SCORE COMPARABILITY AND DIFFERENTIAL ITEM FUNCTIONING

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)

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.