Measurement, 6: 263–275, 2008

ISSN 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366360802497762

Psychology Press
Taylor & Francis Group

COMMENTARIES

Defining Characteristics of Diagnostic Classification Models and the Problem of Retrofitting in Cognitive Diagnostic Assessment

Mark J. Gierl and Ying Cui
Centre for Research in Applied Measurement and Evaluation, University of Alberta

Educational and psychological measurement are undergoing significant development and change as interdisciplinary forces stemming from cognitive science, instructional technology, mathematical statistics, computing science, and educational psychology are shaping both theory and practice. For example, the influence of cognitive psychology on educational measurement, carefully documented almost 20 years ago in Snow and Lohman's (1989) seminal chapter in the 3rd Edition of *Educational Measurement*, has recently become an important source of ideas leading to innovations in cognitive diagnostic assessment (Leighton & Gierl, 2007a). The influence of cognitive principles on assessment practices is also apparent in the development of diagnostic classification models (DCM) as described by André Rupp and Jon Templin in their feature article. These models have at least nine defining features:

> *Diagnostic classification models (DCM) are probabilistic confirmatory multidimensional latent-variable models with a simple or complex loading structure. They*

Correspondence should be addressed to Mark J. Gierl, Professor of Educational Psychology and Canada Research Chair in Educational measurement, Centre for Research in Applied Measurement and Evaluation (CRAME), 6–110 Education North, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada T6G 2G5. E-mail: mark.gierl@ualberta.ca

are suitable for modeling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and noncompensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. Some DCM are further able to handle complex sampling designs for items and respondents as well as heterogeneity due to strategy use.

One promising application of DCM is in the area of cognitive diagnostic assessment in education. DCM can contribute to educational testing in significant ways. For instance, the use of these psychometric and statistical models should increase our understanding of student test performance, given that many educational tests are based on cognitive problem-solving tasks. Because traditional test scores serve only as a coarse indicator of how students think about and solve educational tasks, DCM may overcome this limitation by helping to identify, describe, and report information at a fine grain size on the knowledge, skills, processes, and/or strategies students use to solve test items. DCM may also play a critical role in linking theories of cognition and learning with instruction, particularly when the DCM is guided by a cognitive model. A cognitive model in educational measurement refers to a simplified description of human problem-solving on standardized tasks at some convenient grain size or level of detail in order to facilitate explanation and prediction of students' performance, including their strengths and weaknesses (Leighton & Gierl, 2007b). Instructional decisions are made, in part, on how students think about and solve problems. DCM represents a sophisticated class of models that may allow teachers to make students' thinking overt, so that their cognitive strengths and weaknesses can be identified and evaluated. Instructional interventions can then focus on overcoming weaknesses.

It is worth noting, however, that the successful application of DCM in educational testing will likely come with a price—and this price may be in the form of new test development procedures and practices required to yield data that satisfy the defining characteristics of these models. By implication, this means that *retrofitting* DCM to much of the achievement testing data that is currently available in education is likely to yield unsatisfactory diagnostic classification results. Retrofitting can be described as the addition of a new technology or feature to an older system. Similarly, we might consider cognitive diagnostic retrofitting as the application of a new statistical or psychometric model, such as a DCM, to student response data from an existing testing system that uses traditional test development procedures and practices. We contend that conducting cognitive diagnostic assessment through retrofitting will yield few successful applications, precisely because of the DCM's unique requirements, as outlined by Rupp and Templin.

Due to space limitations, we work with three of the defining characteristics to describe and illustrate why retrofitting DCM to existing educational data will often prove unsuccessful. We also propose one additional defining characteristic for these psychometric and statistical models.

CRITERION 1: THEIR MULTIDIMENSIONAL NATURE

DCM should measure a multidimensional skill set, as the authors note. But can this structure be found using data from existing educational tests? The authors begin by describing a context of use for DCM focusing on educational testing. They present a score report developed by Eunice Jang to illustrate how and where DCM could be used for educational assessment. In describing this exemplary diagnostic score report, the authors state: "This nine-dimensional skill profile was estimated using a rather complex DCM and it is the statistical and substantive properties of the DCM behind such skill profiles that are the focus of the following narrative." This example immediately raises an important issue: the dimensionality of the data. The authors select an example where the test data are characterized by nine dimensions. But rarely, if ever, are educational assessments characterized by a nine-dimensional skill structure. Well-known tests, such as the SAT, are typically characterized by two or three dimensions at most (Cook, Dorans, & Eignor, 1988; Dinoes, Bejar, & Chaffin, 1996; Lawrence & Dorans, 1987). A familiar example of a multidimensional test is the LSAT. This exam measures three dimensions (Douglas, et al., 1999). If the goal is to apply DCM to existing educational tests that possess high-dimensional structures, then our research and practical experience suggests that virtually no data from K–12 educational achievement tests will be suitable.

CRITERION 2: THEIR CONFIRMATORY NATURE

Confirmatory analyses require that the data structure be specified *a priori*. Often, this requirement means that a substantive theory or set of hypotheses is needed to specify the structure of the data in order to direct the psychometric analysis. Ideally, a cognitive model would be developed first to specify the knowledge and skills evaluated on the test and then items would be created to measure these specific cognitive skills. The psychometric analysis, conducted in a confirmatory mode with the DCM, would follow using the cognitive model as a guide and using data collected on the purposefully designed diagnostic items. This order of events—identify cognitive model, develop diagnostic items, conduct confirmatory analysis—provides the analyst with control over how to operationalize the construct, what the underlying data structure should look like, and how the test scores should be interpreted.

These steps cannot be followed with a retrofitting approach because there are neither cognitive model development nor test construction activities. Instead, some type of implicit substantive model is generated *post hoc* by reviewing the existing items, and then these existing items are coded using the *Q-matrix*. Despite the convenience afforded by a retrofitting approach, it is severely limited because there is no guarantee that either an appropriate cognitive model can be identified or an adequate number of items can be located to measure the skills in the cognitive model. Yet, these serious limitations should be expected whenever test development proceeds without an explicit model of test performance, because most educational achievement tests are not intended to promote diagnostic inferences about students' cognitive skills. Consequently, the cognitive analysis of any existing educational test using retrofitting procedures will invariably produce a tenuous fit between the cognitive model and the test data, because the tests were not designed from an explicit cognitive framework, which ultimately leads to inferior data for the DCM psychometric analysis.

CRITERION 8: THE DIAGNOSTIC NATURE OF THE INTERPRETATIONS

Cognitive diagnostic assessment is one step in a much larger undertaking that occurs when a problem is suspected. The results from a cognitive diagnostic assessment provide information about the students' cognitive skills and these results should be reported in a format that is interpretable to students, teachers, parents, and other stakeholders. In other words, this approach to testing is an attempt to highlight the student-by-item interaction, in cognitive terms, where the knowledge, skills, processes, and/or strategies used by a student to respond to items are made explicit. One method of representing this interaction is with a cognitive model (National Research Council, 2001; Pellegrino, 1998, 2002; Pellegrino, Baxter, & Glaser, 1999). Test score inferences anchored to a cognitive model should be more interpretable and meaningful for evaluating and understanding performance, particularly when the items are designed to measure the students' cognitive skills. The diagnostic testing process used to link student performance to the cognitive model begins by specifying a cognitive model and ends by reporting the results using this model. In short, an effective cognitive diagnostic assessment in education must be well integrated into the learning environment, and it must be *specifically designed* to help teachers understand how students think about and solve problems in that learning environment. Rarely, will an existing educational achievement test suit this purpose. Moreover, Rupp and Templin claim that DCM retrofit to assessments originally created to measure coarse, unidimensional constructs (i.e., many educational achievement tests) rarely work well because of problems with parameter convergence and model-data fit.

(PROPOSED) CRITERION 10—THE USE OF PRINCIPLED TEST DESIGN AND ANALYSIS

To overcome the limitations associated with retrofitting data to DCM in the context of cognitive diagnostic assessments in education, we propose a tenth criterion, *principled test design and analysis.* The defining characteristics and, hence, specific data requirements of these psychometric and statistical models warrant, in our opinion, this additional criterion. Principled test design and analysis adopts all of the existing standards of practices in test development. But it also has some additional requirements: a cognitive model must be identified and evaluated, items must be developed to measure the knowledge and skills in the cognitive model, and confirmatory model-based DCM are applied to these data to generate the diagnostic scores and reports. We also contend that a cognitive model of some type will always be needed to develop items and analyze student response data, generate scores, and guide score interpretations for cognitive diagnostic assessments, because this form of testing must be designed to identify and evaluate students' cognitive skills at a fine grain size. This type of specificity cannot be obtained using a retrofitting approach (e.g., coding existing items for cognitive attributes or skills), because items with these specific cognitive characteristics are unlikely to exist on tests developed without a cognitive model and, as a result, the fragile fit between the model and data will often yield weak and, possibly, inaccurate diagnostic inferences, even when the most powerful and sophisticated DCM is used.

REFERENCES

Cook, L. J., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. *Journal of Educational Statistics, 13*, 19–43.

Diones, R., Bejar, I. I., & Chaffin, R. (1996). The dimensionality of responses to SAT analogy items. *ETS Research Report.* Princeton, NJ: ETS, January.

Douglas, J., Kim, H., Roussos, L., Stout, W., & Zhang, J. (1999). *LSAT dimensionality analysis for the December 1991, June 1992, and October 1992*. LSAC Research Report Series. Law School Admission Council, Inc.

Lawrence, I., M., & Dorans, N. J., (1987). *An assessment of the dimensionality of SAT-Mathematical*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC, April.

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

Leighton, J. P., & Gierl, M. J. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*, 3–16.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, and R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Washington, DC: National Academy Press.

Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 49–60). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W. (2002). Understanding how students learn and inferring what they know: Implications for the design of curriculum, instruction, and assessment. In M. J. Smith (Ed.), *NSF K–12 Mathematics and science curriculum and implementation centers conference proceedings* (pp. 76–92). Washington, DC: National Science Foundation and American Geological Institute.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (pp. 307–353). Washington, DC: American Educational Research Association.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed.), pp. 263–331). New York: American Council on Education, Macmillian.

How Binary Skills Obscure the Transition from Non-Mastery to Mastery

Tzur M. Karelitz
Education Development Center

What is the nature of latent predictors that facilitate diagnostic classification? Rupp and Templin (this issue) suggest that these predictors should be multidimensional, categorical variables that can be combined in various ways. Diagnostic Classification Models (DCM) typically use multiple categorical predictors to classify respondents into qualitatively meaningful latent classes and, thus, provide a fine-grained analysis of respondents' strengths and weaknesses. The diagnostic power of DCM is driven by the confirmatory nature of the Q-matrix, which commonly represents item requirements in terms of a set of binary skills. Rupp and Templin note that, apart from a few exceptions, "Most DCM and associated estimation routines allow only for dichotomous latent variables." Binary skills are assumed to represent very simple dimensions in the sense that one either possesses a certain skill (mastery) or not (non-mastery). Such binary skills are useful indicators because it is easy to identify when something is present or absent. There is a slight problem, however. In reality, things are rarely

Correspondence should be addressed to Tzur M. Karelitz, Education Development Center, 55 Chapel Street, Newton, Massachusetts 02458-1060. E-mail: tkarelitz@edc.org