, and G). Finally, at this Y2/*N* benchmark, 59.5% of the total data space was not occupied by any of the models.

The top panel of Figure 4 is a visualization of the complete data space when Y2/*N* ≤ .05. Here, the EIFA region occupied the majority (79.2%) of the space and the bifactor model was not far behind (63.5%). The overlap between these two models was sizeable—over half (51.8%) of all data sets were fit well by both the EIFA and bifactor models. Even at this level of Y2/*N*, however, there were still a few data sets (region C: 3.8%) that were fit by the bifactor model but not the EIFA model.

The DINA, DINO, and unidimensional 3PL models were completely subsumed by the bifactor and EIFA models; that is, there were not any data sets that were fit well by the diagnostic classification or unidimensional models without also being fit well by the bifactor or EIFA models. However, the blue region shows that the DINA model fit 5.1% of all data sets at Y2/*N* ≤ .05, and regions E + I indicate that 2.1% of data sets were fit by the DINA model but not the DINO model. The yellow region shows that the DINO model fit 5.2% of all data sets and regions F + J reveal that 2.3% of data sets were fit by the DINO model but not the DINA model.

The red region indicates that the unidimensional 3PL model fit 2.3% of all data sets at

Y2/*N* ≤ .05. While there was some overlap between the unidimensional and diagnostic classification models, there were still 7 data sets (region G) that were fit by the unidimensional model without being fit by either the DINA or DINO models. Another region of interest is region D, which represents the overlap of all five models. This region occupied 1.1% of the complete data space; that is, 11 data sets in the simulation were fit well by all models. Finally, note that when Y2/*N* ≤ .05, only 17.1% of the complete data space was not occupied by any model.

The bottom panel in Figure 4 displays the total data space when Y2/*N* ≤ .10. Here, the EIFA and bifactor models fit almost every data set, occupying 99.2% and 97.3%, respectively, of the complete data space. Yet there were still 4 data sets (region C) that were fit by the bifactor model but not the EIFA model. At this level of Y2/*N*, the DINA and DINO regions also showed considerable overlap; each of these models occupied over 52% of the data space, but 42.2% (regions D + H) of all data sets were fit by both the DINA and DINO models. In the center of this figure, region D indicates that 228 of all 1,000 data sets were fit by all five models when Y2/*N* ≤ .10. At this relatively high level of Y2/*N*, only 4 data sets were not fit by some model.

Overall, the Y2/*N* results revealed that the FP of the bifactor model approached that of the EIFA model—a model specifically intended to find the solution that best fits the data. The two diagnostic classification models had far lower FP and performed very similarly (though not identically) to one another with regard to Y2/*N*. The most counterintuitive finding is related to the unidimensional 3PL model. This model had an additional free parameter which, according to conventional views of complexity, should have supplied it with a superior ability to capture noise in the random data. And yet, the unidimensional model was, by far, the least inclined to fit well. Possible explanations for this will be discussed later.

Aside from exposing the FP of several widely used IRT models, these results could guide the interpretation of the Y2/*N* statistic. As mentioned earlier, no cutoff criteria have been established for this statistic. The simulation results, particularly those presented in the amoeba plots, indicate that a Y2/*N* cutoff of .01 is probably too low; the DINA, DINO, and unidimensional 3PL models did not fit a single data set at this level of Y2/*N*. At the opposite end of the spectrum, a Y2/*N* cutoff of .10 appears to be too lax—at this benchmark, the EIFA model fit all but 8 of 1,000 data sets, the bifactor model fit all but 27 data sets, and overall, only 4 data sets eluded all models. Perhaps a Y2/*N* cutoff of .05 is more appropriate. At this degree of goodness-of-fit, the more flexible models (EIFA and bifactor) tended to fit around 2/3rds of all data sets, while the less accommodating models (DINA, DINO, and unidimensional 3PL) tended to fit around 1/20th of all data sets. Thus, a Y2/*N* of .05 or lower was somewhat informative with regard to all of the models under investigation.

LD X^2 local dependence index

Table 4 includes for all models the means and standard deviations of the LD X^2 local dependence index, aggregated across all item pairs. As expected, the relatively flexible EIFA and bifactor models were adept at accounting for the local dependence between all item pairs, and the unidimensional 3PL model was typically the least effective model for addressing local dependence. In addressing local dependence, the unidimensional model was handicapped by its meager functional form; the added free parameter did nothing to aid in decreasing the local dependence. This table also confirms that there were inconsequential differences between the converged and non-converged results.

The LD X^2 results for each item pair revealed several notable exceptions to this pattern of results. In 11 of the 21 item pairs, the bifactor model had greater success than the EIFA model in handling the local dependence between items. Specifically, the bifactor model was better at capturing the noise caused by local independence violations in item pairs 2 & 1 through 5 & 4, as well as item pair 7 & 6. It is unsurprising that the bifactor model addressed the local dependence involved with these particular item pairs – the specific factors in this bifactor structure, as illustrated in Figure 1(b), were explicitly constructed to address dependence between Items 1 through 5 (Specific Factor 1) and Items 6 and 7 (Specific Factor 2). What is surprising is that the EIFA model, which allowed all items to load on both factors (excepting the path from Factor 2 to Item 1, which was fixed at zero to identify the model), was unable to account for the local dependence as successfully as the bifactor model in every situation. Perhaps the bifactor model's high FP is in part due to its heightened ability to model specific local dependence noise.

Another counterintuitive result from the LD X^2 analyses was the occasional failure of the (multidimensional) diagnostic classification models to manage local dependence violations as effectively as the undimensional 3PL model. Among item pairs 5 & 2, 6 & 1, 6 & 2, and 7 & 2, the cumulative percentage distribution of LD X^2 in the unidimensional 3PL model overlapped with that of the DINA and DINO models, thereby indicating that all three models did an equally sufficient job of accounting for the local dependence between these items. In item pairs 5 & 1, 6 & 3, 7 & 1, and 7 & 3, the unidimensional model actually surpassed the diagnostic classification models in its capacity to model the local dependence among these particular item pairs. One possible culprit is the presence of the latent attribute variables in the DINA and DINO models. Because these latent factors are discrete, there was a loss of information that would not have occurred had the items been modeled with a continuous latent variable. Thus, in some cases, the higher-order factor that was employed to model the attribute space in the classification models did not perform as well as the single latent dimension that characterized the unidimensional 3PL model. This finding suggests that the multidimensionality that typifies the DINA and DINO models is not particularly well-suited for modeling local dependence between items.

Local dependence violations between certain items are often trivial enough to ignore. Non-ignorable local dependence can be identified by evaluating the absolute magnitude of each of the LD X^2 statistics against some critical value; Houts and Cai (2013) suggest 3.0 as an appropriate criterion. Thus, the column plots in Figure 5 depict for four example item pairs the number and percentage of absolute LD X^2 values less than or equal to 3.0. Plot (a) shows the LD between Items 2 and 1. Here, all four multidimensional models were effective in reducing the LD violations to acceptable LD X^2 levels in approximately 600-680 of the 1,000 data sets, while the unidimensional 3PL model performed expectedly worse. In plot (b), the bifactor and EIFA models were just as well-equipped to diminish the LD between Items 4 and 3 as they were in plot (a). The two diagnostic classification models, however, were only able to produce absolute LD X^2 statistics below 3.0 in approximately 42% of the data sets. This result may seem a bit unexpected—in the DINA and DINO models, Items 4 and 3 were both explained by Attribute 2, so one would anticipate a greater reduction in local dependence. However, Item 3 was also associated with Attribute 1; this cross-loading (or “interaction effect”) seems to have adversely affected the ability to curb the dependence between these items.

Plot (c) of Figure 5 differs from the others in two key ways. First, this plot shows that for item pair 6 and 2, the EIFA structure was more effective than the bifactor model at yielding acceptable LD X^2 statistics. Further, the unidimensional 3PL model outperformed both of the (multidimensional) diagnostic classification models. Overall, the local dependence between Items 6 and 2 was among the most difficult to model; the directional paths of the bifactor and diagnostic structures (as shown in Figure 1) were not arranged in a manner conducive to modeling the residual dependence between these particular items. Despite this fact, however, the bifactor model still outperformed the DINA and DINO models.

The final plot in Figure 5 illustrates the extent of local dependence between Items 7 and 6. The bifactor model, which included a specific factor that was explicitly intended to explain the residual noise generated by this exact item pair, was unsurprisingly adept at addressing this dependence. Over 800 of the 1,000 data sets exhibited absolute LD X^2 values less than or equal to 3.0 when fit with the bifactor model. Of additional interest in plot (d) is the fact that the two diagnostic classification models were almost as successful as the EIFA model with regard to reducing the LD X^2 index to a reasonable value. This is perhaps related to the structure of the diagnostic models, wherein Items 6 and 7 both load on one (and only one) attribute factor.

In general, Figure 5 underscores the high FP of the EIFA model and the low FP of the unidimensional 3PL model. Notice that in all four example item pairs, the EIFA model was able to reduce the LD X^2 values to tolerable levels in approximately 2/3rds of the data sets, while the unidimensional 3PL model consistently addressed the LD in approximately 1/3rd of all data sets. There was some degree of variability in the LD management of the bifactor model, though this structure typically addressed the violations quite effectively. The DINA and DINO models were the most inconsistent, sometimes capturing LD noise nearly as well as the EIFA and bifactor models, yet occasionally functioning even less effectively than the unidimensional model. This outcome occurred because the latent variables are discrete in diagnostic classification models and correspondingly carry less information than the continuous latent variables in standard IRT models (Rupp & Templin, 2008a).

Discussion

It is known that one model may fit the observed data better than another because it has a more flexible functional form or a greater number of estimated parameters (e.g., Collyer, 1985;

Cutting, Bruno, Brady, & Moore, 1992). The present study investigated five IRT models that differed in functional form: an exploratory item factor analytic model, a bifactor model, a deterministic input noisy and-gate model, a deterministic input noisy or-gate model, and a unidimensional model. All multidimensional models were specified to include exactly 20 freely estimated parameters per model and the unidimensional model included 21 free parameters. Thus, the unidimensional model, while simpler in factor structure, was more complex in terms of the number of parameters. All five models were fit to 1,000 data sets that were randomly and uniformly sampled from the complete data space. The models were then compared with respect to two statistics intended for categorical data analysis; the cumulative results of these statistics across all data sets functioned as indicators of each model's inherent propensity to fit any possible data.

Confirmation of hypotheses

Our first hypothesis posited that the EIFA model would exhibit, on average, the highest FP. This prediction was strongly supported by the results. The analyses confirmed that among the candidate structures, the EIFA model had the most pliable functional form. Specifically, the Y2/*N* and LD X^2 results demonstrated that the EIFA model outperformed its competitors in terms of overall model fit and minimization of LD violations. This outcome is unsurprising; the exploratory nature of the EIFA model means that it is exceedingly adaptable to a wide array of data patterns (thereby serving as the realization of "complexity" as defined by Myung, Pitt, and Kim (2005) in our introductory paragraph). This model was included in the study, not to shed new light on the flexibility of an exploratory model, but to serve as a baseline measure of FP.

The second hypothesis predicted that the bifactor model would display higher FP than the two diagnostic classification models. This hypothesis was also confirmed by the results: the bifactor model, relative to the DINA and DINO models, demonstrated a propensity to fit a greater number of random data sets that were uniformly distributed across the entire data space. In fact, as evidenced by the cumulative Y2/*N* and LD X^2 metrics, the bifactor model, when fit to random data, was almost as accommodating as the EIFA model. Moreover, the amoeba plots (Figures 3 and 4) uncovered a small number of data sets that actually fit the bifactor model *better* than the EIFA model. These findings help to explain the growing popularity of the bifactor model—in model comparison studies that rely solely on goodness-of-fit to the observed data, the highly malleable bifactor model will almost always be chosen as the "best" model. The researcher who employs this model runs the risk of overfitting the data.

Our findings also cast some light on the FP of two popular diagnostic classification models. The DINA and DINO models did not display excessive flexibility, which suggests that these models are not as likely as the bifactor and EIFA models to overfit data. In other words, a strong goodness-of-fit indicator deserves more attention when produced by a DINA or DINO model than when produced by a bifactor or EIFA model. Indeed, the aim of diagnostic modeling is classification rather than goodness-of-fit. Rupp and Templin (2008a) noted that when statistical precision is valued over classification, "traditional multidimensional FA or IRT models might be much more appealing, unless, of course, the classifications that result from a [diagnostic classification model] analysis are the aspect of the analysis that is desired most" (p. 231). Further, despite the similarity of the DINA and DINO models, they did not perfectly overlap in their coverage of the complete data space. This indicates that the theoretical difference between these models (i.e., whether item attributes are compensatory or not) causes them to fit well to different data patterns. The decision to employ a DINA rather than DINO model should therefore be based on the theory underlying the test items rather than guided by some atheoretical fit contest.

The importance of functional form

The hypotheses discussed above addressed the superior FP of two particular IRT models. Both the EIFA and bifactor models are characterized by relatively complex functional forms. In each case, the items are modeled using numerous cross-loadings on multiple latent dimensions. It is no wonder that these multifaceted models were able to closely represent a substantial proportion of the random data sets. Far less foreseeable were the outcomes returned by the model with the simplest functional form.

The unidimensional 3PL model consistently demonstrated the weakest FP. The overall model fit results from the Y2/*N* analysis verified that the unidimensional model struggled to recover the univariate and bivariate marginals of the "observed" random data. The LD X^2 results revealed, unsurprisingly, that the unidimensional structure was ineffective with regard to modeling local dependence. Each of the key results indicated weak FP, *despite the fact that the unidimensional 3PL model included an additional free parameter!*

This finding challenges current notions of IRT model complexity. It suggests that model complexity should not be assessed simply by tallying free parameters; discussions of IRT model complexity should also concentrate on the arrangement of the latent variables and structural paths in the model. Measurement researchers should be cautious when using models that are not parsimonious in form (the number of parameters notwithstanding). Models that incorporate multiple latent dimensions, residual factors, cross-loadings, or similar intricacies may have an innate tendency to fit well to any conceivable data, even if such models involve relatively fewer freely estimated parameters. This flexibility makes it impossible to determine whether a model fits well because it truly represents the important trend that exists in the data, or because the model itself has a tendency to accommodate an excessively wide range of data patterns. As Wexler (1978) noted, such flexibility can make a theory (or model) "so weak that there is no way to find evidence either for or against it" (p. 346).

Many before us have also argued that analytic practice would benefit from a better understanding of the limitations of fit indices. Roberts and Pashler (2000) discussed a number of theoretical and practical problems that arise from an overreliance on goodness-of-fit statistics, arguing that "models should not be judged only by how well they fit a data set; there also must be assessment of, and penalty for, flexibility" (p. 362). Our investigation focused on assessing such flexibility by considering functional form, an important aspect of complexity that has not previously been researched in the context of IRT modeling. We found that the functional forms of certain widely used IRT models are associated with a problematic ability to fit well to many diverse patterns of data. Ultimately, IRT models must be falsifiable if they are to be considered as useful representations of theories. Although our results do not establish the absolute falsifiability (or lack thereof) of the various models, they do suggest that it may be difficult to find data patterns that will not be fit well by certain complex IRT models. For that reason, IRT practitioners should de-emphasize good fit if it is obtained using a model with an inherent tendency to fit well.

Limitations

This study was limited primarily by computational issues. The first limitation relates to our representation of the complete data space. We chose to represent this space by sampling from the multinomial simplex; it is certainly possible that a more precise method of generating the complete IRT data space could be developed, and we hope that the psychometric community will make progress in this regard. Due to computational burden, we opted to generate only 1,000 data sets that were randomly sampled from and uniformly distributed across the complete data space. If we had generated more data sets, say, 10,000, then the proxy data space would be even more representative of the actual entirety of the data space. However, this limitation was not debilitating; even with 1,000 random data sets, we were able to demonstrate clear discrepancies between the intrinsic data-fitting capabilities of each model.

The estimation specifications of this study were also limiting. The E-step tolerance of the EM algorithm was relaxed in order to speed up the estimation process. Despite this modification, the random data-fitting procedure was still rather time-consuming. For example, fitting the unidimensional 3PL model to all 1,000 data sets took approximately 30 hours when using a 2.90GHz quad-core processor with 16 GB RAM. If the tolerance between E-step iterations had been left at the default, then estimation would have taken considerably longer.

Furthermore, the estimation process was unable to converge on stable parameter estimates in a sizeable number of data sets (Table 2), despite the considerable increase in EM iterations (20,000 cycles). Perhaps with an even greater number of estimation cycles, a different estimator, additional computing time, or other alterations to the estimation process, the convergence rates would improve. However, the models were fitting random, nonsensical data with no underlying form; in the many cases where data were more noise than signal, one would not expect successful convergence. Thus, while convergence rates may not have been ideal, it is highly unlikely that 100% convergence across all models and data sets would ever be achieved.

Another potential limitation was the analytic approach itself. We followed the same investigative strategy as Preacher (2006), namely, fitting candidate models to a large number of random data sets. While this tactic produced several compelling findings, alternative formulations of the MDL principle may offer deeper insights into these (and other) IRT models, and by circumventing the tedium of fitting 1,000 data sets, they would likely present these insights with far greater efficiency. A particularly promising numerical expression of MDL expression is discussed below.

Future research directions

This line of inquiry opens up a number of topics for future research. First, the hypotheses in this study drew attention to the overly flexible nature of the particular EIFA and bifactor models that were included in the analysis. In the exact EIFA model that was analyzed, the path from Factor 2 to Item 1 was constrained to zero for model identification. The choice to fix this specific path was completely arbitrary. Since this model was to be fit to random data, our reasoning was that one EIFA structure would be as useful as any other. Yet, could it be that the FP exhibited by this EIFA model was elevated (or diminished) by the chosen arrangement of the variables? How might the outcome compare if, for instance, a path had been fixed from Factor 1 instead of from Factor 2? This specification may not alter the FP, but perhaps the EIFA model would occupy a different region of the complete data space. The same type of question arises when considering the bifactor results. Would the findings have shifted if other sets of items had been selected to load on the specific factors? In the future, it would be prudent to compare all combinations of factor loadings in these models. Such an all-encompassing analysis would permit one to make claims about the EIFA and bifactor models on the whole, rather than simply reporting results that are contingent on particular instantiations of these models.

Another direction of future research relates to the Y2/*N* amoeba plots. The various FP regions depicted in these figures exposed several interesting nuances. For example, what sort of data patterns characterize the few data sets that fit the bifactor model better than the EIFA model? Further, the DINA and DINO models fit approximately the same number of data sets, but these two models did not occupy identical regions of the data space. Is it possible to isolate the type of response pattern that tends to fit better to the DINA model than to the DINO model, or vice versa? The Y2/*N* results, especially in the information-theoretic context of "occupying the complete data space," offer ample fodder for future research.

Another important future direction involves the study of various numerical expressions of MDL, especially in the context of IRT. Rather than fitting models to copious random data sets, one could compute a metric such as stochastic information complexity (SIC; Hansen & Yu, 2001; Markon & Kreuger, 2004; Rissanen, 1989):

$$\text{SIC} = -\ln\ f\big(\text{D} \mid \theta^{*}(\text{D})\big) + \frac{1}{2}\ln \left| N \bullet I(\hat{\theta}) \right| . \tag{2}$$

The first term in this equation accounts for goodness-of-fit, where $f(\cdot)$ is the maximum likelihood function of the observed data D. The second term accounts for structural complexity, where $I(\hat{\theta})$ is the determinant of the covariance matrix that results when the Fisher information matrix is used to estimate standard errors. These elements can be obtained using the output from standard IRT software, making SIC especially well-suited for future IRT analyses.

Conclusion

Overall, this report presents a novel outlook on the complexity of IRT models. Information-theoretic analyses demonstrated that the bifactor model has an undesirable tendency to fit any possible data, and that an IRT model with more free parameters but a simpler structure may occupy a much narrower region of the complete data space. These findings establish the MDL principle as a promising methodological tool for understanding the inherent properties of all types of latent variable models. While the present study invoked the MDL principle to expose the vices and virtues of several popular IRT models, we believe that this approach opens up a plethora of new areas of philosophical, theoretical, and practical research in all types of latent variable modeling.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19*(6), 716-723. doi:10.1007/978-1-4612-1694-0_16

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology, 55*:1-15. doi:10.1348/000711002159617

Bentler, P. M., & Weeks, D. G. (1979). Interrelations among models for the analysis of moment structures. *Multivariate Behavioral Research*, *14*(2), 169-186. doi:10.1207/s15327906mbr1402_3

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. doi:10.1007/bf02293801

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261-280. doi:10.1177/014662168801200305

Boulton, D. M., & Wallace, C. S. (1970). A program for numerical classification. *The Computer Journal, 13*(1), 63-69. doi:10.1093/comjnl/13.1.63

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General),* 383-430. doi:10.2307/2982063

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33-57. doi:10.1007/s11336-009-9136-x

Cai, L. (2013). flexMIRT® version 2.00: A numerical engine for flexible multilevel multidimensional item analysis and test scoring [computer software]. Raleigh-Durham, NC: Vector Psychometric Group.

Cai, L., Maydeu-Olivares, A., Coffman, D., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology, 59*, 173-194. doi:10.1348/000711005x66419

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3):265-289. doi:10.2307/1165285

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*(510), 850-866. doi: 10.1080/01621459.2014.934827

Collyer, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception & Psychophysics, 38*(5), 476-481. doi: 10.3758/bf03207179

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General, 121*(3), 364-381. doi: 10.1037//0096-3445.121.3.364

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353. doi:10.1007/bf02295640

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer Science & Business Media.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*(4), 733-760. Retrieved from http://www.jstor.org/stable/24306036

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68*(3), 532-560. doi:10.1207/s15327752jpa6803_5

Grünwald, P. D. (2005). A tutorial introduction to the minimum description length principle. In P. D. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications* (pp. 23-81). Cambridge, MA: MIT Press. arXiv:math/0406077

Grünwald, P. D. (2007). *The Minimum Description Length Principle.* Cambridge, MA: MIT Press.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological),* 83-100. Retrieved from http://www.jstor.org/stable/2984569

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 301-323. doi:10.1111/j.1745-3984.1989.tb00336.x

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association, 96*(454), 746-774.

doi:10.1198/016214501753168398

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*(4), 577-601. doi:10.1007/bf02294609

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 3,* 45-60. doi:10.1007/bf02287965

Houts, C. R., & Cai, L. (2013). *flexMIRT® user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring.* Chapel Hill, NC: Vector Psychometric Group.

Huffman, D. A. (1956). The synthesis of linear sequential coding networks. In C. Cherry (Ed.), *Information Theory* (pp. 77-95). Englewood Cliffs, NJ: Academic Press.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 User's Reference Guide*. Scientific Software International.

Junker, B. W., & Sijstma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258-272. doi:10.1177/01466210122032064

Kolmogorov, A. N. (1963). On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, 369-376. Retrieved from http://www.jstor.org/stable/25049284

Lee, M. D., & Pope, K. J. (2006). Model selection for the rate problem: A comparison of significance testing, Bayesian, and minimum description length statistical inference. *Journal of Mathematical Psychology, 50*(2), 193-202. doi:10.1016/j.jmp.2005.11.010

Li, Z., & Cai, L. (2012, July). *Summed score likelihood based indices for testing latent variable distribution fit in item response theory.* Paper presented at the annual International Meeting of the Psychometric Society, Lincoln, NE. Retrieved from http://www.cse.ucla.edu/downloads/files/SD2-final-4.pdf

Longley, S. L., Calamari, J. E., Wu, K., & Wade, M. (2010). Anxiety as a context for understanding associations between hypochondriasis, obsessive-compulsive, and panic attack symptoms. *Behavior Therapy, 41*, 461-474. doi:10.1016/j.beth.2010.01.002

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.

Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics, 34*, 593-610. doi:10.1007/s10519-004-5587-0

Moody, J. E. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In Moody, J. E., Hanson, S. J., & Lippmann, R. P., (Eds.), *Advances in Neural Information Processing Systems*, 847-854.

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. In Lamberts, K. & Goldstone, R., (Eds.), *Handbook of Cognition*. London, UK: Sage Publications Ltd. doi:10.4135/9781848608177.n19

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64. doi:10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S\text{-}X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289-298. doi:10.1177/0146621603027004004

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*(3). 472-491.

doi:10.1037//0033-295x.109.3.472

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41*(3), 227-259. doi:10.1207/s15327906mbr4103_1

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667-696. doi:10.1080/00273171.2012.715555

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465-471. doi:10.1016/0005-1098(78)90005-5

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics,* 416-431. doi:10.1214/aos/1176346150

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

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

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2015). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98(3),* 1-15. doi:10.1080/00223891.2015.1089249

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*(4), 1151-1172. doi:10.1214/aos/1176346785

Rupp, A. A., & Templin, J. L. (2008a). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6,* 219-262. doi:10.1080/15366360802490866

Rupp, A. A., & Templin, J. L. (2008b). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96. doi:10.1177/0013164407301545

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2). 461-464. doi:10.1214/aos/1176344136

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal. 27*(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Chicago: University of Illinois Press.

Smith, N. A., & Tromble, R. W. (2004). Sampling uniformly from the unit simplex. *Johns Hopkins University, Tech. Rep. 29.* Retrieved from http://www.cs.cmu.edu/~nasmith/papers/smith+tromble.tr04.pdf

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control*, 7(1), 1-22. doi:10.1016/s0019-9958(64)90223-2

Spiegelhalter, D. J., Best, N., Carlin, B. P., & van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. *Research Report, 98-009*. Retrieved from http://www.sph.umn.edu/faculty1/wp-content/uploads/2012/11/rr98-009.pdf

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11,* 287-305. doi:10.1037/1082-989x.11.3.287

Teugels, J. L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *Journal of Multivariate Analysis, 32*(2), 256-268. doi:10.1016/0047-

259x(90)90084-u

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: demonstration of a bifactor model for the Brief Symptom Inventory. *Psychological Assessment*, *24*(1), 101-113. doi:10.1037/a0024712

Tian, W., Cai, L., Thissen, D., & Xin, T. (2013). Numerical differentiation methods for computing error covariance matrices in item response theory modeling: An evaluation and a new proposal. *Educational and Psychological Measurement*, *73*(3), 412-439. doi:10.1177/0013164412465875

Vitányi, P. M., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, *46*(2), 446-464. doi:10.1109/18.825807

Wexler, K. (1978). A review of John R. Anderson's *Language, Memory, and Thought. Cognition*, *6*, 327-351. doi:10.1016/0010-0277(78)90003-3