Analysis of clinical data from cognitive diagnosis modeling framework

de la Torre, J.; van der Ark, L.A.; Rossi, G.

DOI
10.1080/07481756.2017.1327286
Publication date
2018
Document Version
Final published version
Published in
Measurement and Evaluation in Counseling and Development

MEASUREMENT AND EVALUATION IN COUNSELING AND DEVELOPMENT
2018, VOL. 51, NO. 4, 281–296
https://doi.org/10.1080/07481756.2017.1327286

Routledge
Taylor & Francis Group

ASSESSMENT, DEVELOPMENT, AND VALIDATION

Analysis of Clinical Data From a Cognitive Diagnosis Modeling Framework

Jimmy de la Torre[a], L. Andries van der Ark[b], and Gina Rossi[c]

[a]Rutgers, The State University of New Jersey, New Brunswick, NJ, USA; [b]University of Amsterdam, Amsterdam, Netherlands; [c]Vrije Universiteit Brussel (VUB), Brussels, Belgium

ABSTRACT

We propose a general cognitive diagnosis model framework to diagnose mental disorders using item scores obtained from clinical measurement instruments. This framework can be used to validate the extent to which the items measure the specific disorders. The method is illustrated using data obtained with the Dutch version of Millon Clinical Multiaxial Inventory-III.

KEYWORDS

Clinical assessment; Millon Clinical Multiaxial Inventory-III; latent variables; confirmatory factor analysis; multidimensional; exploratory factor analysis

Introduction

The primary goal of studies investigating mental disorders, such as personality and clinical disorders, is to accurately diagnose the respondents' true states. The fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*; American Psychiatric Association [APA], 2013) maintained a categorical classification of separate mental disorders, although they recognize the problematically high comorbidity patterns between categories. Mental disorders do not always fit into the boundaries of a single category, there is high heterogeneity within categories and some symptom domains involve multiple diagnostic categories. Symptoms such as antisocial personality disorder and substance abuse, or anxiety and depression, and so on probably share common underlying vulnerabilities for a larger group of disorders. Clinical reality, for example, points out an extremely high comorbidity of major depression and anxiety disorders. Brown, Campbell, Lehman, Grisham, and Mancill (2001) found that 95% of individuals with major depression (or dysthymia) also had a current or past anxiety disorder. This comorbidity pattern can be understood from the internalizing spectrum (unipolar mood, anxiety, and somatoform disorders), that together with the externalizing spectrum (antisocial and substance abuse disorders) seem to form a superordinate structure (Krueger & Eaton, 2010) for the former *DSM-IV* (APA, 2000) Axis I and Axis II disorders. *DSM-5* has indeed removed the artificial splitting into Axis I, II, and III, however considered it premature to propose alternative definitions for most disorders. With the *DSM-5* models of personality and clinical disorders placed on one Axis, hierarchical models of psychopathology and associated comorbidity patterns, are inherently a part of the necessary further discussion of changes to the nosology of mental disorders in terms of structural organization and the possible subsuming of certain categories into one diagnostic group.

For diagnosing purposes, a critical first step is to identify an appropriate instrument that measures the disorders of interest. These instruments require cut scores that validly differentiate respondents who have the disorders from those who do not. Examples of instruments that have been used for this end are *Diagnostic and Statistical Manual* (*DSM-IV-TR*; APA, 2000) based interviews, such as the Mini-International Neuropsychiatric Interview (MINI; Sheehan et al., 1998), or self-report measures such as the Minnesota Multiphasic Personality Inventory–2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 2001), the Millon Clinical Multiaxial Inventory–III (MCMI-III; Millon, Millon, Davis, & Grossman, 2009), and

CONTACT Jimmy de la Torre j.delatorre@hku.hk The University of Hong Kong, Pokfulam Road, Hong Kong.

the Questionnaire on Personality Traits (usually referred to by the acronym VKP; Duijsens, Eurelings-Bontekoe, & Diekstra, 1996). It is typical for these instruments to contain items that simultaneously measure multiple disorders. Using the same item for multiple scales is a logical choice because disorders are often comorbid and the same behavior or symptom can be an indication for several disorders (e.g., Dolan-Sewell, Krueger, & Shea, 2001; Krueger & Eaton, 2010). Despite potential complications resulting from improper treatment of item overlap such as linear dependency (Pearson, 1897; also see Budescu & Rodgers, 1981; Cohen & Farley, 1979; Hsu, 1992; Rossi, van der Ark, & Sloore, 2007), several authors (e.g., Dahl, 1986; Modestin, 1987; Nurnberg et al., 1991; Nurnberg, Hurt, Feldman & Suth, 1987; Widiger, Hurt, Frances, Clarkin, & Gilmore, 1984) have argued that such items are necessary to maintain content validity. Traditional methods of deriving scale scores involve computing simple or weighted sums. However, because these scores are obtained one scale at a time, the additional information that can be found in the comorbidity of the disorders is implicitly ignored in that responses to other scales are not taken into account. By using a multivariate version of Kelly's (1927) regressed scores, which takes the correlations among the scores into account, Wainer et al. (2001) showed that sum scores can be augmented to obtain more reliable scores. Consequently, as they are, simple sum scores are suboptimal.

In this article, we propose an alternative approach for diagnosing mental disorders. The approach is based on *cognitive diagnosis models* (CDMs; Junker & Sijtsma, 2001; also see de la Torre, 2009a, 2009b, for a didactic discussion of some specific CDMs). It should be noted that this is not the first time CDMs are applied to clinical data. Although most of the developments and applications of CDMs have taken place in educational contexts, these models have sufficient generality to be applied outside the confines of education. For example, CDMs have been used as a tool for providing diagnosis of psychological disorders (Jaeger, Tatsuoka, Berns, & Varadi, 2006; Templin & Henson, 2006). The current article differs from these previous works in that, in addition to fitting a more general CDM, we also investigated whether the general CDM can be replaced by CDMs with simpler formulations. To avoid reference to the underlying cognitive theory implied in the label CDM and for greater generality, some researchers have also referred to CDMs as *diagnostic classification models* (Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). For IRT-based models for cognitive diagnosis, we refer the interested reader to Dimitrov (2007), Dimitrov and Atanasov (2012), and Embretson (1984).

The use of CDMs for clinical data allows for the optimal use of information in diagnosis and can capture the interactions among the disorders. As an additional advantage, with this approach one can examine to which extent a symptom—as described by an item—will be observed given the various combinations of multiple disorders. This feature will be referred to as (item-level) *disorder interaction*. The qualifier item-level is included to stress that the same set of disorders may or may not exhibit interaction depending on the items that measure them and to differentiate this type of interaction from that which can occur when examining the relationship between the disorders.

The remaining sections of the article are structured as follows. The second section provides an introduction of CDMs, including model formulation, validation item-disorder specification, and saturated and reduced model comparison. The third section demonstrates how the approach can be used to diagnose disorders, examine accuracy of item-disorder association, and examine the disorder interaction using data obtained with the Dutch version of the MCMI-III (Rossi, Sloore, & Derksen, 2008). The MCMI-III has been indicated by clinicians as being one of the most frequently used self-report instruments for clinical assessment (Camara, Nathan, & Puente, 2000). For the purposes of this article, 44 items constituting the clinical scales *anxiety* (Scale A), *somatoform* (Scale H), *thought disorder* (Scale SS), and *major depression* (Scale CC) were used in the analysis. The item numbers as they are used in this article, and appear in the MCMI-III manual (Millon et al., 2009) are given in Appendix A. (For a complete item list and their descriptions, we refer the readers to pages 175–179 of the manual.) The probability that a respondent has a particular disorder obtained using CDM is also compared with the respondent's score on the scale measuring the same disorder. The last section discusses the advantages and the disadvantages of the proposed approach and its implications in both patient diagnosis and understanding the nature of disorders.

Cognitive Diagnosis Models

Model Formulation

CDMs are latent variable models developed primarily to provide inferences about the presence or absence of multiple attributes. The different attributes combine to form unique profiles. Traditionally, *attributes* is a generic term used in the educational setting to refer to skills, cognitive processes, or solution strategies. However, there is nothing inherent in the formulation of CDMs that prevents the meaning of the term from being broadened to include other constructs such as mental disorders. Thus, the presence and absence of an attribute can be construed as having and not having the disorder, respectively. For the remainder of this article, the word *disorder* is used in place of *attribute*. Even with a more inclusive meaning, the primary goal of CDM applications remains unchanged—to identify the respondents' disorder profiles based on their item responses.

Several CDMs that range in generality have been introduced in recent years (e.g., de la Torre, 2009a; Henson, Templin, & Willse, 2009; Junker & Sijtsma, 2001; von Davier, 2005). One such model is the *generalized deterministic, input, noisy, and gate* (G-DINA; de la Torre, 2011), a CDM with a general formulation. It should be noted that general CDMs based on different link functions (e.g., G-DINA model and log-linear CDM; Henson et al., 2009) are equivalent to each other in their saturated forms, and, thus, provide identical model-data fit. With appropriate constraints, several specific CDMs have been shown to be special cases of the G-DINA model. Using a general model obviates the need to identify the specific form of the CDM *a priori*. It also allows for the item-level interaction between the disorders to be explored. In addition, within the G-DINA model framework the fits of the full and reduced models can be statistically compared efficiently one item at a time to determine whether simpler models can be used in place of the general models.

Apart from the *saturated* (i.e., general) CDMs such as the G-DINA model, some *specific* (i.e., reduced) CDMs with different assumptions have also been proposed in the literature. For example, the *deterministic, input, noisy, and gate* (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model assumes a *conjunctive* process in that only individual who have *all* the disorders measured by the item are expected to endorse the item; the *deterministic, input, noisy, or gate* (DINO; Templin & Henson, 2006) model assumes a *disjunctive* process in that that individuals who have *at least one* disorder measured by the item are expected endorse the item; and the additive CDM (A-CDM; de la Torre, 2011) assumes that the different disorders measured by the item contribute *independently* to the probability of the item endorsement.

Most CDMs, including the G-DINA model, require a so-called Q-matrix (Tatsuoka, 1983, 1990). A Q-matrix requires the explicit use of a substantive theory in a domain (e.g., which symptoms are associated with which disorders). In typical CDM applications, Q-matrices are constructed a priori, and are based on expert opinions, clinical theories, or results from empirical research. A Q-matrix is a $J \times K$ matrix of 0s and 1s that relate the J items are to the K disorders. The entry at the intersection of the jth row and kth column is 1 if and only if disorder k is measured by item j. Typically, tests used for CDM purposes contain items that measure multiple disorders.

Appendix B shows the Q-matrix for the four scales of MCMI-III used in this article. The entries of this Q-matrix are based on the scale composition of the MCMI-III (Millon et al., 2009, pp. 175–179). Each row of the Q-matrix pertains to a symptom, and each column a disorder. For example, the first row denotes that endorsing Item 1, which describes a lack of physical strength, is indicative of the presence of the second and fourth disorders (i.e., somatoform and major depression) but not the first and third disorders (i.e., anxiety and thought disorder).

For respondent i, disorder profile $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})$ indicates his or her true state with respect to the disorders. If there are four disorders, α_i has four elements. For example, if $\alpha_i = (0000)$, the respondent does not have any of the four disorders and if $\alpha_i = (1110)$, the respondent has the first three disorders. For four disorders, the total number of disorder profiles equals $2^4 = 16$ (Table 1). The disorder profile α_i is not directly observable—otherwise administering inventories would not be necessary—and, therefore, has to be estimated.

Table 1. The 16 Possible Disorder Profiles and Their Associated Disorders.

	Disorder					Disorder			
Disorder Profile Number	A	H	SS	CC	Disorder Profile Number	A	H	SS	CC
1	0	0	0	0	9	1	0	0	0
2	0	0	0	1	10	1	0	0	1
3	0	0	1	0	11	1	0	1	0
4	0	0	1	1	12	1	0	1	1
5	0	1	0	0	13	1	1	0	0
6	0	1	0	1	14	1	1	0	1
7	0	1	1	0	15	1	1	1	0
8	0	1	1	1	16	1	1	1	1

Note. A = anxiety; H = somatoform; SS = thought disorder; CC = major depression.

In the G-DINA model notation, the probability that respondent i, whose disorder profile is α_i, will endorse item j, an indicator for Disorders 2 and 4, for example, can be written as follows:

$$P(X_{ij} = 1|\alpha_i) = \delta_{j0} + \delta_{j2}\alpha_{i2} + \delta_{j4}\alpha_{i4} + \delta_{j24}\alpha_{i2}\alpha_{i4},$$

where δ_{j0} represents the baseline probability (i.e., the probability that respondents who have neither α_2 nor α_4 will endorse item j); δ_{j2} (or δ_{j4}) is the increment in the probability of endorsing the item as a result of the presence of α_2 (or α_4); and δ_{j24} is the interaction effect because of the presence of both α_2 and α_4 (i.e., the change in the probability of endorsement beyond the additive impact of the two disorders). For notational simplification, we suppress Disorders 1 and 3, which are not measured by item j, and denote the endorsement probability of a respondent with both Disorders 2 and 4, Disorder 2 but not Disorder 4, Disorder 4 but not Disorder 2, and neither Disorder 2 or 4 by $P_j(00) = \delta_{j0}$, $P_j(10) = \delta_{j0} + \delta_{j2}$, $P_j(01) = \delta_{j0} + \delta_{j4}$, $P_j(11) = \delta_{j0} + \delta_{j2} + \delta_{j4} + \delta_{j24}$, respectively.

The G-DINA model is said to be saturated if the number of parameters to be estimated is equal to the number of latent groups implied by the number of required disorders for the item. In the above example, the two required disorders imply four latent groups (i.e., 00, 01, 10, and 11). Thus, the model for item j with four parameters (a baseline probability, two main effects, and an interaction effect) is considered saturated. Reduced models are obtained when constraints to the parameters are introduced. For example, the constraint $\delta_{j24} = 0$ produces an additive model. When no deterioration in model fit is evident from imposing the constraints, a reduced model is to be preferred over the saturated model.

The parameters of the G-DINA model can be estimated using marginal maximum likelihood estimation. After calibrating the test, the disorder profile of each respondent can be obtained. For respondent i, the estimated disorder profile is of the form $\hat{\alpha}_i = \{P(\alpha_k = 1|\boldsymbol{X}_i)\}$, where $P(\alpha_k = 1|\boldsymbol{X}_i)$, the kth element of the vector $\hat{\alpha}_i$, is the posterior probability that the disorder k is present in respondent i. These probabilities are *expected a posteriori* estimates, and can be computed as

$$P(\alpha_k = 1|\boldsymbol{X}_i) = \sum_{\forall l:\alpha_{lk}=1} P(\alpha_l|\boldsymbol{X}_i)$$

where $P(\alpha_l|\boldsymbol{X}_i)$ is the posterior probability that respondent i has the disorder profile α_l, and the summation is over all disorder profiles where the kth element of α_l is 1. Probabilities that are close to either 0 or 1 indicate a strong evidence for the absence or presence of a disorder, respectively. For easier interpretation, these probabilities can be converted to either 0 (absent) or 1 (present) based on certain rules. For example, the following rule can be used: The disorder is said to be present if the posterior probability is at least 0.5; otherwise, it is said to be absent. For additional details on DINA-based model estimation procedures, see de la Torre (2009a, 2009b, 2011).

In addition to the item parameter estimates, the marginal maximum likelihood estimation procedure can also provide an estimate of the joint probability distribution of the disorders. That is, it can provide an estimate of the relative frequencies of different disorder combinations in the population. With an estimate

of the disorder distribution, it is possible to examine the relationships (e.g., comorbidities) between the different disorders.

Saturated and Reduced Model Comparison

Model selection is considered one of the most fundamental components of scientific inquiry. For a given data, several CDMs may be deemed plausible. From this set of models, the most appropriate CDM needs to be selected. Although the G-DINA model can provide better model-data fit compared with other reduced models, specific models may be preferred under some circumstances for several reasons. First, reduced models have more straightforward interpretations, and require smaller sample size to be estimated accurately. Second, Occam's razor and the parsimony principle dictate that simpler models be preferred over more complex models when they cannot be distinguished from each. Finally, appropriate reduced models can provide better classification rates than saturated models, particularly when the sample size is small (Rojas, de la Torre, & Olea, 2012).

De la Torre (2011) proposed the use of the Wald test (Buse, 1982) as a statistical way of determining whether one of the interpretable reduced models can be used in place of the saturated G-DINA model. The test has been shown to provide relatively accurate Type I error, particularly with large sample sizes and small number of parameters. In addition, the test has high power to detect when a reduced model is not appropriate (de la Torre & Lee, 2013). It should be noted that the Wald test is carried out at the item level. In addition to greater efficiency, this feature of the Wald test allows for the possibility of using multiple CDMs with a single data.

Data Analysis

The data from the Dutch version of the MCMI-III consisted of the responses of $N = 1{,}210$ Caucasian patients and inmates in Belgium (61% males, 39% females). Response vectors of respondents having disclosure scores (Scale X of the MCMI-III) outside the range 34 to 178 or with random responding (combination of Scale V < 2 and W < 10) were considered invalid and are not included in the sample (Millon et al., 2009). Hence, it can be assumed that the 1,210 respondents in the sample were neither reticent and secretive nor too frank and self-revealing and responded in a valid and consistent manner. Fifty-three percent of the respondents were clinical inpatients, 15% clinical outpatients, and 32% came from a forensic setting (for more details, see Rossi, Elklit, & Simonsen, 2010). The questionnaire used in this analysis contained 44 items from the four scales (A, H, SS, and CC; see Millon, et al., 2009, pp. 175–179), and each of these items measures one to three of the four studied disorders. The number of disorders measured in each scale is summarized in Table 2. The table shows that the selected scales contained 28 items measuring a single disorder, 15 items measuring two disorders, and a single item measuring three disorders. Additionally, it shows that majority of the Scales A and SS items were measuring single disorders, whereas majority of the Scale H and CC items were measuring two disorders. Finally, the table indicates that the number of items per scale ranged from 12 to 18.

Two steps were involved in selecting the CDMs for MCMI-III data. In the first step, the G-DINA model was fitted the data set using the E-M algorithm; in the second step, the Wald test was employed to select the most appropriate CDM for each item. Specifically, three reduced models (i.e., DINA, DINO

Table 2. Number Items in Each Scale Measuring One, Two, and Three Disorders.

	Disorder				
Number of Disorders	A	H	SS	CC	Number of Items
One	9	2	11	6	28
Two	5	9	5	11	15
Three	0	1	1	1	1
Total	14	12	17	18	44

Note. A = anxiety; H = somatoform; SS = thought disorder; CC = major depression.

and A-CDM) were compared against the G-DINA model. A significance level of .05 was used. If none of the reduced CDM is deemed appropriate for a particular item, the G-DINA model is retained; if more than one reduced CDM is deemed acceptable, which did not happen for these CDMs, we choose the model with the largest p value. It should be noted that the Wald test is necessary only when an item measures more than one disorder—when an items measures exactly one disorder, no distinction can be made between general and reduced CDMs.

Based set of CDMs suggested by the Wald test, the item parameters were estimated using the E-M algorithm. The estimation code was written in Ox (Doornik, 2003), and using a 2.50-GHz I5 computer, it took the code less than 25 seconds to estimate the model parameters and classify the respondents. The code can be requested from the first author.

The raw scales were standardized into base rate (BR) scores using the separate gender norms from Millon (1994). BR scores are based on criterion referencing instead of on the more classic norm referencing. Such approach was taken to recognize differences in distribution shapes of the scales. According to Millon et al. (2009) norm referencing (e.g., T scores) is problematic because an invariable percentile of the standardizing sample will seldom correspond to the actual prevalence rate. Therefore, in criterion referencing (e.g., BR scores) the actual prevalence rate of the clinical disorder being measured by the scale is taken into account (Millon et al., 2009).

Diagnosing Respondents

Analyzing the questionnaire response data using the G-DINA model, for each respondent and for each disorder profile (Table 1), the probability that the respondent had the disorder profile was estimated. This resulted in 16 probabilities for each respondent. Because practical diagnosis based on 16 *profile probabilities* for 4 disorders can be inconvenient, the 16 probabilities were aggregated to form 4 *disorder probabilities*.

Table 3 shows the disorder probabilities for anxiety, somatoform, thought disorder, and major depression, for respondents 1 to 10 in the sample (Columns 1, 2, 3, and 4) and the base rate scores for these disorders, which are currently used for diagnosis (Columns, 5, 6, 7, and 8). Note that the disorder probabilities are rounded to two decimals, and no disorder probability equals exactly 0 or 1. Also note that the disorder probabilities are estimates and have a small degree of uncertainty. BR scores greater than 75 (italics) indicate presence of the disorder, and scores greater than 85 (boldface) indicate prominence of the disorder.

Although, in general, the disorder probabilities and the BR scores have similar patterns (i.e., 7 of 10 profile classifications have agreements on at least three disorders), there are also a number of differences. For example, the probabilities that Respondents 4 and 6 have the anxiety disorder were estimated close to 0.00, whereas their BR scores of 75 and 80, respectively, indicated that anxiety is present.

Table 3. Disorder Probabilities and BR Scores for the First 10 Respondents.

	Disorder Probabilities				BR Scores			
Respondent	A	H	SS	CC	A	H	SS	CC
1	1.00	0.97	1.00	1.00	**86**	68	*76*	*77*
2	0.00	0.00	0.00	0.00	40	30	45	20
3	1.00	0.72	0.90	0.38	**97**	65	66	67
4	0.00	0.98	0.14	0.02	*75*	68	61	67
5	1.00	0.35	1.00	1.00	**104**	58	67	64
6	0.01	0.00	0.00	0.00	*80*	0	15	0
7	0.99	0.95	0.27	0.00	**85**	63	59	59
8	0.00	0.01	0.00	0.00	60	60	60	60
9	0.00	0.00	0.00	0.00	0	0	0	0
10	0.95	0.00	0.09	0.00	**88**	30	65	0

Note. A = anxiety; H = somatoform; SS = thought disorder; CC = major depression.

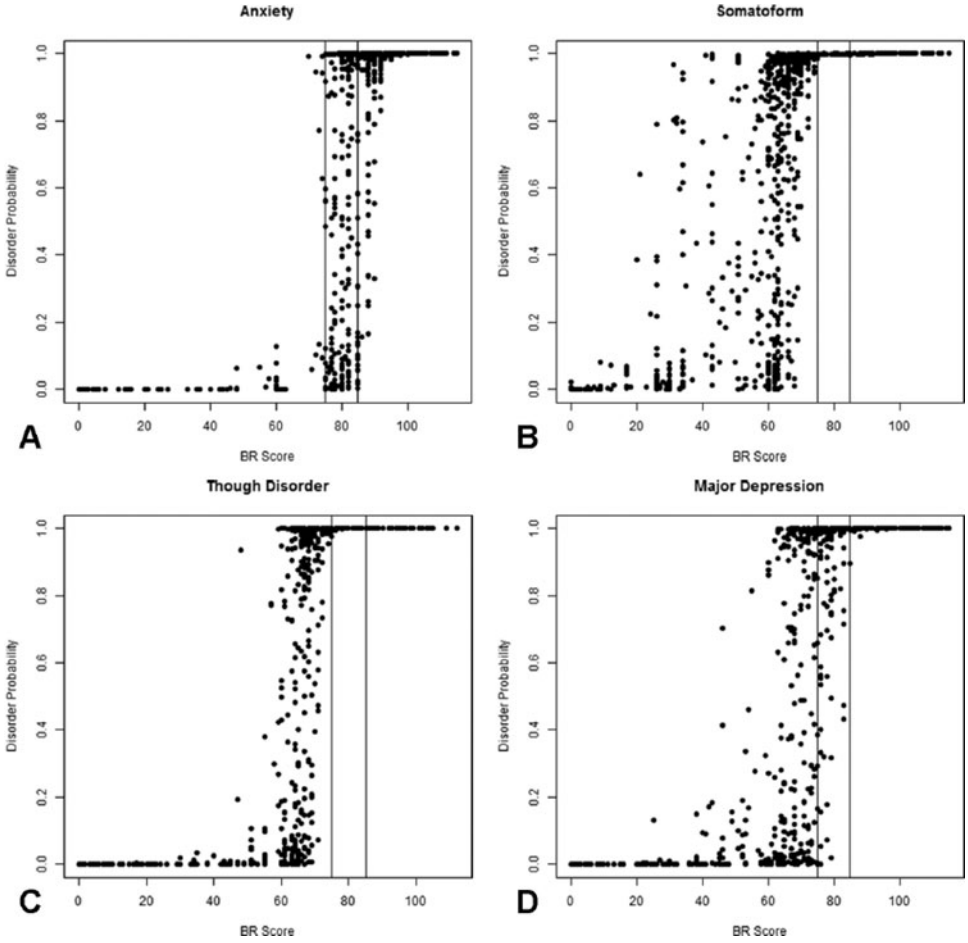

Figure 1. Scatter plots of BR scores (horizontal axes) by disorder probability (vertical axis) for Anxiety (A; top left), Somatoform (B; top right), Thought disorder (C; bottom left), and Major depression (D; bottom right). Vertical lines are BR = 75 and BR = 85.

For the entire sample, the correlation between the logit of the disorder probabilities and the BR scores was .87 for Anxiety, .89 for Somatoform disorder, .87 for Thought disorder, and .93 for Major depression. Figure 1 shows the scatterplots of the disorder probabilities and the BR scores. The cutoff values of the BR scores (i.e., 75 and 85) are represented by vertical lines. Figure 1A (Anxiety) shows that for BR scores less than 70 and greater than 90, the disorder probabilities agree, but the BR scores between 70 and 90 are a gray area, where disorder probabilities may have a different outcome than the BR scores. Figure 1B (Somatoform) shows that the cutoff values of the BR scores are more conservative than the disorder probabilities; every respondent with a BR score greater than 75 has an extremely high disorder probability. Also, high disorder probabilities were found for respondents with BR scores as low as 30. Figure 1C (Thought disorder) also shows that the cutoff values of the BR scores are more conservative than the disorder probabilities; every respondent with a BR score greater than 70 has an extremely high disorder probability. Except for two outlying values the all BR scores less than 60 have a very low disorder probability. Figure 1D (Major Depression) shows that the BR scores are slightly more conservative than the disorder probabilities. Respondents with BR scores as low as 50 can have substantial disorder probabilities.

Even though the correlations between the BR scores and the disorder probabilities are rather high, aggregated percentages show that except for anxiety, the BR scores are more conservative than the disorder probabilities. The cases where respondents are considered to have the disorder using probabilities but not BR scores can be found in the quadrant to the left of 75 and above .5. The expected prevalence

Table 4. Selected CDMs for the Multi-Attribute Items.

Item	Selected Model	Item	Selected Model
1	G-DINA	22	A-CDM
2	A-CDM	25	A-CDM
5	A-CDM	29	A-CDM
9	A-CDM	32	G-DINA
12	G-DINA	33	G-DINA
15	A-CDM	35	A-CDM
16	A-CDM	36	A-CDM
17	A-CDM	38	G-DINA

rates in the sample, according to the CDM are estimated at 53% for anxiety, 49% for somatoform, 52% for thought disorder, and 44% for major depression. Using a BR score 75 as the cutoff, the prevalence rates are 66% for anxiety, 10% for somatoform, 15% for thought disorder, and 35% for major depression. The classification discrepancies between the two diagnostic methods can also be observed in Table 4, where for the first 10 respondents, the probabilities of having the disorders somatoform, thought disorder and major depression indicate that the disorders are likely to be present, whereas the BR scores do not.

For a small subsample ($N = 262$), the diagnosis of a psychiatrist (a single rating on a scale from 0 to 9), which was unrelated to the inventory data, was available. The correlation between the disorder probabilities and the expert opinions were slightly higher than the correlations between the BR scores and the expert opinions. Unfortunately, the correlations were rather small to modest: .05 for somatoform, and between .22 and .28 for anxiety, thought disorder and major depression. These low validity coefficients may be because of unreliable measurement of the diagnosis by the psychiatrist. It is a well-known result from classical test theory (Lord & Novick, 1968) valid inferences can be precluded by low reliability.

Final CDMs for Multi-Attribute Items

As shown on Table 2, 16 of the 44 items in the questionnaire measure more than one disorder. Table 4 shows that none of these 16 items can be fitted using the DINA or DINO models. Rather, 11 items multiple disorders that have additive effects on the item endorsement, in which case, A-CDM was appropriate; five items measure disorders whose effects cannot be described as either conjunctive, disjunctive, or additive, in which case, the G-DINA model was appropriate. It should be noted that for Item 35, the only item which measures three disorders, A-CDM was the selected model. This result indicates that the presence of more attributes does not necessarily preclude the selection of a simpler CDM.

Item Parameter Estimates

Tables 5 to 8 show the parameter estimates of the 44 items obtained from the G-DINA model analysis. The results were grouped based on the number of and specific disorders the items measured. Tables 5, 6, and 7 show the estimates for one-, two-, and three-disorder items, respectively. By examining closely the item parameters estimates, we can gain deeper insights into the relationships between the disorders and their symptoms.

Items measuring a single disorder have only two parameters: $P_j(0)$ and $P_j(1)$. The former is the baseline probability (i.e., the endorsement rate of respondents who do not have the disorder), whereas the latter is the probability of endorsement for respondents who have the disorder measured by the item. The difference between the two probabilities can be viewed as the discrimination index of the item (de la Torre, 2008), and it represents the item's power to distinguish between those who have the disorder from those who do not. Items with low $P_j(0)$ and high $P_j(1)$ can have large discrimination indices, whereas items with either high $P_j(0)$ or low $P_j(1)$ can have small discriminations indices. Under the ideal condition where the discrimination index is 1 (i.e., $P_j(0) = 0$ and $P_j(1) = 1$), endorsement or nonendorsement of an item is a perfect indicator of the presence or absence of the disorder, respectively.

Table 5. Parameter Estimates of Items Measuring One Disorder: Endorsement Rate Conditional on Disorder Status.

		Disorder Status		
Disorder	Item	0—Absent	1—Present	Disc.
A	7	.21	.77	**.56**
	11	.27	.78	**.51**
	23	.13	.61	**.48**
	24	.28	.80	**.52**
	27	.05	.45	**.41**
	31	.19	.70	**.51**
	34	.36	.94	**.58**
	41	.11	.65	**.54**
	43	.07	.48	**.41**
H	3	.03	.22	.19
	6	.05	.29	.25
SS	4	.08	.48	**.40**
	10	.24	.82	**.58**
	13	.18	.77	**.59**
	14	.16	.80	**.64**
	18	.01	.22	.21
	19	.32	.81	**.48**
	20	.06	.32	.26
	26	.07	.30	.22
	30	.07	.58	**.52**
	40	.18	.73	**.56**
	42	.11	.41	.30
CC	8	.19	.90	**.71**
	21	.09	.51	**.42**
	28	.04	.50	**.47**
	37	.02	.57	**.56**
	39	.23	.54	.31
	44	.07	.60	**.53**

Note. A = anxiety; H = somatoform, SS = thought disorder, CC = major depression. Disc. = Discrimination index. Discrimination indices of at least .4 are printed in boldface.

Table 6. Parameter Estimates of Items Measuring Two Disorders: Endorsement Rate Conditional on Disorder Combination.

		Disorder Combination				
Disorders	Item	00	10	01	11	Disc.
A-H	16[a]	.09	.38	.42	.72	**.63**
	33	.14	.72	.45	.88	**.73**
A-SS	12	.16	.73	.65	.95	**.79**
	17[a]	.04	.47	.30	.73	**.69**
A-CC	36[a]	.04	.36	.22	.54	**.49**
H-CC	1	.04	.57	.56	.92	**.89**
	2[a]	.05	.55	.41	.91	**.86**
	9[a]	.04	.47	.41	.83	**.79**
	15[a]	.05	.49	.30	.74	**.69**
	22[a]	.03	.24	.25	.46	**.43**
	25[a]	.08	.33	.42	.67	**.59**
	29[a]	.01	.43	.39	.81	**.80**
SS-CC	5[a]	.20	.51	.54	.85	**.65**
	32	.04	.37	.31	.82	**.79**
	38	.06	.20	.24	.59	**.53**

Note. A = anxiety; H = somatoform; SS = thought disorder; CC = major depression. 00 = both disorders are absent, 10 = only first disorder is present, 01 = only second disorder is present, 11 = both disorders are present. Disc. = Discrimination index. Discrimination indices of at least .4 are printed in boldface.
a = A-CDM.

Table 7. Item 35 Parameter Estimates: Endorsement in Percentage of Conditional on Disorder Combination.

None	H	SS	CC	H-SS	H-CC	SS-CC	H-SS-CC
.09	.23	.27	.45	.41	.59	.63	.77

Note. H = somatoform; SS = thought disorder; CC = major depression.

For discussion purposes, .4 will be used as the minimum index value for an item to be considered discriminating. (We acknowledge that to a large extent this number is arbitrary; however, no commonly accepted rule-of-thumb has been established in the current CDM literature.) This means that respondents who have a particular disorder have at least 40% higher probability to endorse an item compared with those who do not. Based on this cutoff, 9 of 9, 0 of 2, 7 of 11, and 5 of 6 items in Scales A, H, SS, and CC, respectively, are considered discriminating. Scale H has no discriminating item because $P_j(1)$ was low for both items (i.e., less than .3). Compared with respondents who do not have somatoform, those who do are not substantially more likely to endorse Items 3 and 6 whose content is related to difficulties with bodily sensations. For the anxiety, thought disorder, and major depression scales, the most discriminating items are 7, 14, and 8, respectively measuring recurrent (disturbing) thoughts, aimlessness, and depressed feelings. These items have discrimination indices of .56, .64, and .71. On the average, with a mean discrimination index of .45, the 28 single-disorder items as a whole can be considered discriminating.

For items measuring more than one disorder, in addition to endorsement rates for those who have none of the disorders, and those who have all of the disorders, the differential impact of the various disorders with respect to specific symptoms can also be examined. Moreover, whether the presence of multiple disorders results in an interaction for a particular item (i.e., symptom) can also be studied. The 15 items in Table 6 represents all the possible pair-wise combinations of the four disorders, except the somatoform-thought disorder combination which did not appear in the original scale composition. The analog of the discrimination index for two-disorder items, albeit based on incomplete information, is the difference $P_j(11) - P_j(00)$. Based on this index and the cutoff .4, all the 15 items are considered discriminating. The items with the highest and lowest discrimination indices, Item 1 (.89), and Item 22 (.43), both measure the somatoform and major depression combination. Item 1 measures the loss of (physical) strength, and Item 22 disturbances in eating and sleeping patterns.

By computing $P_j(10) - P_j(01)$, we can determine whether the respondents with one of the two disorders have the same probability of exhibiting the symptom described by the items. A positive difference indicates that respondents with only the first disorder are more likely to endorse the item than those who only have the second disorder; a negative difference indicates the opposite. For illustration purposes, we examine the two items with the largest and lowest absolute differences: item 33, which describes worrying (difference = .27), and Item 1, which describes loss of strength (difference = .01). Item 27 measures anxiety and somatoform, whereas Item 1 measures somatoform and major depression. Respondents with anxiety have an endorsement rate of Item 33 that is .27 higher than respondents with somatoform; however, respondents with either somatoform or major depression are equally likely to endorse Item 1.

Based on the Wald test, 5 of the 15 two-disorder items had interaction, and cannot be modeled using the A-CDM. To examine more closely how two disorders interact at the item level, we can compute $P_j(10) + P_j(01) - P_j(00)$ and compare it with $P_j(11)$. Items 12 and 38 will be used to illustrate how the nature of the interaction can be examined in items measuring two disorders. The estimates of $P_j(11)$, $P_j(10) + P_j(01) - P_j(00)$, and the difference between the two estimates for these items are given in Table 8. In Item 12, which describes recurrent thoughts, the presence of both anxiety and somatoform

Table 8. Examining Interaction in Selected Two-Disorder Items.

Disorders	Item	[a]	[b]	[a]-[b]
A-SS	12	.95	1.23	− .28
SS-CC	38	.59	.38	.21

Note. [a] = P(11); [b] = P(10) + P(01) − P(00); [a]-[b] = P(11) + P(00) − P(10) − P(01); A = anxiety; H = somatoform; SS = thought disorder; CC = major depression.

Table 9. The Disorder Profiles and Their Prevalence.

Disorder Profile	Prevalence (%)	Disorder Profile	Prevalence (%)
0000	34.0	0110	4.0
1000	4.3	0101	0.0
0100	3.3	0011	0.4
0010	0.7	1110	5.8
0001	2.4	1101	1.4
1100	1.0	1011	6.8
1010	2.0	0111	2.0
1001	1.2	1111	30.6

Note. For the disorders associated with the disorder profiles, see Table 2. The sum does not add to 100% due to rounding.

resulted in an endorsement rate that was much lower (.95 vs. 1.23) than would be expected by simply adding the effect of each disorder; in Item 38, which measures describes feeling oneself to be worthless, the presence of both thought disorder and major depression tend to reinforce each other resulting in an endorsement rate that was much higher (.59 vs. .38) than what can be accounted for by the two disorders additively. For the remaining items computing $P_j(10) - P_j(00)$ and $P_j(01) - P_j(00)$ will give the incremental effect on item endorsement of possessing a particular disorder. For example, having somatoform will lead to .44 higher endorsement rate of Item 15, which describes sleep problems, whereas having major depression will lead to .25 higher endorsement rate of the same item; having both disorder will result in $.44 + .15 = .69$ higher endorsement rate.

Only one of the 44 items measures more than two disorders, item 35, which describes a lack of pleasure, is an indicator of somatoform disorder, thought disorder, and major depression (see Table 7 for the item parameter estimates). These disorders had an additive effect of the endorsement probabilities. Specifically, possessing somatoform disorder, thought disorder, or major depression led to an increase of .14, .18, or .45 in the endorsement rate compared with the baseline. This indicates that this item a stronger indicator of major depression than either somatoform or thought disorder. Finally, to examine the discrimination of the item, we can compute that respondents who did not have any of the three disorders endorsed the item 9% of the time, whereas respondents who have all the three disorders endorsed the item 77% of the item. The difference of .68 between the two endorsement rates indicates that the item is highly discriminating.

Profile Sizes

The profile sizes in Table 9 show the distribution of the respondents across the disorder profiles, and indicate which combinations of disorders are common and which are rare. The table shows that (approximately) one third (34.0%) of the respondents have estimated disorder profile (0000), another third (30.6%) have estimated disorder profile (1111), and the remaining third have one of the remaining disorder profiles. Respondents in disorder profile (0000) tend to have low endorsement probabilities for all items and this profile may be labeled the "no disorder profile." Respondents in disorder profile (1111) have high endorsement probabilities for all items and have an increased risk to have all four disorders. Disorder order profile (0101) is nonexistent, indicating that an increased risk to have both somatoform and thought disorder without an increased risk to have anxiety and major depression is extremely rare.

Summary and Discussion

We demonstrated that analyzing personality inventories consisting of scales with overlapping items using CDMs yields detailed results for diagnosing respondents, examining the characteristics of items in the inventory, and investigating the underlying structure of the disorders. As a novel approach, several questions may arise in applying the proposed method.

Is the proposed method more complicated than current methods? Although we agree that for most readers the answer is probably "yes," we also believe that in a field as important as clinical psychology only

the best method should be used, irrespective of the difficulty of the method. In addition, the difficulty of the method is not a permanent roadblock because user-friendly software with clear guidelines on the interpretation of the results can be developed to make the application of the method a less arduous task. At the moment, although such software is not available, efforts are underway to make novel and relatively complex methods such as this more accessible.

Is the proposed method better than the current methods? With respect to diagnosing respondents we are inclined to answer "yes" because CDMs acknowledge comorbidity of the disorders. However, results may be blurred by several factors, such as misspecifications in the Q-matrix which can result in less than ideal fit of the CDM to the data. The importance of an accurately specified Q-matrix in supporting valid inferences cannot be overemphasized. For this reason, several researchers are currently examining various methods the Q-matrix specifications can be validated. One such method is the empirical approach developed by de la Torre and Chiu (2016) that can be used in conjunction with the G-DINA model. In addition to ensuring the optimality of the Q-matrix, this method can determine to what extent the empirical data validate the judgments used by clinicians in establishing the associations between the disorders and their symptoms. On the whole, we subscribe to triangulation as a sensible general strategy to check the veracity of the findings. In this regard, we recommend using CDMs in combination with other predictors, such as BR scores, expert opinion, and patients' history to verify the accuracy of the disorder classifications.

With respect to examining characteristics of the items, we believe that CDMs contribute to the large selection of methods based on other psychometric frameworks (i.e., classical test theory and item response theory) that are already at the disposal of the researchers. We advocate that these methods could be employed in the test development process. In addition to psychometric properties, CDMs in their saturated formulations can also provide information pertaining to how multiple disorders interact at the item (i.e., symptom) level. This information can provide researchers and practitioners a better understanding of the potential complex relationships between their disorders and some specific symptoms.

There are several possible reasons why the agreements between the CDM and BR score classifications were less than perfect. First, the discrepancies may be because of the fact that the BR scores do not take into account the comorbidity among disorders. Another reason for the discrepancy can be attributed to the design and aim of the MCMI-III—the inventory is designed mainly as a screening tool for differential diagnosis, rather than to precisely distinguish normal respondents from those who are not, which is better achieved using a clinician's diagnostic evaluation based on all available data. The discrepancies may also be because of how the BR scores are derived. Millon (1994) used a piecewise linear interpolation to transform the raw scores to their corresponding BR anchor points, instead of a precise smooth transformation (Groove & Vrieze, 2009). As Meehl and Roosen (1955) pointed out, prevalences above .5 can result in nonmonotonic BR transformations. Also, maximal classification accuracy may not be achieved unless the tests have been calibrated for local test score distributions (Groove & Vrieze, 2009; Rossi et al., 2007). Finally, the classification discrepancies may have resulted from the misspecifications in the Q-matrix, or the low discriminating items in the inventory. Regardless of the reason, comparing the two diagnoses may help expose weaknesses in one of or both methods, which can potentially lead to their refinement.

Millon (1994) and Millon et al. (2009) considered the positive predictive ratio to be the most useful diagnostic validity statistic, and gave priority to avoid false positives (Rossi & Sloore, 2005), however independent validity studies showed out that the anxiety scale nevertheless has a tendency to over diagnose. Saulsman (2011) examined the construct validity and diagnostic efficiency of the depression and anxiety scales of the MCMI-III and found construct validity for the major depression scale; however, the anxiety scale was more problematic. The scale only showed modest convergence with anxiety measures on worry and panic pathology, but did not correlate with social discomfort/anxiety, and correlated as much with depressive symptoms (which also mainly include worry and panic pathology). Thus, the scale has a poor ability to discriminate depression, and poor diagnostic efficiency statistics. Earlier research by Gibeau and Choca (2005) already pointed out that the anxiety scale shows a tendency to over diagnose and has a particularly low specificity and positive predictive power. However, in clinical settings comorbidity between anxiety and depression disorders is rather a rule than an exception, so it is unlikely

that test instruments will differentiate with a high degree of accuracy. Recent research (e.g., Eaton et al., 2013) formulates a two-dimensional distress-fear liability structure, placing mood and anxiety disorder into one group.

Gibeau and Choca (2005) also demonstrated problems with the sensitivity of thought disorder (this scale has highspecificity), consistent with previous research that pointed out that the MCMI-III may have problems with capturing thought disturbance related to schizophrenia (e.g., Wetzler & Marlowe, 1993). Somatoform and major depression scales have high specificities and reasonable sensitivities at BR 75. It would be interesting to examine to what extent many of these issues can be replicated or addressed if CDM probabilities are used in place of the BR scores.

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

This research was supported in part by the National Science Foundation CAREER Grant No. DRL-0744486. Any opinions expressed herein are those of the authors and do not necessarily represent the views of the National Science Foundation.

Notes on Contributors

Jimmy de la Torre is Professor at the Division of Learning, Development and Diversity at the University of Hong Kong. His primary research interests include item response theory, cognitive diagnosis, Bayesian analysis, and application of diagnostic modeling in various settings.

L. Andries van der Ark is Professor at the Department of Child Development and Education at the University of Amsterdam. His research interests include reliability analysis, nonparametric item response theory, and categorical marginal models.

Gina Rossi is Associate Professor at the Department of Clinical and Life Span Psychology at the Vrije Universiteit Brussel (VUB). Her research focus is personality psychopathology, and assessment of personality (disorders) in mid-adult and older adult populations. She is Associate Editor of Assessment.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders*, (text revision). Washington, DC: Author.

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

Brown, T. A., Campbell, L. A., Lehman, C. L., Grisham, J. R., & Mancill, R. B. (2001). Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample. *Journal of Abnormal Psychology*, *110*, 585–599.

Budescu, D. V., & Rodgers, J. L. (1981). Corrections for spurious influences on correlations between MMPI scales. *Multivariate Behavioral Research*, *16*, 483–497.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *American Statistician*, *36*, 153–157.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (2001). *MMPI-2: Manual for administration and scoring* (Rev. ed.). Minneapolis: University of Minnesota Press.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage implications in professional psychology. *Professional Psychology: Research and Practice*, *31*, 141–154.

Cohen, A., & Farley, F. H. (1979). The common item in measurement: Effect on structure. *Multivariate Behavioral Research*, *14*, 91–108.

Dahl, A. A. (1986). Some aspects of the *DSM–III* personality disorders illustrated by a consecutive sample of hospitalised respondents. *Acta Psychiatrica Scandinavica, 73*(Suppl. 328), 61–67.

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. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*, 163–183.

de la Torre, J. (2009b). 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., & Chiu, C. Y. (2016). A general method of empirical Q-Matrix validation. *Psychometrika*, *81*, 253–273.
de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*, 355–373.
Dimitrov, D. M. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters. *Applied Psychological Measurement*, *31*, 367–387.
Dimitrov, D. M., & Atanasov, D. V. (2012). Conjunctive and disjunctive extensions of the least squares distance model of cognitive diagnosis. *Educational and Psychological Measurement*, *72*, 120–138.
Dolan-Sewell, R. T., Krueger, R. F., & Shea, M. T. (2001). Co-occurrence with syndrome disorders. In W. J. Livesley (Ed.), *Handbook of personality disorders* (pp. 84–104). New York, NY: Guilford Press.
Doornik, J. A. (2003). Object-Oriented Matrix Programming using Ox (Version 3.1) [Computer software]. London, England: Timberlake Consultants Press.
Duijsens, I. J., Eurelings-Bontekoe E. H. M., & Diekstra, R. F. W. (1996). *VKP: Questionnaire on personality traits (ICD-10/DSM-IV Manual Research Version)*. Geneva, Switzerland: World Health Organization.
Eaton, N. R., Krueger, R. F., Markon, K. E., Keyes, K. M., Skodol, A. E., Wall, M., … Grant, B. F. (2013). The structure and predictive validity of the internalizing disorders. *Journal of Abnormal Psychology*, *122*, 86–92.
Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, *49*, 175–186.
Gibeau, P., & Choca, J. (2005). The diagnostic efficiency of the MCMI-III in the detection of Axis I disorders. In R. J. Craig (Ed.), *New directions on interpreting the Millon Clinical Multiaxial Inventory-III (MCMI-III)* (pp. 272–283). Hoboken, NJ: Wiley.
Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–323.
Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.
Hsu, L. (1992). Correcting correlations of personality scales for spurious effects of shared items. *Multivariate Behavioral Research*, *27*, 31–41.
Jaeger, J., Tatsuoka, C., Berns, S. & Varadi, F. (2006). Distinguishing neurocognitive functions using partially ordered classification models. *Schizophrenia Bulletin*, *32*, 679–691.
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.
Kelly, T. L. (1927). *The interpretation of educational measurement*. New York, NY: World Book.
Krueger, R. F., & Eaton, N. R. (2010). Personality traits and the classification of mental disorders: Toward a more complete integration in DSM-5 and an empirical model of psychopathology. *Personality Disorders: Theory, Research, and Treatment*, *1*, 97–118.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Welsley.
Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–216.
Millon, T. (1994). *Manual for the MCMI-III*. Minneapolis, MN: National Computer Systems.
Millon, T., Millon, C., Davis, R., & Grossman, S. (2009). *MCMI-III Manual* (4th ed.). Minneapolis, MN: Pearson Assessments.
Modestin, J. (1987). Quality of interpersonal relationships: The most characteristic *DSM-III* BPD criterion. *Comprehensive Psychiatry*, *28*, 397–342.
Nurnberg, H. G., Hurt, S. W., Feldman, A., & Suth, R. (1987). Efficient diagnosis of borderline personality disorder. *Journal of Personality Disorders*, *1*, 307–315.
Nurnberg, H. G., Raskin, M., Levine, P. E., Pollack, S., Siegel, O., & Prince, R. (1991). Hierarchy of *DSM-III-R* criteria efficiency for the diagnosis of borderline personality disorder. *Journal of Personality Disorders*, *5*, 211–224.
Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, *60*, 489–502.
Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.
Rossi, G., Elklit, A., & Simonsen, E. (2010). Empirical evidence for a four factor framework of personality disorder organization: Multigroup confirmatory factor analysis of the Millon Clinical Multiaxial Inventory–III personality disorders scales across Belgian and Danish data samples. *Journal of Personality Disorders*, *24*, 128–150.
Rossi, G., & Sloore, H. (2005). International uses of the MCMI-III: Does interpretation change? In R. J. Craig (Ed.), *New directions on interpreting the Millon Clinical Multiaxial Inventory-III (MCMI-III)* (pp. 144–161). Hoboken, NJ: Wiley.
Rossi, G., Sloore, H., & Derksen, J. (2008). The adaptation of the MCMI-III in two non-English-speaking countries: State of the art of the Dutch language version. In T. Millon & C. Bloom (Eds.), *The Millon inventories: A practitioner's guide to personalized clinical assessment* (2nd ed., pp. 369–386). New York, NY: Guilford Press.

Rossi, G., van der Ark, L. A., & Sloore, H. (2007). Factor analysis of the Dutch-language version of the MCMI-III. *Journal of Personality Assessment*, *88*, 144–157.

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

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

Saulsman, L. M. (2011). Depression, anxiety, and the MCMI-III: Construct validity and diagnostic efficiency. *Journal of Personality Assessment*, *93*, 76–83.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., . . . Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, *59*, 522–533.

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. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

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

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., III, Rosa, K., Nelson, L., … Thissen, D. (2001). Augmented scores: "Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–388). Mahwah, NJ: Erlbaum.

Wetzler, S., & Marlowe, D. B. (1993). The diagnosis and assessment of depression, mania, and psychosis by self-report. *Journal of Personality Assessment*, *60*, 1–31.

Widiger, T. A., Hurt, S., Frances, A., Clarkin, J., & Gilmore, M. (1984). Diagnostic efficiency and *DSM-III*. *Archives of General Psychiatry*, *41*, 1005–1012.

Appendix A

Mapping of Item Numbers as They Appear in the Current Article and the *MCMI-III Manual* (see Millon et al., 2009, pp. 175–179, for the Item Descriptions).

Item Number		Item Number	
Current	Manual	Current	Manual
1	1	23	108
2	4	24	109
3	11	25	111
4	22	26	117
5	34	27	124
6	37	28	128
7	40	29	130
8	44	30	134
9	55	31	135
10	56	32	142
11	58	33	145
12	61	34	147
13	68	35	148
14	72	36	149
15	74	37	150
16	75	38	151
17	76	39	154
18	78	40	162
19	83	41	164
20	102	42	168
21	104	43	170
22	107	44	171

Note. MCMI-III = Millon Clinical Multiaxial Inventory-III.

Appendix B

Q-Matrix: Composition of Four MCMI-III Scales.

	Disorder					Disorder			
Item Number	A	H	SS	CC	Item Number	A	H	SS	CC
1	0	1	0	1	23	1	0	0	0
2	0	1	0	1	24	1	0	0	0
3	0	1	0	0	25	0	1	0	1
4	0	0	1	0	26	0	0	1	0
5	0	0	1	1	27	1	0	0	0
6	0	1	0	0	28	0	0	0	1
7	1	0	0	0	29	0	1	0	1
8	0	0	0	1	30	0	0	1	0
9	0	1	0	1	31	1	0	0	0
10	0	0	1	0	32	0	0	1	1
11	1	0	0	0	33	1	1	0	0
12	1	0	1	0	34	1	0	0	0
13	0	0	1	0	35	0	1	1	1
14	0	0	1	0	36	1	0	0	1
15	0	1	0	1	37	0	0	0	1
16	1	1	0	0	38	0	0	1	1
17	1	0	1	0	39	0	0	0	1
18	0	0	1	0	40	0	0	1	0
19	0	0	1	0	41	1	0	0	0
20	0	0	1	0	42	0	0	1	0
21	0	0	0	1	43	1	0	0	0
22	0	1	0	1	44	0	0	0	1

Note. A = anxiety; H = somatoform; SS = thought disorder; CC = major depression. 1 = item measures the disorder, 0 = item does not measure the disorder. For the original item numbers, see Appendix A.