

**Theoretical and Practical Considerations for
Implementing Diagnostic Classification Models:
Insights from Simulation-based and Applied Research**

D i s s e r t a t i o n
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)
im Fach **Psychologie**

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät II
der Humboldt-Universität zu Berlin

von

Dipl. Psych. Olga Kunina-Habenicht

Präsident der Humboldt-Universität zu Berlin

Prof. Dr. Christoph Marksches

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II

Prof. Dr. Peter Frensch

Gutachter:

1. Prof. Dr. Matthias Ziegler
2. Prof. Dr. Oliver Wilhelm
3. Prof. Dr. André A. Rupp

Tag der mündlichen Prüfung: 03.06.2010

Table of content

Table of content.....	2
List of abbreviations.....	3
Zusammenfassung.....	4
Abstract	5
Introduction	6
Diagnostic classification models.....	8
Previous applications.....	10
Methodological challenges and recent developments.....	11
Overview over dissertation project	14
Kognitive Diagnosemodelle: Theoretisches Potential und methodische Probleme.....	17
A Practical Illustration of Multidimensional Diagnostic Skills Profiling: Comparing Results from Confirmatory Factor Analysis and Diagnostic Classification Models.....	18
Abstract	19
Sensitivity of Item and Respondent Parameter Estimation to Model Misspecification within a Log-linear Modelling Framework for Diagnostic Classification Models	20
Abstract	21
Added Predictive Value of Multidimensional Proficiency Scores from Diagnostic Classification Models for Global Proficiency Indicators in Elementary-school Mathematics.....	22
Abstract	23
General Discussion.....	24
Main results	25
Theoretical Research Questions Regarding the Robustness of the Fit Assessment for DCMs.....	26
Applied Research Questions Regarding the Use of DCMs for the Diagnostic Assessment Data.....	28
Conclusions and Outlook	30
References	33

List of abbreviations

AIC	Akaike information criterion
BIB	Balanced incomplete block
BIC	Bayesian information criterion
CAT	Computerized adaptive testing
CDM	Cognitive diagnostic models
CFA	Confirmatory factor analysis
CFI	Comparative fit index
DCM	Cognitive diagnostic classification models
DIF	Differential item functioning
DINA	Deterministic inputs noisy and-gate
DINO	Deterministic inputs noisy or-gate
DMA	Diagnostic mathematics assessment
GDM	General diagnostic model
IRT	Item response theory
LLTM	Latent logistic test model
MAD	Mean absolute deviation
NAEP	National assessment of educational progress
NES	National educational standards
NIDA	Noisy-input deterministic-and-gate
NIDO	Noisy-input deterministic-or-gate
PISA	Programme for international student assessment
RMSE	Root mean squared error
RMSEA	Root mean square error of approximation
RUM	Reparameterized unified model
SAT	Scholastic Assessment Test
TIMSS	Third international mathematics and science study
TOEFL	Test of English as a foreign language
WLE	Weighted likelihood estimator

Zusammenfassung

Kognitive Diagnosemodelle (DCMs) sind konfirmatorische probabilistische Modelle mit kategorialen latenten Variablen, die Mehrfachladungsstrukturen erlauben. Sie ermöglichen die Abbildung der Kompetenzen in mehrdimensionalen Profilen, die zur Erstellung informativer Rückmeldungen dienen können. Diese Dissertation untersucht in zwei Anwendungsstudien und einer Simulationsstudie wichtige methodische Aspekte bei der Schätzung der DCMs. In der Arbeit wurde ein neuer Mathematiktest entwickelt basierend auf theoriegeleiteten vorab definierten Q-Matrizen. In den Anwendungsstudien (a) illustrierten wir die Anwendung der DCMs für empirische Daten für den neu entwickelten Mathematiktest, (b) verglichen die DCMs mit konfirmatorischen Faktorenanalysemodellen (CFAs), (c) untersuchten die inkrementelle Validität der mehrdimensionalen Profile und (d) schlugen eine Methode zum Vergleich konkurrierender DCMs vor. Ergebnisse der Anwendungsstudien zeigten, dass die geschätzten DCMs meist einen nicht akzeptablen Modellfit aufwiesen. Zudem fanden wir nur eine vernachlässigbare inkrementelle Validität der mehrdimensionalen Profile nach der Kontrolle der Personenparameter bei der Vorhersage der Mathematiknote. Zusammengekommen sprechen diese Ergebnisse dafür, dass DCMs per se keine zusätzliche Information über die mehrdimensionalen CFA-Modelle hinaus bereitstellen. DCMs erlauben jedoch eine andere Aufbereitung der Information. In der Simulationsstudie wurde die Präzision der Parameterschätzungen in log-linearen DCMs sowie die Sensitivität ausgewählter Indizes der Modellpassung auf verschiedene Formen der Fehlspezifikation der Interaktionsterme oder der Q-Matrix untersucht. Die Ergebnisse der Simulationsstudie zeigen, dass die Parameterwerte für große Stichproben korrekt geschätzt werden, während die Akkuratheit der Parameterschätzungen bei kleineren Stichproben z. T. beeinträchtigt ist. Ein erheblicher Teil der Personen wird in Modellen mit fehlspezifizierten Q-Matrizen falsch klassifiziert.

Abstract

Cognitive diagnostic classification models (DCMs) have been developed to assess the cognitive processes underlying assessment responses. Current dissertation aims to provide theoretical and practical considerations for estimation of DCMs for educational applications by investigating several important underexplored issues. To avoid problems related to retrofitting of DCMs to an already existing data, test construction of the newly mathematics assessment for primary school DMA was based on a-priori defined Q-matrices. In this dissertation we compared DCMs with established psychometric models and investigated the incremental validity of DCMs profiles over traditional IRT scores. Furthermore, we addressed the issue of the verification of the Q-matrix definition. Moreover, we examined the impact of invalid Q-matrix specification on item, respondent parameter recovery, and sensitivity of selected fit measures.

In order to address these issues one simulation study and two empirical studies illustrating applications of several DCMs were conducted. In the first study we have applied DCMs in general diagnostic modelling framework and compared those models to factor analysis models. In the second study we implemented a complex simulation study and investigated the implications of Q-matrix misspecification on parameter recovery and classification accuracy for DCMs in log-linear framework. In the third study we applied results of the simulation study to a practical application based on the data for 2032 students for the DMA.

Presenting arguments for additional gain of DCMs over traditional psychometric models remains challenging. Furthermore, we found only a negligible incremental validity of multivariate proficiency profiles compared to the one-dimensional IRT ability estimate. Findings from the simulation study revealed that invalid Q-matrix specifications led to decreased classification accuracy. Information-based fit indices were sensitive to strong model misspecifications.

I

Introduction

“The purpose of learning is growth, and our minds, unlike our bodies, can continue growing as we continue to live.” (*Mortimer Jerome Adler*)

The primary goal of traditional educational tests is to make inferences about an individual test taker's ability with reference to other test takers in the normative group. In these assessments one-dimensional item response theory (IRT) models (de Ayala, 2009; Embretson & Reise, 2000) are often applied to model the response data, assuming that the observed test results can be sufficiently explained by a single latent dimension. In these studies typically single scores are reported allowing for accomplishing assessment goals, such as a ranked comparison of examinees. Such traditional testing has been criticized for not providing diagnostic information to inform students of their strengths and weaknesses in a specific academic domain (Snow & Lohman, 1989; see also Leighton & Gierl, 2007 and Nichols, Chipman, & Brennan, 1995 for a more general discussion). Ideally diagnostic assessments should not only meet the psychometric standards of current large-scale assessments, but should also provide specific diagnostic information regarding the individual examinees' skills.

Indeed, a recent US-survey has revealed, that there is a significant interest by stakeholders in education to obtain more detailed diagnostic information about the strengths and weaknesses of students' knowledge, skills, and abilities. Teachers have difficulties in interpreting the statistical scales and therefore problems with integrating the survey results in their instructions (Huff & Goodman, 2007).

Applications of cognitive diagnosis assessments aim to provide informative feedback (for an exhaustive review of formative feedback see Shute, 2008) to parents, teachers, and students, which can be used to direct and improve instruction (Embretson, 1991, 1998). Diagnostic assessments allow for testing process to also serve an instructional purpose in addition to the traditional purposes of assessment (Linn, 1990) and can be used to integrate instructions and assessments.

The cognitive diagnostic approach combines cognitive theories with statistical models intended to make inferences about test takers' mastery of tested skills. An assessment designed to evaluate learners' competencies in micro-level skills requires a much finer-grained representation of the construct of interest. Thus, this approach needs to be based on a substantive theory of the construct that describes the processes which learners apply to perform on tasks (Embretson, 1983). Ideally, it also requires clear specifications that delineate the tasks in terms of how they elicit cognitive processes (Embretson & Gorin, 2001; Mislevy, 2007, 2008; Mislevy et al., 2010).

Diagnostic classification models

Cognitive diagnostic classification models (DCMs) have been developed to study and assess the cognitive processes underlying assessment responses. In contrast to summative standards-based assessments, which are used to monitor educational systems by providing information about global proficiencies, cognitive diagnostic assessments seek to provide more fine-grained interpretations to support instruction and learning (Rupp, Templin, & Henson, 2010). Current dissertation focuses on some underexplored methodological issues in DCM context and aims to provide theoretical and practical considerations for estimation of DCMs for educational applications.

In this dissertation we refer to the definition of cognitive diagnostic classification models (DCMs) provided by Rupp and Templin (2008). They differentiate between the terms 'cognitive diagnostic models' (CDMs) and 'cognitive diagnostic classification models' (DCMs). DCMs form a subset of CDMs with statistical formulation and do not include classification approaches like the rule space methodology (e.g. Tatsuoka, 1983, 1995) and attribute hierarchical method (Leighton, Gierl, & Hunka, 2005; Gierl, Leighton, & Hunka, 2007).

In contrast to multidimensional IRT models DCMs are probabilistic confirmatory multidimensional models with categorical latent skills. According to Rupp et al. (2008) DCMs enable multiple criterion-referenced interpretations and feedback for diagnostic purposes that are referenced to a cognitively-grounded theory of response processes at a fine-grained size. Whereas Rupp et al. define DCMs as criteria-referenced models, Henson (2009) disagrees with that statement and claims that DCMs are not necessarily criteria-referenced but norm-referenced models instead. He argues that respondents' classification depends on item selection and person sample and that therefore persons are placed into a master or non-master group without a clear reference to a certain external criterion. Henson (2009) concludes that therefore, one cannot be sure that an individual classified as master will perform as master on some external criterion. Instead, one can only assume that masters will perform better than non-masters based on some defined criterion.

DCMs are also known as 'multiple classification (latent class) models' (e.g., Macready & Dayton, 1977; Maris, 1999), 'restricted latent class models' (e.g., Haertel, 1989, 1990), 'psychometric latent response models' (Maris, 1995), 'IRT based latent class models' (Rousses, Templin, & Henson, 2007), and 'structured IRT models' (e.g., Rupp & Mislevy, 2007).

One necessary requirement for the estimation of DCMs is the specification of cognitive operations or skills that are involved in the solution process of a certain item in a so called Q -

matrix. So far, we lack clear distinctions among terms such as 'attributes', 'processes,' 'skills' or 'subskills' and use these terms as synonyms in the following. The concept of Q-matrix was defined and illustrated by Tatsuo (1983) and is comparable to a weight matrix in the logistic latent test model (LLTM) approach (Fischer, 1973) or a loading matrix in factor analysis. Basically, it is a table that specifies attributes required for successful solving of a task item. Each element q_{jk} in the Q-matrix indicates whether mastery of attribute k is required for correctly solving item j . If attribute k is relevant for the successful solution of the item j then q_{jk} is 1, otherwise it is 0.

Q-matrices in which all items require only one attribute have the so called 'simple loading structure' that is also referred as *between-item multidimensionality*. By contrast, Q-matrices with items needing more than one attribute constitute a 'complex loading structure' that is also called *within-item multidimensionality* (Adams, Wilson, & Wang, 1997). DCMs were designed to handle within-item multidimensionality, thus in most cases Q-matrices with complex loading structure are used for these analyses.

Note that the definition of the Q-matrix represents the cognitive foundations of the particular model. For the specification of the Q-matrix in practice, standard setting procedures with expert panels are needed consisting of item developers, teachers, or domain experts. The use of such expert panels implicitly assumes that (a) experts can determine the attributes needed for correctly solving each item and (b) all relevant attributes were considered by the panel. Any change in the Q-matrix redefines the substantive interpretations of the set of user-specified attributes, even if their labels remain the same. An alternative technique for Q-matrix construction lies in the theoretical derivation of the item classification based on underlying cognitive theories (e. g. for matrix problems assessing figural reasoning as proposed by Embretson, 1998). The development or derivation of one or several competing Q-matrices is a critical and potentially the most challenging step in a DCM analysis (Gorin, 2009). Therefore, it is important that the construction of the Q-matrix is done with care.

In developing the Q-matrix, it is important to keep in mind how the defined skills for the particular assessment interact with each other (Roussos, Templin, & Henson, 2007). In the DCM context it is common to differentiate between *compensatory* and *non-compensatory models*. The difference between these models reflects how latent variables for different attributes are combined to produce the observed responses to test items. Compensatory models are often statistically defined by disjunctive condensation functions and allow that deficiency in one skill can be compensated for by mastery in another skill, while non-compensatory models use conjunctive condensation functions in the model definition

requiring that each relevant skill for the particular item has to be present in order to give a correct response to the certain item. For example, compensatory models are appropriate when attributes represent alternative strategies for solving an item. In this case, successful performance on the item requires that only one of the possible strategies be successfully applied. By contrast, applications of non-compensatory models are suitable when attributes represent different procedural skills or knowledge of certain rules that are all necessary for a successful item solution. Although the choice of condensation rule is crucial and should be provided by the diagnostic setting, the purpose of the assessment, and theoretical considerations about the response process, Wilhelm and Robitzsch (2009) state that in many applications the decision concerning the condensation rule is rather arbitrary.

In the last decades several different DCMs were developed and refined. Reviews by Junker (1999), diBello, Roussos, and Stout (2007), Roussos et al. (2007), and Rupp et al. (2008) provide exhaustive reviews of different DCMs and their statistical properties. Commonly used are non-compensatory Deterministic inputs noisy and-gate (DINA) and Noisy-input deterministic-and-gate (NIDA) models (e.g. Macready et al., 1977, Junker & Sijtsma, 2001; de la Torre, 2009a) as well as their compensatory counterparts Deterministic inputs noisy or-gate (DINO) and Noisy-input deterministic-or-gate (NIDO) models (e.g. Maris, 1999). Hartz (2002) developed the reduced re-parametrized unified model (RUM) which is also known as the 'Fusion model' (see Roussos et al., 2007).

Recently several integrative frameworks for DCMs were proposed, including, for example the flexible family of general diagnostic models (GDM, van Davier, 2005, 2007), the generalized DINA Model (de la Torre, 2008a), as well as the log-linear framework for DCMs (Henson, Templin, & Willse, 2009). Furthermore, several extensions of DCMs for multiple-choice items were suggested in DCM literature (e.g. Bolt & Fu, 2004 for the fusion model; Templin, Henson, Rupp, Jang, & Ahmed, 2009 for DCMs in log-linear framework; de la Torre, 2009b for the DINA model).

Previous applications

Despite their great potential in the educational measurement, the number of successful practical applications of DCMs has remained relatively small. For the domain of mathematics several analyses within the rule-space methodology were carried out for algebra assessments (Tatsuoka, Birenbaum, & Arnold, 1989; Birenbaum, Kelly, Tatsuoka, & Gutvirtz, 1994) and Third International Mathematics and Science Study (TIMSS) (Birenbaum, Tatsuoka, & Yamada, 2004; Tatsuoka, Corter, & Tatsuoka, 2004). So far only a few diagnostic assessments were specifically designed for providing diagnostic feedback (Gorin, 2007). Most

applications of DCMs are add-ons to simulation studies that focus on technical estimations aspects (e.g., de la Torre & Douglas, 2004; Henson et al., 2009). Several researchers retrofitted DCMs to data that was originally used in one-dimensional large-scale assessments such as the Test of English as a Foreign Language (TOEFL) (von Davier, 2005), National Assessment of Educational Progress (NAEP) (Xu & von Davier, 2006, 2008), or Preliminary Scholastic Assessment Test (PSAT) (Hartz, 2002) without discussing additional benefits of DCMs in comparison to traditional psychometric models. As the result of this, estimation problems frequently occur, including problems of non-convergence, highly correlated latent dimensions as well as low scale reliabilities (see also Haberman, 2008).

One example for a carefully conducted DCM application was provided by Jang (2005), who used both quantitative and qualitative approaches to identify nine primary reading comprehension skills by analysing think-aloud verbal protocols for the TOEFL test and fitted the Fusion model to the field test data to estimate skill mastery probabilities (see also Jang, 2009). A dataset on fraction subtraction that is frequently used in DCM literature was originally described by Tatsuoka (1990) and more recently by de la Torre et al. (2004) and Henson et al. (2009), consisting of 20 items involving eight relevant attributes for subtraction of fractions. Furthermore, Templin et al. (2009) estimated a compensatory RUM model applying the log-linear DCM approach for a large-scale diagnostic assessment measuring reading proficiency in English as foreign language using a Q-matrix with 15 multiple choice items and three attributes. Note that estimation of DCMs is not limited to applications in the educational field but can also be applied in the psychiatric domain. For example, Templin and Henson (2006) applied structured DCMs with reduced model complexity to the data from the assessment of pathological gambling disorders.

Methodological challenges and recent developments

In the DCM context some critical methodological issues concerning the item and person parameter estimation need to be elaborated upon more in the future. One serious problem refers to the exponential growth of model complexity with the increase in number of dimensions leading to insufficient robustness of the parameter estimates in complex DCMs. More precise, for K specified attributes 2^K possible answer patterns, also denoted as *latent classes*, exist, assuming that respondents in the same latent class have equal probabilities for a correct response to the certain items. Moreover, currently it remains unknown which test length is required for a sufficient parameter recovery given a specific Q-matrix with a certain number of relevant attributes. In addition, Maris & Bechger (2009) point out the seldom considered issue of equivalence of different DCMs within and across different DCMs.

The second issue that needs to be addressed relates to the comparison of DCMs with the established IRT and confirmatory factor analysis (CFA) models (Kline, 2005; Kaplan, 2000). Sinharay and Haberman recommend using DCMs only if the model approximates data better than more parsimonious and computationally less demanding models. Due to the higher model complexity of DCMs it is important to present strong evidence that these models justify the higher administrative and operation costs for the test development and implementation. Another related issue in this context that has not attracted much attention so far refers to the incremental validity of the multidimensional profiles estimated in the DCM over and above traditional scores (e.g. Sinharay & Haberman, 2009).

The third aspect that requires more attention concerns the examination of the reliability and validity of the Q-matrix specification. At the first glance arguable advantages of DCMs are statistical estimation of student classification and lack of human judgement the locating the cut scores between mastery and non-mastery (Gorin, 2009). However, it is important to make clear that in context of DCMs in spite of apparently clear statistical judgement, the insecurity of the classification is capitalized in the validity and reliability of Q-matrix definition. In the DCM context in most cases additionally clinical judgement in terms of expensive and time-consuming standard setting procedures is required for a reliable Q-matrix definition (for the discussion of clinical and statistical judgement see Meehl (1954) and more recently Dawes, Faust, & Meehl (1989)). The quality of the standard setting procedure (for a general overview see Cizek & Bunch, 2007) has an enormous impact on the reliability of the resulting Q-matrix. Depending on the expertise domain, practical experience, and theoretical focussing of experts involved in the panel, several different relevant attributes and cognitive operations will be likely proposed during the standard setting procedure. Previous studies have shown that reliability of classification in standard setting procedures depends on the composition of the expert panel and used standard setting method (Pant, Rupp, Tiffin-Richards, & Köller, 2009; Tiffin-Richards, Pant, & Köller, 2010). Thus, it might be difficult to achieve a consensus in a heterogeneous panel of experts from different disciplines (e.g. psychologists, didactics, and school teachers) due to different theoretical considerations in different disciplines leading to divergent attribute definitions. Although verification of Q-matrix definition is an essential issue that needs to be considered, in most DCMs applications, Q-matrices per definition are assumed to be correct and therefore the uncertainty regarding attribute definition are not taken into account. The plausibility of Q-matrix definitions is seldom evaluated in practice, certainly because no reliable verification methods are available so far allowing for testing the correctness of the Q-matrix (Gorin, 2009).

Thus, the fourth underexplored question in DCM context refers to the goodness-of-fit evaluation. In contrast to the CFA framework, there are no established global model fit indices or item discrimination indices available for DCMs at the moment. Goodness of fit measures that allow for model evaluation typically involve comparisons of deviations between expected and observed probabilities and calculation of Chi-squared statistics that are used for fit evaluation. However, assumptions for reliable application of Chi-squared statistics are mostly violated for models where a great number of possible answer patterns exist. Thus, estimation of the global model fit indexes can be affected for log-linear DCMs. Recently, several authors suggested calculation of a similar measure on the more aggregated level taking into account the difference between model-based and observed probabilities over all items or over all latent classes (e. g. Templin et al., 2006, Dimitrov, 2007; Henson et al., 2008). In addition, Sinharay, Almond, and Yan (2004), Sinharay (2006), and Sinharay and Almond (2007) introduced several model diagnostics for so called 'Bayesian networks' when estimating DCMs in the Bayesian framework. Furthermore, de la Torre (2008c) discussed several techniques for comparing models and assessing goodness of fit. Moreover, first attempts were made to address the problem of Q-Matrix validation for the DINA Model by de la Torre (2008b). Despite of the efforts to address the issue of the model fit evaluation, very little is known about the sensitivity of the proposed model fit measures toward invalid Q-matrix specification.

Another major issue of linking and equating for DCMs is worth studying and has not attracted much attention so far. Linking and equating techniques are typically embedded in longitudinal data designs and allow for adjustment of item difficulty in different test booklets, different samples, or different repeated measures over several time points (for general overview see Kolen & Brennan, 2004). Recently, Xu et al. (2008) investigated the efficiency of several linking procedures for the GDM. However, in particular the question concerning the sufficient number of shared or common items, so called "anchor items", that facilitate the equation process, is investigated only insufficiently for DCMs so far.

Furthermore, the quantification of the reliability of multidimensional proficiency profiles for DCMs is an area of ongoing research (e. g. Templin, 2009; Templin & Henson, 2009). Note, that the transition of the reliability concept from traditional psychometric models is not directly possible because of different conceptualizations of true scores in models with continuous and categorical latent variables (Rupp & Templin, 2009).

Moreover, first efforts were made to explore the possibilities of computerized adaptive testing (CAT, for a general overview see van der Linden & Glas, 2000) for DCMs in several

simulation studies (e. g. McGlohen, 2004; Finkelman, Kim, Roussos, & Verschoor, 2008; Cheng, 2009). CAT is a special approach to the assessment of latent traits in which the selection of the test items that are presented to the examinee is based on the responses the examinee gave on previously administered items (Wainer, 2000). The main advantage of CAT lies in its capability to increase measurement efficiency substantially by reducing the number of presented items in comparison with a conventional test with a fixed number of items in a fixed order.

Overview over dissertation project

The current dissertation project differs from previous DCM applications by investigating and integrating several important underexplored issues in the DCM context. First, in this project DCMs were not retrofitted to any existing data but instead the test construction of the newly mathematics assessment was based on a-priori defined Q-matrices. Second, we compared DCMs with established psychometric models and addressed the validity issue of the DCM profiles as well as verification of the Q-matrix definition. Third, we examined the impact of the invalid Q-matrix specification on item, respondent parameter recovery, and the sensitivity of the selected fit measures in a complex simulation study.

The new diagnostic assessment DMA in the current dissertation was developed based on didactic theories for teaching mathematics in primary school. It was supposed to allow for detailed feedback on basic arithmetic competencies (i.e., technical counting skills in addition, subtraction, multiplication, division; modelling skills in word problems) for students in grades 3 and 4. Contrary to most previous applications, in this dissertation the item construction was based on theoretically a-priori defined Q-Matrices differentiating between the basic skills “addition”, “subtraction”, “multiplication”, “division”, “executing inverse operation”, “executing carry over”, “solving word problems”, and “converting measurement units”. These attributes were partly based on the approach proposed by Carpenter, Fennema, Franke, Levi and Empson (1999) as well as on didactic reports of typical conceptual mistakes children make in basic arithmetic tasks. For each item the complex loading structure of the required attributes for the successful solution was represented via the Q-matrix.

We have piloted DMA in the sample of 241 children in third and 223 children in fourth grade from six German primary schools in April 2008. After item revision process we conducted a main field test with a sample of 2032 students in fourth grade from 46 schools in different German federal states using a complex balanced multi-matrix design with 21 booklets. In addition to DMA some examinees additionally worked on an established

mathematical test assessing German national based standards in mathematics in primary school and/or on a verbal cognitive ability test.

The current cumulative dissertation contains four manuscripts. First manuscript in chapter two provides a short introduction to DCMs and a general overview over the dissertation project. Chapters 3, 4, and 5 present design and main findings of three studies addressing the following five research questions organized into two distinct subsets:

Theoretical Research Questions Regarding the Robustness of the Fit Assessment for DCMs

- 1) How biased and efficient are the item parameter recovery and the respondent classification under different conditions of model misspecification for core DCMs?
- 2) How do different global and local indices for the model-data fit evaluation perform under different conditions of model misspecification?

Applied Research Questions Regarding the Use of DCMs for the Diagnostic Assessment Data

- 3) Can data for the newly developed diagnostic assessment be successfully scaled with DCMs in order to create reliable multivariate attribute profiles for each student?
- 4) How similar are the resulting multivariate attribute profiles in comparison with the results of corresponding multidimensional CFA models?
- 5) Can the resulting multidimensional attribute profiles based on the best fitting DCM explain additional variance in the prediction of school grades in mathematics and performance in a national standard based assessment in Germany for mathematics over and above unidimensional proficiency scores in mathematics?

In the first study presented in chapter three we have applied several different DCMs within the GDM framework to pilot data for DMA and compared this model to the corresponding multidimensional CFA model. This study focussed on the third and fourth research questions and investigated the estimation accuracy and model fit of DCMs. Moreover, in this study we addressed the question whether or not DCMs can provide additional information beyond the established psychometric models.

In the second study in chapter four we implemented a complex simulation study with 32 conditions and investigated parameter recovery and classification accuracy for correct and wrongly defined log-linear DCMs. In addition we tested the sensitivity of the information-based fit indices and item discrimination fit indices toward model misspecification. With respect to model misspecification, we differentiated between two invalid definitions of

interaction effects and two types of Q-matrix misspecification. The basic manipulated test design factors included among others the number of respondents (1,000; 10,000), attributes (3; 5), and items (25; 50).

This simulation study addressed the first two research questions and differs from previous simulation studies in several aspects. First, in contrast to previous simulation studies, we have generated several random misspecified Q-matrices to be able to generalize the impact of model misspecification. Additionally, we have proposed extensions for two item discrimination indexes. Previous studies either introduced new model fit indices or investigated the impact of Q-matrix misspecification on accuracy of parameter recovery and examinees' classification. To our knowledge our study is the first study that combines both research questions and investigates the sensitivity of the proposed model fit measures toward two different types of model misspecification.

In chapter five we present a third study where we applied results of the simulation study to data from the main field test study for DMA. One important aim of this study was to illustrate a successful application of DCMs in log-linear framework by considering model fit measures suggested in the simulation study. The second goal of the study was to investigate the added predictive value of multidimensional profiles based on DCMs over and above traditional IRT scores in the (a) prediction of school grades in mathematics and (b) prediction of the performance in a standards-based large-scale assessment of mathematics.

The dissertation is concluded by the discussion of main results and a critical evaluation of implications of these findings for future theoretical and practical work.

II

Kognitive Diagnosemodelle: Theoretisches Potential und methodische Probleme

Olga Kunina-Habenicht

Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt Universität zu Berlin

Oliver Wilhelm

Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt Universität zu Berlin

Franziska Matthes

Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt Universität zu Berlin

André A. Rupp

Department of Measurement, Statistics, and Evaluation (EDMS), University of Maryland

Manuscript in press in *Zeitschrift für Pädagogik*

III

A Practical Illustration of Multidimensional Diagnostic Skills Profiling: Comparing Results from Confirmatory Factor Analysis and Diagnostic Classification Models

Olga Kunina-Habenicht

Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt Universität zu Berlin

André A. Rupp

Department of Measurement, Statistics, and Evaluation (EDMS), University of Maryland

Oliver Wilhelm

Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt Universität zu Berlin

Manuscript published in *Studies in Educational Evaluation*, 35, 64-70.

Abstract

In recent years there has been an increasing international interest in fine-grained diagnostic inferences on multiple skills for formative purposes. A successful provision of such inferences that support meaningful instructional decision-making requires (a) careful diagnostic assessment design coupled with (b) empirical support for the structure of the assessment grounded in multidimensional scaling models. This paper investigates the degree to which multidimensional skills profiles of children can be reliably estimated with confirmatory factor analysis models, which result in continuous skill profiles, and diagnostic classification models, which result in discrete skill profiles. The data come from a newly developed diagnostic assessment of arithmetic skills in elementary school that was specifically designed to tap multiple skills at different levels of definitional grain size.

IV

Sensitivity of Item and Respondent Parameter Estimation to Model Misspecification within a Log-linear Modelling Framework for Diagnostic Classification Models

Olga Kunina-Habenicht

Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt Universität zu Berlin

André A. Rupp

Department of Measurement, Statistics, and Evaluation (EDMS), University of Maryland

Oliver Wilhelm

Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt Universität zu Berlin

Manuscript submitted for publication

Abstract

In the current study we investigated parameter recovery and classification accuracy for correct and misspecified log-linear cognitive diagnostic classification models (DCMs) and tested the sensitivity of information-based fit indices and item discrimination fit indices toward model misspecification. The basic manipulated test design factors included the number of respondents (1,000; 10,000), attributes (3; 5), and items (25; 50). With regard to model misspecification, we differentiated between misspecification of interaction effects and two types of Q-matrix misspecifications. Results showed that parameter recovery for correctly specified models was consistent for intercepts and main effects and sufficient enough for interaction terms for samples with 10,000 respondents. By contrast, the estimation of interaction effects was seriously impaired in samples with 1,000 examinees. While the misspecification of interaction effects had little impact on classification accuracy, invalid Q-matrix specifications led to notably decreased classification accuracy. Information-based fit indices were sensitive to strong model misspecifications, whereas neither of the two item discrimination indices allowed for accurate detection of items with incorrect Q-matrix entries. Nevertheless, one of the item-fit indices can be meaningfully aggregated and used as a global fit measure. We conclude with some recommendations for the estimation and data-model fit evaluation of log-linear DCMs.

V

**Added Predictive Value of Multidimensional Proficiency Scores from
Diagnostic Classification Models for Global Proficiency Indicators in
Elementary-school Mathematics**

Olga Kunina-Habenicht

Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt Universität zu Berlin

Oliver Wilhelm

Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt Universität zu Berlin

André A. Rupp

Department of Measurement, Statistics, and Evaluation (EDMS), University of Maryland

Manuscript submitted for publication

Abstract

Diagnostic classification models (DCMs) hold great potential for applications in formative assessment by providing multivariate profiles on multiple attributes. Recently, Henson, Templin, and Willse (2008) proposed an integrative log-linear modelling framework covering core DCMs. In the current study six competing DCMs with simple and complex loading structure were estimated for data from a newly developed diagnostic mathematics assessment in elementary school for the 4th grade for 2032 students. Estimated DCMs as well as the corresponding unidimensional item response theory (IRT) model were compared in terms of model fit. It was investigated whether multidimensional profiles based on DCMs have an added predictive value over and above traditional IRT scores for school grades in mathematics and estimated unidimensional proficiency scores on a standards-based large-scale assessment of mathematics. Results revealed a negligible incremental validity of multivariate proficiency profiles compared to the unidimensional proficiency estimate from the simpler IRT models.

VI

General Discussion

Main results

In the current dissertation project a new diagnostic assessment for arithmetic skills DMA was developed, which allowed for detailed feedback on basic arithmetic competencies for students in grades three and four. Based on data for a pilot and main study for DMA this dissertation project addressed the following five research questions organized into two distinct subsets:

Theoretical Research Questions Regarding the Robustness of the Fit Assessment for DCMs

1. How biased and efficient are the item parameter recovery and the respondent classification under different conditions of model misspecification for core DCMs?
2. How do different global and local indices for the model-data fit evaluation perform under different conditions of model misspecification?

Applied Research Questions Regarding the Use of DCMs for the Diagnostic Assessment Data

3. Can data for the newly developed diagnostic assessment be successfully scaled with DCMs in order to create reliable multivariate attribute profiles for each student?
4. How similar are the resulting multivariate attribute profiles in comparison with the results of corresponding multidimensional CFA models?
5. Can the resulting multidimensional attribute profiles based on the best fitting DCM explain additional variance in the prediction of school grades in mathematics and performance in a national standard based assessment in Germany for mathematics over and above unidimensional proficiency scores in mathematics?

In order to answer these research questions we conducted one simulation study and two empirical studies illustrating applications of DCMs. In the first study, which was designed to answer the third and fourth research questions, we applied several different DCMs within the GDM framework to pilot data for a newly developed assessment of arithmetic for primary school and compared the fit of these models to CFA models.

In the second study, which was designed to answer the first two research questions, we implemented a complex simulation study and investigated the implications of the Q-matrix misspecification on the parameter recovery and the classification accuracy for DCMs in the log-linear framework. Additionally, we suggested two item discrimination parameters and tested their sensitivity toward two different forms of model misspecification - invalid definition of the Q-matrix and interaction levels.

In the third study, which was designed to answer the fourth and fifth research questions, again we applied a variety of DCMs and CFA models to the field data for $N = 2032$ students. One important aim of this study was to illustrate a successful application of the log-linear DCMs by considering the model fit measures suggested in the simulation study. A second unique aim of this particular study was to investigate the added predictive value of the multidimensional profiles based on DCMs over and above the traditional unidimensional proficiency scores in - (a) the prediction of school grades in mathematics and (b) the prediction of the performance in a standards-based large-scale assessment of mathematics.

Detailed discussion of the findings for each of these empirical studies can be found in the corresponding manuscripts that were presented in the previous chapters of this dissertation. Therefore, in this section we discuss main findings of the dissertation with regard to the five research questions outlined earlier. In the last section, implications of the main findings for future theoretical and practical work are critically evaluated.

Theoretical Research Questions Regarding the Robustness of the Fit Assessment for DCMs

To answer the *first research question*, using a complex simulation study we found that the item parameter estimation was sufficiently accurate in the true data generating models for conditions with $N = 10,000$. For items measuring one or two attributes, logit-values for intercepts and main effects were estimated accurately and efficiently even in conditions with $N = 1,000$. However, the estimation of two-way-interaction effects was unreliable and parameter recovery for the main effect and interaction effect parameters dropped dramatically for items that measured three attributes for cases with $N=1,000$. Results of the simulations study revealed no clear answer to the question what test length is required for a sufficient parameter recovery given a specific Q-matrix with a certain number of relevant attributes. In particular, parameter recovery for the logit-parameters for interaction terms was impaired in conditions with small samples in which the number of items did not exceed the number of possible latent classes.

The unreliable parameter recovery of estimates for interaction terms logit-parameters for small samples in the simulation study illustrates that the robustness of item parameters can be substantively impaired for complex DCMs, especially in applications with small sample sizes. Interestingly, despite the interaction misspecification and unstable parameter estimates for logit-values for interaction terms, the item response function $P(X|Q, \alpha)$ seems to be estimated quite robust. This result was captured in the MAD index, indicating the sufficient recovery of the item response function for models with interaction misspecification.

With respect to the respondent classification accuracy, misspecification of the interaction effect terms did not have a big impact. This finding partly corresponds to the previous findings of Rupp, Templin & Henson (2010) and suggests that parsimonious DCMs neglecting interaction-terms can be used to reduce model complexity without a substantial loss of precision in classification accuracy. Comparable results were also reported for parsimonious DCMs such as the higher-order latent trait models proposed by de la Torre & Douglas (2004) or Templin, Henson, Templin, and Roussos (2008). Usage of these statistically less demanding DCMs with reduced parameter space is especially advisable for DCM applications with small sample sizes.

By contrast, classification rates were substantially lower for models with a variety of invalid Q-matrices indicating that random permutation of 0s and 1s in the Q-matrix by a degree of 30% and /or a misspecification of the number of dimensions led to significant impairment of respondents' classifications. Since wrong definitions of Q-matrix lead to an arbitrary setting of cut-off points between mastery and non-mastery as well as wrong classification of respondents, we suggested one method for validation of Q-matrix definitions using item discrimination indices that is discussed next.

To answer the *second research question*, we investigated the sensitivity of several model fit measures to model violations. Evaluating the correctness of Q-matrices is a crucial issue in the DCM context because the specification of the Q-matrix reflects theoretical hypotheses about the cognitive processes that are involved in responding to diagnostic assessment items. This requires establishing robust and sensitive model fit measures that indicate absolute and relative global model fit as well as adequate item fit statistics allowing for detection of misspecification in the Q-matrix on the item level.

In the simulation study we proposed two candidates allowing for Q-matrix verification on the item level by extending the cognitive diagnostic index at the attribute level proposed by Henson, Roussos, Douglas, and He (2009) as well as the MAD_j index, which reflects the mean deviation between the observed and expected item response probabilities for all possible latent classes. Contrary to our predictions, however, both proposed indices did not allow for a reliable detection of invalid items in misspecified Q-matrices. Nevertheless, the mean of MAD values appears to be an appropriate global index for the model fit evaluation over all items. Moreover, we found that AIC and BIC information indices were sensitive to strong model violations in models with invalid Q-matrices. Thus, in this dissertation we introduced a promising model fit measure that can be easily calculated once the logit-parameters for the main and interaction effects are estimated allowing for global model fit evaluation.

Applied Research Questions Regarding the Use of DCMs for the Diagnostic Assessment Data

With regards to the *third and fourth research questions*, we found that the CFA model provided a better fit to the data than the DCM that was chosen as a discrete counterpart to the CFA in terms of the measurement framework specific fit indices. Correlation patterns of the continuous latent skill variables in the CFA model and the discrete latent skill variables in the DCM showed that both latent variable models supported similar conclusions. Furthermore, comparison of CFA models and DCMs revealed very high correlations between person scores for all attributes.

In the third study we estimated six competing DCMs in the log-linear framework and illustrated how competing DCMs can be compared in terms of goodness-of-fit considering the information criteria AIC and BIC as well as the MAD index. Consistent with previous findings from the simulation study, BIC tended to prefer parsimonious four-dimensional DCM while the AIC preferred the higher-dimensional model with six dimensions. When comparing DCMs in the log-linear framework to the traditional IRT model in the third study, our results did not provide a clear answer to the question whether the unidimensional or the best fitting four-dimensional DCM fits the data best. AIC and BIC revealed a better model fit for the one-dimensional model whereas the CFA model yielded a better fit for the four-dimensional model. Correlations between the one-dimensional WLE score and the ability estimates for the four dimensions ranged from .61 to .81, indicating that the postulated attributes in the DCM and IRT models are similar but are not completely exchangeable and reflect slightly different aspects of the mathematical ability.

Due to high correlations between the attributes in first and third studies, it remains unclear whether the postulated attributes for the DMA form separate theoretical constructs and can be statistically distinguished in the empirical data. The fact that the postulated attributes (e.g. attributes 'addition' and 'subtraction') in the Q-matrices, which was the basis for the test construction for the DMA, cannot be separated as distinguishable dimensions in the empirical data is consistent with previous applications that have shown that data may not provide information as fine as suggested by the underlying cognitive theory (Thissen-Roe, Hunt, & Minstrell, 2004; Katz, Attali, Rijmen, & Williamson, 2008).

Nevertheless, it is possible that these attributes are less strongly correlated for students in the selected sub-samples (e.g. children with dyscalculia problems) who have difficulties only in specific attributes. Thus, it is worth to compare the DCM models for the complete sample with the results of DCM analysis for subgroups of certain interest in order to address the issue of measurement invariance in the DCM context for different achievement groups. Related

methodological issues were extensively investigated in the CFA (e. g. Meredith, 1993, Vandenberg & Lance, 2000) and in IRT framework (in terms of differential item functioning, e. g. Holland & Wainer, 1993, Zumbo, 1999).

Almost identical correlation pattern for CFA and GDM models as well as almost perfect correlation between person's estimates from the DCM and factors scores for the CFA models in the first study are consistent with the results reported by Habermann, von Davier, and Lee, 2008 who compared DCMs with multidimensional IRT models. These results support the view of DCMs as discrete alternatives to traditional multidimensional latent variable models from the CFA (e.g., McDonald, 1999) or the IRT (e.g., Ackerman, Gierl, & Walker, 2003). In this sense discrete variables in DCMs seem to reflect an approximation of the underlying continuous variables based on the statistical foundation of the tetrachoric correlations of the latent variables. From statistical perspective dichotomization of continuous variables goes along with reduction or loss of information (MacCallum, Zhang, Preacher, & Rucker, 2002). However, previous studies have shown that teachers have difficulties interpreting continuous variables or claim that these values are not diagnostically relevant (Huff & Goodman, 2007). Thus, one important message of the analyses from the pilot data was that the DCM does not provide any "new" information beyond the multidimensional CFA model. It provides "different" information - namely, a direct representation of possible conditional skill relationships and a classification of students that is potentially useful diagnostic information for teachers and parents.

Providing informative feedback based on the estimated multidimensional profiles is one of main proposed advantages of DCMs. However, a related problematic issue concerns the lack of studies demonstrating the validity of these multidimensional scores estimated in the DCM over the traditional IRT parameters (e.g. Sinharay & Haberman, 2009). That issue was addressed in the *fifth research question* by using the DCM scores as additional predictors in a hierarchical linear regression model in the third study. We found only a negligible incremental predictive validity for these profile scores for prediction of mathematics grade and performance on an established mathematical assessment over traditional IRT parameter estimates. Thus, it remains challenging to show that there is a non-trivial practical benefit from using multidimensional proficiency profiles over and above traditional IRT person parameters. Previous studies have shown that relevant diagnostic information can also result from the traditional one-dimensional and multidimensional IRT models (see Wainer, Sheehan, & Wang, 2000; Walker & Beretvas, 2003). Since the reliability of multidimensional profiles is seldom reported, it is important to keep in mind that an unreliable multidimensional

profile is certainly worse than a reliable one-dimensional profile that carries similar meaning (Rupp & Templin, 2009). Furthermore, if the incremental validity of such profiles over traditional scores is not ensured, it is reasonable to rely on scores based on models with lower dimensionality and parameter space.

Conclusions and Outlook

I conclude with a critical evaluation of implications of the discussed findings for future theoretical and practical work. The current dissertation project illustrated a successful application of DCMs to empirical data and focuses on some of the important methodological issues in the DCM context such as the added value of multidimensional profiles over traditional IRT scores. To avoid problems related to retrofitting DCMs to data from non-diagnostic tests, we developed a new diagnostic test which construction was based on a-priori defined Q-matrices. One limitation of this project pertains to the definitions of Q-matrices that were mainly orientated on mathematical operations of the tasks rather than on deeper underlying cognitive operations. It would be desirable to develop one or more alternative Q-matrices that are based on recent didactic theories rather than on the mathematical surface characteristics. This would allow to compare competing models via information criteria or the MAD index as suggested in the third study. In this assessment we intentionally did not consider complex tasks requiring core mathematical competencies like 'argumentation', 'problem-solving', or 'communication' due to the fact that it appears difficult to define reliable attribute sub-skills or underlying cognitive operations for such complex tasks, partly because the distinction between the proposed attributes and general cognitive abilities including fluid intelligence, declarative knowledge, and reading comprehension is rather challenging. Nevertheless, an attempt to define a reliable Q-matrix via a standard-setting procedure with didactic experts and school teachers for tasks involving such processes would be an interesting and apparently ambitious study.

The implemented simulation study extended previous studies investigating the impact of model misspecification with respect to several aspects. First, we applied DCMs in the recently proposed integrative log-linear framework and estimated interaction terms in addition to main effects. Second, in contrast to previous simulation studies where only few specific invalid Q-matrices were used, we have generated several random misspecified Q-matrices to be able to generalize the impact of model misspecification. Third, we distinguished between two forms of model misspecification and proposed extensions for two item discrimination indexes. Previous studies either introduced new model fit indices or investigated the impact of Q-matrix misspecification on accuracy of parameter recovery and examinees' classification. To

our knowledge our study is the first study that combines both research questions and investigates the sensitivity of the proposed model fit measures toward two different types of model misspecification. Moreover, we provided evidence that the MAD index, that can be easily calculated once the logit-parameters for main and interaction effects were estimated, allows for global model fit evaluation. Finally, in the last empirical study we illustrated a method for comparison of competing DCMs by considering the information on AIC, BIC, and the MAD index.

As reported in the simulation study, samples with at least 1,000 respondents are required for sufficient estimation of the interaction terms for DCMs in the log-linear framework. Such sizable data sets are most likely present in the large-scale assessments rather than in formative assessments. However, there is no reason to expect additional results that would go beyond the findings based on the established one-dimensional models from retrofitting of DCMs to the large scale data for which item development and selection that was optimized for one-dimensional models (Habermann & van Davier, 2007). Thus, we face the dilemma that the implementation of the original aim of DCMs – namely, providing diagnostic feedback in formative and class room assessment – remains challenging because typically in these settings only small sample sizes are available. Application of DCMs is only realistic in these fields if the diagnostic assessment is already implemented at scale of a school district or state.

In the future it would be worthwhile to extend the simulation study by addressing the impact of the Q-matrix specification on the reliability of the ability estimates. Moreover, in the current simulation study only balanced randomized Q-matrices were used. In further studies it would be interesting to consider underspecified and overspecified Q-matrices as well and to extend the simulation design by defining different sets of logit-parameters also including scenarios for non-compensatory DCMs where interaction terms have higher impact than main effects. I foresee that in contrast to the implemented study where main effects had a higher impact than the interaction effects, misspecification of interaction levels will likely lead to the impairment of parameter recovery in case of non-compensatory DCMs.

As an extension of the third study it would be appealing to investigate the consistency of the ability profiles by considering different sub-samples and different subgroups of items that are supposed to measure the same ability. I expect that in this case classification of students will largely depend on the item selection, item difficulty, and general ability level in the sub-sample, because item difficulty has an immediate effect on the attribute mastery and thus on the distribution of latent classes, which in turn may lead to a different classification of the respondents in two assessments where different items are supposed to assess the same ability.

In comparison to the established psychometric approaches such as IRT and factor analysis models, research on DCMs is in its infancy in the areas of Q-matrix validation and goodness-of-fit evaluation. In the past many technical aspects of DCMs have been discussed. Addressing these technical questions was necessary to ensure the identifiability of DCMs. On the other hand it is also essential to answer the question whether DCMs can be applied in an interdisciplinary setting (e.g. formative assessment) where alternative established IRT and CFA models have already proven to be useful (Henson, 2009). Therefore, applications presenting the superiority of DCMs in comparison with the established models are needed in the future. Application of DCMs gains additional importance when the classifications from DCMs reflect degrees of mastery whose meaning is grounded in an underlying cognitive theory (Rupp & Templin, 2009). However, I share the viewpoint of Sinharay and Haberman (2009) that such theories for multiple separable dimensions do not currently exist. Although several technical and methodological issues discussed above still need to be solved in the DCM framework, DCMs represent a promising and important methodological approach potentially allowing for providing informative feedback to students and teachers.

VII

References

- Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37-53.
- Adams, R., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Birenbaum, M., Kelly, A. E., Tatsuoka, K.K., & Gutvitz, Y. (1994). Attribute-mastery patterns from rule space as the basis for student models in algebra. *International Journal of Human-Computer Studies*, 40, 497-508.
- Birenbaum, M., Tatsuoka, C., & Yamada, Y. (2004). Diagnostic assessment in TIMMS-R: Between countries and within country comparisons of eight graders' mathematics performance. *Studies in Educational Evaluation*, 30, 151-173.
- Bolt, D. & Fu, J. (2004). *A Polytomous Extension of the Fusion Model and Its Bayesian Parameter Estimation*. Paper presented at the Annual meeting of the National Council on Measurement (NCME) in Education, San Diego, CA, April, 2004.
- Carpenter, T.P., Fennema, E., Franke, M.L., Empson, S.B., & Levi, L.W. (1999). *Children's mathematics: Cognitively Guided Instruction*. Portsmouth, NH: Heinemann.
- Cheng, Y. (2009). When Cognitive Diagnosis Meets Computerized Adaptive Testing: CD-CAT. *Psychometrika*, 74(4), 619-632.
- Cizek, G.J. & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668-1674.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.
- diBello, L., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, Psychometrics) (pp. 979-1027). Amsterdam, Netherlands: Elsevier.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- de la Torre (2008a). *The generalized DINA model framework*. Unpublished manuscript. State University of New Jersey.
- de la Torre (2008b). An empirically based method of Q-Matrix Validation for the DINA Model: Development and Applications. *Journal of Educational Measurement*, 45, pp. 343-362.

- de la Torre (2008c). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, 73(4), 595–624.
- de la Torre, J. (2009a). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115-130.
- de la Torre, J. (2009b). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement*, 33(3), 163-183.
- 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.
- Embretson, S. (1983) Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179-197.
- Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 37, 359–374.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 380-396.
- Embretson, S. & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38 (4,), 343-368 .
- Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.
- Finkelman, M., Kim, W., Roussos, L., Verschoor, A.J. (2008). A Binary Programming Approach to Automated Test Assembly for Cognitive Diagnosis Models. (Research Report 2008-1). Arnheim: Cito.
- Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.
- Gierl, M. J., Leighton, J.P., & Hunka, S.M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 173–201). New York: Cambridge University Press.
- Gorin, J. S. (2007). Test construction and diagnostic testing. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 173–201). New York: Cambridge University Press.
- Gorin, J. S. (2009). Diagnostic classification model: Are they necessary? Commentary on Rupp and Templin (2008). *Measurement: Interdisciplinary Research and Perspectives*, 7, 30–33.

- Haberman, S. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33, 204-229.
- 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., 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.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26(4), 301-323.
- Haertel, E. H. (1990). Continuous and discrete latent structure models for item response data. *Psychometrika*, 55, 477-494.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.
- Henson, R. (2009). Diagnostic classification models: Thoughts and future directions. *Measurement: Interdisciplinary Research and Perspectives*, 7, 34-36.
- 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.
- Henson, R. A., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute level discrimination indices. *Applied Psychological Measurement*, 32, 275-288.
- Holland P. W. & Wainer, H. (1993). (Eds.) *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates Publishers
- Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). Cambridge: Cambridge University Press.
- Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG-TOEFL* (Unpublished dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign.
- Jang, E.E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application LanguEdge assessment. *Language Testing*, 26, 31-73.

- Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment*. Unpublished manuscript. Accessed January, 22, 2010, from <http://www.stat.cmu.edu/~brian/nrc/cfa>
- 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.
- Kaplan, D. (2000). *Structural equation modelling: Foundations and extensions*. London: Sage, Advanced Quantitative Techniques in the Social Sciences series.
- Katz, I., Attali, Y., Rijmen, F., & Williamson, D.M. (2008). *ETS's iSkills assessment: Measurement of information and communication literacy*. Paper presented at the 23th annual conference of the society for industrial and organizational psychology. San Francisco, CA.
- Kline, R. B. (2005) *Principles and Practice of Structural Equation Modeling*. New York: The Guilford Press.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking. Methods and practices* (2nd ed.). New York: Springer.
- Leighton, J. P., & Gierl, M. J., & Hunka, S. M. (2005). The skill hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205-236.
- Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and practice*. Cambridge: Cambridge University Press.
- Linn, R. (1990). Diagnostic testing. In N. Frederiksen, R. L. Glasser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (p.453-486). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99-120.
- Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*(4), 523-547.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187-212.
- Maris, G. & Bechger, T. (2009): Equivalent diagnostic classification models. *Measurement: Interdisciplinary Research and Perspectives, 7*, 41-46.
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7* (1), 19-40.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

- McGlohen, M. K. (2004). *The Application of a Cognitive Diagnosis Model via an Analysis of a Large-Scale Assessment and a Computerized Adaptive Testing Administration*. (Unpublished dissertation). University of Texas at Austin.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543.
- Mislevy, R. J. (2007). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 257-305). Portsmouth, NH: Greenwood Publishing Group.
- Mislevy, R. J. (2008). *How cognitive science challenges the educational measurement tradition*. (Unpublished manuscript). University of Maryland.
- Mislevy, R. J., Behrens, J. T., Bennett, R. E., Demark, S. F., Frezzo, D. C., Levy, R., Robinson, D.H., Rutstein, D.W., Shute, V.J., Stanley, K, & Winters, F. I. (2010). On the Roles of External Knowledge Representations in Assessment Design. *The Journal of Technology, Learning, and Assessment*, 8, 1-55.
- Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation*, 35, 95-101.
- Roussos, L., diBello, L. V., Stout, W., Hartz, S., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnosis system. In J. P. Leighton, & Gierl, M. J. (Ed.), *Cognitively diagnostic assessment for education: Theory and practice* (pp. 275-318). Thousand Oaks, CA: SAGE.
- Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, 44, 293-311.
- Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and applications* (pp. 205-241). Cambridge: Cambridge University Press.
- Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6, 219-262.

- Rupp, A. A., & Templin, J. (2009). The (un)usual suspects? A measurement community in search of its identity. *Measurement: Interdisciplinary Research & Perspectives*, 7, 115-121.
- Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.
- Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, 78(1), 153–189.
- Sinharay, S. (2006). Model Diagnostics for Bayesian Networks. *Journal of Educational and Behavioral Statistics*, 31(1), 1–33.
- Sinharay, S. & Almond, R. G. (2007). Assessing Fit of Cognitive Diagnostic Models: A Case Study. *Educational and Psychological Measurement*, 67(2), 239-257.
- Sinharay, S., Almond, R. G., & Yan, D. (2004). *Assessing Fit of Models With Discrete Proficiency Variables in Educational Assessment*. (ETS Research Report RP-04-07). Princeton, NJ: Educational Testing Service.
- Sinharay, S. & Haberman, S. J. (2009). How much can we reliably know about what examinees know? *Measurement: Interdisciplinary Research and Perspectives* 7, S. 46–49.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–332). New York: Macmillan.
- 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. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-360). Hillsdale, NJ: Erlbaum.
- Tatsuoka, K. K., Birenbaum, M., & Arnold, J. (1989). On the stability of students' rules of operation for solving arithmetic problems. *Journal of Educational Measurement*, 26(4), 351-361.
- Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, 41(4), 901–926.

- Templin, J. (2009). Measuring the Reliability of Diagnostic Model Examinee Estimates. paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA., April.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287-305.
- Templin, J. & Henson, R. (2009). *Practical issues in using diagnostic estimates: Measuring the reliability and validity of diagnostic estimates*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA., April.
- Templin, J., Henson, R., Rupp, A., Jang, E., & Ahmed, M. (2009). *Cognitive diagnosis models for nominal response data*. Manuscript submitted for publication.
- Templin, J., Henson, R., Templin, S., & Roussos, L. (2008). Robustness of unidimensional hierarchical modeling of discrete attribute association in cognitive diagnosis models. *Applied Psychological Measurement*, 32, 559-574.
- Thissen-Roe, A., Hunt, E., & Minstrell, I. (2004). The DIAGNOSER project. Combining assessment and learning. *Behavior research methods, instruments, & computers*, 36, 234-240.
- Tiffin-Richards, S. P., Pant H. A. & Köller, O. (2010). *Setting Standards for English Foreign Language Assessment: Methodology, Validation and a Degree of Arbitrariness*. Manuscript submitted for publications.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4-69.
- van der Linden, W. J., & Glas, C.A.W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer.
- 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.
- von Davier, M. (2007). *Hierarchical general diagnostic models*. (ETS Research Report, RR-07-19). ETS: Princeton, NJ.
- Wainer, H. (2000). *Computerized adaptive testing: A primer (2nd ed.)*. Mahwah: Lawrence Erlbaum Associates.
- Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement*, 37(2), 113-140.

- Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*, 255–275.
- Wilhelm, O. & Robitzsch, A. (2009). Have cognitive diagnostic models delivered their good? Some substantial and methodological concerns. *Measurement: Interdisciplinary Research and Perspectives, 7*, p. 53–57.
- Xu, X. & von Davier, M. (2006). *Cognitive Diagnosis for NAEP proficiency data* (ETS Research Report 06-08). Princeton, NJ: Educational Testing Service.
- Xu, X. & von Davier, M. (2008). *Linking in the general diagnostic model* (ETS Research Report RP-08-08). Princeton, NJ: Educational Testing Service.
- Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.