Proper Use of Assessment Results from Common Core State Standards

Executive Report for the Maryland State Department of Education

By the MARC Team

(Bob Lissitz, Hong Jiao, Ming Li, Daniel Yangsup Lee and Yoonjeong Kang)

INTRODUCTION TO THE REPORT

Since the Common Core State Standards (CCSS) were released in 2010, nearly every U. S. state has formally adopted these standards in mathematics and English language arts (ELA) and many joined one of two Consortia to develop and implement common tests. For some states, consortium tests will be used to replace their state tests for high stakes accountability decisions such as grade promotion and teacher evaluation. In addition, to go with the new summative assessments, innovative uses of formative assessments have been recommended to improve feedback systems for teachers and students and to improve educational achievement.

As a member of the consortium of Partnership of Assessment Readiness for College and Careers (PARCC), the state of Maryland has recently adopted the CCSS and will begin operational PARCC testing based on these new standards in the school year 2014-15. Together with these standardized tests is the implementation of computerized diagnostic systems using formative assessments to provide teachers with diagnostic information on students' (especially those at relatively low performance levels) strengths and weaknesses, and these formative assessments are set to be field tested in early 2015 and available for use in the 2015-16 school year.

Given the high stakes associated with the use of scores from standardized tests as well as the need for proper diagnosis of student learning, we believe two issues need to be addressed:

one is test fairness and the other is the adequacy of the diagnostic assessment system. In terms of test fairness, test score comparability across states is a major concern that requires measurement invariance to ensure valid test score inferences and comparisons across states. In this situation, multiple-group item response theory (IRT) models for differential item functioning (DIF) detection are highly recommended. For the diagnostic assessments, Cognitive Diagnostic Models (CDMs) are suggested to provide students and teachers with more accurate and useful information on student learning. Therefore, based on our literature review, this document focuses on a discussion of three fundamental concerns: (1) score comparability and DIF for multiple groups, (2) the selection of software packages for multiple-group IRT analysis, and (3) CDMs. We hope this review helps stake holders of Maryland make informed model choices under CCSS so that valid inferences about effects of educational input and student learning based on test scores can be achieved. We look forward to working with the State of Maryland next year as these important issues come front and center.

SCORE COMPARABILITY AND DIFFERENTIAL ITEM FUNCTIONING

Given the high stakes associated with the use of test scores, it is important to collect validity evidence to support such usage of test scores. Without validity, any inferences made from a measure are meaningless (Zumbo, 1999). Under the CCSS, tests developed by each consortium are based on the same common core standards; however, states within one consortium may adopt different curriculum and instruction and student populations from different states could be very diverse. As acknowledged by PARCC, test score comparability across states is an important issue to be addressed.

In this part, we will discuss briefly methods for detecting DIF items across multiple groups as well as multiple-group IRT models for dealing with DIF. DIF analysis is very important because it can be used in validating test score inferences and group comparisons. In 2013, NAEP reported for the first time the test results of grades 4 and 8 students in reading, mathematics, and science for several large states including California, Florida, Illinois, New York, and Texas, as well as national averages and results for different demographic groups. Based on the results, progress in student learning was evaluated. It can be imagined that if DIF analysis had not been conducted, any further conclusions or inferences made would be meaningless. As a member of CCSS, the state of Maryland will inevitably use results from the common assessments for various purposes. To score students' performance fairly, measurement invariance has to be investigated.

Measurement Invariance

When a test is administered to students from potentially different populations, item invariance is of primary concern as it directly affects test fairness and validity. Item invariance indicates that

an item (ultimately the test) performs in the same manner for each subgroup of examinee population. When an IRT model is employed for test construction, an implicit assumption is that item invariance or measurement invariance (MI) should hold across subgroups of the examinee populations. In practice, MI is often tacitly assumed rather than investigated (Borsboom, 2006). Conceptually, MI holds when a measure utilized under different conditions yields the same observed scores for individuals with identical attributes being measured (Meade & Bauer, 2007). The establishment of MI across subgroups is a logical prerequisite to conducting valid, sensible cross-group comparisons, and violation of MI may lead to biased measurements of ability which may make further inference unreliable and invalid. Given the validity issue brought about by the violation of MI, investigation of MI has become increasingly important.

In IRT literature, the violation of MI is termed as differential item functioning (DIF). An item exhibits DIF when students with the same ability level but from different groups have different probabilities of answering an item correctly. DIF analysis is an important step in the process of test development. It is often conducted based on grouping variables such as gender, ethnicity, disability, and language groups to assure that a test is fair for all students belonging to different groups. Literature and technical reports for state tests well documented such psychometric practice. For example, Zenisky, Hambleton and Robin (2004) explored gender DIF in a larger-scale science assessment. Abedi, Leon and Kao (2008) examined performance differences between students with disabilities and students without disabilities using DIF analyses in a high-stakes reading assessment. Further, Laitusis, Maneckshana, Monfils and Ahlgrim-Delzell (2009) examined DIF by three disability groups on an on-demand performance assessment (serving as an alternate assessment for the standardized statewide assessment test) for students with severe cognitive impairments. For the PARCC consortium tests, DIF analysis

should be conducted across states as well because students from different states are suspected to differ so greatly and the possibility of the presence of DIF items could be very high.

Methods for DIF Detection

A large number of methods have been developed and adapted for identifying DIF. A lot of DIF detection is determined by comparing item parameter estimates across two or more groups of examinees. Some of the commonly used approaches, such as the Mantel-Haenszel (MH; Mantel & Haenszel, 1959; Holland & Thayer, 1988), logistic regression, and loglinear analysis, are based on statistical models developed for categorical data. MH method is used for dichotomously scored items and is powerful. It has statistical test and effect size and can be implemented in SAS and SPSS. MH DIF detection method is widely used in state tests. Other existing DIF methods for multiple groups are the generalized MH method which is used with polytomous items (Zwick, Donoghue & Grima, 1993) and the generalized Lord's chi-square test. Kim, Cohen and Park (1995) presented a method for detection of DIF in multiple groups for a 2PL model with dichotomously scored items and unidmensionality assumed. The method is closely related to Lord's chi-square for comparing vectors of item parameters estimated in two groups. Magis, Raiche, Beland and Gerard (2011) proposed a generalized logistic regression procedure to detect DIF in the presence of more than two groups of respondents. SIBTEST (Shealy & Stout, 1993) is another method that standardizes the two groups of interest to have a common distribution of the latent trait and then estimates the expected difference in scores between the groups.

Other techniques in use are based on differences in specific IRFs between the groups of interest (Millsap & Everson, 1993). These methods assume a specific form of the function relating the probability of a correct response to an item and ability. One approach to DIF

detection that has shown promise is the IRT likelihood ratio test (Thissen, Steinberg & Wainer, 1988) which allows for the comparison of model fit, assuming equality of the parameter estimates for the item in question across the reference and focal groups (compact model), with the model fit when this constraint is relaxed and differences for the item parameters across the groups are allowed (augmented model).

It should be noted that the MH and SIBTEST methods are examples of non-parametric DIF detection procedures that are used in the classical test theory framework. Methods such as Lord's chi-square method, differences in specific IRFs, and the likelihood ratio test are IRT-based procedures which are very useful in detecting DIF in test practice (Roussos & Stout, 1996).

Instead of the simultaneous DIF method, a stepwise DIF analysis using polytomous responses of multiple groups was proposed by Muraki (1999) which include: Step 1: no DIF model is assumed by treating an entire data set as a single group; Step 2: only slope parameters (item discrimination) are separately estimated for each group, while the set of item location parameters are assumed to be common among subgroups and are estimated as a single group; and Step 3: the slope group parameters estimated at the second step are fixed and only item location parameters are separately estimated for each group. For all steps, sets of category parameters are assumed to be common among groups and they are estimated as a single group. In this study, the multiple-group partial credit model was used as the multiple-group IRT model for both the simulated data sets and the National Assessment of Educational Progress (NAEP) data, and the non-uniform and uniform DIF items were successfully detected. One limitation with the study, however, is that only two groups were used for comparison.

Yao and Li (2010) developed a DIF detection procedure in multidimensional IRT framework to flag only those items that have adverse DIF. According to the authors, it is very important to have a procedure to investigate the cause of DIF and detect only adverse DIF because adverse DIF resulted from nuisance dimensions tends to lower construct validity.

Multiple-Group IRT Models

In performing DIF analysis, it is often seen that separate comparisons are made between a single reference group and each of multiple focal groups. The drawbacks of conducting these separate tests of DIF were discussed by Penfield (2001) in terms of type I error rate, power as well as the substantial time and computing resources required. The undesired qualities from doing separate analyses, according to Penfield, can potentially be avoided by using the procedure that simultaneously assesses DIF across all groups. With the simultaneous procedure, estimation of both item parameters and examinee population distribution parameters can be performed, and one way to achieve the simultaneous analysis is through using multiple-group IRT models. Having been developed for traditional IRT models (e.g., Bock & Zimowski, 1997; Muraki, 1999), one of the practical applications of the multiple-group IRT models is the study of DIF.

Multiple-group IRT methods can be either unidimensional or multidimensional. In terms of the unidimensional IRT models, Rasch model, two-parameter logistic (2-PL) model, and three-parameter logistic model (3-PL) are three standard models for binary (dichotomous) data. The unidimensional models for polytomous data include partial credit model and graded response model. When DIF analyses are operationalized with unidimensional models, two important premises are that (1) the target ability is statistically extracted from the item responses, and (2) the groups in question are anchored over this target ability – not just placed on the same

standardized scale – before comparing item parameters (Camilli, 1992). That is, in all groups the probability of a correct response must be a function of the identical target ability, which, in fact, offers great simplicity and ease of interpretability. According to Camilli, however, a number of weaknesses of conducting DIF analysis using unidimensional IRT models exist. For example, undimensional models may estimate a composite of the underlying abilities if there are more than one ability contributes, which makes the comparison of item parameters invalid because groups cannot be anchored over a single target ability. In many cases, an item measures several abilities rather than a single latent construct. Even within a single content area such as math, the content of an item may measure multiple skills. For example, a math word problem may require both reading and math abilities. In such situation, multidimensional IRT models may be considered more realistic than unidimensional models. Multidimensional IRT models may provide a useful mechanism to aid in the interpretation of DIF because the models are prudent to isolate and interpret secondary abilities (Camilli, 1992). Additionally, since these models can be used to estimate an examinee's ability on several dimensions simultaneously, enhanced diagnostic information can be provided.

Most of the multidimensional IRT models are derived by generalizing their unidimensional counterparts to multidimensional models (Lin, 2008) which include, for example, the multidimensional 1-PL, 2-PL and 3-PL models for binary data, and the multidimensional partial credit model and graded response model for polytomous data. There are two types of multidimensional IRT models: the compensatory models (Lord & Novick, 1968; McDonald, 1967; Reckase, 1985, 1995) and the noncompensatory models (Embretson, 1984; Sympson, 1978). The multidimensional compensatory models indicate that a low ability on one dimension can be overcome or compensated for by a high ability on another dimension. Though

this property may be reasonable for some items, it may not be realistic for others. For example, if an item requires the ability to read in order to understand it, then high math ability cannot compensate for a lack of reading skills. For items of this type, the non-compensatory models may be more reasonable where being low on one ability cannot be compensated for by being high on another ability in order to have a high probability of success on an item. This means one must demonstrate proficiency in all abilities in order to give a correct response.

Among the various multidimensional multiple-group IRT models, one model that may offer particular advantages in the study of DIF is the multiple-group testlet model (Li, Bolt & Fu, 2006), which is an extension from the general testlet model developed based on the standard unidimensional IRT models by adding an item-testlet interaction effect parameter (Bradlow, Wainer, & Wang, 1999). Testlets refers to groups of items based on a common stimulus, such as items based on a common reading passage in a reading comprehension test. When tests are made up of testlets, standard IRT models are often not appropriate due to the local dependence among items from a testlet. Therefore, several IRT models for testlets have been proposed (Bradlow, et al., 1999; Li et al., 2006; Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2002). Testlets provide contexts or situations in assessing skills and abilities, so they have become an increasingly popular and attractive way of designing education tests (Li et al., 2006). Due to their multidimensional nature, testlet models offer great potential in understanding individual differences in test performance and item functioning (Li et al., 2006). For example, multiple-group testlet models can be applied to identify sources of DIF related to testlet factor and/or item specific factors. The multiple-group testlet model developed by Li et al. is a 2PL model. In a recent study by Jeon, Rijmen, and Rabe-Hesketh (2013), a generalization of the multiple-group bifactor model was used for assessing DIF for testlet-based tests. By including group-specific

difficulty parameters, the model can be used to assess DIF for testlet-based tests. Their generalized model encompasses testlet and second-order models for binary and polytomous responses. Through a simulation study, it is shown that ignoring between-group differences in the structure of the multivariate latent space can result in substantially biased estimates of DIF.

Common Scale in Multiple-Group Analysis

In IRT applications, item parameters calibrated using more than two independent groups must be expressed on a common metric or scale. When data from a single group are calibrated (i.e., the parameters of the model are estimated), the θ scale is identified by stipulating a population distribution for θ (Baker, 1990). For example, one can specify that examinees are sampled from a normal population distribution with mean θ equal to 0 and standard deviation equal to 1. This specification identifies the scale for θ, which in turn identifies the scale for the item parameters. In a multiple-group situation, one must make scale of the latent variable common across groups so that the item parameters can be estimated with respect to this scale and then tested for equivalence.

Two approaches can be used for developing a common metric in a multiple-group situation: separate calibration and concurrent calibration. Specifically, separate calibration is the process of estimating the parameters separately for each group and then linking or rescaling the parameters onto the common scale. To put these estimates on the same scale, usually one group is selected as the reference group and the scale of this group is used as the base scale, to which the parameters estimated in other groups are transformed (Lin, 2008). Concurrent calibration is to estimate the parameters simultaneously for all groups and constrain the parameters of the same item to be equal across groups (Lin, 2008). One can make the θ scale common across groups by

using an anchor test (i.e., a set of test items that are constrained to have the same parameters between groups). According to Kolen and Brennan (2004), one prominent advantage of concurrent calibration is that it estimates parameters for all groups at one time, and there is no need for linking. Mislevy (1987) and Bock and Zimowski (1996) described the multigroup IRT procedures for concurrently estimating item and ability parameters for all groups using the maximum marginal likelihood (MML) method. During the process of estimation, the item parameters are estimated over all groups whereas the ability distribution is estimated separately for each group so that they can be different when the groups are nonequivalent.

It should be mentioned that in the unidimensional IRT models the scale of parameters is determined only up to a linear transformation to eliminate the scale indeterminacy (i.e., the origin and scale unit can be set arbitrarily at any value without changing the fit of the model) in item response models (Hambleton, Swaminathan, & Rogers, 1991). The scale indeterminacy problem in multidimensional models is more complicated than that in unidimensional models. In addition to the indeterminacy of origin and scale, multidimensional models have an additional indeterminacy, rotation indeterminacy, which can be rotated in the ability space without changing the model fit (Li & Lissitz, 2000). When reference and focal groups are calibrated separately, their multidimensional estimates (if there are several dimensions) may be on different scales. With using multidimensional multiple group IRT procedure, the groups will be put on the same scale based on the non-DIF common items (Yao, 2010).

Implications of DIF Analysis in Multiple-Group IRT

It is important that tests developed by PARCC to be used by the State of Maryland be free of DIF before high stakes decisions for students and teachers could be made based on test scores. If

items on a test were found to have DIF for certain groups or across states, caution should be excised in test score use. Given the high expense in creating some items such as technology enhanced innovative items, other approaches such as multiple-group IRT models could be considered as an option to still use expensive items flagged with DIF and maintain fairness across states.

References

Abedi, J., Leon, S., & Kao, J. (2008). *Examining differential distractor functioning in reading assessments for students with disabilities*. (CSE Report No. 743). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Baker, F. B. (1990). Some observations on the metric of BILOG results. *Applied Psychological Measurement*, *14*, 139-150.

Bock, R. D.,& Zimowski, M. (1996). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Burket, G. R. (2002). *PARDUX. Version 6.1 [Computer software]*. Monterey, CA: CTB/McGraw-Hill.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16,*129-147.

Laitusis, C. C., Maneckshana, B., Monfils , L., & Ahlgrim-Delzell, L. (2009). Differential item functioning comparisons on a performance-based alternate assessment for students with severe cognitive impairments, autism and orthopedic impairments. *Journal of Applied Testing Technology, 10 (2)*. Retrieved from http://www.testpublishers.org/assets/documents/Differential_Item_article_6.pdf

Davey, T., Oshima, T., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*,*29*, 405-416.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika*, *49*, 175-186.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Holland, W. P., & Thayer, D. T. (1988). Differential item performance and the Mantel_Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: LEA.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics, 38,* 32-60.

Kim, S. H., Cohen, A. S., & Park, T. H. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, *8*, 291-312.

Kolen, M., & Brenna, R. (2004). *Test equating, Scaling, and Linking: Methods and Practices*. Springer Science+Business Media, Inc.

Laitusis, C. C., Maneckshana, B., Monfils, L., & Ahlgrim-Delzell, L. (2009). *Differential item functioning comparisons on a performance-based alternate assessment for students with severe cognitive impairments, autism and orthopedic impairments.* Retrieved from www.jattjournal.com/index.php/atp/article/view/35/21

Li, Y., Bolt, D. M., & Fu, J. (2006). A multiple-group testlet model and its application to DIF assessment. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Li, Y., & Lissitz, R. W. (2000) An Evaluation of Multidimensional IRT Equating Methods by Assessing the Accuracy of Transforming Parameters onto a Target Test Metric. *Applied Psychological Measurement*, *24*, 115-138.

Lin, P. (2008). *IRT vs. factor analysis approaches in analyzing multigroup multidimensional binary data: The effect of structural orthogonality, and the equivalence in test structure, item difficulty, and examinee groups* (Unpublished doctoral dissertation). University of Maryland, College Park.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Welsley.

Magis, D., Raiche, G., Beland, S., & Gerard, P. (2011). A logistic regression procedure to detect differential item functioning among multiple groups. Unpublished manuscript.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *The Journal of National Cancer Institute*, *22*, 719-748.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, *14*, 611-635.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Mislevy, R. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11*, 81-91.

Mislevy, R. J., & Bock, R.D. (1990). *BILOG 3* (2nd ed.). Mooresville, IN: Scientific Software Inc.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement*, *36*, 217-232.

Muraki, E., & Bock, R. (1991). *PARSCALE: Parametric Scaling of Rating Data*. Chicago: Scientific Software International, Inc.

Penfield, R. D. (2001). Assessing differential item functioning across multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, 14, 235-259.

Reckase, M. D. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice*, *14*, 12-14.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 410-412.

Reckase, M. D., & Martineau, J. (2004). *The vertical scaling of science achievement tests.* Paper commissioned by the Committee on Test Design for K-12 Science Achievement, Center for Education, National Research Council.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*, 215-230.

Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing*, *3*, 365-384.

Shealy, R., & Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF, *Psychometrika*, *58*, 159-194.

Sympson, J. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82-98). Minneapolis: University of Minnesota.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity*, (pp. 147-169). Hillsdale, NJ: Erlbaum.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory. An analog for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice*, (pp. 245-270). Boston, MA: Kluwer-Nijhoff.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and application. *Applied Psychological Measurement*, *26*, 109-128.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide.* Princeton, NJ: Educational Testing Service.

Yao, L. (2010). *BMIRTII: Bayesian Multivariate Item Response Theory [Computer program]*. Defense Manpower Data Center, Monterey, CA.

Yao, L. (2004). Bayesian Multivariate Item Response Theory and BMIRT software. *2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]*. Alexandria, VA: American Statistical Association.

Yao, L. (2003). *BMIRT: Bayesian Multivariate Item Response Theory. [Computer software]*. Monterey, CA: CTB/McGraw-Hill.

Yao, L., & Li, F. (2010, May). "A DIF Detection Procedure in Multidimensional Framework and Its Applications." Paper presented at the annual meetings of the National Council on Measurement in Education, Denver, CO.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item-writing practices. *Educational Assessment*, *9*, 61-78.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]*. Chicago: Scientific Software International.

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 (Ontario), Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233-251.

SOFTWARE PACKAGES FOR MULTIPLE GROUP IRT ANALYSIS AND ACCURACY OF PARAMETER ESTIMATES

1. Multiple group IRT analysis

(1) Application of multiple group IRT analysis
- Different groups with different test forms
  - Nonequivalent group horizontal equating of test forms
  - Vertical equating of test forms
- Different groups with the same test form
  - DIF analysis & DRIFT analysis
  - Concurrent calibration of item parameters
  - Estimation and comparison of latent ability across groups

(2) Main issues for multiple group IRT analysis
- Model identification
  - Given that scale for latent ability in any IRT model is arbitrary, one must assign the metric of latent ability during calibration.
  - In a multiple group IRT analysis, many programs assign the metric of ability distribution by setting the mean and standard deviation of the ability distribution for one group (i.e., reference group) at mean of 0 and standard deviation of 1 while the mean and standard deviation of the ability distribution for the other groups (i.e., focal group) are freely estimated.
  - In addition, item parameters need to be constrained for model identification purposes. In general, for 1PL model at least one difficulty parameter across groups needs to be constrained while for 2 PL model, at least both one discrimination- and one difficulty parameter need to be constrained across groups.
- Common scale or metric across different groups (Commonality)
  - Another issue that arises in multiple group IRT analysis is commonality. Commonality can be achieved when item and ability parameter estimates are on the same scale or metric across different groups.
  - Given that different groups have different ability levels, item parameters or ability parameters from separate IRT analysis can be on different scales. If item and ability parameter estimates are not on the common scale, comparisons across groups may not be meaningful.
  - In general, a set of anchor items is selected and used to achieve commonality across different groups during calibration.

(3) Available software programs for multiple group IRT analysis
- Commercial: BILOG-MG, MULTILOG, IRTPRO, flexMIRT, and Mplus
- Noncommercial: BMIRT and FLIRT (R package).

2. Review of IRT software programs for multiple group IRT analysis

(1) Purpose of review

- Given that different software programs employ different defaults and/or options for model identification and commonality, it is important to understand what kinds of defaults and/or options each software program adopts for resolving the two main issues.
- The review of IRT software programs provides information particularly regarding these two main issues (model identification and commonality).
- The review focuses on use of software programs for multiple group IRT analysis in the context where the same test form is administered to different groups.
- Lastly, this review focuses on seven software programs for multiple group IRT analysis which include BILOG-MG, MULTILOG, IRTPRO, flexMIRT, Mplus, BMIRT, and FLIRT (R package).

(2) Review summary

- Constraint on ability distribution for model identification purpose
  - BILOG-MG, IRTPRO, flexMIRT, and BMIRT fix the ability distribution of reference group (i.e., first group) to mean of 0 and standard deviation of 1 while mean and standard deviation of ability distribution for second group are freely estimated by default.
  - In MULTILOG, reference group is the second group and thus the ability distribution of the second group is fixed to mean of 0 and standard deviation of 1 by default. For the first group, standard deviation of the ability distribution is fixed to 1 while mean of ability distribution is freely estimated. The MULTILOG does not allow the switch of the first and second group. Users may do so manually by reversing the data for the two groups, such that the second group becomes the first group.
  - Mplus fixes first factor loading (i.e., discrimination parameter) to 1 across groups by default while means and standard deviation of the ability distribution for all groups are freely estimated.
  - FLIRT fixes the mean of the ability distribution for first group to 1 while mean of second group and standard deviations of both groups are freely estimated.
  - The defaults for model identification can be overridden easily by users with IRTPRO, flexMIRT, Mplus, BMIRT within the programs while the other programs are not flexible.

- Constraint on item parameter for model identification and commonality purposes
  - For multiple group IRT analysis, BILOG-MG, MULTILOG, and FLIRT automatically assume that all item parameters are the same for different groups and thus all items are treated as anchor items by default. Thus these programs provide only one set of item parameter by default.
  - With IRTPRO, flexMIRT, Mplus, and BMIRT, users can select anchor items and use them to achieve commonality. That is, these programs do not assume that all item parameters are the same for different groups.

- With BILOG-MG and MULTILOG, users may trick the program to obtain group specific parameter estimates when employing multiple group IRT analysis. Users can do so by treating the same non-anchor items as different items across groups and change the file format accordingly.

▪ Model availability & data format

- BILOG-MG and MULTILOG can estimate multiple group unidimensional IRT models only while the other software programs can estimate multiple group IRT model for both unidimensional and multidimensional models. Specific IRT models that each software program can estimate are presented in Appendix I.
- Different software programs use different data formats. Particularly when BILOG-MG or MULTILOG is used for group-specific calibration, the non-anchor items should be treated as different test items and construct data file accordingly. In addition, flexMIRT requires separate group data files while the other programs require one data file for multiple group IRT analysis. More detailed information can be found in Appendix I.

(3) Conclusions- Choice of software programs for multiple group IRT analysis

▪ When the same test form is administered to different groups (i.e., different states), equality of all item parameters may not be tenable. In this case, use of multiple group IRT model is one approach to take group heterogeneity into account.

▪ Seven software programs (BILOG-MG, MULTILOG, IRTPRO, MPLUS, flexMIRT, FLIRT (R package), BMIRT) are reviewed in this document and we found that IRTPRO, MPLUS, and flexMIRT seem to be more flexible software programs for multiple group analyses in terms of model identification and commonality. With these three programs, users can easily put item parameters on a common scale across groups.

Multiple group Rasch model analysis with simulated data

(1) Purpose of analysis

▪ The purpose of the analysis was to (1) examine the impact of different choices of IRT models (single vs. multiple group) on item parameters and ability parameters and (2) evaluate available multiple group IRT software programs for multiple group IRT analysis.

(2) Data simulation

▪ For the full scaled simulation study, the IRT models used in PARCC assessment will be selected.

▪ Manipulated factors in full scaled simulation study have 9 factors:

- Number of groups (2 groups; reference group vs. focal group)
- Sample size (3 levels: 1,000, 2,000, and 4,000 per group)
- Number of total items (50 items)
- Number of anchor items (15items)
- Number of DIF items (2 levels: 5 items (10%) and 10 items (20%))
- Magnitude of DIF (2 levels: 0.5 and 1.0)

- Type of DIF (2 levels: uniform vs. non-uniform)
- Mean ability difference between two groups (0.5; Group1=0 vs. Group2= - 0.5)
- Standard deviation difference in ability distribution across two groups (2 levels: 0 and 0.5).

▪ As a preliminary simulation, one data set following the RASCH model was simulated under the following conditions and analyzed:

- Number of groups (2 groups)
- Sample size (4,000 per group)
- Number of total items (50 items)
- Number of anchor items (15items)
- Number of DIF items (5 items=10%)
- Magnitude of DIF (0.5)
- Type of DIF (uniform)
- Mean ability difference between two groups (0.5)
- Standard deviation difference in ability distribution across two groups (0=no difference).

(3) Preliminary Data Analysis

▪ The software program *R* (R Development Core Team, 2010) was used to generate item response data following the Rasch model. For the reference group, the true item difficulty parameters and ability distribution were generated from a normal distribution with mean of 0 and standard deviation of 1. So this is the matched case between ability distribution and item difficulty parameters. The DIF item parameters for the focal group were simulated by subtracting 0.5 from each reference group item's difficulty parameter. Generated item difficulty parameters for both groups are presented in Table 1 in Appendix II.

▪ With simulated data, three different Rasch analyses were performed (single group, multiple group with equality constraint on all items, vs. multiple group analysis with equality constraint on anchor items) using seven different software programs.

▪ For model identification, the ability distribution for the reference group was set at mean of 0 and standard deviation of 1 while the ability distribution for the focal group was freely estimated.

▪ To obtain group specific item parameters with BILOG-MG and MULTILOG, we changed the data file such that the same non-anchor items treated as different items across groups.

▪ For the MULTILOG program, we manually reversed the data for two groups in data file so that reference group matched other software programs.

(4) Results

▪ When ignoring group heterogeneity in item parameters and using a single group IRT analysis for calibration, the results revealed that the single group IRT analysis produced large amount of biases of both item- and ability parameter estimates.

▪ When comparing two multiple group IRT analyses, the results provided similar results. This result might be because both the number of DIF item and the magnitude of DIF were relatively small.

▪ From the software program perspective, BILOG-MG, IRTPRO, Mplus, and flexMIRT provided similar results across three analyses in terms of item- and ability parameter estimates. MULTILOG, FLIRT, and BMIRT provided dissimilar results from BILOG-MG, IRTPRO, Mplus, and flexMIRT. The factors affecting these results will be examined in full scaled simulation study.
▪ The bias values of the item and ability parameter estimates are presented in Table 2 and 3 in Appendix II.

(5) Implication

▪ When the same test form is administered to different groups (e.g., states) like assessment for Common Core State Standards, it is typically the case that group homogeneity cannot hold. Therefore, it is important to take such group heterogeneity into account during the process of calibration and ability estimation.
▪ The preliminary results support that ignoring group heterogeneity and using single group IRT analysis may produce biases in the item and ability parameter estimates and thus may not provide fair results and accurate measures of students' achievement. The results support that multiple group IRT analysis produced lower biases in item and ability parameter estimates.
▪ Multiple group IRT model could be one useful approach for analyzing item response data from different populations by taking group heterogeneity into account.
▪ In the full scaled simulation study, the IRT models used in PARCC assessment will be chosen for the simulation. Rigorous examination will be conducted to (1) investigate the consequences of ignoring group heterogeneity in item and ability parameter estimates under various conditions, and (2) evaluate available multiple group IRT software programs with a large number of replicated data set.

Appendix I.

Software Programs for Multiple Group IRT Analysis

BILOG-MG (Michele Zimowski, Eiji Muraki, Robert Mislevy & Darrell Bock)			
Multi-Group IRT Model	Model Identification	Commonality	Note
Fits ▪ Multiple group 1PL, 2PL, and 3PL models to binary data	▪ It automatically fixes the mean and standard deviation of the ability distribution for first group (mean of 0 and standard deviation of1) while mean and standard deviation of the ability distribution for second group are freely estimated. ▪ It constrains all item parameters to be equal for different groups by default when employing multiple group IRT analysis.	▪ By putting equality constraints on all item parameters, it assumes that item and ability parameter estimates are on the same scale or metric. ▪ In case where users want to use a set of anchor items to achieve commonality, user should use the trick. User can do so by changing data file format. That is, users should treat the same non-anchor items as different items across groups and change the file format accordingly.	▪ BILOG-MG provides Graphical User Interface(GUI). ▪ BILOG-MG can handles omits and not-presented data ▪ For DIF analysis, BILOG-MG estimate item parameters separately for each group. Then the item parameters are adjusted under the constraint that the mean thresholds of the groups to be equal. However, BILOG-MG implements DIF/DRIFT analyses for only 1PL model. ▪ Obtaining group-specific item parameter estimates is not easy because users need to change data format outside of the BILOG-MG program.
MULTILOG (David Thissen)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
Fits ▪ Multiple-group 1PL, 2PL, or 3PL to binary data ▪ Multiple-group graded response model, nominal model, and multiple choice model to polytomous data	▪ It automatically fixes the mean and standard deviation of ability distribution for second group (mean of 0 and standard deviation of 1) while for first group, standard deviation of the ability distribution is fixed to one and mean of the ability distribution is freely estimated. ▪ It constrains all item parameters to be equal for different groups by default when employing multiple group IRT analysis.	▪ Similar to BILOG-MG, by putting equality constraints on all item parameters, it assumes that item and ability parameter estimates are on the same scale or metric. ▪ In case where users want to use a set of anchor items to achieve commonality, user should use the trick. User can do so by changing data file format. That is, users should treat the same non-anchor items as different items across groups and change the file format accordingly.	▪ It provides GUI. ▪ With common item method, it needs to be run twice for DIF analysis. ▪ Like BILOG-MG, obtaining group-specific item parameter estimates is not easy because users need to change data format outside of the MULTILOG program. ▪ It provides DIF analysis.
IRTPRO (Cai, Thissen & du Toit, 2011)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
Fits ▪ Multi-group 1PL, 2PL, or 3PL logistic model to binary data ▪ Multi-group rating	▪ It automatically fixes the mean and standard deviation of the ability distribution for first group (mean of 0 and standard deviation of one) while mean	▪ There is no default for commonality. ▪ It is easy for users to select the anchor items and constrain parameters of the anchor items	▪ It provides GUI. ▪ It provides DIF or drift analysis for various IRT models including unidimensional and multidimensional model.

scale model, partial credit model, generalized partial credit, graded response model, or nominal model to polytomous data ▪ Multi-group bifactor model ▪ multidimensional model of the above models	and standard deviation of the ability distribution for second group are freely estimated. ▪ It is easy for users to change default option for the model identification. ▪ Users have an option for choosing a set of anchor items and putting equality constraints on anchor items across groups.	to be equal across groups for achieving commonality.	▪ Generally, it is easy to run IRT models.
flexMIRT, Version 2.0 (Cai,2013)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
Fits ▪ multiple-group 1PL, 2PL, and 3PL logistic model ▪ multi-group generalized partial credit model, partial credit model, and rating scale model ▪ multi-group bifactor model ▪ multilevel and multidimensional models of all the preceding models	▪ It automatically fixes the mean and standard deviation of the ability distribution for first group (mean of 0 and standard deviation of 1) while mean and standard deviation of the ability distribution for second group are freely estimated. ▪ It is easy for users to change default option for the model identification. ▪ Users have an option for choosing a set of anchor items and putting equality constraints on anchor items across groups.	▪ There is no default for commonality. ▪ It is easy for users to select the anchor items and constrain parameters of the anchor items to be equal across groups for achieving commonality.	▪ It provides DIF or drift analysis for various IRT models including unidimensional and multidimensional model. ▪ For multiple group IRT analysis, data for each group should be stored as individual files.
Mplus (Muthén, & Muthén, 1998-2014)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
Fits ▪ multiple group1PL and 2PL logistic model to binary data ▪ multi-group Samejima's graded response model to polytomous data ▪ multilevel and multidimensional models of all the preceding models	▪ It automatically fixes the first factor loading (i.e., discrimination parameter) to one across groups by default while means and standard deviations of the ability distributions for all groups are freely estimated. ▪ It is easy for users to change default option for the model identification. That is, users can fix the mean and standard deviation of the ability distribution for first group while mean and standard deviation of the ability distribution for second group are freely estimated. ▪ Users have an option for choosing a set of anchor items and putting equality constraints on anchor items across groups.	▪ There is no default for commonality. ▪ It is easy for users to select the anchor items and constrain parameters of the anchor items to be equal across groups for achieving commonality.	▪ It uses different parameterization from other IRT software programs. For difficulty parameters, users can obtain by dividing threshold parameters in Mplus output by corresponding factor loading. ▪ It provides DIF or drift analysis for 1PL, 2PL, and graded response model. ▪ 3PL model cannot be estimated with Mplus.

BMIRT (Lihua Yao, 2003,2010)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
Fits ▪ multiple-group 1PL, 2PL, or 3PL logistic model to binary data ▪ multiple-group generalized two-parameter partial credit model, the testlet model, the graded response model, and the higher-order IRT model. ▪ multidimensional model of the above models	▪ It automatically fixes the mean and standard deviation of the ability distribution for first group (mean of 0 and standard deviation of 1) while mean and standard deviation of the ability distribution for second group are freely estimated. ▪ Users have an option for choosing a set of anchor items and putting equality constraints on anchor items across groups.	▪ There is no default for commonality. ▪ It is easy for users to select the anchor items and constrain parameters of the anchor items to be equal across groups for achieving commonality.	▪ It is non-commercial program. Users need to email to author and then download them program. ▪ It uses Bayesian estimation. Users need to specify information on parameters of the priors and proposals in code. So users who don't know Bayesian estimation, it would not be easy to use BMIRT software. ▪ For IRT analysis, users need to install and specify the path JAVA environmental variable on computer ▪ The program code is quite different from other IRT program.
FLIRT (Jeon, Rijman, & Rabe-Hesketh, 2014)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
Fits ▪ multiple-group 1PL, 2PL, and bifactor model ▪ multidimensional models of all the preceding models	▪ It fixes the mean of the ability distribution for first group (mean of 0) while mean of the ability distribution for second group is freely estimated. Standard deviations of the ability distributions for all groups are estimated. ▪ It constrains all item parameters to be equal for different groups by default when employing multiple group IRT analysis.	▪ By putting equality constraints on all item parameters, it assumes that item and ability parameter estimates are on the same scale or metric. ▪ According to author, she will going to add an option for putting equality constraints on anchor items soon.	▪ Item covariates or person covariates can be included in IRT model. ▪ It provides DIF analysis. ▪ With the current version, FLIRT cannot be used for 3PL and polytomous IRT models. ▪ Installation process is complicated. "FLIRT" package requires the Matlab Compiler Runtime (MCR) and it should be downloaded and installed before installing "FLIRT" package.

Appendix II.

Simulation Study of Multiple Group IRT Analysis

Table 1

True Difficulty Parameter in Reference and Focal Group

Item #	Difficulty Parameter (Reference Group)	Difficulty Parameter (Focal Group)	Item #	Difficulty Parameter (Reference Group)	Difficulty Parameter (Focal Group)
1	-0.136	-0.136	26	**0.234***	**0.734***
2	-0.041	-0.041	27	**0.593***	**1.093***
3	1.011	1.011	28	**2.001***	**2.501***
4	-0.158	-0.158	29	**-1.837***	**-1.337***
5	-2.157	-2.157	30	**-0.862***	**-0.362***
6	0.499	0.499	31	1.583	1.583
7	-0.755	-0.755	32	0.155	0.155
8	0.779	0.779	33	-0.275	-0.275
9	0.755	0.755	34	0.788	0.788
10	-1.100	-1.100	35	-0.223	-0.223
11	0.167	0.167	36	1.392	1.392
12	-0.029	-0.029	37	-0.489	-0.489
13	1.876	1.876	38	0.137	0.137
14	0.245	0.245	39	0.004	0.004
15	0.702	0.702	40	-0.727	-0.727
16	-0.015	-0.015	41	-0.721	-0.721
17	-0.143	-0.143	42	-0.191	-0.191
18	0.321	0.321	43	1.335	1.335
19	0.122	0.122	44	0.356	0.356
20	-0.595	-0.595	45	0.843	0.843
21	-0.442	-0.442	46	0.775	0.775
22	0.291	0.291	47	0.080	0.080
23	0.724	0.724	48	-0.673	-0.673
24	0.460	0.460	49	1.836	1.836
25	0.185	0.185	50	-0.207	-0.207

Note.

* indicate DIF items.

Table 2
Comparison of Bias of Difficulty Parameter Estimates across Three Analyses

Software	Single Group Analysis[1]	Multiple Group Analysis- Equality constraint on all items[1]	Multiple Group Analysis- Equality constraint on anchor items
	Average of bias across all items (SD)		
BILOG-MG	0.154 (0.163)	-0.042 (0.151)	R: 0.008 (0.033) F: 0.031 (0.039)
MULTILOG	<0.001 (7.223)	<0.001 (0.080)	R: <0.001 (0.067) F: -0.091 (0.159)
IRTPRO	0.281 (0.078)	0.007 (0.078)	R: 0.005 (0.030) F: 0.021 (0.038)
Mplus	0.262(0.086)	0.002 (0.078)	R: 0.001(0.034) F: 0.015 (0.044)
flexMIRT	0.280 (0.078)	0.006 (0.078)	R: 0.003 (0.030) F: 0.019 (0.038)
FLIRT	-0.617 (1.692)	-0.338 (1.693)	-
BMIRT	0.073 (0.072)	-0.072 (0.077)	R: 0.006(0.006) F: 0.131(0.053)

Note.
True item parameters for focal group were used to calculate bias in single group analysis and multiple group analysis with equality constraint on all items.

Table 3
Comparison of Bias of Ability Estimates across Three Analyses

Software	Single Group Analysis[1]	Multiple Group Analysis- Equality constraint on all items	Multiple Group Analysis- Equality constraint on anchor items
		Average of bias across all ability (SD)	
BILOG-MG	R: 0.177(0.325) F: 0.332 (0.524)	R: 0.002(0.325) F: 0.156(0.524)	R: 0.002 (0.325) F: 0.074 (0.336)
MULTILOG	R: -0.531(0.346) F: -0.501 (0.352)	R: -0.002 (0.317) F: -0.038 (0.324)	R: -0.041 (0.317) F: -0.042 (0.324)
IRTPRO	R: 0.248(0.313) F: 0.261 (0.324)	R: 0.001(0.313) F: -0.043(0.325)	R: 0.001 (0.313) F: 0.017 (0.325)
Mplus	R: 0.243(0.318) F: 0.278 (0.328)	R: 0.004(0.316) F: -0.026(0.326)	R: 0.003 (0.316) F: 0.035(0.326)
flexMIRT	R: 0.247(0.313) F: 0.260 (0.324)	R: <0.001(0.313) F: -0.045(0.325)	R: <0.001 (0.313) F: 0.015 (0.325)
FLIRT	R: 0.238 (0.317) F: 0.263 (0.326)	R: 0.007 (0.316) F: 0.512 (0.325)	-
BMIRT	R: 0.050(0.326) F: 0.066 (0.336)	R: -0.073 (0.326) F: -0.120 (0.338)	R: -0.008 (0.329) F: 0.161 (0.341)

References

Cai, L. (2013). flexMIRT® version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software].Lincolnwood, IL: Scientific Software International.

Jeon, M., Rijmen, F. and Rabe-Hesketh, S. (2014). flirt: Flexible Item Response Theory Modeling. R package version 1.13. http://www.mathworks.com/products/compiler/mcr/

Muthén, B., & Muthén, L. K. (1998-2014). MPLUS (Version 7.11). [Computer Software]. Los Angeles, CA: Muthén & Muthén.

Thissen, D. (1991). MULTILOG: Multiple category item analysis and test scoring using item response theory [Computer software]. Chicago: Scientific Software International

Yao, L. (2010). BMIRTII: Bayesian multivariate item response theory—second version. [Computer software]. Monterey, CA: *www.BMIRT.com*.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.

COGNITIVE DIAGNOSTIC MODELS

Introduction

The Council of Chief State School Officers (CCSSO) announced in 2006, as part of the ongoing effort to improve educational achievement across states, to establish the Formative Assessments for Students (FAST) State Collaborative on Assessment and Students Standards (SCASS), to help push forward teaching and learning through innovative uses of formative assessments. The initiative has turned to experts together with Education Testing Service to come up with practical tools for diagnosis and improvement. In addition, FAST SCASS has turned to participating states such as Maryland as a member of the Partnership of Assessment Readiness for College and Careers (PARCC) to implement useful formative assessment tools in all classrooms.

Under the Race to the Top program, Maryland schools recently adopted the common-core state standards and will begin operational PARCC test based on these new standards starting in the 2014-15 school year. A new feature being implemented to go in tandem with the new standardized tests are computer adaptive diagnostic assessments called K-1 formative tools. For more information, the reader is referred to the PARCC website which provides additional detail on the upcoming development of this system (http://www.parcconline.org/non-summative-assessments). These assessments will help teachers target students, especially those at low ability levels, to focus on areas that may be hindering them from progressing. Such assessments would also help teachers spend less time planning for interventional activities.

In 2013, PARCC, on behalf of the state of Florida, issued a procurement of services for the establishment of diagnostic and formative assessment tools, to go with the new summative assessments to give additional attention to students who are behind relative to their peers. According to the proposal for a computerized diagnostic system, the mathematics portion of the diagnostic assessment design is required to diagnose students based on the skills within the

mathematics domains as described by the Common-Core State Standards, to inform teachers of students that need additional work on specific sub-domains.

In response, PARCC has set out to implement a diagnostic system that will address these requirements. Their goal is to develop easy to use and understand online diagnostic assessments that will allow for immediate diagnostic results. For more detailed information on the goals of this new innovative diagnostic assessment, one may refer to the PARCC documentation on K-1 diagnostic assessments. These formative assessments are set to be field tested in early 2015 and available for use in the 2015-2016 school year.

Behind these ambitious goals to implement computerized diagnostic assessments are the theories that would allow for proper diagnosis of students' learning using Cognitive Diagnostic Models (CDM). CDMs are the building blocks to diagnostic assessments which can be used by teachers and students by giving them fine-grained information to help teachers differentiate instructional needs, dynamically group students according to mastery of skills, assign individualized assignments and develop individualized intervention plans based on these diagnostic profiles. For example, homework assignments for a mathematics class could be individualized by assigning problems that heavily focus on their weaknesses according to the skills that students do not master based on the diagnostic information provided by the formative assessment. Teachers could also group students according to the skills they lack and have sub-lessons for these students as a way of differentiated instruction. Essentially, CDMs output information about students' strengths and weaknesses that could be used to help teachers and students focus on the skills that they really need help on.

The purpose of this document is to provide some theoretical background on these diagnostic models by introducing some technical terminology and then providing an overview of

some models that could be used in practice. The goal is to prime the way for more analyses regarding these types of diagnostic assessments, in an overall effort to provide background information about some commonly studied diagnostic models that might be useful for the stakeholders of such tests in Maryland. Ultimately, it is hoped that this review will shed light on which models will serve best for giving students and teachers the more accurate and useful information that is in line with the definition of formative assessments as proposed by the CCSSO, PARCC, and Maryland public schools.

Basic Terminology of Cognitive Diagnostic Models (CDM)

The following is an overview of the basic terminology used in cognitive diagnostic models. This report is not an all-inclusive review of all CDMs. Some more theoretically complex models will be mentioned but not elaborated upon due to their impracticalities in practice. Before discussing the models in detail, a few common characteristics about the models will be described, including, compensatory vs. non-compensatory, latent trait vs. latent class models, their handling of slipping and guessing parameters, and finally, the specification of the skills or what is otherwise known as Q-matrix specification.

A skill is referred to as a single latent attribute that a person possesses. For example, within the context of mathematics, a skill would be being able to add two digit numbers. A skill could be more complex than simple addition. For example, a skill in mathematics could also be "solving linear equations", which is a far more complex skill and would include the skill of adding two digit numbers. This level of complexity and relationship between skills needs to be defined when specifying what is called a Q-matrix before any actual analysis can even take place.

Compensatory and non-compensatory models. CDMs can be compensatory, non-compensatory, or both. These terms refer to how skills are related to modeling the probability of a correct response. In Item Response Theory, the probability of a correct response to a certain item by a certain student is modeled by using:

1) Information about the student based on their responses to the items in the overall test (referred to a person's ability), and

2) Information about the characteristics of the item (difficulty of the item based on how others perform on the item)

The probability of a correct response to an item is given by a student's ability and difficulty of the item. This relationship is modeled by:

$$P(X = 1|\theta) = \frac{\exp(\theta_i - b_j)}{1+\exp(\theta_i - b_j)},$$

where P is the probability of a correct response, θ_i is student i's ability and b_j is the difficulty of the item.

CDMs specify which skills are necessary to have in order to increase the probability of a correct response on the item. In unidimensional IRT, the skill might be general mathematics ability. But CDMs allow for a more specific specification of skills that might allow one to assess the skills within mathematics ability that might lead to a correct response. Some items might need just one skill within mathematics ability. Others might require several skills. For example, a mathematics question in the context of a word problem would require the student to know how to read, synthesize information, and have some basic mathematics skills to correctly answer the item. In this light non-compensatory models assume that the student must have mastered all the skills within the item in order to get the item correct.

Non-compensatory. The example that was just described is a type of non-compensatory model because one can imagine that if any one of the skills described were missing (say the student could not read), then the student will be less likely to get the item correct. In essence, a high competency in one skill cannot "compensate" for the lack or low competency of other skills. If the student lacks a single skill needed to answer the item correctly, the student is likely to get the item incorrect. Non-compensatory models are sometimes referred to conjunctive, meaning that all skills need to be mastered for a high probability of success on an item. In the example of a mathematics test that requires high level of language proficiency, having high reading ability will not compensate for the lack of basic math skills or vice versa, having high math skills will not compensate for low reading ability if the student cannot understand the question.

Compensatory. Sometimes, certain skills compensate for the lack of others. Models that allow for this type of relationship are called compensatory models. Compensatory models imply that the student's possession of one skill can compensate for the low proficiency or the lack of the other skills that are required for answering the item correctly. In the context of education, compensatory modeling is much more studied in the summative assessment setting for subdomain score reporting.. Such compensatory modeling is referred to as disjunctive.

Latent class and latent trait model based CDMs. CDMs have been developed along two different frameworks: latent trait models and latent class models. The main difference between the two frameworks is that whether the latent attributes are measured on a continuum or as discrete categories.

Latent class models. For latent class model based CDMs, the diagnostic information is specified at either dichotomous levels of mastery or non-mastery of skills or polytomous levels such as basic, proficient, and advanced. For the latent trait model based CDMs, the diagnostic information is provided along a continuous scale. Often a cut point or multiple cut values could be used to indicate different levels of proficiency. While the former provides grouping information but no information is available about how far a student is away from a target while latter provides more fine-grained information and quantifies the relative distance between a student's current stand and how far the student is from an achievement target.

Latent-class CDMs make explicit the probability of a correct response on an item based on an student's mastered attribute profile, which are known as the latent classes. The different combinations of attributes that have been mastered differentiate different latent classes. If for example, a test contains items that measure two different attributes, then there will be 4 different possible latent class memberships: students who have mastered neither attributes, students who have mastered all attributes, and two classes representing that only one attribute has been mastered. In essence, the number of classifications that a student can fall into is finite.
The first CDMs were developed from this framework. The outcomes from these models can inform whether a student has mastered a certain skill or not. In this regard, a disadvantage of this latent class approach is that students are only classified into a certain classification profile indicating mastery or non-mastery of each specific skill. Without detailed information about how much of the skill a student possesses and whether a student is close to the master threshold or still far away from the threshold, it is still not informative enough to provide guidance for instructions to make up the deficiency.

Latent trait models. Latent trait models allow for estimation for each skill. This allows for more specific information, like how much of the skill a student possesses. In this regard, latent trait models are more useful in providing diagnostic feedback to teachers about planning a targeted remedial instruction. For example, a teacher could devise different remedial instructions to adapt to the needs of students who might have different levels of skill proficiency as we do expect that students who are about to master the skills need different amount of remedial instructions compared with those students who are still far away from the mastery of a latent skill.

Slipping and guessing. Different CDMs treat slipping and guessing parameters distinctively. Slipping refers to an instance where a student of supposedly high ability answers an easy item incorrectly. Guessing is the opposite: a student with low ability answers a difficult item correctly. Some CDMs model these parameters at the item level, others model them at the attribute level, and yet some CDMs take a combined perspective at both item and attribute level.

Q-matrix. CDMs function through what are called Q-matrices. These matrices contain rows for each item and columns for all the possible attributes found in a test. For each cell, a 1 indicates that the item measures the skill. The cell contains a zero if the item does not measure the skill. Cells can also contain polytomous values if one wishes to specify the degree of mastery of the attribute measured by the item. As a simple example, figure 1 shows some sample items.

item 1: 4+1

item 2: 4*3-2

item 3: (4/2)-1

Figure 1. Sample items for a mathematics test assessing basic algebra arithmetic skills.

These items require the following skills: addition, subtraction, multiplication and division. A Q-matrix for this simple example would look like the following:

Table 1. Sample Q-matrix corresponding to figure 1.

	Addition	**Subtraction**	**Multiplication**	**Division**
Item 1	1	0	0	0
Item 2	0	1	1	0
Item 3	0	1	0	1

Specifying the Q-matrix is analogous to specifying the loadings in a factor analysis framework. Models use Q-matrices together with response data to produce student profiles about which skills individual students may or may not possess. Therefore, specification of the Q-matrix is one of the most important parts of the CDM application, as misspecification can lead to inaccurate, incorrect, and/or meaningless diagnosis (Rupp & Templin, 2008; DeCarlo, 2011).

Constructing a Q-matrix is as complicated as the type of items that make up the exam. For the simple example above, a simple inspection of items would be sufficient. Other methods include multiple rater inspection or iterative procedures based on item parameters (Rupp, Templin, & Henson, 2010). However, one can imagine that the Q-matrix construction can become complicated and will bring in some subjectivity when items become complex. For example, to get a sense of how complex it may become, sometimes the attributes may be hierarchically related, meaning that some skills will be pre-requisites of others. These Q-matrices are usually constructed by content experts based on their knowledge of the skills they believe are

required to answer an item correctly. Because of the subjective nature of this task, several advanced methods have been developed to properly specify Q-matrices that are complex (DiBello, Stout, and Roussos, 1995; de la Torre, 2008; deCarlo, 2011). For example, some models that capture skills not specified by the Q-matrix have been developed (MLTM-D; Embretson, & Yang, 2012).

Incompleteness of a Q-matrix refers to the assumption that there exist additional strategies not represented by the Q-matrix, which allow for a correct response on an item. In some instances, a Q-matrix does not capture all the possible strategies that a student might take to correctly answer an item. Some models will assume that all the possible strategies are accounted for, while others allow for this assumption to be relaxed. Such models allow for more realistic situations, especially when items are complex and can be solved in many different ways.

Cognitive Diagnostic Models (CDM)

The models discussed in this section will be explained with minimal technical details. For mathematical formulas of these models, the reader is directed to the appendix B at the end of the report (Rule Space Model and Restricted Latent Class model not provided). All the models discussed share the common goal to diagnose students, taking on different approaches according to the type of information desired. Models developed earlier are specific to certain contexts while more recent models are generalized, flexible versions of these earlier models, which can be specified to become these earlier models. The advantage of these flexible models is that they fit the realities of most testing situations that many of the simpler models cannot address. However, these models are oftentimes more difficult to estimate, they require very large sample sizes, and may sometimes be difficult to interpret. In reality, simpler models may suffice because they

provide reasonably simpler output that allows one to draw meaningful interpretations of complex data structures if certain assumptions met. Many operational CDMs today still use simpler models despite the recent development of more sophisticated models.

Latent class models. The following models fall under the category of latent class models. For a more technical and comprehensive list of these models, reader can refer to a thorough review of these latent class models by Roussos, Templin and Henson (2010) and work by Rupp, Henson, and Templin (2010).

Rule space. One of the earliest classification models of non-compensatory cognitive diagnostic methods was introduced by Tatsuoka in 1984 called the Rule Space model (RSM; Tatsuoka, 1985, K. Tatsuoka & M. Tatsuoka, 1989). The model can be used with both dichotomous and polytomously scored items, but is limited to dichotomous classifications, making it a type of latent class CDM model. RSM first uses traditional IRT to estimate ability and item parameters. It then uses this information, along with a reduced Q-matrix (reduced in the sense that certain dependencies among skills allow for a smaller number of permissible attribute profiles) to compute an expected latent variable corresponding to the expected score pattern from the reduced Q-matrix.

Slipping and guessing issues are considered as aberrant response patterns that are addressed by an atypical/caution index for each person. This index is essentially a residual value that shows how far away from the expected set of responses a student response falls. Large negative values of this index mean more incorrect responses than expected (slipping), whereas large positive values mean more correct responses than expected (guessing).

RSA can handle around 20 attributes and require sample sizes of approximately 1000 to 2000 student. A crucial disadvantage of the RSA is that because probabilities are only applied in separate steps, fits statistics that use common probability distributions to evaluate the entirety of the model do not exist. Also, use of the model is stymied due to the lack of easily accessible software. Many of the latent class models for CDM purposes will have similar properties to this model with some slight modifications. Software currently available for estimation using the rule space model can be done by BUGLIB.

RLCM. Latent class CDMs are considered restricted versions of the traditional latent class model (RLCM; Haertel, 1984, 1990). These models are restricted in the sense that the Q-matrix is limited to a certain number of response patterns. LCMs model the probability of a correct response on an item based on the response to the item, the student's response data across items, and the probability of a student being in a certain latent class. Levels of mastery are not continuous, but are instead modeled by vectors of 1's and 0's representing mastery and non-mastery of skills. The model says that if a student has mastered all the skills required by an item, then he/she will have a high probability of answering the item correctly, where as if the student has not mastered all the skills, this probability will be small. This probability stays the say no matter how many skills or which particular skills have not been mastered. This model has more recently referred to as the Deterministic input, noisy "and" gate mode (DINA; Junker and Sitjma, 2001), which will be subsequently discussed. Other latent class models have been derived from this framework.

DINA and DINO. The Deterministic Input; Noisy "And" Gate model (DINA) (Hartel, 1990; Junker & Sitjsma 2001) is a non-compensatory, conjunctive CDM. It allows for dichotomous and polytomous response variables but only dichotomous attribute variables. The

model is used to deterministically determine the items a student will get correct and incorrect, giving a high probability of a correct response only if all the required skills have been mastered. In other words, the probability of success in an item depends on having all the required skills within the item. Those who lack at least one skill within the item are classified the same regardless of which combinations of skills they have and lack. The DINA model allows for slipping and guessing, and the model allows for only one possible strategy to obtain a correct response. The probability of a slip is assumed to be zero. The Deterministic Input; Noisy "Or" Gate Model (DINO) (Templin & Henson, 2006, Templin 2006) is the compensatory analog of the DINA model. Extensions of the DINA/DINO model include the higher-order DINA (HO-DINA; de la Torre & Douglas, 2004), which allows for a hierarchical structure for the items, and the multi-strategy DINA (MS-DINA; de la Torre & Douglas, 2005), which allow for multiple strategies to solve the items. Software currently available for estimation of the DINA and DINO model include: DCM and LCDM in Mplus, DCM in R, DINA in Ox, MDLTM, BUGLIB, and AHM.

NIDA and NIDO. The Noisy Inputs, Deterministic "And" gate model (Maris, 1999; Junker & Sijtsma, 2001) is an extension of the DINA model. Like the DINA model, it is a non-compensatory conjunctive model, which allows for both polytomous and dichotomous responses, but only produces dichotomous attribute levels. Unlike the DINA model, the NIDA model differentiates between students who lack different combinations of skills required to answer an item correctly. For instance, one would assume that a student who possesses more skills within the item would have a higher probability of answering the item than an student who possesses fewer skills. Also, NIDA models the slipping and guessing parameters at the attribute level, relaxing difficulty levels across attributes, allowing for more fine grained profiles among those

who may not have all skills mastered but different combinations of them. The Noisy Input, deterministic-or-gate model (NIDO) (Templin, 2006) is the compensatory analog to the NIDA model. Estimation using NIDA/NIDO model can be done using the Mplus extension called DCM.

RUM. The Reperameterized Unified Model (RUM; Hartz, 2002), also known as the *fusion model*, further relaxes the restrictions made by the NIDA model. Specifically, it relaxes the constraint at the item level, allowing both different slipping and guessing parameters at the item and skill level. Essentially, the model captures the idea that the same skills within items can vary in difficulty. The RUM model is also explicitly acknowledges that the Q-matrix may not be complete, allowing for different strategies to solve the item correctly. Though these differences could certainly be specified in DINA/NIDA models, the RUM allows for a more parsimonious handling of such situations. RUM can be non-compensatory (Dibello et al., 1995; Hartz, 2002) as well as compensatory (Hartz, 2002; Templin, 2006).

Two types of RUMs exist. The full RUM and reduced RUM. Though both similar in regards to capturing differences at the item and skill level, the full RUM also addresses any model misfits that arise from an incomplete Q-matrix. This improves the chances of the data fitting a model, even with an incomplete matrix. This model, however, has been proven to be difficult to estimate with small sample sizes and number of items due to the extra parameters in the model.

General Latent Class DCMs. In addition to the models just described exist models that are more flexible in their parameterization, essentially allowing for the unification of the models that were just described. These models include the multiple classification latent class model (MCLCM**;** Maris, 1999), loglinear cognitice diagnostic model (LCDM; Henson, Templin,

Willse, 2009), and general diagnostic model (GDM; von Davier, 2005; Xu & von Davier, 2006), which will be briefly discussed. These complex models are oftentimes difficult to estimate due to their complexity. They require large sample sizes and many items as the number of skills and complexities within the items increase. As a result, some models, like the MCLCM, have been rarely used in operational settings and sometimes even programs to handle such models have not been developed with success.

LCDM. The log-linear cognitive diagnostic model (Henson, Templin, Willse, 2009) is a flexible model that allows relationships between categorical variables to be modeled using a latent class model. Using binary item response data, it uses the log-odds of a correct response, although it is denoted as a log-linear function. The LCDM is a generalization of the RUM, reduced RUM, DINA and DINO and can take on compensatory or non-compensatory effects. The model also provides insight to a more appropriate model for some items. The LCDM provides a model that can describe the functionality between attribute mastery and the item response probability without having to restrict the model as being conjunctive or disjunctive. In addition, it has the benefit of estimating attribute mastery while providing empirical information regarding a more reasonable model. An extension called LCDM on Mplus can be used to run estimations using the LCMD model.

GDM. Another more generalizable model that can be modified into the simpler models is the General Diagnostic Model (von Davier & Yamamoto, 2007; von Davier, 2005, 2007). The model allows for both compensatory and non-compensatory CDMs. GDM can be specified to include a variety of latent class models as well as models that include latent dimensions. It includes higher-level interactions, allowing for higher-order latent ability models. It can also be formulated to permit binary as well as ordered polytomous item responses. In this regard, the

LCDM is a specific version of the GDM. Data from large-scale tests such as the NAEP and TOEFL have been analyzed under GDM, but have not succeeded within the hierarchical framework from which the data were derived (von Davier, 2005; Xu & von Davier, 2008). Little information about the bias and precision of item or skill parameter estimates and classification accuracy for respondents has been made publicly available.

Latent-trait models. Another alternative to CDMs are latent-trait models like multidimensional item response models (MIRT) that capture information about the multidimensional nature within items to make appropriate diagnostics. MIRT models can provide more information about the different skills by representing skills in a continuum rather than discrete levels of competence. Like latent class approaches, latent-trait models can also be specified to assume compensatory or non-compensatory relationships among skills within items, although most studies have focused on the compensatory models because they are easier to estimate (Stout, 2007). For a more comprehensive review of latent-trait models, the reader is referred to the article by Stout (2007).

LLTM. The linear logistic test model (LLTM) (Fischer, 1973) is a unidimensional IRT model where skills are mapped onto the predicted difficulty levels of items for diagnosis. Student latent traits can be used to obtain the probability of solving items that contain certain skills when organized on a single dimension. Item difficulty is modeled by the skill structure within the item as well as the relative difficulties of the skills by means of weights. The model does not provide explicit diagnostic information, but rather, uses a Q-matrix to identify the skills that affect the difficulty of the item.

MIRT-C. The compensatory multidimensional item response theory model first introduced by Reckase and McKinley (1991), and then popularized by Adams, Wilson, and

Wang (1996, 1997) using the Rasch version is a type of compensatory latent trait model. Low ability level on one trait can be compensated by high values in other traits. The MIRT-C assumes a complete Q-matrix, which allows for less parameters to be estimated at the expense of a less than realistic situation in which alternative problem solving strategies are not considered.

MIRT-NC. In some instances, skills required for a correct response to an item are undeniably non-compensatory. For this reason, the non-compensatory multidimensional item response theory model (MIRT-NC; Sympson, 1978) was introduced. The model shows a multiplicative relationship that accounts for an overall low probability of a correct response to an item when at least one skill is low. Like the compensatory model, the model assumes a complete Q-matrix. In certain situations, this assumption may not be desired, such as when there exist multiple alternative strategies to obtain a correct response on an item. Later models (MLTM; Embretson, 1985, 1997) allow for a relaxation of this assumption of a complete Q-matrix.

MLTM. The multicomponent latent trait model (MLTM; Embretson, 1985, 1997; Whitely, 1980) is an extension of the MIRT-NC. Unlike the MIRT-NC model, MLTM allows for alternative strategies that are part of the individual skill level, meaning that the Q-matrix is allowed to be incomplete. For example, the model might be specified to allow an examinee to employ one of two strategies to correctly answer an item, one where all the skills required for the item are successfully executed, and the other where at least one skill is not successfully executed and the examinee guesses the answer correctly. In general, MLTM allows for multiple cognitive strategies that can be represented within the model. Examinees are assumed to go through the same order of strategy application. The order of each strategy application and the skills for each strategy are crucial to the model (DiBello, Roussos, & Stout 2007).

GLTM. The general component latent trait model (GLTM; Embretson, 1985, 1997) is a combination of the MLTM for a single strategy and the LLTM. The model estimates item and person parameters in the same was as MLTM, but adopts the complex attribute structure that influence the item difficulty of an item by defining the difficulty parameter similarly to the LLTM. The GLTM can be specified to model the LLTM and MLTM.

MLTM-D. The MLTM-D (Embretson and Yang, 2012) is a non-compensatory IRT model. It allows for a hierarchical relationship between components at one level and attributes nested within components at the second levels. It utilizes two matrices, a Q-matrix, used to represent components in one level, and a C-matrix, used to define the nested attributes within components. The model is a generalization of the MLTM and the GLTM model and can be applicable to measures of broad traits, such as achievement tests, in which the component structure varies between items. The number of parameters depends on how the exam structure is defined. The MLTM-D allows for a relatively small number of item parameters when attribute structures are defined within components because difficulty parameters can be represented by linear combinations of attributes and control variables that can be estimated using weights that estimate the impact of these attributes (Embretson & Yang, 2012). The MLTM-D is applicable for large-scale assessments with items that vary in the types of cognitive operations and skills needed to obtaining solutions, such as in end-of-year high stakes mathematics assessments that contain a broad range of concepts.

Applications of Diagnostic Models

MLTM-D with High-Stakes Mathematics Data (Embretson & Yang, 2012). The MLTM-D has been applied to a large-scale high-stakes test of 8th grade mathematics achievement. The test

consists of 86 multiple choice items scaled with a 3-PL logistic IRT model. Competency levels were evaluated with respect to proficiency categories set by experts based on analysis of items. The test contains four standard areas: (1) Number/computation, (2) Algebra, (3) Geometry, and (4) data analysis. Each standard contains benchmarks, and benchmarks contain indicators (25 total indicators) that define the skills needed to master the benchmark. Fitting the MLTM-D resulted in better fit than the DINA and LCDM models. In addition, students with similar overall scores near the mathematics competency borderline were diagnosed to provide diagnosis at the standard level, showing which standards students need remedial instruction on. In addition, the analysis provided the level of mastery within each standard.

GDM in Language Testing Data (von Davier, 2005). The GDM was applied to an Internet-based TOEFL test containing partial credit data for two forms, A and B, and with two sections, reading and listening. Four skills were identified for reading and four for listening, which were used to construct four distinct Q-matrices (Form A and Form B reading and Form A and Form B writing). Another Q-matrix was constructed for a joint analysis of the reading the listening to see if the two sections, reading and listening, needed to be analyzed separately. This matrix contained eight skills. All Q-matrices were retrofitted to existing tests. The GDM model was compared to a 2-PL IRT and two-dimensional 2-PL IRT model (for the aggregate data) for fit. A 2-PL ability parameter calibration was used as a benchmark for comparison of classifications from the diagnostic model. The analyses showed similarities between the skills across test forms. It also showed the need to clearly separate skills when making items used specifically for diagnostic test developments. Most specific information, such as examples of how students were classified was not provided.

Some Concluding Remarks

This report is by no means exhaustive of all the CDM literature that is currently available. The models that were presented here were given as examples as a basis to build on the more applicable and useful models that could be used for further research on diagnostic assessments with the new upcoming Maryland State Assessments under PARCC. We hope to build further evidence for using these models and find grounds for introducing them to educators.

Only a few of the many CDMs developed were mentioned in this report. For a more complete and comprehensive list of the models, the reader is directed to Roussos, Templin, and Henson (2007), Stout (2007), DiBello, Roussos, and Stout (2007), and Rupp, Henson, and Templin (2010).

References

Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory in practice* (pp. 143–166). Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles Policy and Practice*, *5*, 7-73.

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

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343-362.

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

de la Torre, J., & Douglas, J. A. (2005). *Modeling multiple strategies in cognitive diagnosis*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Montréal, QC, Canada (April).

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood based classification techniques. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Erlbaum: Hillsdale.

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. InC. R. Rao & S. Sinharay (Eds.),

Handbook of Statistics, Volume 26, Psychometrics (pp. 979–1030). Amsterdam, The Netherlands: Elsevier.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika*, *49*, 175–186.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometric s* (pp. 195-218). New York: Academic Press.

Embretson, S. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ: Erlbaum.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. L. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305-321). New York: Springer.

Embretson, S. E., Yang, X. (2012). A multicomponent latent trait model for diagnosis. *Psychometrika*, *78*, 14-36.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.

Formative Assessment For Students and Teachers (FAST) SCASS (2012). Distinguishing Formative Assessment from other Educational Assessment Labels. Published paper on CCSSO website. http://www.ccsso.org/Documents/FASTLabels.pdf

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement*, *8*, 333–346.

Haertel, E. H. (1990). Continuous and discrete latent structure models of item response data. *Psychometrika*, *55*, 477–494.

Hartz, S. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practice. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.

Henson, R. A., Templin, J., & Willse, J. (in press). Defining a family of cognitive diagnosis models, *Psychometrika*.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

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

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361–373.

Roussos, L., DiBello, L., Henson, R., Jang, E., & Templin, J. (2010). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S.E. Embretson (Ed.), *Measuring psychological constructs: advances in model-based approaches*. Washington: American Psychological Association.

Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*, 78–96.

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

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement*, *44*, 313–324.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D.J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference (pp. 82 98). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, *12*, 55–73.

Tatsuoka, M. M., & Tatsuoka, K K (1989). Rule space. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (pp. 217- 220). New York: Wiley.

Templin, J. L. (2006). Generalized linear mixed proficiency models for cognitive diagnosis. Manuscript under review.

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

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

von Davier, M. (2007). *Hierarchical general diagnostic models* (Research Report No. 07–19). Princeton, NJ: Educational Testing Service.

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

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*, A19-A9A.

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

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data* (Research Report RR-08-27). Princeton, NJ: Educational Testing Service.

Appendix I.

Table of Models Discussed in the Report.

	Compensatory	**Non-compensatory**
Latent Class models	DINA RSM NIDA LCDM GDM RUM	DINA NIDO LCDM GDM RUM RLCM
Latent Trait models	LLTM MIRT-C	MIRT-NC GLTM MLTM-D

Appendix B.

In the following formulas, the subscripts representing items are denoted by subscript *i*, persons by the subscript *j*, and attributes by the subscript *k*. In general, the following variables are defined as follows:

θ_j is the person parameter (unidimensional or multidimensional)
β_i is the difficulty of the item
η_k is the contribution of attribute *k* to the difficulty of the item
K is the number of dimensions of θ_j
q_{ij} is the score of item *i* on attribute *k* in the cognitive complexity of items

Latent class models

Deterministic input, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001)

$$P(X_{ij} = 1 \mid \xi_{ij}, s_i, g_i) = (1 - s_i)^{\xi_{ij}} g_i^{(1-\xi_{ij})}$$

where s_j is the probability slipping (possesses the skill, but gets item incorrect), g_i is the probability of guessing (does not possess the skill, but gets the item correct), and ξ_{ij} is the latent variable indicator, indicating whether examinee *j* has mastered all required attributes $(\xi_{ij} = 1)$ or has not $(\xi_{ij} = 0)$ for item *i*. Then,

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{jk}^{q_{ik}}$$

α_{jk} are latent vectors representing knowledge states. The probability of a response is only high when a student has mastered all required attributes.

Deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006)

$$P(X_{ij} = 1 \mid \omega_{ij}, s_i, g_i) = (1 - s_i)^{\omega_{ij}} g_i^{(1-\omega_{ij})}$$

Instead of using ξ_{ij} from the DINA model, the DINO model replaces this variable with ω_{ij} and is defined as follows:

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{jk})^{q_{ik}},$$

indicating whether examinee *j* has mastered at least one of the required attributes for item *i*. If examinee has mastered one or more of the attributes in item *i*, then $\omega_{ij} = 1$, otherwise $\omega_{ij} = 0$.

Noisy inputs, deterministic "and" gate model (NIDA; Maris 1999)

$$P(X_{tj}=1\,|\xi_{tj},s_t,g_t)=\prod_{k=1}^{K}\left[(1-s_k)^{\alpha_{jk}}g_k^{1-\alpha_{jk}}\right]^{q_{ik}}$$

where s_k and g_k are guessing and slipping probabilities defined as:

$s_k = P(\eta_{tjk}=0|\alpha_{jk}=1,q_{ik}=1)$,
$g_k = P((\eta_{tjk}=1|\alpha_{jk}=0,q_{ik}=1)$,
$P(\eta_{tjk}=1|\alpha_{jk}=a,q_{ik}=0)\equiv 1$, regardless of the value of a.

Noisy inputs, deterministic "or" gate model (NIDO; Templin & Henson, 2006)

$$P(X_{tj}=1|\,\alpha_{tj})=\frac{1}{[1+\exp\left(\sum_{k=1}^{K}(\tau_k+\beta_k\alpha_{tk})q_{jk}\right)]}$$

where τ_k represents lack of mastery of attribute *k*, and β_k is mastery of attribute *k*.

Reparameterized Unified Model (RUM; DiBello et al., 1995; Hartz, 2002)

$$P(X_{tj}=1\,|\alpha_j,\theta_j)=\pi_i^*\prod_{k=1}^{K}r_{ik}^{*(1-\alpha_{jk})q_{ik}}\left(P_{\beta_i}(\theta_j)\right),$$

where $P_{\beta_i}(\theta_j)=\frac{1}{1+\exp[-1.7[\theta_j-(-\beta_i)]]}$

and $\pi_i^*=\prod_{k=1}^{K}\pi_{ik}^{q_{ik}}$ and $r_{ik}^*=\frac{r_{ik}}{\pi_{ik}}$,

with π_i^* being the probability that examinee will correctly execute all mastered required skills for item *i* and r_{ik}^* is a reduction for each non-mastered skill.

Log-linear cognitive diagnosis model (LCDM; Henson et al., 2009)
For binary item response data, the probability of a correct response is given by:

$$P(X_{tj}=1\,|\alpha_j,q_i)=\frac{1}{(1+\exp(-1(\lambda_i^T h(\alpha_j,q_i)-\pi_i)))}$$

λ_i^T is as vector of weights for item *i* and $h(\alpha_j,q_i)$ are linear combinations of attributes in the item and attribute possession, and π_i is the probability of a correct response for examinees who have not mastered any attributes. The model obtains generality from the $h(\alpha_j,q_i)$ term, which can be a single term, α_{jk}, and interactions of terms depending on q_i.

For compensatory model:

$$h(\alpha_j, q_i) = \lambda_1 \alpha_{j1} + \lambda_2 \alpha_{j2} + \cdots + \lambda_K \alpha_{jK}$$

and for the non-compensatory model:

$$h(\alpha_j, q_i) = \lambda_1 \alpha_{j1} \alpha_{j2} \alpha_{jK}$$

General Diagnostic Model (GDM; von Davier & Yamamoto, 2004, 2007; von Davier, 2005, 2008)

$$P(X_{ij} = x | \alpha_j, q_i) = \frac{1}{(1 + \exp(-1(\beta_{ix} + \lambda_{ix}^T h(\alpha_j, q_i))))}$$

π_{ix} is an intercept term, λ_{ix}^T is a vector of weights for item i where λ_{ix} specifies the impact of attribute *k* for category *x* in item *i*. Like LCDM, generality is obtained from the $h(\alpha_j, q_i)$ to be other models. LCDM can be specified using the GDM.

Latent trait models

Linear Logistic Test Model (LLTM; Fischer, 1973)

$$P(X_{ij} = 1 | \theta_j, q_i, \eta) = \frac{\exp(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k + \eta_0)}{1 + \exp(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k + \eta_0)}$$

where q_i is the score of item *i* on attribute *k*, and η_k is the weight of attribute *k* on the item difficulty.

Compensatory multidimensional item response theory (MIRT-C; Bock and Aitkin, 1981)
The logistic form of the MIRT-C model is given by:

$$P(X_{ij} = 1 | \theta_j, \lambda_i, \pi_i) = \frac{1}{(1 + \exp(-1.7(\beta_i + \sum_{m=1}^{M} \lambda_{im}\theta_{jm})))}$$

β_i is the difficulty, θ_{jm} is examinee *j*'s ability level on latent dimension *m*, and λ_{im} is the weight of dimension *m* on item *i*.

Non-compensatory multidimensional item response theory (MIRT-NC; Sympson, 1978)

$$P(X_{ij} = 1 | \theta_j, \beta_i) = \sum_{m=1}^{M} \frac{1}{(1 + \exp(-1.7(\theta_{jm} - \beta_{im})))}$$

β_{im} is the difficulty of item on component *m*, θ_{jm} is examinee j's ability level on dimension *m*.

General component latent trait model (GLTM; Embretson 1985, 1997)

Similar to the MIRT-NC, except β_{im} is replaced by:

$$\beta_{im} = \sum_{k} \eta_{mk} q_{ikm} + \eta_{m0}$$

so that,

$$P(X_{ij} = 1 | \theta_j, \beta_i) = \sum_{m=1}^{M} \frac{1}{(1+\exp(-1.7(\theta_{jm} - \sum_k \eta_{mk} q_{ikm} + \eta_{m0})))}$$

Multicomponent latent trait model for diagnosis (MLTM-D; Embretson & Yang 2012)

The probability that student *j* solves item *i* is given by:

$$P(X_{ij} = 1) = \prod_{m=1}^{M} P_{ijm}^{c_{im}}$$

and

$$P(X_{ij} = 1 | \theta_{jm}, q_{im}, \eta_m) = \frac{\exp(1.7(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}))}{1 + \exp(1.7(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}))},$$

where θ_{jm} is the ability for subject *j* on component m, q_{imk} is the score for attribute *k* in component *m* for item *i*, η_{mk} is the weight for attribute *k* on component *m*, η_{m0} is the intercept for component *m*, and c_{im} is a binary variable indicating involvement of component *m* in item *i*.