In general, this review paper summarizes the key components in the application of the DCMs. It provides practitioners with some important information regarding conceptual framework, utility, application settings, and parameter estimation of the DCMs. This paper is a good addition to the review literature on diagnostic modeling.

ACKNOWLEDGMENTS

The author would like to express his deepest gratitude to his father Jinkuan Jiao for his love, inspiration, and encouragement.

REFERENCES

Anozie, N. O., & Junker, B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. Presented at the Workshop on Educational Data Mining at the 21st National Conference on Artificial Intelligence (AAAI 2006). Boston, USA. July 16–17, 2006.

Fu, J., & Li, Y. (2007, April). An integrated review of cognitively diagnostic psychometric models. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Henson, R. A., & Templin, J. (2007, April). Large-scale language assessment using cognitive diagnosis models. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Henson, R. A., Templin, J., & Willse, J. (2007, April). Defining a family of cognitive diagnosis models. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Junker, B. W. (2007). Some issues and applications in cognitive diagnosis and educational data mining (PowerPoint). Keynote presentation to the 2007 Annual Meeting of the Psychometric Society, Tokyo, Japan.

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*(4), 219–262

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (Research Report No. RR-06-08). Princeton, NJ: Educational Testing Service.

Some Notes on the Reinvention of Latent Structure Models as Diagnostic Classification Models

Matthias von Davier

Educational Testing Service
Princeton University, Princeton, New Jersey

INTRODUCTION

Rupp & Templin (2008) have done a remarkable job with their paper. The overview tries and to a large extent succeeds in being a balanced one—one that does not only include developments

Correspondence should be addressed to Matthias von Davier, Princeton University, ETS, MS 02-T, Princeton, NJ 08541. E-mail: mvondavier@ets.org

originating from one school of thought. Given the complexity of the field and the speed of publication of developments, this balance is quite an accomplishment. Understandably, other reviews of cognitive diagnosis models (DiBello, Roussos & Stout, 2007; Fu, 2007) also struggle with the inclusion of all recent developments. Yet, others approached the issue by deliberately providing only a comparison of a selective subset of cognitive diagnostic models (von Davier, DiBello, & Yamamoto, 2006). Apart from model comparisons, there are many precautions that should be considered when applying latent variable models with multidimensional skill profiles, some of which have been discussed in Haberman and von Davier (2007). I will take advantage of the opportunity provided here to point out a few issues, one being that there are models mislabeled as diagnostic, which deal with linear decompositions of item difficulties rather than estimating multidimensional skill variables. Then, I will discuss the issue that there are many new names for essentially well-known models for multiple simultanous classifications. I will also comment on models that combine multiple skills deterministically, thus mapping multiple skill patterns onto the same conditional probabilities. I will then suggest additional recent and not so recent literature on more exploratory approaches used for diagnosing strategy differences and close the commentary with some final remarks.

DIAGNOSTIC MODELS VS. LINEAR CONSTRAINTS

As in other reviews of cognitive diagnosis models, Rupp & Templin's paper also (mis-)classifies as cognitive diagnosis models approaches that constrain item parameters. The most well-known, constrained version of the unidimensional Rasch model is the linear logistic test model (LLTM: Fischer, 1973). Note that the LLTM does *not* provide multidimensional skill profiles or any other cognitive information about examinees apart from an overall unidimensional ability estimate equivalent to the Rasch model. Therefore, the LLTM should *not* be considered a cognitive diagnosis model. The LLTM is much rather an *item-diagnostic* model that decomposes (read: constrains!) item difficulties into fewer components, while the ability variable is still defined as in the Rasch model. The set of item features in the LLTM may be interpreted as attributes or factors driving item difficulty. However, this set of item features does not vary over examinees, so there is no multidimensional skill variable in approaches like the LLTM. This does not limit in any way the immense contribution of the LLTM to educational measurement, but it does provide a clearer distinction between item feature models and models that diagnose/measure/assume multiple person features or skill variables. Cognitive diagostic models provide multiple ability components, whereas item-difficulty decompositions maintain a *single* (*unidimensional*) skill variable. For example, linearly constrained difficulties can be used in unidimensional two-parameter logistic (2PL) or three-parameter logistic (3PL) item response theory (IRT) and are used in the LLTM, a constrained unidimensional Rasch model.

ON THE TERM "DIAGNOSTIC CLASSIFICATION MODELS"

Many names are attached to currently discussed models for cognitive diagnosis. As Rupp & Templin (2008) point out, the roots of the currently discussed or rediscovered models go back three decades or more. I chose the term *general diagnostic model* (GDM; von Davier & Yamamoto,

2004a; von Davier, 2005) for my developments, which are no exception here. A new name for a family of models may help to distinguish the approach, but may also be counterproductive in helping to understand relationships to existing approaches. Researchers have pointed out that the ordinal version of the GDM is (only) a multidimensional, mixture distribution, confirmatory, discrete version of a generalized partial credit model. Even though this is an accurate description, it is not as handy as the invented name. The same holds for most other names such as cognitive diagnosis models, skill profile models, cognitively diagnostic models, and other variations and for more specific models such as NIDA, DINA, NIDO, etc.

I'd like to suggest a radical step: let us go back to the term latent structure analysis (Goodman, 1974; Haberman, 1979; Lazarsfeld & Henry, 1968), and use this term for all models discussed here. The disadvantage of not inventing new names is that we cannot use this technique to distinguish our developments from others, whereas the advantage is, in my view, a much bigger one. New generations of researchers using these models will not get confused, and we will not work in subcultures or isolated schools who all develop their own vocabulary and work only within the confines of one approach, while missing the opportunity to learn about other applications of essentially the same model family.

In that regard, again I'd like to praise Rupp & Templin (2008); they do a good job pointing out that the models share more in common than what may distinguish them. The foundation of all of the models can be easily expressed in terms of a slight variation of the latent class model equation. Let $X_1, \ldots X_{\mathrm{I}}$ denote the observed response variables, with responses $x_i \in X_i$, and let $A = \{a_1, \ldots, a_K\}$ be a discrete latent variable. Let P denote a discrete probability measure over $P(A$, and assume that the conditional response probabilities $P(x_1, \ldots, xI|a_k)$ are well defined for all a_k. At this point, the elements of A are not (yet) specified in any way. They may be real numbers or vectors or nominally distinguishable elements (latent classes) of any type. Then,

$$P(x_1, \ldots, x_I) = \sum_{a \in A} P(a) p(x_1, \ldots, x_I \mid a)$$

defines a probability measure over $X = \prod_{i=1}^{I} X_i$, that is, the space of all possible response patterns. This equation is fundamental to all diagnostic models. Very often, there will be additional assumptions about the specific shape of the discrete latent class variable A, as well as other, additional assumptions simplifying the form of the conditional probabilities. For example, many approaches assume

$$p(x_1, \ldots, x_I \mid a) = \prod_{i=1}^{I} p(x_i \mid a),$$

that is, local independence of response variables given latent class a. In addition, one may assume that the latent variable takes on a specific form. In the vast majority of diagnostic models, A is broken down into a product space of several mastery/nonmastery variables $A = \{0,1\}^J$ and $a = (b_1, \ldots b_J)$, with $b_j \in \{0,1\}$. In this case, we speak of a as representing J binary latent skills, each of which can take on two states. Nevertheless, the basic structure is

one of a latent class model with 2^J latent classes. More general approaches (von Davier, 2005; Fu, 2005) allow more than two levels, so that we may see latent structures of the form

$$A = \prod_{j=1}^{J} \Omega_j$$

with polytomous $\Omega_j = \{0,1, \ldots ,o_j\}$. Haberman, von Davier, & Lee (2008) compare performance and estimation of such polytomous discrete models with multidimensional IRT (MIRT) models and find that models with surprisingly small o_j are already quite competitive with MIRT models.

SOME NOTES ON DETERMINISTIC INPUT (DI) MODELS

The differences between diagnostic models become evident only at the level of defining the $P(x_i|b_1, \ldots b_J)$, that is, at the level of specifying how the joint set of $b_1, \ldots b_J$ affects the response probabilities. Common to most diagnostic models is a design matrix (often called Q-matrix) that specifies which b_i are involved at all. Hence, the term confirmatory, since this design matrix specifies what skills are involved in responses to which items, as a design matrix in confirmatory factor models does. In addition, assumptions are put in place of how the skills involved work together. The appendix in the Rupp & Templin (2008) paper gives an overview of the different model equations.[1]

A comparison of approaches using the model-based conditional probabilities shows that the DINA and DINO are actually quite restrictive, which can be illustrated by the following example. Suppose there is a test with ten dichotomous items. Suppose further that there are three skills ($A1$,$A2$,$A3$) with two levels, mastery (1) and nonmastery (0). Each skill is measured by a different subset of items, with one pure item for each skill, and there is one item requiring all three skills. The conditional probabilities of this item are given in Table 1 for the DINA and DINO, using g for guessing and s for the "slipping" probability, as defined in the tabulation in Rupp and Templin (2008). The entries p_{A1A2A3} denote the probabilities of incorrect (X = 0) and correct (X = 1) responses, that is $p(X = x \mid q = (1,1,1), A = (A1, A2, A3))$ for the general case.

The latent structure of this three-skill model has $8 = 2^3$ elements, each of which represents a skill profile and potentially maps to a distinct probability of a response to an item. The unconstrained latent structure model with eight latent classes has 87 parameters, while diagnostic models, being constrained latent class models, typically have fewer parameters. Ten dichotomous items can be responded to in $1024 = 2^{10}$ different ways, so that the saturated model for this case has 1,023 parameters. When comparing the general latent class case with the DINA and DINO models, there are eight conditional probabilities for $P(x = 1 \mid a1, a2, a3)$, in the general case, and only two probabilities for the DINA and DINO.

[1]Note that in some cases the constraints listed in the table on the slope parameters γ may not be what the original authors had in mind, since these parameters are often assumed to be nonnegative.

TABLE 1
Probabilities under Deterministic Input (DI) and the General Case for an Item with Three Skills Required and All Skill Combinations.

p(X = 0)			*p(X = 1)*			*Skills*		
DINA	*DINO*	*general*	*DINA*	*DINO*	*general*	*A1*	*A2*	*A3*
1-g	1-g	$1-p_{000}$	g	g	P_{000}	0	0	0
1-g	s	$1-p_{100}$	g	1-s	p_{100}	1	0	0
1-g	s	$1-p_{010}$	g	1-s	p_{010}	0	1	0
1-g	s	$1-p_{110}$	g	1-s	p_{110}	1	1	0
1-g	s	$1-p_{001}$	g	1-s	p_{001}	0	0	1
1-g	s	$1-p_{101}$	g	1-s	p_{101}	1	0	1
1-g	s	$1-p_{011}$	g	1-s	p_{011}	0	1	1
s	s	$1-p_{111}$	1-s	1-s	p_{111}	1	1	1

This comparison shows that the specific choice of one or the other type of constraints can have a profound effect. Compared to the general case, the DINA and DINO provide very parsimonious models, however, potentially at the cost of model-data fit. Although space restrictions prevent detailed discussions, note that most of the other models discussed in Rupp & Templin (2008) are closer to the general latent class case, and less restrictive than the DI* models, which arrive at this level of conformity across conditional probabilities by the mapping of all skill vector components on only two possible values.

LATENT STRUCTURES AND LATENT TRAITS

Rupp and Templin's (2008) review briefly discusses mixture distribution IRT models (Yamamoto, 1989; Mislevy & Verhelst, 1990, Rost, 1990; Kelderman & Macready, 1990) and the use of these models for diagnostic classifications. Mixture IRT models have indeed been used to identify strategy differences (see, for example, Kelderman et. al., 1990; Rost & von Davier, 1993; Embretson, 2007; Rijkes & Kelderman, 2007) by modeling the latent structure jointly as a combination of a continuous latent trait and additional classifications. Mixtures of IRT models have been developed for the dichotomous Rasch model (Rost, 1990), for ordinal constrained polytomous Rasch models (such as the rating scale model and the dispersion model; von Davier & Rost, 1995), as well as for the 2PL and 3PL IRT models (Mislevy & Verhelst, 1990), and the generalized partial credit model (von Davier & Yamamoto, 2004b). These models have been used to identify groups of examinees that change their response strategies induced by speeded tests (Yamamoto & Everson, 1997; Bolt, Cohen, & Wollack, 2002; Boughton & Yamamoto, 2007) as well as induced by surface features or cognitive demands of items (Rost et al., 1993; Rijkes & Kelderman, 2007).

The defining difference to models termed diagnostic classification models is that mixture IRT models tend to be used in an exploratory manner, uncovering differences between groups based on systematic response tendencies, whereas diagnostic models take a confirmatory approach (von Davier & Rost, 2006; von Davier & Yamamoto, 2007). A notable exception to

this rule is the Saltus model (Wilson, 1989, Draney & Wilson, 2007), which can be understood as a mixture IRT model with a confirmatory structure that reflects shifts in item difficulties based on stage-like theories of development. The distinction between exploratory mixture IRT and confirmatory "diagnostic" analysis leads to the suggestion that ordinary IRT, and potentially MIRT, should be applied in the first steps of data analysis. This should be done to investigate whether the hidden structure in the item response data can be explained by a single latent trait. A careful analysis with a confirmatory approach involving a few skills should be attempted only if there is indication that latent groupings with respect to more than one variable seem necessary to fit the data.

HOW TO TRY OUT THESE APPROACHES

A simple way to begin trying out some of the diagnostic models discussed in Rupp & Templin (2008) is to use datasets in conjunction with diagnostic models involving only small numbers of skills. A general purpose latent class analysis program like LEM (Vermunt, 1997) or more recently LatentGOLD (Vermunt & Magidson, 2005) can be used for those cases by implementing the skill model as a constrained latent structure. Alternatively, models that share many characteristics with the ones discussed in Rupp and Templin (2008) can obviously be specified in general latent variable frameworks like the ones by DeBoeck & Wilson (2004), Skrondal & Rabe-Hesketh (2004), or Moustaki (2003).

Other researchers have provided special purpose packages or add-ons to existing software: De La Torre and Douglas (2004) provided an add-on to a computational programming language (OX), and Templin & Hanson (2006) presented a front-end to MPLUS that enables specification of some approaches discussed in the Rupp and Templin (2008) paper. Some of the fusion model or reparameterized unified model (RUM) approaches discussed by Rupp and Templin (2008) can be estimated with special purpose software commercially available as Arpeggio suite. The GDM and latent class analysis, as well as IRT, discrete MIRT, and mixture IRT models can be estimated using the stand-alone software multidimensional discrete latent trait models (mdltm) (von Davier, 2005), which is available free of charge for research and teaching purposes upon request from Educational Testing Service (ETS) (mailto: mvondavier@ets.org).

CONCLUSION

Rupp and Templin (2008) have done a great job explaining the different approaches, the rationales for their use, and the differences between these models. They maneuver the landscape of diagnostic models with great skill and present readers with a choice of models that gives rise to the expectation of gaining deeper insight into a student's strengths and weaknesses. My impression is that this paper will contribute to the ongoing discussions about the appropriate use of multiple skill models. There is a host of models being discussed in the literature, and Rupp & Templin (2008) have made a great case for the potential and expected utility of these approaches. What we often lack is a strong theory that provides hypotheses about complex skill domains. Examples of assessments put together using items that tap into multiple skills are scarce, as traditional approaches to item development use task characteristics that involve only

one ability variable. The full potential of using constrained latent structure models for skill diagnosis will become evident once more assessments are built to explicitly measure multiple skills.

REFERENCES

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. Journal of Educational Measurement, 39(4), 331–348.

Boughton, K., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier and C. H. Carstensen (Eds.) *Multivariate and mixture distribution Rasch models*. New York: Springer.

DeBoeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer: New York.

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

Draney, K., & Wilson, M. (2007). Application of the saltus model to stage-like data: Some applications and current developments. In M. von Davier and C. H. Carstensen (Eds.) *Multivariate and mixture distribution Rasch models*. New York: Springer.

Embretson, S. (2007). Mixed Rasch models for measurement in cognitive psychology. In M. von Davier and C. H. Carstensen (Eds.) *Multivariate and mixture distribution Rasch models*. New York: Springer.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359–374.

Fu, J. (2005). *A polytomous extension of the fusion model and its bayesian parameter estimation*. Unpublished dissertation.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 61, 215–231.

Haberman, S. J. (1979). *Qualitative data analysis* (Vols. 1 & 2). New York, Academic Press.

Haberman, S. J., & von Davier, M. (2007). A Note on models for cognitive diagnosis. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26), Psychometrics. 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. RR-08-45. ETS Research Report.

Kelderman, H., & Macready, G. B. (1990). The use of log-linear models for assessing differential item functioning across manifest and latent examinee groups. Journal of Educational Measurement, 27, 307–327.

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

Mislevy, R. J., & Verhelst, N. D. (1990). Modeling item responses when different subjects employ different solution strategies. Psychometrika, 55(2), 195–215.

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. British Journal of Mathematical and Statistical Psychology, 56, 337–357.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. Applied Psychological Measurement, 14, 271–282.

Rost, J., & von Davier, M. (1993). Measuring different traits in different populations with the same items. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric methodology*. Proceedings of the 7th European meeting of the Psychometric Society in Trier. Stuttgart: Gustav Fischer Verlag.

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*(4), 219–262

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

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

Vermunt, J. K. (1997). *LEM1.0: Ageneral program for the analysis of categorical data*. Tilburg, The Netherlands: Tilburg University.

Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 user's guide*. Belmont, MA: Statistical Innovations.

von Davier, M. (2005) A general diagnostic model applied to language testing data. ETS Research Report RR-05-16.

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., & DiBello, L., & Yamamoto, K. (2008). Reporting test outcomes using models for cognitive diagnostics. In J. Hartig, E. Klieme, & D. Leutner (Eds.). *Assessment of competencies in educational contexts* (pp. 151–176). Cambridge, MA: Hogrefe & Huber Publishers.

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In: C. R. Rao & S. Sinharay (Eds.). *Handbook of statistics*, Vol. 26. Psychometrics. Amsterdam: Elsevier–North Holland.

von Davier, M., & Yamamoto, K. (2004a). A class of models for cognitive diagnosis. Paper presented at the invitational ETS Spearman Conference, Philadelphia, PA.

von Davier, M., & Yamamoto, K. (2004b). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. Applied Psychological Measurement, 28(6), 389–406.

von Davier, M., & Yamamoto, K. (2007) Mixture distribution and HYBRID Rasch models. In M. von Davier and C. H. Carstensen (Eds.) *Multivariate and mixture distribution Rasch models*. New York: Springer.

Wilson, M. (1989). Saltus: A psychometric model for discontinuity in cognitive development. Psychological Bulletin, 105, 276–289.

Yamamoto, K. Y. (1989). *HYBRID model of IRT and latent class model* (ETS Research Report RR-89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K. Y., & Everson, H. T. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). Muenster, Germany: Waxmann.