Article

Educational and Psychological Measurement
2018, Vol. 78(3) 357–383

DOI: 10.1177/0013164416685599
journals.sagepub.com/home/epm
SAGE

Retrofitting Diagnostic Classification Models to Responses From IRT-Based Assessment Forms

Ren Liu[1], Anne Corinne Huggins-Manley[1], and Okan Bulut[2]

Abstract

Developing a diagnostic tool within the diagnostic measurement framework is the optimal approach to obtain multidimensional and classification-based feedback on examinees. However, end users may seek to obtain diagnostic feedback from existing item responses to assessments that have been designed under either the classical test theory or item response theory frameworks. Retrofitting diagnostic classification models to existing assessments designed under other psychometric frameworks could be a plausible approach to obtain more actionable scores or understand more about the constructs themselves. This study (a) discusses the possibility and problems of retrofitting, (b) proposes a step-by-step retrofitting framework, and (c) explores the information one can gain from retrofitting through an empirical application example. While retrofitting may not always be an ideal approach to diagnostic measurement, this article aims to invite discussions through presenting the possibility, challenges, process, and product of retrofitting.

Keywords

diagnostic measurement, diagnostic classification modeling, item response theory, retrofit, dimensionality, test development

[1]University of Florida, Gainesville, FL, USA
[2]University of Alberta, Edmonton, Alberta, Canada

Corresponding Author:
Ren Liu, School of Human Development and Organizational Studies in Education, Gainesville, FL 32611, USA.
Email: liurenking@ufl.edu

The current landscape in educational measurement appears to be shifting away from creating and administering assessments that result in a single score for a large content domain toward assessments that can provide mastery decisions for multiple smaller domains. Researchers have urged the measurement field to maximize educational benefits from assessments (de la Torre & Minchen, 2014; Nichols, 1994; Nichols, Chipman, & Brennan, 2012), and practitioners have expressed their interest in using assessment results to better inform interventions (Bennett, 1999; Huff & Goodman, 2007; National Research Council, 2010).

In response to those demands, diagnostic measurement has become increasingly popular. Diagnostic measurement is an effective framework that provides multidimensional, classification-based feedback about examinees (Rupp, Templin, & Henson, 2010). Developing a measurement tool within the diagnostic measurement framework is no doubt the exemplary approach to obtain such information. However, there are contexts where examinees have already provided item responses to assessments that are designed under the classical test theory (CTT) or item response theory (IRT) frameworks, but end users are interested in obtaining diagnostic feedback from existing item responses. Retrofitting diagnostic measurement models has been used as a plausible approach to obtain such information (e.g., Jang, Dunlop, Wagner, Kim, & Gu, 2013; Kim, 2014; Li, Hunter, & Lei, 2016). These models are called diagnostic classification models (DCMs), which are a class of probabilistic, confirmatory, multidimensional latent class models (Rupp et al., 2010).

While retrofitting DCMs to existing assessments has the potential to produce greater information regarding examinee's latent traits, this process goes directly counter to the inferences attempted to be made from the original assessment. Under unidimensional IRT, for example, retrofitting requires practitioners to extract multidimensionality from assessments developed to remove any multidimensionality. As shown consistently in Gierl and Cui (2008), Haberman (2008), Sinharay (2010), Sinharay and Haberman (2008), and Sinharay (2014), test construction practices for unidimensional assessments almost always explicitly seek to maximize reliability of the total score. This means that items with high discriminations (e.g., point-biserial correlations or a-parameters) are often selected in the assessment, while those with the possibility of inducing multidimensionality are left out. Thus, retrofitting multidimensional DCMs can introduce a conundrum with respect to dimensionality that may not be easily resolved. However, retrofitting provides a way to attempt to reap the benefits of DCM in the current landscape in which not many tests have been designed to assess multidimensional skills, and it will be a number of years before that situation changes given the time intensive nature of developing such assessments. Therefore, it is possible that retrofitting may be a primary source of DCM applications for the near future until the test construction processes for multidimensional assessments become more ingrained in practice. We support the notion that retrofitting should not be encouraged as a standard approach to a measurement endeavor and not all assessments are suitable for retrofitting. But it also may be justifiable to recognize the struggle between the urging needs of diagnostic information and limited resources to develop and administer new diagnostic assessments.

The purpose of this study is to (a) discuss the possibility and challenges of retrofitting, (b) propose a step-by-step retrofitting framework, and (c) explore the information one can gain from retrofitting through an application example. We aim to present this study from a comprehensive perspective on retrofitting, inform end users of the challenges and opportunities in retrofitting, and leave the decision of whether to retrofit under specific assessment contexts open to discussion.

Possibility and Challenges of Retrofitting

In this study, we define ''retrofitting'' as the practice of fitting DCMs to responses obtained from assessments that are not designed under the diagnostic measurement framework. The primary difference between diagnostic and nondiagnostic frameworks is that the former specifies multiple categorical traits and classifies examinees on each trait (e.g., possessing or not possessing a cognitive latent trait; mastery or non-mastery in a content domain), while the latter specifies one or multiple continuous trait(s) and assigns scores to examinees on the trait continuum or continua (e.g., estimating a score for each examinee that falls on a standardized scale similar to *z* scores). A commonly cited example of assigning scores on elementary-level math ability (e.g., Kunina-Habenicht, Rupp, & Wilhelm, 2009) can be used for a brief conceptual description of these two frameworks. In nondiagnostic frameworks, math ability could be represented through a unidimensional continuous latent trait (i.e., math ability) or multiple continuous traits (e.g., addition, subtraction, and multiplication skills) in the assessment, and scores are provided to order examinees on the latent trait continua. In the diagnostic measurement framework, math ability must be represented through multiple categorical traits (e.g., addition, subtraction, and multiplication skills) and examinees are classified into two or more categories (e.g., has mastered or has not mastered) on each latent trait based the estimated probability of mastery of each trait.

It is critical that we treat different psychometric frameworks as mere lens through which we come to understand examinees. In other words, the product of these frameworks per se should not be considered legitimate and encompassing evaluations of examinees. Thissen's (2016) demonstrates the inappropriateness of asking definitive questions about IRT model fit and about data dimensionality because there is no model that perfectly fits a data set nor any data set that reflects pure unidimensionality. A model is considered ''fit'' or an assessment is called ''unidimensional'' when one sets a cut-off value and compares statistics against that value (Ho, 2016), but the truth (i.e., examinee characteristics in this case) per se is much more complex than what an assessment framework or a statistical model could cover. The interaction between items and examinees involves many factors such as cultural diversity, learning styles, and many individual differences that are in human nature. Instead of offering a definitive answer on whether an assessment is unidimensional or multidimensional or whether it should provide scores or classifications, the intended purpose of the assessment should rather inform psychometric decisions (Thissen, 2016).

With that in mind, one might argue that retrofitting might be plausible for two reasons. First, fitting DCMs to existing responses could fulfill the purpose of obtaining skill-level information about examinees, which is not directly available under nondiagnostic measurement frameworks. Second, just because an assessment has been developed under a unidimensional framework does not mean that during test development skill-level considerations were ignored. In fact, content development for any assessment often involves the theoretical breakdown of a larger construct into subdomains. Those subdomains could be considered as multiple dimensions when retrofitting DCMs.

In sum, the possibility of retrofitting is inherent due to the complexity of human characteristics and the resulting fact that a unidimensional trait can always be broken down theoretically into smaller attributes.

The general concept of retrofitting has been described as ‘‘the addition of a new technology or feature to an older system’’ (Gierl, Roberts, Alves, & Gotzmann, 2009). Although no study has focused on the methodology of retrofitting DCMs or evaluated what information could be gained from this highly debatable practice, retrofitting has been used widely as an add-on to simulation studies addressing different research questions in diagnostic measurement (e.g., Chen & de la Torre, 2013; Chen, de la Torre, & Zhang, 2013; Cui, Gierl, & Chang, 2012; de la Torre, 2009; de la Torre & Douglas, 2004; Henson, Templin, & Willse, 2009; Hou, de la Torre, & Nandakumar, 2014; Ravand, 2016; Templin & Bradshaw, 2013; von Davier, 2005, 2007, 2014). Practitioners have also applied DCMs to existing nondiagnostic assessments to provide diagnostic feedback in specific content areas (e.g., Jang, 2009; Jang et al., 2013; Kim, 2014; Lee & Sawaki, 2009; Li & Suen, 2013; Li et al., 2015; Ravand, Barati, & Widhiarso, 2012). Either in methodological or application studies, retrofitting has fulfilled its purposes in a variety of contexts.

Despite its benefits, some researchers argue that retrofitting is unlikely to produce satisfactory results from the perspective of either assessment design (e.g., DiBello & Stout, 2007; Gierl, 2007; Gierl & Cui, 2008) or statistical quality (Buck & Tatsuoka, 1998; Svetina, Gorin, & Tatsuoka, 2011; VanderVeen et al., 2007).

Three major problems have been highlighted in terms of assessment design. First, the power of diagnostic measurement cannot be realized without a principled assessment design (Nichols, Kobrin, Lai, & Koepfler, 2016). The cognitive theories that should underlie the diagnostic assessment design are missing; thus, it is often difficult to support the attribute specifications in DCMs either theoretically or empirically. Second, it is more problematic to retrofit DCMs to unidimensional assessments than to multidimensional ones because of the conflicting goals between those two dimensionality assumptions (de la Torre & Karelitz, 2009; Gierl & Cui, 2008). When items are developed for a unidimensional assessment, content, and measurement experts often work together to revise or remove items that introduce multidimensionality to obtain a more reliable total score. As a result, multidimensional diagnostic information is difficult to obtain. Sinharay and Haberman (2008) and Sinharay (2014) have highlighted the issue of dimensionality in their work on extracting subscores from unidimensional tests. They found that assessments overwhelmingly fail to produce valuable subscores in classical test theory context due to low reliabilities of, and high correlations among, subscores (Sinharay, 2010). Third, one may argue that some attributes have not been measured sufficiently for producing reliable results (Templin & Bradshaw, 2013), or that some attributes are always measured together resulting in confounding (Tatsuoka, 1995). Unfortunately, these three issues cannot be easily resolved because the items have already been developed.

In terms of statistical quality, at least three problems have been identified, including (a) attributes identified from the single trait are often highly correlated (VanderVeen et al., 2007), (b) correlations between an attribute and total score may be low or negative (Buck & Tatsuoka, 1998; Svetina et al., 2011), and (c) model fit indices may show poor results (Gierl & Cui, 2008). Fortunately, these three problems can be investigated empirically to assess the degree to which they occurred, allowing the retrofitting team to make informed decisions about their particular application.

None of these problems are trivial when one tries to obtain accurate and reliable information from retrofitting. As stated before, there is no doubt that developing an assessment under the diagnostic framework is a better way to obtain diagnostic information than retrofitting. One needs to exercise caution when retrofitting, which should only be considered when it is not possible to develop a diagnostic assessment. Also, retrofitting may be best used for low-stakes decisions or purely for the purpose of learning more about the construct. We continue this article by proposing a methodological approach that can minimize the problems and maximize the benefits of retrofitting within such contexts.

Boosting the Suboptimal Approach: A Framework for Retrofitting

Similar to developing a diagnostic assessment, retrofitting is an iterative process that encompasses core procedures in operational testing practices. A unique feature of retrofitting is that the items have already been developed and that responses have been collected from examinees. The iteration in retrofitting is thus about revising the specification of attributes and the relationships between attributes and items, instead of revising the items themselves. Based on a review of published retrofitting examples cited above and on experience in retrofitting, we propose that retrofitting consists of four stages as shown in Figure 1: gathering information, specifying attributes and attribute–item relationships, modeling item responses, and interpreting results. The following sections present this process and discuss decisions one should make when retrofitting is used.

Gathering Information

Retrofitting begins with gathering information about the assessment, end users, and item responses. It not only involves necessary groundwork for specifying attributes and attribute–item relationships, but also prepares information that is critical for any inferences one intends to make after retrofitting.

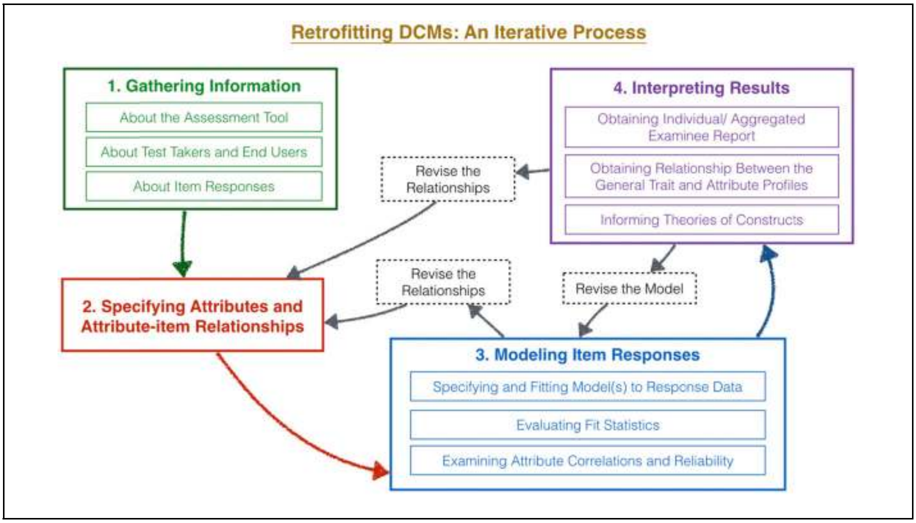
Retrofitting DCMs: An Iterative Process
1. Gathering Information
About the Assessment Tool
About Test Takers and End Users
About Item Responses
2. Specifying Attributes and Attribute-item Relationships
3. Modeling Item Responses
Specifying and Fitting Model(s) to Response Data
Evaluating Fit Statistics
Examining Attribute Correlations and Reliability
4. Interpreting Results
Obtaining Individual/ Aggregated Examinee Report
Obtaining Relationship Between the General Trait and Attribute Profiles
Informing Theories of Constructs
Revise the Relationships
Revise the Relationships
Revise the Model

Figure 1. The iterative process of retrofitting DCMs.

At a minimum, three aspects of information must be gathered about the assessment: design framework, dimensionality, and parameter estimates. First, the design framework of the assessment is helpful for separating the subskills that are underlying the overall latent trait. For example, if the assessment has been developed under the evidence-centered design framework (Mislevy & Haertel, 2006), it would be helpful to obtain the student, evidence, task, and assembly models to understand the underlying structure and motivations behind content, item, and other assessment development choices. The attributes and/or behaviors identified in the assessment conceptual framework could help practitioners explore possible underlying attributes that need to be specified. Attributes in science domains (e.g., mathematics) may be more distinguishable at the item level than non-science domains (e.g., reading comprehension; Rupp et al., 2010), possibly making assessments in the former domains easier to retrofit. Second, one must also explore the intended dimensionality of the assessment. If the assessment is intended to be multidimensional, it may not be appropriate to simply borrow the original item-to-dimension specifications into the DCMs. A review of such decisions is probably still warranted. If the assessment is intended to be unidimensional, retrofitting may be more difficult. Third, one must obtain the item parameters to understand the item characteristics. For example, if the assessment has been developed under the unidimensional framework and previously analyzed using the two-parameter IRT model, one could examine the discrimination and difficulty parameters of each item (e.g., parameter estimates obtained from a technical report on that particular assessment). An understanding of item statistics is necessary before connecting items to attributes because item statistics influence the retrofitting process in a variety of ways. For example, if item difficulty values are widely spread across the latent continuum, this can impact the type of DCM one might want to use for retrofitting. In many DCMs, item difficulty parameter estimates are not allowed to vary across latent classes and hence misclassifications may be likely to occur. In fact, most DCMs only model between-class differences and each class has the same item difficulty parameters (Choi, 2010). We further discuss options for this type of item difficulty spread in the Discussion section.

In addition to understanding the assessment, one needs to gather information about test takers and other end users of the assessment. Ultimately, if one is retrofitting then it is implied that the results will be used in some way different from the original intended use of the assessment. Hence, as test use is critical to evaluating the validity of any measurement analysis output (Kane, 2013), it is important to go back and redefine what the new assessment output and scores will be used for.

Finally, one should explore the item responses collected that are going to be retrofitted. This includes examining descriptive statistics of the responses on each item as well as score distributions. If scores were ever reported in performance categories (e.g., pass/fail decisions in licensure tests), information should be gathered about the cut-scores used to create those categories so that they can be applied to the current data set and the frequencies of examinees in each category can be explored. Comparing these unidimensional classifications based on cut-scores to the multidimensional classifications obtained from retrofitting can help evaluate whether retrofitting provides additional information about examinees. One should also explore the dimensionality of the data set, which may or may not be the same as the intended dimensionality when developing the assessment. For an assessment on which a single score is reported, one could directly assess dimensionality through software programs such as the DIM-pack (DIMTEST, DETECT, and HAC; Stout, Froelich, & Gao, 2001). One could also use the Haberman (2008) approach to compare the reliability or assess the reduction in mean squared error of unidimensional and multidimensional scores to help decide whether subscores from each dimension provide meaningful and reliable information beyond total scores. If the data set is approximately unidimensional, retrofitting may not be appropriate. If there is more than one dimension, the number of dimensions may inform the number of attributes one wants to specify in the next stage.

Specifying Attributes and Attribute–Item Relationships

When developing a diagnostic tool, attributes are specified before developing items. Retrofitting is different in a way that it specifies attributes from developed items. Based on the research that is currently published in DCM, attributes are usually specified as ordinal variables, which could be either dichotomous (e.g., has mastered or has not mastered) or polytomous (e.g., has fully mastered, has partially mastered, and has not mastered). We limit this discussion to dichotomous attributes for simplicity, although polytomous attributes can also be specified (e.g., Chen & de la Torre, 2013; von Davier, 2005). In the dichotomous case, if one specifies A attributes, examinees are classified into one of the 2^A attribute profiles, also known as latent classes (Rupp et al., 2010). Each profile is a unique mastery pattern, denoted as an A-length vector $\alpha_c = \{\alpha_1, \alpha_2, \ldots, \alpha_A\}$. Each examinee is assigned either a value of 1 for mastery or 0 for non-mastery for each attribute. For example, examinees with attribute profile $\{1, 0, 1\}$ have mastered the first and third attributes (α_1 and α_3), but not the second attribute (α_2).

Specifying attributes from existing items involves at least three important considerations. First, an item could measure one or more attribute(s), although attributes that are measured in isolation are likely to have higher classification accuracy (Madison & Bradshaw, 2015). If two attributes are measured together all the time, we suggest combining those two attributes into one larger attribute when retrofitting, because items cannot be further revised. Second, one could hypothesize the relationships among attributes. Attributes could be hypothesized as either correlated or sequentially ordered (Templin & Bradshaw, 2014). By default, if attributes are directly entered into the next stage of fitting DCMs, they are considered as correlated. In other contexts, some attributes may depend on each other. This dependency, also known as an attribute hierarchy (Leighton, Gierl, & Hunka, 2004), means that one attribute can only be mastered after the mastery of another attribute (Liu & Huggins-Manley, 2017). One could specify an attribute hierarchy either at the time of specifying the attributes based on theory or after analyzing the model fitting results (Liu, Huggins-Manley, & Bradshaw, 2016). Third, the total number of attributes should be fairly small. A large number of attributes would not be only difficult to estimate within a given test length, but also difficult to interpret (Embretson & Yang, 2013; Tatsuoka et al., 2016; Xu & Zhang, 2016). de la Torre and Minchen (2014) recommended a rule of thumb for the maximum number of attributes as 10. However, 10 attributes may be difficult to handle because examinees will be classified into $2^{10} = 1{,}024$ possible latent classes, if no attribute hierarchy is assumed. Generally, most above-mentioned retrofitting studies specify three to five attributes; thus, examinees are classified into one of as few as $2^3 = 8$ or as many as $2^5 = 32$ attribute profiles, unless an attribute hierarchy is assumed.

After each item is associated with one or more attributes, attribute–item relationships need to be contained in an item-by-attribute incidence matrix known as a Q-matrix (Tatsuoka, 1983). In a Q-matrix, each entry q_{ia} equals to 1 if item i measures attribute a, and 0 otherwise. When designing the Q-matrix, it is necessary to make sure that each attribute is measured enough times, no matter if it is measured in isolation or with other attributes (Madison & Bradshaw, 2015). When retrofitting, attributes that are not measured enough times may be either (a) combined with another similar attribute or (b) retained for the completion of the Q-matrix but not to be interpreted. In addition, it is not necessary that all items need to be specified in the Q-matrix for retrofitting. Instead, one could select specific items that are associated with attributes of interest.

A correct specification of the Q-matrix is essential because the misspecification of the Q-matrix can result in model-data misfit and bias in examinee classifications (Kunina-Habenicht, Rupp, & Wilhelm, 2012; Rupp & Templin, 2008a, 2008b). To ensure that the hypothesized Q-matrix is supported by item responses, one may consider validating the Q-matrix using the methods proposed in de la Torre (2008) or de la Torre and Chiu (2016), which can help identify and correct misspecified entries in the Q-matrix. Ideally, the construction of the final Q-matrix should take both theory and empirical item responses into consideration.

Modeling Item Responses

After attributes and attribute–item relationships are specified, it is time to fit DCMs to the item responses. Since the early 1980s, more than 20 DCMs have been proposed to model various hypotheses about the relationships among attributes or to satisfy different assessment needs. Readers can refer to Ma, Iaconangelo, and de la Torre (2016) for the relationships as well as suggestions for selection, among different models. It is common to first fit more than one DCM to the item responses and then select the final model based on fit statistics (e.g., Lei & Li, 2016). The following sections discuss steps of fitting and evaluating DCMs.

Specifying and Fitting Model(s) to Response Data. Among all DCMs, we briefly present five models that may be useful in retrofitting. Although other models may also be available for retrofitting, we present these five because they represent most common assumptions about the relationship between attributes and items. These models are the generalized DINA (G-DINA; de la Torre, 2011) model, the additive CDM (A-CDM; de la Torre, 2011), the deterministic inputs, noisy ''and'' gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) model, the deterministic input, noisy ''or'' gate (DINO; Templin & Henson, 2006) model, and the higher-order DINA (HO-DINA; de la Torre & Douglas, 2004) model.

The G-DINA model is a general model that subsumes most DCMs including the A-CDM, DINA, and DINO models. Let $j = 1, \ldots, J$ index items and $k = 1, \ldots, K$ index attributes, and K_j^* is the number of required attributes for item j. $\boldsymbol{\alpha}_{lj}^*$ is the reduced attribute vector indexing the attributes required for item j. The G-DINA models the conditional probability that an examinee with attribute profile α_{lj}^* scores a 1 on item j as:

$$g\left[P\left(\boldsymbol{\alpha}_{lj}^*\right)\right] = \lambda_{j0} + \sum_{k=1}^{K_j^*} \lambda_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \lambda_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \lambda_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \quad (1)$$

where $g\left[P\left(\boldsymbol{\alpha}_{lj}^*\right)\right]$ is $P\left(\boldsymbol{\alpha}_{lj}^*\right)$, $\log\left[P\left(\boldsymbol{\alpha}_{lj}^*\right)\right]$ and $\text{logit}\left[P\left(\boldsymbol{\alpha}_{lj}^*\right)\right]$ when the identity, log and logit links are used, respectively. λ_{j0} is the intercept for item j, λ_{jk} is the main effect of attribute k, $\lambda_{jkk'}$ is the two-way interactional effect of attribute k and k', and $\lambda_{j12\ldots K_j^*}$ is the higher-order interaction among $\alpha_1, \ldots, \alpha_{K_j^*}$. Equation 1 shows that the G-DINA model is a saturated model because it includes all possible effects of attributes.

The A-CDM assumes that attributes do not interact in their effects on items. For example, on an assessment measuring both depression and anxiety, removing interaction terms assumes that the attributes of depression and anxiety do not interact in their relationship to item responses. The A-CDM is the G-DINA model without interaction terms. The item response function (IRF) in A-CDM is as follows:

$$g\left[P\left(\boldsymbol{\alpha}_{lj}^{*}\right)\right] = \lambda_{j0} + \sum_{k=1}^{K_j^*} \lambda_{jk}\alpha_{lk}. \tag{2}$$

The DINA model assumes that an item can be correctly answered only if all the required attributes on that item are mastered. It is also known as a noncompensatory model, reflecting the assumption that a deficit in one attribute cannot be compensated for by a surplus in another attribute. It constrains every effect to zero except the intercept and highest order interaction among the attributes. Its IRF can be written as

$$P\left(\boldsymbol{\alpha}_{lj}^{*}\right) = \lambda_{j0} + \lambda_{j12\ldots K_i^*} \prod_{k=1}^{K_j^*} \alpha_{lk}. \tag{3}$$

The DINO model assumes that a deficit in one attribute can be compensated by a surplus in another attribute. Therefore, it is known as a compensatory model. It constrains absolute values of all the main and interaction effect parameters in the G-DINA model to different signs by the order of interactions. Its IRF can be formulated as:

$$P\left(\boldsymbol{\alpha}_{lj}^{*}\right) = \lambda_{j0} + \lambda_{jk}\alpha_{lk}, \tag{4}$$

where $\lambda_{jk} = (-1)\lambda_{jk'k''} = \cdots = (-1)^{K_j^*+1}\lambda_{j12\ldots K_j^*}$. Under the DINO model, the probability of a correct response is the same for an examinee mastering one required attribute on item j and an examinee mastering all required attributes on item j.

The HO-DINA model assumes that the attributes share a continuous higher-order factor, denoted as θ. It uses a two-level model where the lower-level is a DCM and the higher-level is an IRT model. The HO-DINA model defines the probability of an examinee e's mastery status as a function of attribute k, given his/her higher-order latent trait θ as

$$P(\boldsymbol{\alpha}_{ek}|\theta_e) = \frac{exp(\tau_k\theta_e - \beta_k)}{1 + exp(\tau_k\theta_e - \beta_k)}, \tag{5}$$

where τ_k and β_k are the discrimination and difficulty parameters, respectively, of attribute k. The HO-DINA model might be promising for retrofitting because the higher-order factor resembles the overall unidimensional score.

In practice, one could fit a general DCM (e.g., the G-DINA model) to the data if no prior hypothesis is made and sample sizes allow (J. Templin, personal communication, January 15, 2016). This modeling approach would allow for free estimation of all parameters associated with any possible relationships between attributes and item responses. If there are prior hypotheses about the effects of attribute relationship on items, one could fit both the selected model that reflects those hypotheses and a general DCM to the data. A comparison of fit indices between the selected and general models would help determine if those hypotheses are supported in item responses. The next section presents several types of available fit indices that have been used in evaluating DCMs.

Evaluating Fit Indices. Acceptable fit is a prerequisite to studying model results, but is insufficient for describing a useful retrofitting. Three types of fit indices should be examined, including model fit, item fit, and person fit. We provide a brief introduction to some fit statistics below, and refer readers to Lei and Li (2016) and Hu, Miller, Huggins-Manley, and Chen (2016) for a detailed discussion on performance of fit indices in choosing appropriate DCMs.

For model fit statistics presented below, a smaller value indicates better fit. Absolute fit indices are used to determine the adequacy of fit of a single model. Relative fit indices investigate whether two models fit significantly different from each other. At least six indices could be used when evaluating the absolute fit: the limited information fit statistics (Joe & Maydeu-Olivares, 2010), the maximum of all pairwise χ^2 statistics ($maxX^2$, Chen et al., 2013), the mean absolute deviation between observed and expected correlations (MADcor; DiBello, Roussos, & Stout, 2007), the standardized root mean square root of squared residuals (SRMSR; Maydeu-Olivares & Joe, 2014), the mean of absolute deviations of residual covariances (Res; McDonald & Mok, 1995), and mean of absolute values of Q_3 statistic (Yen, 1984). Absolute model fit indices are commonly compared with some rule-of-thumb criteria to determine the level of acceptable fit. For example, SRMSR smaller than .05 indicates a substantively negligible amount of misfit, thus is considered good fit (Maydeu-Olivares, 2013). For relative fit, one could compare the -2 log likelihood (-2LL; Neyman & Pearson, 1992), Akaike Information Criterion (AIC; Akaike, 1987), Bayesian Information Criterion (BIC; Schwarz, 1978), or Consistent AIC (CAIC; Bozdogan, 1987). Likelihood ratio tests can be used to test the significance of the difference between the fit of the saturated model and that of nested models.

To evaluate item fit, the root mean square error of approximation (RMSEA; von Davier, 2005) can be computed. Generally, RMSEA smaller than .05 is considered a good fit (Kunina-Habenicht et al., 2009).

To evaluate person fit, the probability of responding aberrantly (ρ; Liu, Douglas, & Henson, 2009) can be computed. It represents the magnitude of an examinee's response deviating from the estimated attribute profile. For example, a ''spuriously high scorer'' is an examinee who has mastered none of the required attributes but has correctly answered many items, while a ''spuriously low scorer'' is an examinee who has mastered all the required attributes but has incorrectly answered many items. Generalized likelihood ratio tests are commonly performed at $\alpha = .05$ to test the significance of the difference between the desired response pattern and the actual response pattern of an examinee. In Liu et al. (2009), around 8% to 14% of N = 2,922 examinees were flagged as spurious scorers. This may be used as a reference for future applications. Examinees showing large misfit may be at higher risk for model-based misclassification given that the model does not well represent their item responses. In this case, it may be that their attribute profiles should not be reported.

In summary, the choice of the final DCM should involve both theory and empirical fit statistics. If the fit statistics of the best fitting DCM are not satisfactory, retrofitting may be unlikely to yield accurate and reliable classifications of examinees.

Examining Attribute Correlations and Reliability. For attribute profiles to provide meaningful diagnostic information, they must be distinct from the unidimensional latent trait and have adequate reliability (Sinharay, 2014). Therefore, attribute correlations and reliability under the DCM should be evaluated before interpreting results.

First, the tetrachoric correlations between each pair of attributes provide information about the distribution of the attribute profiles (Hartz, 2002; Templin & Henson, 2006). Most often, attributes obtained from retrofitting to an assessment that was originally designed to be unidimensional are expected to be highly correlated. If the correlations among attributes are close to or equal to 1, it logically follows that examinees will overwhelmingly fall into either the all-mastery or none-mastery profile. In this case, the assessment is essentially unidimensional, and the classifications along multiple attributes are most likely to provide misleading information about examinees. As stated by Sinharay (2010), although a unidimensional trait can always be broken down into smaller attributes, it is very likely that those attributes are too correlated to provide useful information.

Second, the tetrachoric correlations can be used to compute reliability using the method developed by Templin and Bradshaw (2013). That method can be conceptualized as the consistency of examinee classifications across different administrations and mirrors test–retest reliability in IRT (Templin & Bradshaw, 2013). Similar to other psychometric frameworks, reliability is one of the most important prerequisites of validity (Kane, 2013). Unreliable attribute classifications can ultimately lead to invalid inferences from the DCM results.

Interpreting Results

The primary outcome from retrofitting is examinee profiles. Suppose three attributes are specified when one retrofits to an elementary math ability test: addition, subtraction, and multiplication. Two hypothetical individual diagnostic reports are presented in Figure 2. In this example, two examinees have received the same overall score (i.e., 325) with different attribute mastery decisions and a different probability of mastery on each attribute. As mentioned in Rupp et al. (2010), statistically, examinees have greater than or equal to a 0.5 probability of mastery on an attribute are classified as mastery on that attribute, and less than 0.5 as nonmastery. Examinees

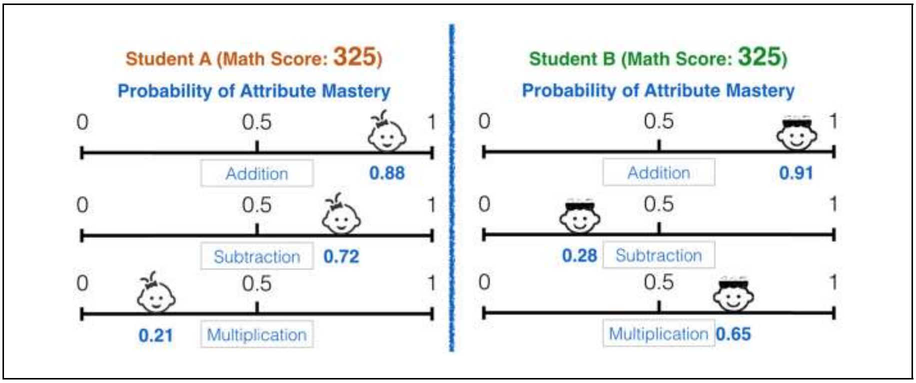

Figure 2. Example individual diagnostic reports from retrofitting.

that near the 0.5 cut point may be classified as ‘‘undecided’’ because the assessment does not provide enough information about them (Rupp et al., 2010). Although both students in Figure 2 have received the same total test score, student A has mastered addition and subtraction, but not multiplication, while student B has mastered addition and multiplication, but not subtraction.

Providing diagnostic feedback to examinees along with their single, unidimensional test scores may help individuals identify their attribute-level strengths and weaknesses. In addition to obtaining individual diagnostic feedback, one may also (a) aggregate individual information to the classroom, school, or state level for instruction planning or low-stakes decision making (e.g., Ketterlin-Geller & Yovanoff, 2009); (b) infer a learning sequence from the number of examinees in each attribute profile (Templin & Bradshaw, 2014); (c) use the correlation among attributes to learn more about theories of constructs (Templin & Henson, 2006); and (d) investigate the relationship between examinees’ single, unidimensional test scores (e.g., θ_e) and their categorical diagnostic profiles ($\boldsymbol{\alpha}_e$) to see if some attributes are more difficult to master than others. The first three types of results could be obtained through regular DCM practices, but the last one is unique to retrofitting. We present these four types of results through an empirical example in the next section.

An Empirical Example of Retrofitting

To help readers understand the retrofitting framework, an empirical example is presented in this section. The purpose of the example is not to examine whether this specific assessment or dataset could be used for retrofitting; instead, it aims to demonstrate possible outcomes that one can gain from the methodological framework of retrofitting.

Response data were collected from 422 examinees who participated in a 51-item mock TOEFL (Test of English as a Foreign Language) listening test. This mock

Table 1. Q-matrix in DCM and a, b, c Parameters in IRT.

	DCM_Q			IRT_Easy			IRT_Medium			IRT_Hard		
Item	*A*1	*A*2	*A*3	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
1	1	0	0	0.59	−1.04	0.26	0.42	−0.76	0.29	0.84	0.3	0.03
2	1	0	0	0.88	0.06	0.05	0.54	0.64	0.13	0.92	1.3	0.12
3	0	0	1	1.07	−1.19	0.05	0.17	0.89	0.01	0.46	1.26	0.04
4	1	0	0	0.72	−0.6	0.05	0.96	1.14	0.19	0.38	0.29	0.28
5	0	1	0	0.87	−1.14	0.24	1.2	−0.4	0.07	0.71	0.54	0.02
6	1	0	0	0.95	−0.71	0.12	0.86	−0.27	0.24	1.06	1.09	0.04
7	0	1	1	0.55	−1.28	0.27	0.69	1.07	0.14	1.12	0.66	0.02
8	0	1	1	0.52	−0.92	0.06	1.09	0.24	0.05	0.71	1.29	0.03
9	0	0	1	1.05	−1.32	0.01	1.12	0.11	0.21	0.58	1.05	0.13
10	0	1	0	0.39	−0.9	0.2	0.47	0.32	0.08	0.91	1.24	0.02
11	0	1	1	1.1	−1	0.02	0.13	−1.01	0.16	1.06	1.09	0.1
12	1	0	0	1	−1.1	0.08	0.43	−0.98	0.01	0.27	1.88	0.28
13	0	1	0	0.78	−0.67	0.04	0.72	−0.81	0.02	0.69	1.57	0.04
14	1	0	0	0.95	−1.1	0.08	0.57	0.05	0.1	0.78	0.8	0.05
15	1	0	1	0.69	−0.84	0.08	0.88	0.11	0.06	0.85	0.76	0.04
16	0	1	0	0.85	−0.05	0.02	1.1	−0.6	0.11	0.59	1.2	0.01
17	0	0	1	0.75	−0.11	0.02	0.32	0.02	0.13	0.43	0.84	0
Summary[a]	7	7	7	0.81	−0.82	0.10	0.69	−0.01	0.12	0.73	1.01	0.07

[a] The “Summary” row in “DCM_Q” columns refers to the total number of items measuring an attribute. The “Summary” row in “IRT” columns refers to the mean value of a parameter of all items.

TOEFL test was developed by a testing-prep company in Nanjing, China. The items were developed following the TOEFL 2000 framework (Douglas, Jamieson, Nissan, & Turner, 2000) and the assessment content, format, and length were approximating the original TOEFL listening. This mock test has three sections with 17 items in each section. The sections were designed to measure the same construct but have different difficulty levels. The test was administered to a larger sample of 6,250 examinees and a single score was reported for each examinee. The three-parameter logistic (3PL; Birnbaum, 1968) model was used for item calibration. The *a* (discrimination), *b* (difficulty), *c* (guessing) parameters for each item on the three sections are shown in the right columns of Table 1. The means of each parameter in the ‘‘summary’’ row show that the *a* and *c* parameters in the three tests have similar values, while the values of the *b* parameter differ across the three banks.

The primary purpose of this mock test was to predict students’ TOEFL scores and place them in different test-prep classes. However, after scores are obtained, students want to identify the specific listening skills that they need to improve, and teachers are interested in gaining more information to facilitate instruction. Retrofitting is therefore considered a possible approach to provide this kind of formative information.

Table 2. Absolute Model Fit.

		$maxX^2$	*MADcor*	*SRMSR*	100^*Res	MQ_3
Easy	G-DINA	5.597	0.034	0.042	0.718	0.046
	A-CDM	4.516	0.034	0.042	0.705	0.046
	DINA	4.495	0.035	0.043	0.726	0.045
	DINO	4.458	0.036	0.043	0.751	0.045
	HO-DINA	4.398	0.035	0.044	0.737	0.045
Medium	G-DINA	47.236	0.038	0.046	0.893	0.041
	A-CDM	3.823	0.031	0.039	0.741	0.042
	DINA	3.721	0.033	0.040	0.776	0.042
	DINO	3.667	0.033	0.040	0.775	0.042
	HO-DINA	3.579	0.033	0.040	0.778	0.042
Hard	G-DINA	5.669	0.034	0.043	0.792	0.042
	A-CDM	4.907	0.033	0.042	0.780	0.042
	DINA	5.134	0.034	0.042	0.793	0.042
	DINO	5.231	0.034	0.042	0.791	0.041
	HO-DINA	5.442	0.035	0.043	0.814	0.043

This specific assessment may be well-suited for retrofitting because three skills were identified under the overall listening ability in the original assessment development framework. Based on Douglas et al. (2000), the three skills are (a) basic comprehension, (b) pragmatic understanding, and (c) connecting information and understanding organization. During the retrofitting process, we worked alongside content developers of the mock TOEFL test to link items to the three skills. The final Q-matrix is specified in the left column of Table 1. In the final Q-matrix, each skill is measured seven times within 17 items. Please note that we are treating the three sections as distinct item banks because the skills bear different meanings (in terms of difficulty) across the three banks, although they share the same name. The five DCMs that were introduced in the previous section were fit to the responses. R (R Core Team, 2016) and the ''CDM'' package (Robitzsch, Kiefer, George, & Uenlue, 2016) were used for retrofitting. ''ROFDM'' (Liu, 2016) was used for producing outputs. In the following sections, we briefly evaluate fit statistics and focus on presenting the information that one can obtain from retrofitting.

Fit Statistics, Attribute Correlations, and Reliability

Table 2 presents the results for absolute fit. Overall, models within each item bank had a similar fit. All models had SRMSR values smaller than 0.05, indicating acceptable fit. Models fit better to data in the medium item bank than in the easy or hard item banks. Table 3 presents results for relative fit. The relative fit statistics were also very similar across models. Across the three item banks, DINA model was the best fit according to AIC, and the HO-DINA model was the best fit according to BIC and CAIC. One of the reasons that the HO-DINA model had better relative fit according

Table 3. Relative Model Fit.

		−2LL	AIC	BIC	CAIC	NP
	G-DINA	8617.03	8715.03	8913.23	8962.23	49
	A-CDM	8620.02	8710.02	8892.05	8937.05	45
Easy	DINA	8624.76	8706.76	8872.60	8913.60	41
	DINO	8634.33	8716.33	8882.17	8923.17	41
	HO-DINA	8627.12	8707.12	8868.92	8908.92	40
	G-DINA	9353.86	9451.86	9650.07	9699.07	49
	A-CDM	9359.68	9449.68	9631.70	9676.70	45
Medium	DINA	9365.31	9447.31	9613.15	9654.15	41
	DINO	9367.39	9449.39	9615.24	9656.24	41
	HO-DINA	9369.32	9449.32	9611.12	9651.12	40
	G-DINA	9302.91	9400.91	9599.11	9648.11	49
	A-CDM	9307.85	9397.85	9579.88	9624.88	45
Hard	DINA	9314.15	9396.15	9561.99	9602.99	41
	DINO	9316.17	9398.17	9564.01	9605.01	41
	HO-DINA	9316.66	9396.66	9558.46	9598.46	40

Note. NP is the number of parameters.

Table 4. Likelihood Ratio Test Among Models.

	Model1	Model2	χ^2	*df*	*p*
	A-CDM	G-DINA	2.992	4	.559
Easy	DINA	G-DINA	7.729	8	.460
	DINO	G-DINA	17.300	8	.027
	A-CDM	G-DINA	5.816	4	.213
Medium	DINA	G-DINA	11.447	8	.178
	DINO	G-DINA	13.531	8	.095
	A-CDM	G-DINA	4.942	4	.293
Hard	DINA	G-DINA	11.237	8	.189
	DINO	G-DINA	13.259	8	.103

to BIC and CAIC is that these two indices impose a large penalty for highly parameterized models. The HO-DINA model had only 40 parameters, the least among five models. Table 4 presents the likelihood ratio tests between the saturated G-DINA model and each nested model. Results show that only the DINO model in the easy item bank fit significantly worse than the saturated G-DINA model at α = .05. To conclude, all five models fit similarly well to the data.

Figure 3 presents RMSEA values for item fit. In our dataset, item fit seems adequate given all RMSEA values were below .05, except for the G-DINA model in the medium item bank. Generally, the spread of the RMSEA was smaller in the hard item bank and the largest in the easy item bank. Across models, the G-DINA model had

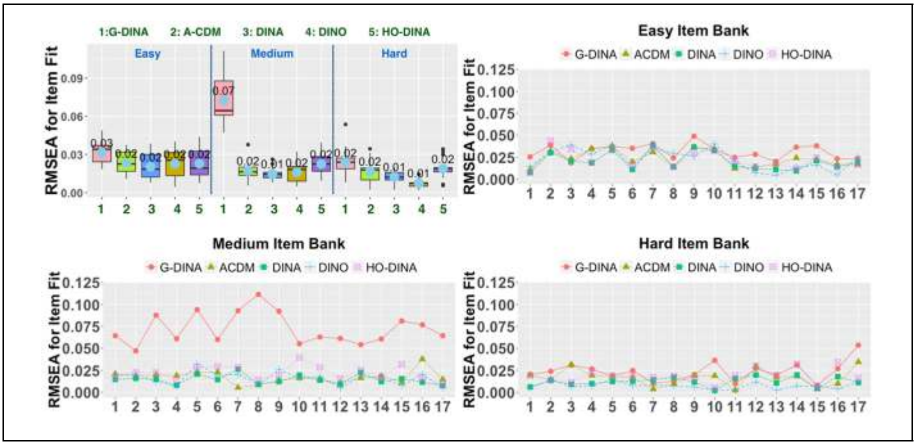

Figure 3. Item fit statistics.

Note. The “RMSEA” in von Davier (2005) is computed differently from the “RMSEA” often used in structural equation modeling.

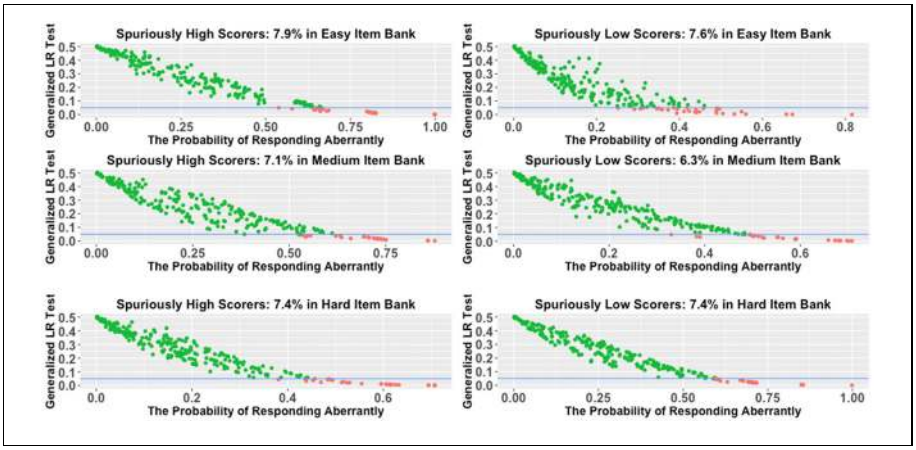

Figure 4. Person fit statistics.

the worst item fit across three item banks, especially in the medium item bank. DINA had the best item fit in the easy and medium item banks, and DINO had the best item fit in the hard item bank. When an item did not fit adequately, one can consider dropping it from retrofitting. Ultimately, we did not drop any items as all item fit adequately in this dataset, but in other datasets that would have been further considered.

Figure 4 presents person fit results. Due to limitations of space, we only show person fit statistics under the HO-DINA model, given that it was the most parsimonious and showed adequate model fit. Across three item banks, about 7% of examinees were spurious high scorers and 7% were spurious low scorers at α = .05. This

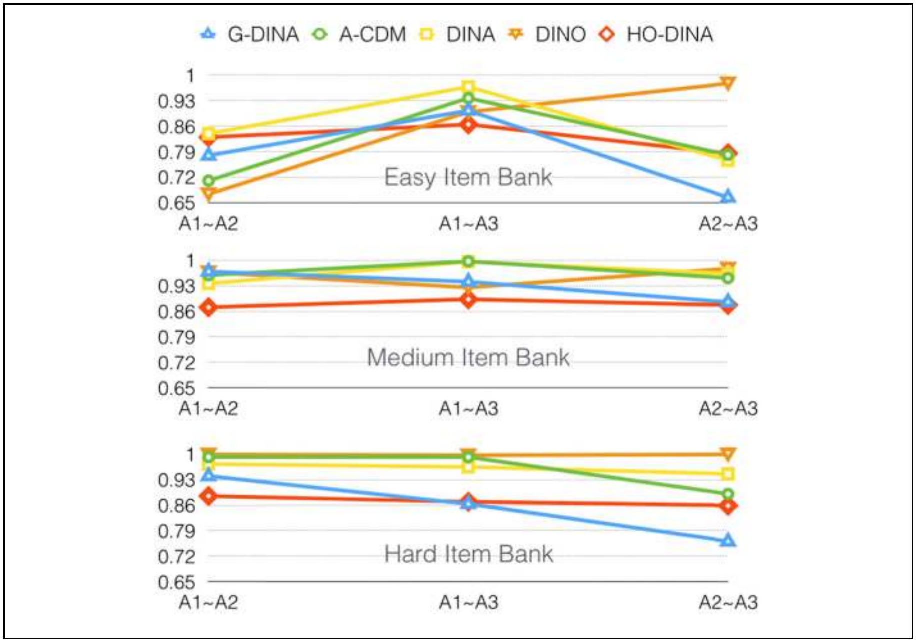

Figure 5. Tetrachoric correlations among attributes.

percentage is lower than the results from an assessment designed under the DCM framework in Liu et al. (2009). Hence, for this assessment and dataset, person fit seems adequate.

Figure 5 shows the tetrachoric correlations in each item bank across the five DCMs. Although different models did not all agree with each other, the correlations between each pair of attributes were mostly above .70 across all item banks. This is expected as we retrofitted to a test that was developed under a unidimensional structure. We hypothesized that attributes were less correlated in the HO-DINA model partly because that model incorporated a higher-order latent trait. Reliability was the highest in the medium item bank and the lowest in the easy item bank and this finding was consistent across models. For example, under the HO-DINA model, the reliability for the easy, medium, and hard banks were .71, .81, and.73, respectively.

Overall, the attribute correlations, the reliability estimates, and the majority of the fit indices showed adequate results in this dataset. However, in other retrofitting applications, this may or may not occur. Evaluating those indices is necessary but insufficient to inform users whether retrofitting could provide useful information.

Results Interpretation

As discussed in the above framework, one of the most important pieces of information one could gain from retrofitting is examinee classification on each of the

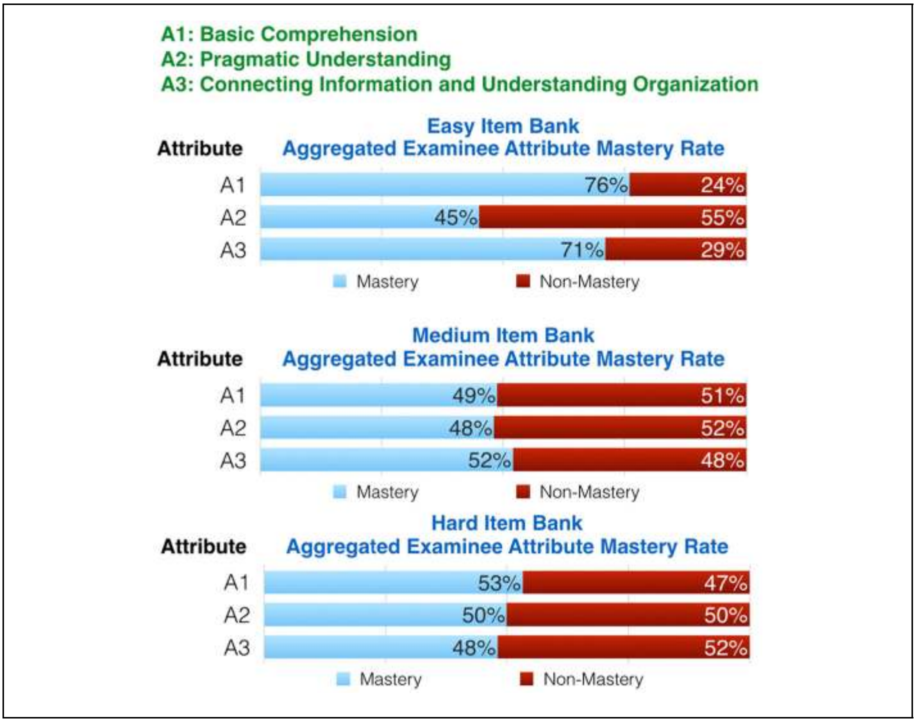

Figure 6. Aggregated examinee attribute mastery rate.

attributes. The information includes both a probability of mastery and mastery decisions on each attribute, similar to what we have shown in Figure 2. In the remainder of this section, we present the four pieces of information that were mentioned in the retrofitting framework.

First, beyond individual-level information, one could aggregate mastery status across a large cohort of examinees. Figure 6 shows the aggregated examinee attribute mastery rate in each item bank. In the easy item bank, most examinees have mastered α_1 and α_3. In the medium and hard item banks, around half of the examinees were classified as having mastered each attribute.

Second, the frequency of examinees in each attribute pattern may also be informative to end users. We present this information in Figure 7. Across three banks, most examinees were classified either as nonmastery on each attribute ($\{0, 0, 0\}$) or all-mastery ($\{1, 1, 1\}$). This should be expected in most retrofitting contexts in which the test was originally developed under the unidimensional framework. However, there were some examinees classified as mastering only one or two of the attributes. For example, in the easy item bank, around 20% of examinees were classified as mastering α_1 and α_3 (i.e., $\{1, 0, 1\}$).

Third, one could examine whether some pairs of attributes are more correlated than others using the information presented in Figure 5. For example, α_1 and α_3 are

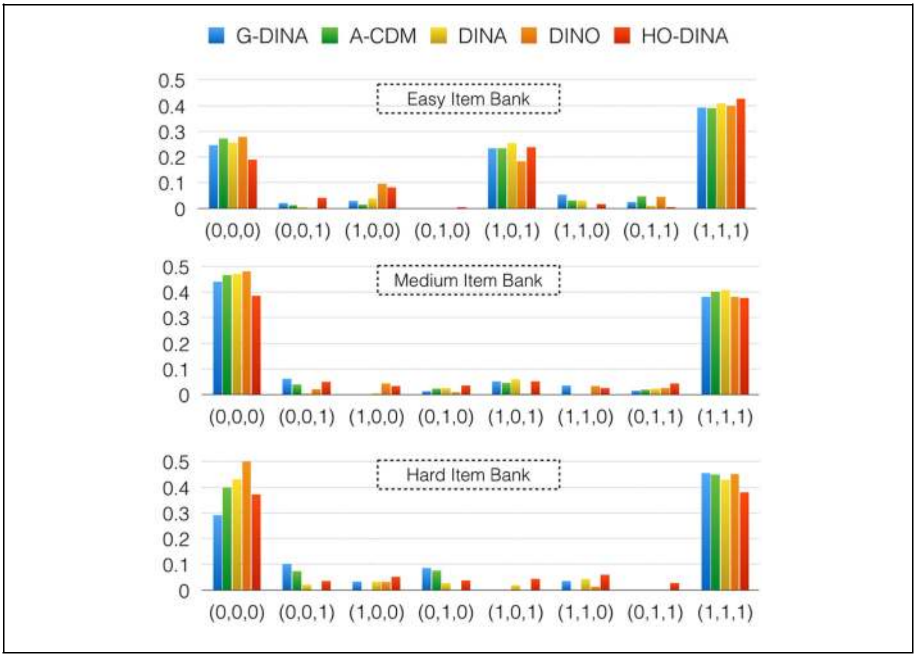

Figure 7. Aggregated examinee attribute pattern frequency.

more correlated than the other two pairs in the easy item bank. This means that examinees who have mastered α_1 more likely have also mastered α_3 instead of α_2. This information can be cycled back for learning, instruction, and future assessment development (Rupp et al., 2010).

The last piece of information we present here is a unique product one can gain from retrofitting. It is the relationship between attribute profile in the DCM ($\boldsymbol{\alpha}$) and the general ability ($\hat{\theta}$) estimated under the IRT. This information is presented in Figure 8. In all item banks, as the number of mastered attributes increased, $\hat{\theta}$ increased. However, the trend is clearer in the medium item bank than in the easy or hard item banks. This information can be used in several ways by end users. For example, the alignment of $\boldsymbol{\alpha}$ and $\hat{\theta}$ in the medium item bank can be used to hypothesize a sequence of attribute difficulty as follows: $\alpha_3 \rightarrow \alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_1\alpha_3 \rightarrow \alpha_1\alpha_2 \rightarrow \alpha_2\alpha_3 \rightarrow \alpha_1\alpha_2\alpha_3$. For the examinees in the easy and hard item bank, it may not be easy to perfectly associate their mastery profiles to their $\hat{\theta}$ scores. However, this exemplifies the value of retrofitting which is to provide nuanced attribute mastery information when examinees obtain similar $\hat{\theta}$ scores.

Conclusion

One core issue of validity lies within the usefulness of an assessment tool and the evidence that can be provided for those uses (Kane, 2013). The overall value of

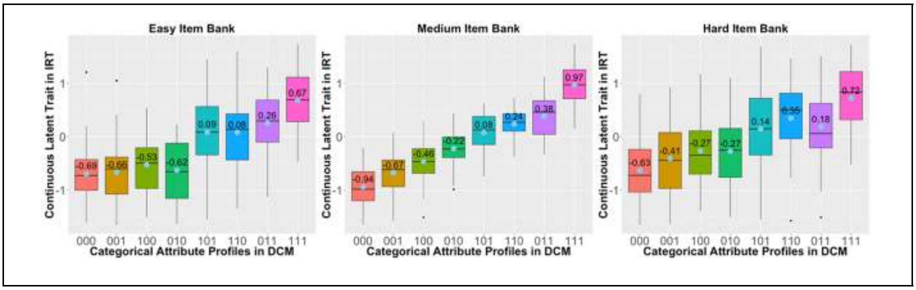

Figure 8. Relationship between attribute profiles in DCM and the general ability in IRT.

retrofitting, assuming that the results are accurate and reliable, is the increased utility for some applications. However, we want to restate that retrofitting should not be encouraged as an ideal approach. Furthermore, results obtained from assessments without a sound development framework should not be used for any high-stakes decision making. We posit that retrofitting is not always successful and useful, but can be in particular situations. For example, some high-stakes licensure tests (e.g., The United States Medical Licensing Examination; USMLE) provide both a single score with pass/fail decisions to examinees and subscores for subcontent areas because those tests must cover multiple subcontent areas within the large domain to ensure that an examinee with a ‘‘pass’’ has mastered all the required knowledge and skills to practice (Greenburg, Case, Golden, & Melnick, 1997; Raymond, Kahraman, Swygert, & Balog, 2011). For those tests, retrofitting may be successful because there are distinct subcontent areas that each item is intended to measure (e.g., anatomy and epidemiology on the USMLE) and factor analyses on data from that test display multidimensionality (e.g., De Champlain, Swygert, Swanson, & Boulet, 2006). It may also be useful because the mastery decisions on each subcontent area could inform examinees of what they need to learn and improve. In sum, we propose that retrofitting may be considered when (a) it is impossible to develop a diagnostic test due to practical constraints, (b) the purpose is to obtain formative information that supports learning or informs the theory of construct, and (c) underlying multiple skills or behaviors were clearly specified during item development and supported through dimensionality assessment.

There are many obstacles one may face when retrofitting. For example, it is possible that items measuring the same set of attributes can have different levels of difficulty. As mentioned by Choi (2010), if we ignore the discrepant levels of difficulty in retrofitting, classification accuracy may decrease. There are three approaches to accommodate varied item difficulty that are currently available. The first approach is to treat item difficulty as a continuous variable and directly incorporate it into the modeling process using the diagnostic classification mixture Rasch model (Choi, 2010). The second approach is to treat item difficulty as an ordinal variable (e.g., easy, medium, and difficult) and define the required attribute level of mastery using polytomous DCMs (e.g., Chen & de la Torre, 2013). The third approach is to use the multicomponent latent trait model for diagnosis (MLTM-D; Embretson & Yang, 2013), which involves both a continuous latent trait and multiple categorical attributes in one model. Embretson and Yang (2013) has demonstrated that he MLTM-D works well for heterogeneous item pools such as the licensure tests mentioned above. Beyond item difficulty, other item and test statistics in IRT may affect the classification accuracy of retrofitting. Future research on these topics would be helpful for practitioners to evaluate the potential of retrofitting.

Beyond providing information about specific examinees, retrofitting may be applied to some large-scale standardized assessments not for the purpose of obtaining information about examinees, but for using the large amount of data available from standardized assessments to understand more about the construct domains themselves. Indeed, the process of diagnostic measurement can provide a wealth of information about construct and learning theories (Rupp et al., 2010). Retrofitting DCMs into data from these assessments could help to refine theories about how these domains are interrelated, learned by students, and more.

An alternative approach of exploring the possibility of retrofitting is to investigate the mathematical equivalence between IRT models and DCMs. Raykov and Marcoulides (2015) have shown that CTT and IRT are mathematically equivalent approaches with some constraints. Future research demonstrating the mathematical relationships between DCMs and IRT models would be very helpful.

We argue that retrofitting should not be seen as a strong alternative to principled assessment development practices. While retrofitting presents a feasible approach to gain more actionable information from existing assessments of other psychometric frameworks under certain circumstances, much caution is needed to use and interpret DCMs appropriately. At a bare minimum, persons performing retrofitting should be able to fully and clearly explain to end users the quality of information obtained from the retrofitting and the potential risks of using it in various ways. However, as the interest in obtaining information on multidimensional attributes increases, it is recognized that assessments that are developed under the diagnostic framework will better support learning, teaching, and decision making.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317-332.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice*, *18*(3), 5-12.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental rest scores* (Part 5, pp. 397-479). Reading, MA: Addison Wesley.

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

Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*, 119-157. doi:10.1177/026553229801500201

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*, 419-437.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123-140.

Choi, H.-J. (2010). *A model that combines diagnostic classification assessment with mixture item response theory models* (Unpublished doctoral dissertation). University of Georgia, Athens.

Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*, 19-38.

De Champlain, A., Swygert, K., Swanson, D. B., & Boulet, J. R. (2006). Assessing the underlying structure of the United States Medical Licensing Examination Step 2 test of clinical skills using confirmatory factor analysis. *Academic Medicine*, *81*(10), S17-S20.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343-362.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115-130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

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

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement*, *46*, 450-469.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*(2), 89-97.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253-273. doi:10.1007/s11336-015-9467-8

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 979-1030). Amsterdam, Netherlands: Elsevier.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, *44*, 285-291.

Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework.* Princeton, NJ: Educational Testing Service.

Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika*, *78*, 14-36.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement*, *44*, 325-340.

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement*, *6*, 263-268.

Gierl, M. J., Roberts, M., Alves, C., & Gotzmann, A. (2009, April). *Using judgments from content specialists to develop cognitive models for diagnostic assessments*. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.

Greenburg, A. G., Case, S. M., Golden, G. S., & Melnick, D. E. (1997). Core clinical content of Step 2 of the USMLE: Using surgery as an example. In A. J. J. A. Scherpbier, C. P. M. van der Vleuten, J. J. Rethans, & A. F. W. van der Steeg (Eds.), *Advances in medical education* (pp. 34-36). Dordrecht, Netherlands: Springer.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204-229.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301-321.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practice* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191-210.

Ho, A. D. (2016). The new (educational) statistics properties of scales that matter. *Journal of Educational and Behavioral Statistics*, *41*, 94-99.

Hou, L., de la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*, 98-125.

Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, *16*, 119-141. doi: 10.1080/15305058.2015.1133627

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). Cambridge, England: Cambridge University Press.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion Model to LanguEdge assessment. *Language Testing*, *26*(1), 31-73.

Jang, E. E., Dunlop, M., Wagner, M., Kim, Y. H., & Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. *Language Learning*, *63*, 400-436.

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*, 393-419.

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Ketterlin-Geller, L. R., & Yovanoff, P. (2009). Diagnostic assessments in mathematics to support instructional decision making. *Practical Assessment, Research & Evaluation*, *14*(16), 1-11.

Kim, A. Y. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*, 227-258. doi:10.1177/0265532214558457

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, *35*(2), 64-70.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59-81.

Lee, Y.-W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*, 239-263.

Lei, P. W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement*, *40*, 405-417.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*(3), 205-237.

Li, H., & Suen, H. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, *18*(1), 1-25.

Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, *33*(3), 391-409.

Liu, R. (2016). *R Output Functions for Diagnostic Measurement (ROFDM)* [Computer software manual; Version 1.0]. Retrieved from https://sites.google.com/site/renliustats/dcm

Liu, R., & Huggins-Manley, A. C. (2016). The specification of attribute structures and its effects on classification accuracy in diagnostic test design. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (pp. 243-254). New York, NY: Springer.

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement*, *77*(2), 220-240.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, *33*, 579-598.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*(3), 200-217.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics*, *2*(2), 99-120.

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*, 491-511.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. Measurement. *Interdisciplinary Research and Perspectives*, *11*(3), 71-101.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*, 305-328.

McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*, 23-40.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20.

National Research Council. (2010). *State assessment systems: Exploring best practices and innovations: Summary of two workshops* (Alexandra Beatty, rapporteur. Committee on Best Practices for State Assessment Systems: Improving assessment while revisiting standards. Center for Education, Division of Behavioral and Social Sciences and Education). Washington, DC: National Academies Press.

Neyman, J., & Pearson, E. S. (1992). On the problem of the most efficient tests of statistical hypotheses. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics* (pp. 73-108). New York, NY: Springer.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64*, 575-603.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (2012). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

Nichols, P. D., Kobrin, J. L., Lai, M., & Koepfler, J. (2016). The role of theories of learning and cognition in assessment design and development. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 297-327). Chichester, England: Wiley-Blackwell.

R Core Team. (2016). R (Version 3.3) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 782-799.

Ravand, H., Barati, H., & Widhiarso, W. (2012). Exploring diagnostic capacity of a high stakes reading comprehension test: A pedagogical demonstration. *Iranian Journal of Language Testing*, *3*(1), 12-37.

Raykov, T., & Marcoulides, G. A. (2015). On the relationship between classical test theory and item response theory from one to the other and back. *Educational and Psychological Measurement*, *76*, 325-338.

Raymond, M. R., Kahraman, N., Swygert, K. A., & Balog, K. P. (2011). Evaluating construct equivalence and criterion-related validity for repeat examinees on a standardized patient examination. *Academic Medicine*, *86*, 1253-1259.

Robitzsch, A., Kiefer, T., George, A.C., & Uenlue, A. (2016). *CDM: Cognitive diagnosis modeling* [Computer software manual; R package version 5.0.0]. Retrieved from http://CRAN.R-project.org/package=CDM/

Rupp, A. A., & Templin, J. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*, 78-96.

Rupp, A. A., & Templin, J. L. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 219-262.

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

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

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150-174.

Sinharay, S., & Haberman, S. J. (2008). *Reporting subscores: A survey* (ETS Research Memorandum No. 08-18). Princeton, NJ: Educational Testing Service.

Sinharay, S. (2014). Analysis of added value of subscores with respect to classification. *Journal of Educational Measurement*, *51*, 212-222.

Stout, W., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357-375). New York, NY: Springer.

Svetina, D., Gorin, J. S., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing*, *11*(1), 1-23.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345-354.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

Tatsuoka, C., Clements, D. H., Sarama, J., Izsak, A., Orril, C. H., de la Torre, J., & Khasanova, E. (2016). Developing workable attributes for psychometric models based on the Q-matrix. *Journal for Research in Mathematics Education*. Retrieved from http://www.personal.psu.edu/users/h/h/hht1/JRME_2016_04.pdf

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*, 251-275.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*, 317-339.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287-305.

Thissen, D. (2016). Bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics*, *41*(1), 81-89.

VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading section of the SAT Reasoning Test. In D. S. McNamara (Ed.), *Reading comprehension strategies* (pp. 137-171). Mahwah, NJ: Erlbaum.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2007). *Hierarchical general diagnostic models* (ETS Research Report No. RR-07-19). Princeton, NJ: Educational Testing Service.

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series*, *2014*(2), 1-13.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, *81*, 625-649. doi:10.1007/s11336-015-9471-z

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125-145.