2020 ■ Vol. 16 ■ no. 5

Using Diagnostic Classification Models in Psychological Rating Scales

Ren Liu [a] and Dexin Shi [b]

[a]University of California, Merced
[b]University of South Carolina

Abstract ■ The purpose of this article is to demonstrate how to use diagnostic classification models (DCMs) in psychological rating scales and reflect on how DCMs differ from classical test theory (CTT) and item response theory (IRT) scoring approaches in terms of assumptions and results. DCMs, a viral topic in today's psychometric world, has thrived in educational assessment. However, many researchers and practitioners are uncertain how DCMs could be used in psychological rating scales and what information they could provide. This article presents an example of applying CTT, IRT, and DCM scoring approaches with data from 10,000 respondents on an operational personality rating scale.

Keywords ■ diagnostic classification model, psychological rating scales, item response theory, classical test theory.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

rliu45@ucmerced.edu

10.20982/tqmp.16.5.p442

Introduction

In the 1980s, Haertel introduced a restricted latent class model to classify individuals with respect to their possession of a set of skills or attributes (Haertel, 1989). The major assumption behind that model is that the latent traits are treated as categorical rather than continuous variables. Similar model developments can be found in even earlier literature such as Dayton and Macready (1976) and Macready and Dayton (1977). However, Haertel's model is commonly recognized as the first model of its kind, which was later referred to as the family of cognitive diagnosis models (e.g. Templin & Henson, 2006) or diagnostic classification models (DCMs; e.g. Rupp, Templin, & Henson, 2010). Later, Haertel's model was named the "deterministic inputs, noisy, and gate" (DINA) model in Junker and Sijtsma (2001) and remained one of the most widely discussed models in the family of DCMs. Since then, more than 30 DCMs have been introduced, most of which were built to analyze data in educational assessment. However, DCMs, aiming to classify individuals into latent classes, are also apparent candidates for many psychological rating scales that aim to assign individuals with attributes. For example, the arguably most famous personality rating scale: Myers-Briggs Type Indicator (MBTI; Myers, McCaulley, Quenk, & Hammer, 2003) aims to classify each individual with one of 16 possible personality types. Those 16 personality types are all possible combinations of four binary latent traits: introversion/extroversion, sensing/intuition, feeling/thinking, perceiving/judging. However, little is known regarding how to use DCMs in such tests and how DCM results are different from results of traditional scoring approaches such as classical test theory (CTT) and item response theory (IRT).

This article discusses how to use DCMs in psychological rating scales and provides an empirical example of using DCM, CTT, and IRT to score the same dataset. We first review the fundamental characteristics of DCMs, and DCMs that are possible for scoring item data from psychological rating scales. Next, we describe an operational test data and discuss the results from three psychometric approaches. Finally, we point out some caveats of using DCMs in psychological rating scales. Note that the purpose of this article is not to compare the mathematical differences between psychometric approaches. Instead, the focus is on how to use DCMs in psychological assessment empirically and discuss what information they can provide. Previously in this journal, George and Robitzsch (2015) provided a tutorial on how to use a R package to fit DCMs to correct/incorrect responses in educational settings. Different from that tutorial, the current article focuses on DCMs for psychological rating scales and comparing between information that different scoring approaches provide.

Latent Variables in DCMs

DCMs refer to a class of multidimensional models expressing the relationship between item responses and multiple categorical latent traits. One of the most commonly used examples in DCM introductory materials is an elementary math test where the goal is to decide whether students have mastered four skills: addition, subtraction, multiplication, and division (e.g. Rupp et al., 2010). If the test is scored with the CTT approach, each student gets a total summated score (e.g., an 85 out of 100) and an observed score for each of the four skills. If the test is scored with a multidimensional IRT model, each student gets an estimated latent score (e.g., -0.36) for each skill. Using a DCM, the "score" for each student is an estimated profile of mastery/non-mastery status of each skill. This poses a very first question of the legitimacy of using DCMs in educational and psychological tests: are latent traits binary/categorical? Since the concept of "latent traits" is an immaterial construct, a more accurate question might be: is that a good idea we consider latent traits as binary/categorical variables? A major advantage of doing so is that one could trade the accuracy of locating examinees on a single trait continuum with that of roughly grouping them based on their mastery status of multiple latent traits. Obtaining the probability that an individual masters multiple latent traits is intuitively more informative than obtaining information on only one trait. In addition, since the reliability is conceptualized as the accuracy of group classifications, DCMs were demonstrated to have higher reliability than IRT models with similar test lengths (e.g. Templin & Bradshaw, 2013).

Despite its clear advantages in providing multidimensional classifications, using DCMs to model the relationship between item responses and categorical traits could pose potential problems. When multiple latent traits are specified in one test, it is not uncommon to see that those traits are highly correlated. As a result, it may be hard to extract multidimensional information from a possibly unidimensional dataset. For example, Liu, Huggins-Manley, and Bulut (2018) fit a general DCM to three datasets where the correlation between traits were around .80-.90. The bar chart of examinees' latent trait profiles in each dataset display a "U-shape" curve where most examinees were classified as none-masters or all-masters at the two ends while few examinees were classified into those partial mastery profiles. In such cases, even if the model fits the data well, most examinees are essentially categorized into one of the two groups (i.e., none-masters and all-masters), providing little useful information. Besides the collinearity issue, another potential problem of DCMs is the large number of parameters when there is a complex item-trait loading structure. A simple loading structure means that an item only measures one latent trait, whereas a complex structure means that an item measures more than one trait. For a test measuring four latent traits, an item can be associated with as little as two parameters and as much as 16 parameters. Such a large number of parameters could induce problems of model under-identification (e.g. Gu & Xu, 2018; Xu & Zhang, 2016) and overfitting (e.g. von Davier, 2018). For example, Templin and Bradshaw (2014) showed that a two-parameter IRT model fit a dataset better with a smaller number of parameters comparing to a general DCM with much more parameters.

The discussion of whether latent traits should be modeled under DCMs as categorical variables could continue, but there is probably not a definite answer. Although DCMs are going viral these days, it is necessary to see the fact that DCMs can fulfill certain scoring purposes as an alternative psychometric approach, but they are not "better" comparing to IRT or CTT approaches. Now let us introduce DCMs that are theoretically available for scoring psychological rating scales.

DCMs for Psychological Rating Scales

Before implementing a DCM, we need to (1) specify latent traits, and (2) specify which items measure which traits. For $k = 1, 2, \ldots, K$ latent traits (also known as *attributes*), there are 2^K possible attribute possession patterns (also called *attribute profiles*), where each attribute profile can be represented by a vector $\boldsymbol{\alpha} = (\alpha_1,\ \alpha_2, \ldots,\ \alpha_K)$. For example, there are four attributes in the MBTI (i.e., $K = 4$): (1) extroversion/introversion, (2) sensing/intuition, (3) thinking/feeling, and (4) judging/perceiving. These four attributes form the 16 personality types ($2^4 = 16$). Each attribute takes on a value of 0 or 1 representing the two opposite conditions of that attribute. For example, $\alpha_1 = 0$ for extroversion and $\alpha_1 = 1$ for introversion, following the presentation sequence in the previous sentence. With such specification, an examinee with $\boldsymbol{\alpha} = (1, 0, 0, 1)$ is assigned with the "ISTP" type. The 0s and 1s can be arbitrarily assigned to either condition since the possession of one condition is the non-possession of the other. The information of which items measure which attributes are contained in an item-by-attribute incidence matrix called a Q-matrix (Tatsuoka, 1983, 4). In a Q-matrix, an entry $q_{i,k} = 1$ when item i measures attribute k, and $q_{i,k} = 0$ otherwise.

To model the relationship between examinees' attribute possession and their item responses, DCMs for both dichotomous and polytomous items have been developed in the literature. Both item types are common in psychological assessments. For example, an item could ask examinees to choose whether they prefer to work (1) in a private setting or (2) in a group setting. Examinees select option 1 and 2, assigned with a score of 0 and 1 on this item, respectively. On the other hand, an item could also ask whether examinees strongly disagree, disagree, agree, or strongly agree with a statement: "I prefer working in a group setting". Such items are commonly scored in a polytomous fashion.

For dichotomous items, the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009) has become the boilerplate model over the years because it is the most general form of DCMs, accommodating many earlier DCMs. The LCDM defines the probability of individuals in attribute profile c scoring a "1" on item i as

$$P(X_i = 1|\boldsymbol{\alpha}_c) = \frac{exp\left[\lambda_{0,i} + \lambda_i^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)\right]}{1 + exp\left[\lambda_{0,i} + \lambda_i^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)\right]}, \tag{1}$$

where $\lambda_{0,i}$ is the intercept associated with item i, and $\lambda_i^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)$ index all the main effects and higher-order interaction effects of the $k = 1, \ldots K$ attributes associated with item i, which can be expressed as

$$\begin{aligned}&\sum_{k=1}^{K} \lambda_{1,i,k}\left(\alpha_{c,k} q_{i,k}\right) + \\ &\sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} \lambda_{2,i,k,k'}\left(\alpha_{c,k}\alpha_{c,k'} q_{i,k} q_{i,k'}\right) + \\ &\ldots\end{aligned}$$

For polytomous items, the nominal response diagnostic model (NRDM; Templin, Henson, Rupp, Jang, & Ahmed, 2008) is the most general model. Let $m = 0, 1, 2, \ldots, M-1$ index response options. The NRDM defines the probability of individuals in attribute profile c selecting response option m on item i as

$$P(X_i = m|\boldsymbol{\alpha}_c) = \frac{exp\left[\lambda_{0,i,m} + \lambda_{i,m}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)\right]}{\sum_{m=0}^{M-1} exp\left[\lambda_{0,i,m} + \lambda_{i,m}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)\right]}, \tag{2}$$

where $\lambda_{0,i,m}$ is the intercept parameter associated with option m on item i, and $\lambda_{i,m}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)$ index all the main effects and higher-order interaction effects of the k attributes associated with option m on item i, which can be expressed as

$$\begin{aligned}&\sum_{k=1}^{K} \lambda_{1,i,k,m}\left(\alpha_{c,k} q_{i,k}\right) + \\ &\sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} \lambda_{2,i,k,k',m}\left(\alpha_{c,k}\alpha_{c,k'} q_{i,k} q_{i,k'}\right) + \\ &\ldots\end{aligned}$$

Equations 1 and 2 demonstrate the same approach of splitting the effects of attributes on items into three parts: the intercept, main effects and interaction effects. The main difference is that a subscript m is added to all the effects in Equation 2. In other words, instead of splitting the effects of attributes at the item level, Equation 2 models the effects of attributes for each response option on each item. When there are only two response options (i.e., $m = 0,\ 1$), Equation 2 becomes Equation 1.

In many current psychological tests, each item only measures one attribute (i.e., items have a simple loading structure). This means that each item in the LCDM and each response option in the NRDM has only two parameters: an intercept parameter and a main effect parameter associated with the related attribute. All the higher-order interactions in both models are fixed to 0. Without interactions, most dichotomous DCMs are mathematically equivalent to the LCDM. We can rewrite the LCDM in Equation 1 for simple-structure items as

$$P(X_i = 1|\boldsymbol{\alpha}_c) = \frac{exp\left[\lambda_{0,i} + \lambda_{1,i,k}\left(\alpha_{c,k} q_{i,k}\right)\right]}{1 + exp\left[\lambda_{0,i} + \lambda_{1,i,k}\left(\alpha_{c,k} q_{i,k}\right)\right]}, \tag{3}$$

where $\lambda_{1,i,k}$ is the main effect associated with attribute k on item i, and $\alpha_{c,k}$ is a binary indicator representing whether examinees in attribute profile c possess attribute k (i.e., $\alpha_{c,k} = 0$ or 1). Similarly, we can rewrite the NRDM in Equation 2 for simple-structure items as

$$P(X_i = m|\boldsymbol{\alpha}_c) = \frac{exp\left[\lambda_{0,i,m} + \lambda_{1,i,k,m}\left(\alpha_{c,k} q_{i,k}\right)\right]}{\sum_{m=0}^{M-1} exp\left[\lambda_{0,i,m} + \lambda_{1,i,k,m}\left(\alpha_{c,k} q_{i,k}\right)\right]}, \tag{4}$$

where $\lambda_{1,i,k,m}$ is the main effect associated with attribute k on response option m of item i. Using the taxonomy of IRT polytomous models, the NRDM adopts a "divide-by-total" approach (Thissen & Steinberg, 1986) where the probability of selecting a particular response option is modeled as the effect of a particular response option divided by the sum of such effects of each response option. For instructional purposes, let us break down the summation symbol on the denominator of Equation 4. On item i with five response options ($M = 5$): 0, 1, 2, 3, and 4, the probability of selecting response option 3 is expressed as

$$P(X_i = 3|\boldsymbol{\alpha}_c) = \frac{exp[\lambda_{0,i,3} + \lambda_{1,i,k,3}(\alpha_{c,k}q_{i,k})]}{\begin{array}{c} exp[\lambda_{0,i,0} + \lambda_{1,i,k,0}(\alpha_{c,k}q_{i,k})] + \\ exp[\lambda_{0,i,1} + \lambda_{1,i,k,1}(\alpha_{c,k}q_{i,k})] + \\ exp[\lambda_{0,i,2} + \lambda_{1,i,k,2}(\alpha_{c,k}q_{i,k})] + \\ exp[\lambda_{0,i,3} + \lambda_{1,i,k,3}(\alpha_{c,k}q_{i,k})] + \\ exp[\lambda_{0,i,4} + \lambda_{1,i,k,4}(\alpha_{c,k}q_{i,k})] \end{array}}. \quad (5)$$

If a test has 40 such items, we need to estimate 40 (items) × 5 (response options) × 2 (intercept and main effect) = 400 parameters under the NRDM, which is a lot. To address that problem, (Liu & Jiang, 2018, 2019) proposed three smaller DCMs for ordinal item responses: the rating scale diagnostic model (RSDM), the ordinal response diagnostic model (ORDM), and the modified ordinal response diagnostic model (MORDM). These three models are constrained versions of the NRDM with fewer parameters that need to be freely estimated.

The RSDM reduces the number of parameters through constraining the parameters of the same response option across items measuring the same attribute to be the same. The relationship between the NRDM and the RSDM in the DCM context is analogous to that between the nominal response model (NRM; Bock, 1972) and the rating scale model (RSM; Andrich, 1978) in the IRT context. For simple-structure items, the original RSDM can be simplified as:

$$P(X_i = m|\boldsymbol{\alpha}_c) = \frac{exp\left[\lambda_{0,i} + \lambda_{0,m,k}q_{i,k} + \left(\lambda_{1,i} + \lambda_{1,m,k}q_{i,k}\right)\alpha_{c,k}q_{i,k}\right]}{\sum_{m=0}^{M-1} exp\left[\lambda_{0,i} + \lambda_{0,m,k}q_{i,k} + \left(\lambda_{1,i} + \lambda_{1,m,k}q_{i,k}\right)\alpha_{c,k}q_{i,k}\right]}. \quad (6)$$

Comparing Equation 6 to Equation 4 shows that the intercept parameter for response option m on item i: $\lambda_{0,i,m}$ is broken down into a parameter shared across all response options of item i: $\lambda_{0,i}$ and a parameter for response option m shared across all items measuring attribute k: $\lambda_{0,m,k}$. The same breakdown also applies to the main effect parameters. As we did with the NRDM in Equation 5, we rewrite the RSDM in Equation 6 with respect to the probability of selecting response option 3 on an item with five response options as:

$$P(X_i = 3|\boldsymbol{\alpha}_c) = \frac{exp\left[\lambda_{0,i} + \lambda_{0,3,k}q_{i,k} + (\lambda_{1,i} + \lambda_{1,3,k}q_{i,k})\alpha_{c,k}q_{i,k}\right]}{\begin{array}{c} exp\left[\lambda_{0,i} + \lambda_{0,0,k}q_{i,k} + (\lambda_{1,i} + \lambda_{1,0,k}q_{i,k})\alpha_{c,k}q_{i,k}\right] + \\ exp\left[\lambda_{0,i} + \lambda_{0,1,k}q_{i,k} + (\lambda_{1,i} + \lambda_{1,1,k}q_{i,k})\alpha_{c,k}q_{i,k}\right] + \\ exp\left[\lambda_{0,i} + \lambda_{0,2,k}q_{i,k} + (\lambda_{1,i} + \lambda_{1,2,k}q_{i,k})\alpha_{c,k}q_{i,k}\right] + \\ exp\left[\lambda_{0,i} + \lambda_{0,3,k}q_{i,k} + (\lambda_{1,i} + \lambda_{1,3,k}q_{i,k})\alpha_{c,k}q_{i,k}\right] + \\ exp\left[\lambda_{0,i} + \lambda_{0,4,k}q_{i,k} + (\lambda_{1,i} + \lambda_{1,4,k}q_{i,k})\alpha_{c,k}q_{i,k}\right] \end{array}}. \quad (7)$$

If the above 40-item example is about measuring four attributes, we only need to estimate [40 (item-level parameters) + 4 (attributes) × 5(response options)] × 2 (intercept and main effect) =120 parameters under the RSDM, a significant reduction from 400.

Different from the RSDM, the ORDM constrains only the main effect parameters in the NRDM, through an approach analogous to constraining the NRM to arrive at the generalized partial credit model (GPCM; Muraki, 1992) in the IRT context. For simple-structure items, the original ORDM can be simplified as:

$$P(X_i = m|\boldsymbol{\alpha}_c) = \frac{exp\sum_{m=0}^{m}\left[\lambda_{0,i,m} + \lambda_{1,i,k}\left(\alpha_{c,k}q_{i,k}\right)\right]}{\sum_{s}^{M-1} exp\sum_{m=0}^{s}\left[\lambda_{0,i,m} + \lambda_{1,i,k}\left(\alpha_{c,k}q_{i,k}\right)\right]}, \quad (8)$$

Comparing Equation 8 to the simplified NRDM in Equation 4 shows that the main effect parameter for response option m on item i: $\lambda_{1,i,k,m}$ loses subscript m and it is replaced by $(m+1)\times\lambda_{1,i,k}$. As we did with the NRDM and the RSDM, we rewrite the ORDM in Equation 8 with respect to the probability of selecting response option 3 on an item with five response options as:

$$P(X_i=3|\alpha_c)=\frac{exp\left[\lambda_{0,i,0}+\lambda_{0,i,1}+\lambda_{0,i,2}+\lambda_{0,i,3}+4\times\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})\right]}{\begin{array}{c}exp\left[\lambda_{0,i,0}+\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})\right]+\\ exp\left[\lambda_{0,i,0}+\lambda_{0,i,1}+2\times\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})\right]+\\ exp\left[\lambda_{0,i,0}+\lambda_{0,i,1}+\lambda_{0,i,2}+3\times\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})\right]+\\ exp\left[\lambda_{0,i,0}+\lambda_{0,i,1}+\lambda_{0,i,2}+\lambda_{0,i,3}+4\times\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})\right]+\\ exp\left[\lambda_{0,i,0}+\lambda_{0,i,1}+\lambda_{0,i,2}+\lambda_{0,i,3}+\lambda_{0,i,4}+5\times\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})\right]\end{array}}. \tag{9}$$

For the 40-item example discussed above, we only need to estimate [40 (items/intercept) × 5 (response options/intercept)] + 40 (main effect) = 240 parameters under the ORDM.

The MORDM is a combination of the RSDM and the ORDM, requiring the smallest number of parameters among all current polytomous DCMs. For simple-structure items, the original MORDM can be simplified as:

$$P(X_i=m|\boldsymbol{\alpha}_c)=\frac{exp\sum_{m=0}^{m}\left[\lambda_{0,i}+\lambda_{0,m,k}q_{i,k}+\lambda_{1,i,k}\left(\alpha_{c,k}q_{i,k}\right)\right]}{\sum_{s}^{M-1}exp\sum_{m=0}^{s}\left[\lambda_{0,i}+\lambda_{0,m,k}q_{i,k}+\lambda_{1,i,k}\left(\alpha_{c,k}q_{i,k}\right)\right]}. \tag{10}$$

Comparing Equation 10 to the simplified ORDM in Equation 8 shows that the intercept parameter for response option m on item i: $\lambda_{0,i,m}$ is broken down to $\lambda_{0,i}$ and $\lambda_{0,m,k}$, just as what we did in Equation 6 with the RSDM. As we did with the NRDM, RSDM and the ORDM, we rewrite the MORDM in Equation 10 with respect to the probability of selecting response option 3 on an item with five response options as:

$$P(X_i=3|\boldsymbol{\alpha}_c)=\frac{exp\{\lambda_{0,0,k}q_{i,k}+\lambda_{0,1,k}q_{i,k}+\lambda_{0,2,k}q_{i,k}+\lambda_{0,3,k}q_{i,k}+4\times[\lambda_{0,i}+\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})]\}}{\begin{array}{c}exp\left[\lambda_{0,0,k}q_{i,k}+\lambda_{0,i}+\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})\right]+\\ exp\{\lambda_{0,0,k}q_{i,k}+\lambda_{0,1,k}q_{i,k}+2\times[\lambda_{0,i}+\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})]\}+\\ exp\{\lambda_{0,0,k}q_{i,k}+\lambda_{0,1,k}q_{i,k}+\lambda_{0,2,k}q_{i,k}+3\times[\lambda_{0,i}+\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})]\}+\\ exp\{\lambda_{0,0,k}q_{i,k}+\lambda_{0,1,k}q_{i,k}+\lambda_{0,2,k}q_{i,k}+\lambda_{0,3,k}q_{i,k}+4\times[\lambda_{0,i}+\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})]\}+\\ exp\{\lambda_{0,0,k}q_{i,k}+\lambda_{0,1,k}q_{i,k}+\lambda_{0,2,k}q_{i,k}+\lambda_{0,3,k}q_{i,k}+\lambda_{0,4,k}q_{i,k}+5\times[\lambda_{0,i}+\lambda_{1,i,k}(\alpha_{c,k}q_{i,k})]\}\end{array}}. \tag{11}$$

For the 40-item example discussed above, we only need to estimate 40 (items) × 2 (item-level intercept and main effect) + 4 (attributes/intercept) × 5 (response options/intercept) = 100 parameters under the MORDM.

To sum up, the RSDM, the ORDM, and the MORDM are constrained versions of the NRDM. Although it is tempting to figure out which model is better, it would be reckless to provide a blanket statement. Instead, one could compare model fit indices for a particular dataset and choose a parsimonious model that fits adequately. In addition to the four polytomous DCMs, other polytomous DCMs have been proposed in the literature. A brief review of those models

Table 1 ■ The **Q**-matrix for the 16Personalities Test

	α_1	α_2	α_3	α_4	α_5
Item	(Mind)	(Energy)	(Nature)	(Tactics)	(Identity)
1	1	0	0	0	0
2	1	0	0	0	0
...					
12	1	0	0	0	0
13	0	1	0	0	0
14	0	1	0	0	0
...					
24	0	1	0	0	0
25	0	0	1	0	0
26	0	0	1	0	0
...					
36	0	0	1	0	0
37	0	0	0	1	0
38	0	0	0	1	0
...					
48	0	0	0	1	0
49	0	0	0	0	1
50	0	0	0	0	1
...					
60	0	0	0	0	1

can be found in Liu and Jiang (2018, 2019).

To implement DCMs, one could use Bayesian software programs such as JAGS (Plummer, 2003) or Stan (Carpenter et al., 2017) and program from scratch. One could also use non-Bayesian-specific programs such as the "CDM" R package (Robitzsch, Kiefer, George, & Uenlue, 2019), the "GDINA" R package (Ma & de la Torre, 2019), FlexMIRT (Cai, 2017), Latent Gold (Vermunt & Magidson, 2016), and Mplus (Muthén & Muthén, 2019). Among the five, the first two R packages are free, require light programming, and provide informative outputs in an easy way. A review of those two packages can be found in Rupp and van Rijn (2018).

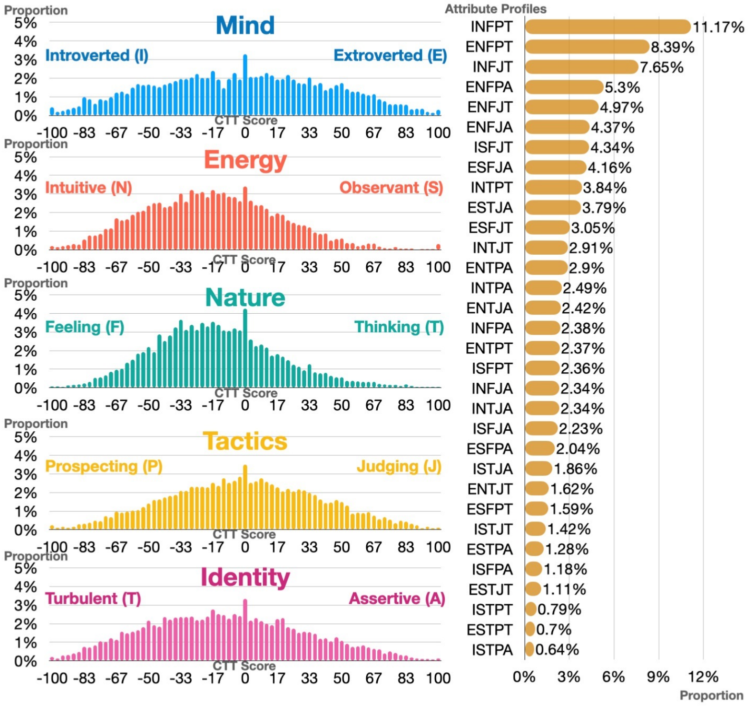
Scoring an Operational Psychological Test: CTT, IRT, and DCM

In this section, we examine how DCM results are different from results of CTT and IRT scoring approaches. To do that, we apply the three approaches to an operational personality test dataset.

The data used in this study is provided by NERIS Analytics Limited in London, United Kingdom. It contains $N = 10,000$ examinees' item responses on a personality test called "16Personalities" (https://www.16personalities.com/). The theoretical framework behind this test is the "Big five factor model" (also known as the "OCEAN" model; Rothmann & Coetzer, 2003) where individuals are characterized according to five factors: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (Goldberg, 1993). Those five factors are described as five binary attributes in 16Personalities: mind (extroverted/introverted), energy (observant/intuitive), nature (thinking/feeling), tactics (judging/prospecting) and identity (assertive/turbulent). Details about the five attributes can be found at https://www.16personalities.com/articles/our-theory.

In order to measure those five attributes, the test consists of 60 items where each attribute is measured independently by 12 items. The item-attribute relationship is specified in a 60-by-5 Q-matrix displayed in Table 1. The 1s and 0s in Table 1 are assigned following the sequence of the two conditions of each attribute mentioned in the paragraph above. For example, $\alpha_1 = 1$ and 0 represent "extroverted" and "introverted", respectively. Each item has seven response options ranging from "disagree" to "agree" which represent the degree that the respondent endorses the item stem.

Since model comparison with DCMs is not the interest of this study, we only fit one model: the NRDM - the most general polytomous DCM to the dataset. In practice, one could fit more than one model (e.g., the RSDM and the ORDM) and select a final candidate based on model fit indices and the principle of parsimony. For CTT, we computed the summated scores for each attribute and classified examinees based on the original test design. For IRT, Figure 1 ■ Distributions of examinee scores and classifications under the CTT framework.



we fit a monotonic polynomial generalized partial credit model with the Q-matrix specification (Falk & Cai, 2016; da Silva, Liu, Huggins-Manley, & Bazán, 2018; Reckase, 2009) and obtained five latent scores for each examinee, one for each attribute.

CTT: The Original Scoring Framework

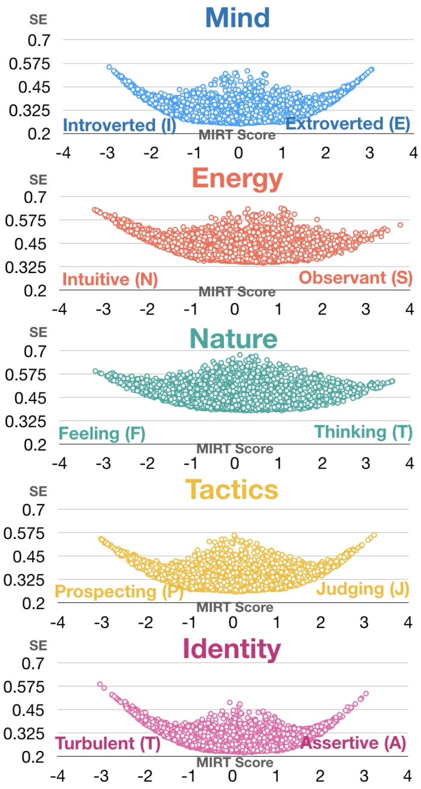
Originally, the test was developed and scored under the CTT framework where each item is scored between -3 and 3 for the seven response options. As a result, the raw score range for each attribute is $\pm 3 \times 12$ = [-36, 36]. In the original scoring framework, examinees' raw scores are multiplied by 2.78 to create a score range between -100 and 100. An examinee with a subscore greater than 0 is categorized to the $\alpha = 1$ category on each attribute. For example, an examinee with a raw score of 20 on α_1: "mind" has a reported score of 53.4 and is classified into the "extroverted" category.

Examinees' scores and attribute profiles under the CTT framework are displayed in Figure 1. Five by-attribute histograms are on the left-hand side, examinee scores are on the x-axis and the examinee proportions are on the y-axis. The attribute profile bar chart is on the right-hand side. Overall, the scores for each attribute seem normally distributed. The mean/median values for "mind" and "tactics" were close to 0, while those for "energy", "nature", and "identity" were slightly below 0. This echoes with the bar chart which shows that "INFP-T", "ENFP-T", "INFJ-T" were the most frequent attribute profiles in this dataset.

MIRT-Q: Leaping into the Latent World

Shifting gears from CTT to IRT or DCM acknowledges that attributes are latent and can only be approximated through observable indicators. We fit a multidimensional generalized partial credit model with the Q-matrix specification using the "mirt" R package (Chalmers, 2012). The model was able to converge at the threshold of 10^{-6} for maximum parameter change and showed marginal fit to the data according to absolute fit indices. For example, the root mean square error of approximation (RMSEA) value was 0.063 and the standardized root mean squared residual (SRMSR) value was 0.098. However, the estimates of examinees' latent traits had large standard errors. As shown in Figure 2, most standard errors were between 0.3 and 0.6, despite the large sample size of the dataset (i.e., $N = 10,000$). Among the five attributes, "energy" and "nature" seem to have larger standard errors for their parameter estimates than the other three attributes.

Figure 2 ■ The estimates and standard errors of examinee scores under the MIRT framework.

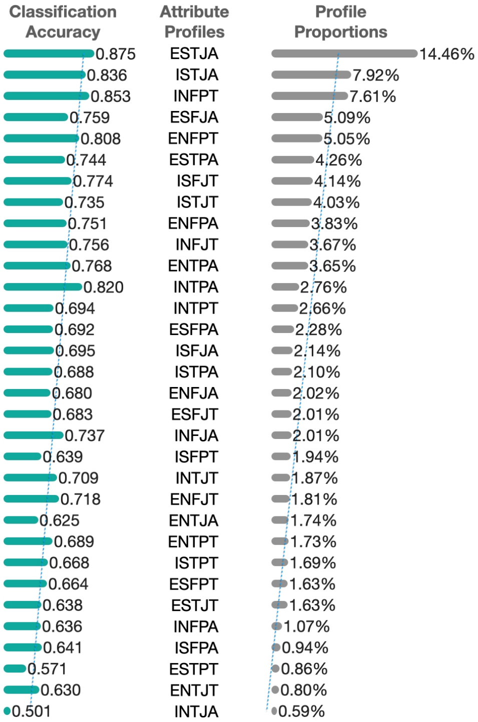


DCM: Categorical Latent Traits

One of the main differences between MIRT and DCM is that the latent traits are assumed to be continuous for the former and categorical for the latter. We fit the NRDM to the dataset using the "GDINA" R package. Whenever a DCM is fit to a dataset, one could either put a monotonicity constraint on the model or not. Under a simple-structure NRDM, a monotonicity constraint means that: 1) when examinees possess an attribute associated with a certain item, the probability of them selecting a higher response category on that item should be greater than or at least be equal to the probability of selecting its adjacent lower response category; and 2) when examinees don't possess an attribute associated with a certain item, the probability of them selecting a higher response category on that item should be smaller than the probability of selecting its adjacent lower response category. de la Torre and Sorrel (2017) demonstrated that imposing monotonicity constraints improves classification accuracy. Recent DCM studies (e.g. Ma, 2019a, 2019b) also impose the monotonicity constraints in model-fitting. In this study, we fit models both with and without the monotonicity constraints. Results showed that the NRDM with the monotonicity constraints produced better fit and higher classification accuracy. Therefore, the analysis continued with results from the NRDM with the monotonicity constraints.

The NRDM showed marginal fit to the dataset with RMSEA = 0.060 and SRMSR = 0.085. A comparison of the results between the NRDM and the MIRT model is discussed in the next section. For now, let us look at two pieces of information from the DCM scores displayed in Figure 3: 1) proportions of examinees that were classified with each attribute profile, and 2) classification accuracy associated with each attribute profile. For the first piece of information, "11111", "01111", and "00000" were the most frequently occuring profiles. Specifically, 30% of examinees were classified with one of those three profiles. This echoes with the "U-shape" curve problem mentioned in a previous section, meaning that most people were classified into none-possession ("00000") or all-possession ("11111") categories. Intuitively, some examinees may be misclassified. However, we don't know examinees' "true" attribute profiles. We can only evaluate whether the classifications are reliable. In DCM, we may conceptualize reliability as classification accuracy (Wang, Song, Chen, Meng, & Ding, 2015;

Figure 3 ■ Proportions and classification accuracy of examinee attribute profiles under the DCM framework.



Iaconangelo, 2017). Wang et al. (2015) computed the classification accuracy as the agreement between the observed examinee classifications (using estimation methods such as the maximum or expected a posterior) and the expected examinee classifications (using their individual likelihood functions). The general trend in Figure 3 shows that attribute profiles with more examinees were associated with higher classification accuracy. The profile "11111", containing 14.46% of the examinees, was associated with the highest classification accuracy (i.e., 0.875) among all attribute profiles. At the attribute-level, the classification accuracy was high: 0.960, 0.949, 0.924, 0.942, and 0.954, for each attribute respectively. The test-level classification accuracy was 0.763.

Examining the Results Together

The three measurement frameworks that can be utilized in psychological tests are simply different. One is not better than the other. Theoretically, the major difference between CTT and MIRT-Q/DCM is whether we consider the measured traits as observed or latent variables. The major difference between MIRT-Q and DCM is whether we conceptualize the measured traits as continuous or categorical variables. The purpose of examining the results from CTT, MIRT-Q and DCM together is to better understand the differences between them that are manifested empirically.

Figure 4 displays the relationship between CTT and DCM scores. The five by-attribute scatter plots on the left-hand side describes the relationship between an examinee's summated subscore under the CTT framework (on the x-axis) and the marginal probability of possessing an attribute under the DCM framework (on the y-axis). A high probability (close to 1) on the y-axis represents that we are more certain that an examinee is in the $\alpha = 1$ category, while a low probability (close to 0) represents that we are more certain that an examinee is in the $\alpha = 0$ category. Probabilities close to 0.5, meaning that we have less certainty, are not ideal. The general trend in the scatter plots shows an S-shaped curve where examinees who had lower summated scores were more likely to be classified into the $\alpha = 0$ category and examinees who had higher summated scores were more likely to be classified into the $\alpha = 1$ category. However, the S-shapes were not perfect in those five scatter plots. In a perfect relationship between CTT and DCM scores, the S-shape is a thin curved line. The five scatter plots all had wide body part in the center of the S-shape (i.e., scores that are close to 0 on the x-axis). Among the five, the first attribute: "mind" had the thinnest body, while "energy" and "nature" had the widest body. In addition, the centers of the body for "energy", "nature", and "iden-

Figure 4 ■ Relationship between scores under CTT and DCM frameworks.

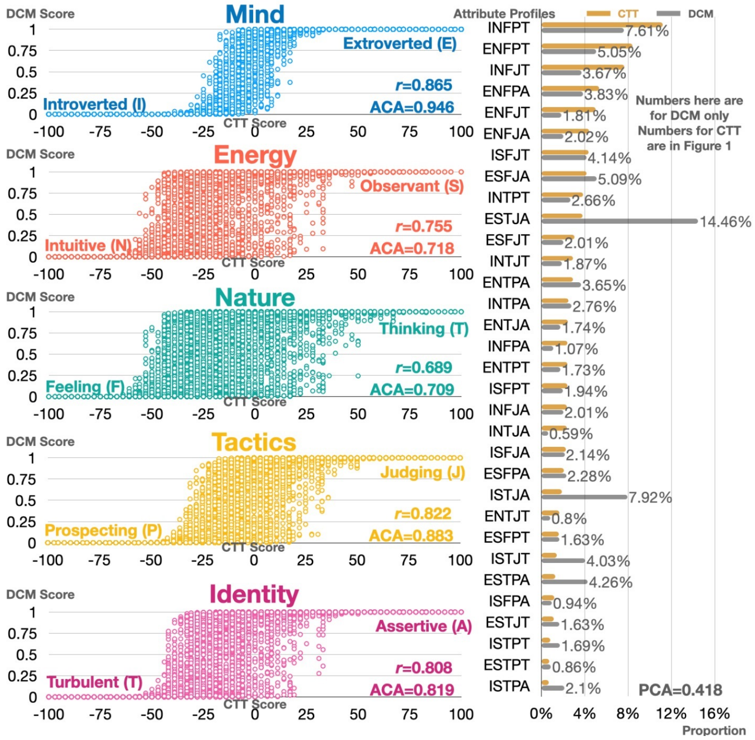


Table 2 ■ Model fit Information for the MIRT Model and the DCM

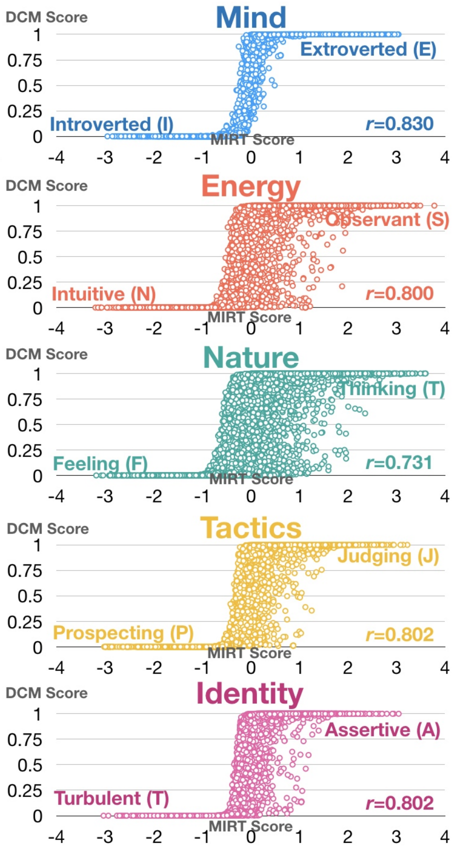
Item	RMSEA	SRMSR	AIC	BIC
MIRT	0.063	0.098	2080529	2083629
DCM	0.060	0.085	2117940	2123355

tity” were around $x = -20$, while those of the other two scatter plots were around $x = 0$. This matches the location of the peak of the raw score distribution of each attribute displayed in Figure 1.

In order to quantify such relationship at the continuous level, we computed: 1) Pearson’s correlation coefficient (r) which assumes a linear relationship between two variables; and 2) the coefficient for a simple logistic regression between CTT scores and DCM classifications. Although the latter seemed to be a more appropriate way to model the S-shaped relationship, it missed the purpose of examining the relationship between two continuous variables. Moreover, the marginal probabilities may seem close to 0 in the graphs, but they were still continuous variables that took on a particular non-zero value. Therefore, we present the r at the bottom-right corner of each scatter plot. As expected, the correlation was highest for “mind” ($r = 0.865$) and lowest for “energy” ($r = 0.755$) and “nature” ($r = 0.689$). In addition to quantifying the relationship at the continuous level, we computed the attribute classification agreement (ACA) and profile classification agreement (PCA) rates between CTT and DCM classifications. “Mind” had the largest ACA where 94.6% of examinees were classified into the same category. “Energy” and “nature” had lower ACA where around 71% of examinees were classified into the same category. On the right-hand side of Figure 4, attribute profile proportions are displayed for DCM and CTT in the same order as that in Figure 1. If we use the CTT classifications as a benchmark, DCM “over-classified” examinees into mostly four profiles: “11111”, “01111”, “01110”, “11101”. In addition to the “U-shape” issue mentioned previously, it seems that more examinees are classified into the “1” category on “energy”, “nature”, and “identity” under the DCM. Overall, 41.8% of examinees were classified into the same attribute profiles.

Table 2 compares the absolute and relative model fit

Figure 5 ■ Relationship between scores under MIRT and DCM frameworks.



information between MIRT and DCM. DCM fit the data slightly better in terms of absolute model fit (i.e., RMSEA and SRMSR) and MIRT fit the data slightly better in terms of relative model fit (i.e., AIC and BIC). However, those differences were small. Another thing to consider is that, DCM estimated 751 parameters (including 720 item parameters and 31 structural parameters), almost twice as the number of parameters associated with the MIRT estimation: 420.

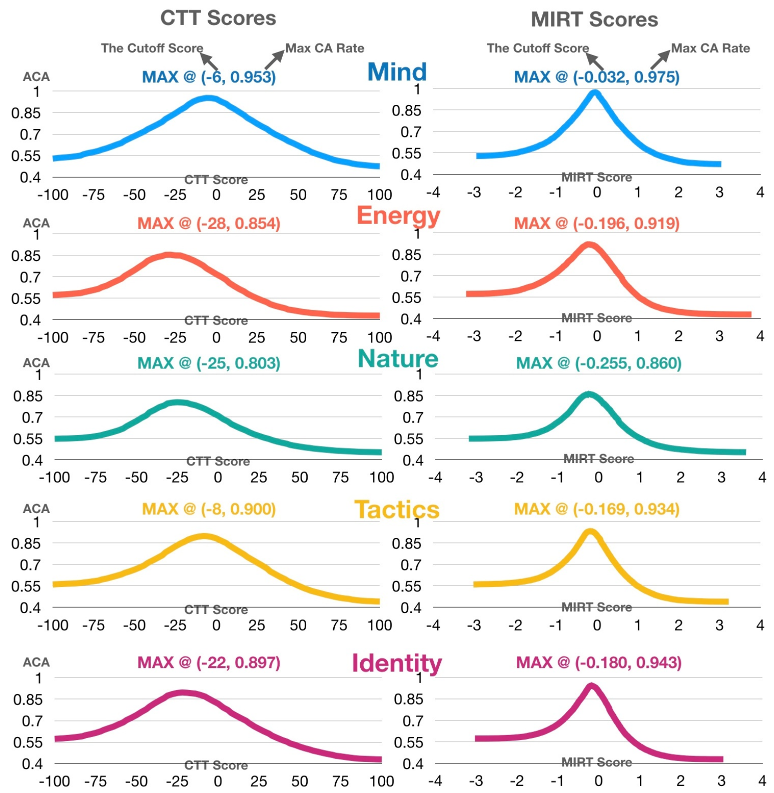
Figure 5 displays the relationship between MIRT and DCM scores. Comparing to Figure 4, we can see that the body part of the scatter plots is thinner in each graph. Moreover, the center of the dots is around $x = 0$ for each attribute, regardless of the raw score distribution. Pearson's r was calculated between the MIRT and DCM scores for each attribute. This information is listed at the bottom-right corner of each by-attribute graph. Surprisingly, on three attributes: "mind", "tactics", and "identity", the correlations between MIRT and DCM scores were lower than those between CTT and DCM scores.

As a final set of comparisons, Figure 6 displays the ACA rates between the DCM and different score cutoffs on the CTT and MIRT score scales. The peak of each curve represents the maximum ACA and its associated CTT or MIRT scores. Overall, the ACA between MIRT and DCM could be higher than that between CTT and DCM when appropriate cutoffs were chosen. The possible maximum ACA rates between CTT and DCM were between 0.80 and 0.95 across the five attributes, while those numbers were between 0.86 and 0.98 for the relationship between MIRT and DCM. The cutoffs associated with the maximum ACAs on the MIRT scales were close to 0. In CTT, the cutoffs associated with the maximum ACAs were closer to 0 for "mind" and "tactics" and further from 0 for the other three attributes. In addition, the MIRT curves were more peaked than CTT curves because individuals were more clustered around the peak under the MIRT.

Score Reporting

Reporting scores to end users may be one of the most important outcomes of psychometric practices. Let us use

Figure 6 ■ Classification agreement rates between DCM classifications and different levels of CTT/MIRT score cutoffs.



scores from examinees A and B as an example to illustrate the different scores that examinees may obtain under different scoring frameworks. Both examinees A and B were classified into the "ESTJA" group (i.e., profile "11111") under the DCM framework. However, under the CTT framework, examinee A was classified into the "ENFJT" group (i.e., profile "10010") due to scores of 11, -5, -4, 7, and -2, while examinee B was still classified into the "ESTJA" group (i.e., profile "11111") due to scores of 83, 70, 67, 74, and 79. We can see that examinees' classifications between the CTT and DCM frameworks were more likely to agree with each other when their scores were further away from the CTT cut-offs (i.e., more extreme on the latent traits). Providing examinee B with classifications from both frameworks may be fine, but examinee A would be confused if they were provided with different classifications from the two frameworks, with one mainly based on content experts' judgment (under CTT) and the other mainly based on statistical probabilities (under DCM). Although one could use approaches such as the relative diagnostic profile (Liu, Qian, Luo, & Woo, 2018) to reconcile both scores, it would be important to use a principled assessment design framework during test development and be consistent with one scoring model that best meets the purpose of the test.

Discussion

> *All models are wrong, and the value of any model is only to the extent to which it supports the purpose for which it was built.*
> —George E. P. Box (Box, 1979)

> *Models should not be true, but it is important that they are applicable, and whether they are applicable for any given purpose must of course be investigated. This also means that a model is never accepted finally, only on trial.*
> —George Rasch (Rasch, 1960)

These two quotes from decades ago still shed light on today's psychometrics. They remind us, that the models we develop to fit the data represent our limited and simplified theory of the construct. The development of IRT models and DCMs in recent years offers a rich pool of tools to analyze examinees' item responses. Comparing to tools in the world (of CTT) without latent construct, the tools in the IRT/DCM world offer unparallel vantages such as bringing more information about items and examinees and easier equating of scores between test forms. However, IRT/DCM are not approaches that are superior than CTT. They are mathematically comparable, as demonstrated in Kohli, Koran, and Henn (2015), Raykov and Marcoulides (2016), and Takane and De Leeuw (1987). The fundamental difference between the three frameworks are different theories of the existence/non-existence of latent variables and whether the underlying variable is assumed to be normal or categorical. As such, the choice of the scoring approach probably should hinge on the intended use of the test.

For psychological tests that bear the purpose of diagnosing or classifying individuals with certain behaviors, DCMs are promising candidates for scoring because they are created for classifying examinees into groups. However, it is necessary to point out two caveats here when applying DCMs in practice. First, for the dataset used in this paper, a MIRT model with half the number of parameters of a DCM fit the dataset similar to the DCM. This issue may be partially alleviated through using smaller models such as the RSDM introduced above. But as multidimensional models, DCMs will have many parameters which require a relatively large sample size to estimate. Second, standard-setting results by expert panel may or may not agree with the classification decisions made by DCMs. If we arbitrarily set a cut-off other than 0.5 for the marginal probability of possession and use the new cut-off to reclassify individuals, the meaning of the latent classes is altered. As a result, the concept of the criterion-referenced tests may be not applicable if the test data is scored under DCMs. This could raise issues of how to interpret and use the DCM classifications. For example, an individual in the dataset with an observed score of -51 on “energy” was classified as “intuitive” (because the score is less than 0) under CTT, but as “observant” ($\alpha = 1$) under a DCM. Although it is understandable that the scores are under different frameworks, one may find it difficult to justify why an individual who consistently endorses the more “intuitive” side of the “energy” items ends up being an “observant”. In this dataset, the DCM classification cut-off (i.e., the marginal probability of 0.5) for each attribute closely aligns with the center of the raw score distribution of each attribute. Therefore, for a sample of scores where the mean is not around the center of the observed scale (e.g., examinee scores on “energy”, “nature” and “identity”), the interpretation of DCM classifications may be difficult when comparing to the observed score scale. Despite the caveats, DCMs provide exciting alternatives to traditional psychometrics which mostly aim to order examinees on a continuum. Nowadays, as researchers and practitioners are becoming more interested in obtaining actionable feedback on multiple characteristics of people from rating scales, DCMs are expected to play a greater role in psychological measurement.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. doi:10.1007/BF02291411

Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics* (pp. 201–236). Washington: Academic Press.

Cai, L. (2017). Flexmirt: Flexible multilevel multidimensional item analysis and test scoring [computer software] (Version 3.51). Chapel Hill, NC: Vector Psychometric Group.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1), 1–38.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*(6), 1–29.

da Silva, M. A., Liu, R., Huggins-Manley, A. C., & Bazán, J. L. (2018). Incorporating the q-matrix into multidimensional item response theory models. *Educational and Psychological Measurement*, *79*(4), 665–687.

Dayton, C. M., & Macready, G. B. (1976). A probabilistic model for validation of behavioral hierarchies. *Psychometrika*, *41*(2), 189–204.

de la Torre, J., & Sorrel, M. A. (2017). *Attribute classification accuracy improvement: Monotonicity constraints on the g-dina model*. Switzerland: In The International Meeting of the Psychometric Society. Zurich.

Falk, C. F., & Cai, L. (2016). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika*, *81*, 434–460.

George, A. C., & Robitzsch, A. (2015). Cognitive diagnosis models in r: A didactic. *The Quantitative Methods for Psychology*, *11*(3), 189–205.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American psychologist*, *48*(1), 26–29.

Gu, Y., & Xu, G. (2018). The sufficient and necessary condition for the identifiability and estimability of the dina model. *Psychometrika*, *111*, 1–16.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301–321.

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*(2), 191–210.

Iaconangelo, C. (2017). *Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models. (unpublished doctoral dissertation)*. New Brunswick, NJ: Rutgers University.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272.

Kohli, N., Koran, J., & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and psychological measurement*, *75*(3), 389–405.

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from irt-based assessment forms. *Educational and psychological measurement*, *78*(3), 357–383.

Liu, R., & Jiang, Z. (2018). Diagnostic classification models for ordinal item responses. *Frontiers in psychology*, *9*, 1–100.

Liu, R., & Jiang, Z. (2019). A general diagnostic classification model for rating scales. *Behavior research methods*, *44*, 1–18.

Liu, R., Qian, H., Luo, X., & Woo, A. (2018). Relative diagnostic profile: A subscore reporting framework. *Educational and Psychological Measurement*, *78*(6), 1072–1088.

Ma, W. (2019a). A diagnostic tree model for polytomous responses with multiple strategies. *British Journal of Mathematical and Statistical Psychology*, *72*(1), 61–82.

Ma, W. (2019b). Evaluating the fit of sequential g-dina model using limited-information measures. *Applied Psychological Measurement*, *44*(3), 167–181.

Ma, W., & de la Torre, J. (2019). Gdina: The generalized dina model framework (Version 2.5). Retrieved from http://CRAN.R-project.org/package=GDINA

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

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muthén, L. K., & Muthén, B. O. (2019). *Mplus (version 8.3) [computer software]*. Los Angeles, CA: Muthén & Muthén.

Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (2003). *Mbti manual: A guide to the development and use of the myers-briggs type indicator, 3rd*. Palo Alto, CA: Consulting Psychologists Press.

Plummer, M. ((2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *3rd international workshop on distributed statistical computing* (Vol. 124, p. 125).

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and psychological measurement*, *76*(2), 325–338.

Reckase, M. (2009). *Multidimensional item response theory*. Berlin: Springer.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2019). Cdm: Cognitive diagnosis modeling (Version 7.3). Retrieved from http://CRAN.R-project.org/package=CDM

Rothmann, S., & Coetzer, E. P. (2003). The big five personality dimensions and job performance. *SA Journal of Industrial Psychology*, *29*(1), 68–74.

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

Rupp, A. A., & van Rijn, P. W. (2018). Gdina and cdm packages in r. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 71–77.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.

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

Templin, J. L., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275.

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

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

Templin, J. L., Henson, R. A., Rupp, A. A., Jang, E., & Ahmed, M. (2008). Cognitive diagnosis models for nominal response data. In *Annual meeting of the National Council on Measurement in Education*, New York.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577.

Vermunt, J. K., & Magidson, J. (2016). *Technical guide for latent gold 5.1: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations Inc.

von Davier, M. (2018). Diagnosing diagnostic models: From von neumann's elephant to model equivalencies and network psychometrics. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 59–70.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*, 457–476.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, *81*(3), 625–649.

Citation

Liu, R., & Shi, D. (2020). Using diagnostic classification models in psychological rating scales. *The Quantitative Methods for Psychology*, *16*(5), 442–456. doi:10.20982/tqmp.16.5.p442

Received: 05/07/2020 ∼ Accepted: 30/09/2020