PSYCHOMETRIKA—VOL. 79, NO. 2, 340–346
APRIL 2014
DOI: 10.1007/S11336-013-9363-Z

# HIERARCHICAL DIAGNOSTIC CLASSIFICATION MODELS MORPHING INTO UNIDIMENSIONAL 'DIAGNOSTIC' CLASSIFICATION MODELS—A COMMENTARY

MATTHIAS VON DAVIER

EDUCATIONAL TESTING SERVICE

SHELBY J. HABERMAN

EDUCATIONAL TESTING SERVICE

This commentary addresses the modeling and final analytical path taken, as well as the terminology used, in the paper "Hierarchical diagnostic classification models: a family of models for estimating and testing attribute hierarchies" by Templin and Bradshaw (Psychometrika, doi:10.1007/s11336-013-9362-0, 2013). It raises several issues concerning use of cognitive diagnostic models that either assume attribute hierarchies or assume a certain form of attribute interactions. The issues raised are illustrated with examples, and references are provided for further examination.

Key words: latent structure model, latent class analysis, diagnostic models, Guttman scaling, hierarchical models.

## 1. Introduction

The aim of diagnostic models is to provide more information about a person's attributes such as skills and abilities than what a single test score can provide. While there are a few examples of diagnostic models that allow binary and ordinal attribute levels (e.g. von Davier, 2005), most diagnostic models distinguish only binary attributes that are typically coded as 'attribute-mastery = 1' and 'non-mastery = 0' (e.g. Rupp, Templin, & Henson, 2010). Not only do diagnostic models aim at identifying more than one (binary) skill or attribute variable, they also seek to tell us whether these multiple attributes function in a compensatory, non-compensatory, or conjunctive manner. Attributes are assumed to operate in a compensatory manner if an increase in the number of required attributes (typically) yields an increased probability of success. More specifically, if two attribute patterns differ only in the mastery of one (required) attribute, the pattern with the larger number of required attributes results in a higher probability of success on an item. Conjunctive attributes, on the other hand, do not increase the probability of success with every increase in the number of required attributes, unless all required attributes are mastered. In other words, if a conjunctive attribute model applies, it does not matter whether a person does not master any attributes, or if a person has all but one required attribute, the probability of a success is the same (and typically low), unless a person has all required attributes for an item, in which case the 'high' probability of success is obtained (e.g. Junker & Sijtsma, 2001). Attempts to distinguish between these different types of structural assumptions appear to be based on test content as well as on empirical evidence. We will provide some evidence that these distinctions may be less clear than desirable.

We do not necessarily object to complex models. Our combined research experience dates back as far as the 1970s and is based on a background of mixture distribution and latent structure

Requests for reprints should be sent to Matthias von Davier, Educational Testing Service, Princeton, NJ, USA. E-mail: mvondavier@ets.org

models (e.g., Haberman, 1977; von Davier & Rost, 1995). Our more direct experience with what is nowadays called diagnostic classification models spans about 9 years (Haberman & von Davier, 2006; von Davier 2005, 2013).

But in these years, there have been few occasions where the specific promise of diagnostic models has been fulfilled. On a regular basis, in our experience, either less complex models (von Davier, 2005) or alternative specifications of models (Haberman, von Davier, & Lee, 2008) fit data as well or even better, or compensatory models fit data simulated using a conjunctive models almost as well as the generating model and provide almost identical attribute classifications (von Davier, 2013).The analysis presented in the Templin and Bradshaw (2013) article is no exception: At a first glance, the model introduced in that article appears to be a promising special case of the log-linear cognitive diagnosis model (LCDM) (Henson, Templin, & Willse, 2009). The proposed constraints of the attribute pattern distribution are used to set some probabilities to zero, a customary approach in restricted latent structure models. At a second glance, however, the real data example used in the paper shows that simpler models—for example, the unidimensional two-parameter logistic item response theory (2PL IRT) model, as well as ordered latent class models, can do the job and do it better (see Table 7 in Templin & Bradshaw, 2013). More specifically, results in Table 7 show that these customary unidimensional models are achieving preferable results over the hierarchical LCDM when looking at information criteria.

The following subsections develop some ideas that may be useful for further research on whether diagnostic models in general, and models that allow hierarchical ordering of attributes in particular, are suitable tools to provide added analytical value.

## 2. Some Clarification on Nomenclature

The HDCM presented by Templin and Bradshaw (2013) is a constrained version of the LCDM (Henson et al., 2009). In that sense, it has little to do with the *hierarchical general diagnostic model* (HGDM) by von Davier (2007, 2010), which extends results by Vermunt (2003) and provides multilevel extensions for diagnostic models. While the term *hierarchical* is sometimes used to imply an ordering of contributing factors, it often refers to a kind of data collection organized in multiple hierarchical levels, and the terms *hierarchical model* and *multilevel model* are often used interchangeably (Hox, 2002; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2001).

While there is some common ground in that the LCDM, HDCM, and HGDM are all based on the GDM, the HDCM and the HGDM address very different types of hierarchies. As Rupp et al. (2010) pointed out, the LCDM was developed based on the general diagnostic model (GDM; von Davier 2005, 2007), and the relationship of the HGDM to the GDM is obvious. Also, all of the above are special cases of latent class models (Lazarsfeld, 1950; Lazarsfeld & Henry, 1968), as pointed out in various places (e.g. Haberman et al., 2008; Maris, 1999; von Davier 2005, 2009).

## 3. Hierarchies of Binary Variables

Let us say that we have three binary attributes $A, B, C \in \{0, 1\}$. The idea of a hierarchical relationship among binary variables is simple to put into words, and examples exist in abundance: The outcome of the first variable precedes the outcome of the second. More formally, we may write $A \preccurlyeq B$. Typically, the hierarchy assumes that one outcome of the second variable (1, or mastery in our context) completely determines the outcome of the first (also mastery, i.e. 1), while the other outcome, if observed on of the first (0, non-mastery) variable, completely determines the outcome of the second (also 0). While this sounds confusing, it becomes trivial if

TABLE 1.
Binary tree for $A \preccurlyeq B$.

| $A$ | 0 | | 1 | |
|---|---|---|---|---|
| $B$ | 0 | 1 | 0 | 1 |
| $P(A, B)$ | $p_{00}$ | 0 | $p_{10}$ | $p_{11}$ |

TABLE 2.
Binary tree for $A \preccurlyeq B \preccurlyeq C$.

| $A$ | 0 | | | | 1 | | | |
|---|---|---|---|---|---|---|---|---|
| $B$ | 0 | | 1 | | 0 | | 1 | |
| $C$ | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| $P(A, B, C)$ | $p_{000}$ | 0 | 0 | 0 | $p_{100}$ | 0 | $p_{110}$ | $p_{111}$ |

we look at the contingency table for $A \preccurlyeq B \preccurlyeq C$. Tables 1 and 2 illustrate the distribution of the possible combination of attributes, expressed as nonzero probability entries in a binary tree, if the variables are related in what Leighton and Gierl (2007) called a *linear hierarchy* (see also Templin & Bradshaw, 2013).

For two variables, there are three possible attribute combinations if $A \preccurlyeq B$. We obtain three nonzero probabilities and the only impossible combination has a vanishing probability of $P(A = 0, B = 1) = 0$. Why there are only four nonzero entries for $A \preccurlyeq B \preccurlyeq C$ becomes clear when looking at Table 2. The only additional non-vanishing probability for three variables is $P(A = 1, B = 1, C = 1) \neq 0$, because both $A$ and $B$ have to be 1 in order for $C = 1$ to be possible.

As a consequence, we have four attribute patterns $(0, 0, 0)$, $(1, 0, 0)$, $(1, 1, 0)$, and $(1, 1, 1)$, that carry nonzero probabilities if the three binary variables form a 'linear hierarchy'. Note that this is a perfect Guttman pattern (Guttman, 1950). This is true no matter whether we look at observed response variables in a Guttman scale or at a number of assumed latent binary variables: We have to observe a 1 on $A$ in order to potentially observe a 1 on $B$, and we have to observe 1 on $A$ and $B$ in order to potentially observe a 1 on $C$. More importantly, note that exactly this pattern is evident in the LCDM results depicted in Figure 1 in Templin and Bradshaw (2013) so that—even without using the HDCM—an underlying Guttman ordering of the latent attributes could have been suspected.

In consequence, we can enumerate the binary attribute patterns that carry positive probabilities in a 'linear hierarchy' or Guttman ordering so that

$$\begin{aligned} 0^* &= (0, 0, 0), \\ 1^* &= (1, 0, 0), \\ 2^* &= (1, 1, 0), \\ 3^* &= (1, 1, 1), \end{aligned}$$

because there is a one-to-one mapping between the number of attributes that are mastered and the attribute pattern in this case. In other words, in a linear hierarchy there is no information gained by knowing which attribute pattern was observed given that we know how many attributes are mastered (= "1").

In a nutshell: A *linear hierarchy* is a misnomer because the hierarchical attribute patterns that are observed are exactly those that do not violate the Guttman assumption. In that case, the pattern of attributes is not informative, telling only how many attributes are mastered. To be very clear: In this case the underlying attribute model is a *deterministic unidimensional order*, and using multiple attributes only obscures the validity of a much simpler model.

## 4. Within Attribute Hierarchies, Conjunctive Attributes Do Not Exist

An insight that may have been lost in the paper by Templin and Bradshaw (2013) is that the number of conjunctive parameters available through the LCDM is severely limited when assuming hierarchical attributes and restricting the attribute space in the way proposed in the paper.

The problem is simple: A hierarchical order implies that the higher-order attribute can only be mastered if a lower-order attribute was mastered. Therefore, the conjunctive functioning cannot be distinguished from the compensatory functioning of attributes, because the higher-order attribute can never be present in a person not having the lower-order attribute.

In addition, under the assumption of attribute hierarchies, items are not allowed or at least are implausible if they require only the higher-order attribute but not include the lower-order attribute, for—by definition of the hierarchy—the higher-order attribute cannot be present without the lower-order attribute.

If attributes always appear together, then the LCDM reduces to a compensatory model for these attributes, and hence a special case of the GDM (von Davier, 2005) applies.

## 5. Are Conjunctive Attributes Uniquely Defined?

We learned in the subsections above that conjunctive attributes either do not exist or only exist in some reduced sense if attributes are combined in a hierarchical manner, for the attribute distribution essentially reduces to a number of permissible Guttman patterns.

Recent research has shown that diagnostic models face additional issues. DeCarlo (2011) has shown that in models involving conjunctive attributes, there are cases where not all attribute patterns can be identified or distinguished. A key to understanding why this may be the case could be an adaptation of a recent result by van der Linden (2012) to the case of binary latent variables.

In addition, there is growing evidence that compensatory models can be expected to fit data better than conjunctive models: As one example, Rojas and de la Torre (2012) have shown in a simulation that a G-DINA (general deterministic-input noisy-and) model typically fits data better than the DINA model does. This result implies that the inclusion of partially or fully compensatory attribute effects in the model improves model data fit. Also, von Davier (2013) showed that even for a dataset that was simulated using the conjunctive DINA model, a compensatory model achieves very similar model data fit and produces almost identical attribute classifications.

For the underlying cause of these findings, we may want to look beyond our realm and try to find inspiration from research in other domains that also study conjunctive models. More generally, the use of (approximate) Boolean functions

$$f : \{0, 1\}^k \to [0, 1]$$

was suggested by von Davier, Xu, and Yamamoto (2011) as a means to approximate conjunctive diagnostic models by additive or linear (compensatory) functions. It turns out that these types of linear approximations to Boolean functions are used for this purpose in cryptography and computational complexity (Hermelin & Nyberg, 2012; Tsai, 1996), where nonlinear (conjunctive multiplicative, etc.) functions such as algorithms for encrypting data are approximated successfully by additive (compensatory) Boolean functions. Why this could be of value can be seen when studying the function

$$\xi = \prod_k a_k^{q_{ik}}$$

used in the DINA model (Junker & Sijtsma, 2001) to combine multiple attributes. This attribute function is a Boolean (multiplicative) function that maps the multidimensional binary attribute vector onto a single binary variable. In the equation above, the $a_k \in \{0, 1\}$ are the binary attribute variables (attributes) and the $q_{ik} \in \{0, 1\}$ are the Q-matrix entries for item $i$. An application of approximations of this binary function as suggested by von Davier et al. (2011) could lead to a better understanding why compensatory models appear to be able to fit data generated by conjunctive models.

Finally, von Davier (2011, 2013) presented mathematical proofs and empirical evidence for two model equivalencies that show that the conjunctive DINA model can be recast into a compensatory modeling framework, the linear GDM, by defining attribute space and Q-matrix projection functions. An extension of this result to other models such as the LCDM is straightforward. With this result it becomes questionable that any model involving the assumption of conjunctive attributes is uniquely defined, because multiple attribute space and Q-matrix mappings exist that produce the exact same model-based conditional response probabilities without the assumption of conjunctive attributes.

## 6. Unidimensional Diagnostic Classification Models (DCMs) Are a Misnomer

It was shown in the previous sections of the commentary that attribute hierarchies—if they indeed exist—can be formulated in simpler Guttman sets that collect attribute subsets into unidimensional variables. The paper by Templin and Bradshaw (2013) does so in the case of a linear hierarchy, which we showed here to be nothing but a Guttman scale, i.e. a set of strictly ordered latent classes.

The paper by Templin and Bradshaw (2013) came full circle. It started out by suggesting a model with multiple attributes to replace the customary unidimensional analytical approach. The intermediate step is a model with attribute-pattern constraints that shows for the data that a model with only four strictly ordered attribute patterns with non-zero probabilities fits almost as well as the model with an unconstrained attribute space. Then we are introduced to unidimensional DCMs, that is, located latent class models, and it is indeed shown that these (just like the 2PL) fit the data better than the originally assumed model with multiple attributes. While the authors reference constrained, located latent class models (e.g., Lindsay, Clogg, & Grego, 1991) as being used, the unnecessary introduction of a new name "unidimensional diagnostic classification model" (UDCM) obscures the fact that the original intent failed to fit a model with multiple attributes.

In the paper, the hierarchical, multiple attribute models are abandoned after it turns out that they do not fit the data better than simpler approaches. While HDCMs are still found in the tables, the performance of the HDCM in terms of criteria such as the Bayesian Information Criterion (BIC) (Schwarz, 1978) is not better than what unidimensional models can provide. The located latent class models (LLCM; e.g. Formann, 1989), called UDCM by Templin and Bradshaw (2013) paper, use four and five levels and outperform the HDCM. They perform not substantially better than the 2PL, the difference being due to the fact that likely Templin and Bradshaw (2013) did not use the assumption of a constrained ability distribution for the LLCM. This result is not unexpected. More specifically, multiple classification latent class models with four or five levels per attribute variable (von Davier, 2005) do perform almost as well as a multidimensional item response theory model with multivariate normal ability distribution (Haberman et al., 2008). Along the same lines, de Leeuw and Verhelst (1986), Lindsay et al. (1991), as well as Follmann (1988) have shown that the located latent class model produces model-data fit and item parameter estimates that are equivalent to the conditional Rasch model.

## 7. Conclusion

What do we learn? In contrast to what Templin and Bradshaw (2013) appear to suggest, we should start with the simplest possible model, rather than with a potentially overly complex model (Haberman & von Davier, 2006; Haberman et al., 2008; von Davier, 2009). Then, only if obvious indication of misfit is found, should we go to a less parsimonious model. This is a longstanding tradition, and scholars from William of Occam to Albert Einstein provide some of the well-known references to support this suggestion.

Most of all, we should try not to invent new names for old models in order to avoid confusion for ourselves and for future generations of researchers.

### References

DeCarlo, L.T. (2011). On the analysis of fraction subtraction data: the DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.

de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, *11*, 183–196.

Follmann, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika*, *53*, 553–562.

Formann, A.K. (1989). Constrained latent class models: some further applications. *British Journal of Mathematical & Statistical Psychology*, *42*, 37–54.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stouffer et al. (Eds.), *The American soldier: Vol. 4. Measurement and prediction*. New York: Wiley.

Haberman, S.J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, *5*, 815–841.

Haberman, S.J., & von Davier, M. (2006). Some notes on models for cognitively based attribute diagnosis. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1031–1038). Amsterdam: Elsevier.

Haberman, S.J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions* (Research Report 08-45). Princeton, NJ: Educational Testing Service.

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.

Hermelin, M., & Nyberg, K. (2012). Multidimensional linear distinguishing attacks and Boolean functions. *Cryptography and Communications*, *4*(1), 47–64.

Hox, J. (2002). *Multilevel analysis: techniques and applications*. Mahwah: Erlbaum.

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.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Lazarsfeld, P.F. (1950). The logical and mathematical foundations of latent structure analysis. In S.A. Stouffer et al. (Eds.), *Measurement and prediction* (pp. 362–472). Princeton: Princeton University Press.

Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Leighton, J.P. & Gierl, M.J. (Eds.) (2007). *Cognitive diagnostic assessment for education. Theory and applications*. Cambridge: Cambridge University Press.

Lindsay, B., Clogg, C.C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*, 96–107.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.

Raudenbush, S.W., & Bryk, A.S. (2001). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.

Rojas, G., & de la Torre, J. (2012). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Unpublished manuscript, Rutgers University, New Brunswick, NJ.

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

Schwarz, G.E. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Templin, J.L., & Bradshaw, L. (2013). Hierarchical diagnostic classification models: a family of models for estimating and testing attribute hierarchies. *Psychometrika*. doi:10.1007/s11336-013-9362-0.

Tsai, S.C. (1996). Lower bounds on representing Boolean functions as polynomials in $Z_m$. *SIAM Journal on Discrete Mathematics*, *9*, 96–101.

van der Linden, W. (2012). On compensation in multidimensional response modeling. *Psychometrika*, *77*(1), 21–30.

Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*(1), 213–239.

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

von Davier, M. (2007). *Hierarchical general diagnostic models* (Research Report 07-19). Princeton, NJ: Educational Testing Service.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement*, *7*(1), 67–74.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling*, *52*(1), 8–28. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2010/02_vonDavier.pdf.

von Davier, M. (2011). *Equivalency of the DINA model and a constrained general diagnostic model* (Research Report 11-37). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-11-37.pdf.

von Davier, M. (2013). The DINA model as a constrained general diagnostic model—two variants of a model equivalency. *British Journal of Mathematical & Statistical Psychology*. doi:10.1111/bmsp.12003.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models—foundations, recent developments and applications* (pp. 371–379). New York: Springer.

von Davier, M., Xu, X., & Yamamoto, K. (2011, July). *Why diagnostic models are latent class models—or the other way around*? Paper presented at the many faces of latent class analysis symposium, 75th international meeting of the Psychometric Society, Hong Kong, China.

*Manuscript Received: 11 JUN 2013*
*Published Online Date: 30 JAN 2014*