**ABSTRACT**

Using a national physicians’ exam dataset, this study identified the classes of latent DIF with Mixture Rasch Model (MRM) and assessed the important characteristics that caused the DIF classes overall as well as in each latent class with Classification and Regression Tree (CART). It was aimed to provide an explorative data-driven approach in verifying that demographic and environment characteristics interactively act together well predicted latent class membership. It was also anticipated that the finding of this study could help test developers identify the sources that may cause latent DIF in real-world situations.

***Keywords:*** latent DIF; mixture Rasch model; classification and regression tree

**Exploring Demographic and Environment Characteristics in Explaining Latent DIF Using Mixture Rasch and Classification and Regression Tree**

**Introduction**

Differential item functioning (DIF), a phenomenon that individuals with the same level of ability from different groups perform differing probabilities of answering the item correctly (Cho et al., 2016; De Boeck et al., 2011; Hambleton et al, 1991; Scheuneman, 1979; Zieky, 2016), has been considered as an important psychometric feature which is closely related with test fairness (AERA, APA, & NCME, 2014). Multiple statistical methods have been developed for detection of DIF, such as comparison of item parameters (Thissen et al., 1988, 1993), comparison of areas between item response functions (Raju, 1990), different ordinal logistic models (French & Miller, 1996; Miller & Spray, 1993) and logistic discriminant function analysis (LDFA; Spray & Miller, 1994), among others. Typically, DIF is investigated by examining the relationship based on manifest examinee characteristics. Traditional manifest characteristics, such as the demographic variables sex, race, or ethnicity (Cohen & Bolt, 2005; Millsap, 2011), and non-traditional manifest variables such as grade, social economic status have all been used as predictors of latent class membership (Chen & Jiao, 2014). However, the manifest groups are set as known groups and assumed to be homogeneous in most previous research. Researchers have argued its inappropriateness in presetting the manifest variable directly for understanding the root cause of DIF (Chen & Jiao, 2014; Cho & Cohen, 2010; Cohen & Bolt, 2005; De Boeck et al., 2011). Heterogeneity could exist in the pre-determined classification of variables, which means DIF existing in one manifest group does not necessarily indicate all examinees in this group are consistently advantaged or disadvantaged. An interaction of multiple manifest variables rather than one single manifest variable can also be a cause of DIF. Furthermore, from a data-driven approach, research supporting the idea that either the traditional or the non-traditional manifest DIF approach derives from a focus on the research interest, not on the real reason causing DIF (Cohen & Bolt, 2005; Samuelsen, 2005). Additionally, due to the reliability or validity reasons (De Boeck et al., 2011), these pre-determined manifest groups could not accurately or fully predict DIF in the latent class membership (Chen & Jiao, 2014; Cohen & Bolt, 2005; de Ayala et al., 2002; Samuelsen, 2005). For instance, Chen & Jiao (2014) pointed out the limitation in numbers of the manifest variables explored in explaining DIF and used not only background but also cognitive variables.

A latent approach to DIF analysis has therefore been considered and proved useful in uncovering the true group membership. Instead of setting the group membership of interest a priori, latent DIF detects DIF among the unknown hidden examinee groups rather than the manifest groups (Cho & Cohen, 2010; Cohen & Bolt, 2005; De Boeck et al., 2011; McLachlan & Peel, 2000). It classifies the examinees into the distinct latent classes based on the differences in item response patterns (de Ayala et al., 2002; Samuelsen, 2005) and the qualitative heterogeneity among examine groups is easily captured (Cohen & Bolt, 2005). One of the useful tools to investigate the latent DIF is mixture Rasch model (MRM; de Ayala et al., 2002; Kelderman & Macready, 1990; Mislevy & Verhelst, 1990; Rost, 1990). MRM assumes that different parameter values may apply to different latent classes of examinees in that population. Examinees in MRM are characterized by both a latent ability parameter as well as a parameter indicating latent class membership. Conditional probabilities depend on a continuous latent ability, along with a categorical variable that distinguishes examinee latent class based on the higher posterior probability of being a latent class member (Chen & Jiao, 2014). Latent DIF, therefore, provides a mechanism for modeling DIF regarding to the latent populations of examinees as opposed to manifest populations.

A latent DIF analysis with examinees’ characteristics is useful for explaining DIF, in addition to detecting DIF between latent groups (Cho et al., 2016). Categorical and Regression Tree (CART) is a data-driven statistical approach that explores the relationship between the outcome variables and more than two independent variables altogether with no need to prespecify the group interactions (Breiman et al.,1984; Ma, 2018, *p*. 6). It can be divided into two types of trees: Classification Tree (CT) when the dependent variable is categorical and Regression Tree (RT) when it is continuous. It can explore how individual characteristics interactively affect the dependent variables and formulate the relationship based on the pieces collected; it can also lead to variable importance assessment that will be useful to understand the predictive power of explanatory variables. Through recursive partitioning, it progressively divides the individuals into the smaller groups with increasing degree of homogeneity in the dependent variable within each group (Ma, 2018, *p*. 22). The measure that guides the splitting is called impurity. It measures the degree to which cases in a group belong to different categories (values) of the dependent variable (Breiman et al., 1984). With application of CART, variables with levels of importance that impact the outcomes can be obviously reflected in the nodes of the trees. The closer the variables to the root node, the more important level the variables indicate in the tree. The variables in the terminal nodes are the ones with higher homogeneous feature than any other variables in the tree.

An example with a physicians’ exam dataset was demonstrated in this study with the combined statistical techniques latent DIF and CART. With a two-step approach, this study initially employed MRM to determine the number of latent classes for all the examinees based on the higher posterior probability of being a latent class member, and then it explored the interrelationship between the physicians’ demographic and environment characteristics and the latent groups that were identified in the first step with CART. Variables of importance that impacted the latent classes and outcome scores in each latent class could also be displayed in the trees. It was aimed to provide a deeper understanding of the reasons that caused the latent DIF as well as the score differences in each latent class. It was hypothesized that physicians within the same latent class shared the similar characteristics of demographics as well as environment features.

To aid our understanding of the methods, a few research questions were explored with the dataset provided in the study:

RQ1: How many latent classes were there in the physicians’ performances based on the examinees’ response patterns?

RQ2: What were the important physicians’ demographic and environment characteristics that caused the latent DIF and the outcome scores in each latent DIF group?

RQ3: How did these characteristics interact with each other in terms of the outcome variables?

**Methods**

***Data***

Data from a medical certification board’s examination for physicians in 2017 was used for this study. The exam is for renewal of their certificate following at least 10 years in practice and is designed to measure the single construct of medical knowledge and clinical decision-making skills in the medical specialty. There were two forms of the tests: form A and form B, each consisting of 6,755 and 6,641 examinees. Specifically, there were 8001 Whites (54.3%), 1667 Asian (11.3%), 898 Black or African American (6.1%), 109 American Indian or Alaska Native (.7%), 73 Native Hawaiian or Other Pacific Islander (.5%), 660 Others (4.5%) for the race. For ethnicity, there were 6862 non-Hispanic (46.6%), 3757 Not Hispanic or Latino (25.5%) and 789 Hispanic or Latino (5.4%). For sex, it was 7, 348 males (54.9%) and 6, 048 females (45.1%). The mean of the number of years in practice for all the physicians were 18.84 years (*SD* = 9.35). The number of the items in the examination was 260 and all of them were included into the MRM analysis. Identical to the standard Rasch model, in the mixture Rasch model, the exam items were constructed and calibrated under the assumption of unidimensionality.

***Procedures***

Survey from the examinations’ registration questionnaire included information about the physicians’ demographic and environment characteristics, which is a mandatory component of examination registration and completed 3 to 4 months prior to the examination day (Peterson et al., 2019). The demographic characteristics were physicians’ races, ethnicities, and years of practice. The environment characteristics included physicians’ scope of practice, practice organization, site size, location and so on. All participating physicians were extensively informed about the goals of the study and assured that their data would be used for scientific purposes only.

***Measures***

The topics are related to scope of practice, practice organization, location, ownership, size, and provider specialty mix, which help understand practice organization and staffing, how physicians use electronic health records, and scope of practice. The items relevant with the scope of practice are, for example, “Do you currently provide any direct patient care?” and “Are you currently practicing outpatient continuity care”. The items relevant with the practice ownership are, for example, “Which of the following best describes your role in the ownership of your principal practice?” and the responses are categorized as “no official ownership stake (0); sole owner (1); partial owner or shareholder (2); self-employed as a contractor (including locums) and others (3)”. The item relevant with the practice size is “Which of the following describes your principal practice size?” and the four options are “solo practice (0); 2-5 providers (1); 6-20 providers (2); >20 providers (3)”. Also, the item relevant with the physician specialty mix of the principal practice is “which of the following describes the physician specialty mix of your principal practice?” and the responses are “family medicine only (0); primary care specialty mix (Family Medicine, Internal Medicine and/or Pediatrics (1); multiple specialties (not only primary care) (2)”. This study was approved by the Institutional Review Board from the American Association of Family Physicians.

The standardization of training required to practice medicine may impact the ability to identify the sources of DIF through physicians’ demographic but also the environment characteristics they had. Therefore, we sought to explore sources of DIF based on the two directions (Peterson et al., 2019). Following several years of conducting a traditional DIF analysis, few items are found to exhibit construct irrelevant DIF. For example, items regarding the sickle cell anemia are often found to benefit African American physicians more than the reference group of white physicians, but this is because many African-American physicians have patient panels composed of African-American patients, for whom sickle cell anemia is a prevalent disease. Although this item exhibits DIF, it is related to the measurement construct of medical knowledge based on the characteristics of the physician’s patients.

***Analysis***

This study examined the number of latent DIF classes and then explored the relationship between the physicians’ characteristics and their latent groups. The first step involved identifying the number of latent classes through the exploratory MRM analysis (Rost, 1990), which estimates dichotomously scored item responses using the Rasch model but incorporates multiple latent examinee populations. The mixRasch (Willse, 2009) package in software R was used for this purpose. Five latent classification models were constructed, and model fit was compared on a basis of three criteria: Akaike information criterion (AIC; Akiaike, 1974), Baysian information criterion (BIC; Schwarz, 1978) and Consistent Akaike information criterion (CAIC; Bozdogan, 1987). The latent classification model with the lower values indicated the better model fit to the data. To guarantee the consistency of the results, MRM analyses were conducted separately for both form A and form B and the results could be compared. The percentage of examinees for each latent group and whether any mean differences existed across the groups were provided; the mean or percentage differences of each explorative internal or external characteristic were also examined.

*MRM Model*

As an integration of Rasch model (Rasch, 1960, 1980) and latent class analysis (Dayton, 1999), MRM is a representation of the probability of an item score regarding item difficulty and ability parameters, conditional on discrete latent classes. The unconditional probability of response vectors in the MRM is:

and the conditional probability of a correct response in the MRM is:

where a response vector is represented by x = (x1, …, xI). is the difficulty parameter for item *i* conditional on latent class c (c =1, …, C), and is the ability parameter for an examinee *j* in latent class c. represents the probability of a correct response to item *i* by an examinee *j* with ability in latent class c. The proportions of examinees in latent classes are summed as one. More detailed explanation about the MRM equation could be found out in Chen & Jiao (2014).

*CART Method*

We followed up with the CART analysis to link the relationships between the latent class groups and the physicians’ demographics as well as the environment characteristics. The physicians’ demographics include the race and ethnicity; and their environment characteristics are, for instance, the questions related with the principal practice size as well as the physicians’ patient characteristics. The variables are scaled dichotomously (1: Yes; 0: No), ordinally (e.g. 1. very dissatisfied to 5. very satisfied) or continuously (the higher value, the higher measure of the scale). CART analysis screens the important variables that impact the outcome and has them shown in the tree. It performs the binary splitting from the top (called the root node) with each explanatory variable being screened in the model. The principle *impurity* measures the degree to which students in a node vary in an outcome measure. The smaller the impurity, the more homogeneous is for the node (Breiman et al., 1984). Through comparing the impurity of the root node with the sum of impurities of its child nodes, we can get the reduction in impurity. The explanatory variable that yields the largest reduction in impurity is selected for performing the first split and becomes the parental (root) node. The resulting child nodes are markedly different in the outcome measure. Through the repetitive partition, the examinees are classified into smaller and smaller nodes. Nodes that cannot be split further become the terminal nodes. The variables with a higher importance level therefore are on the top and variables with less importance is to the bottom of the tree. If the variable shows up more than once in the tree, it is not only an indication of importance but also an indication of interactions between the independent variables with the outcome in the tree.

CART analyses were also conducted to explore the important characteristics that contribute to the physicians’ scores in each latent class. A node of *N* = 50 was decided as the minimum size of the terminal node. A complete list of the variables that were selected in the CART exploratory analyses were shown in Table 3.

**Results**

To find the measurement model with the best fit, the number of latent groups were detected with an exploratory approach based on model fit indices. In form A, AIC identified a 5-class solution since the number was the smallest among all while BIC and CAIC both identified a 4-class solution. Similar conclusion was found for form B as well, which was a demonstration of the reliability of the result. Given the fact that AIC tended to select a model with more latent classes in the MRM or mixture IRT models (Alexeev et al., 2011), therefore, a 4-class solution was determined as the most appropriate number of the latent classes on the basis of lower values of BIC and CAIC. Table 1 provided the AIC, BIC and CAIC values of the estimated models for both form A and form B.

Table 2 provided the test score distribution with means and standard deviations in each latent class. In general, the range of the ability scores for the four latent groups was from 170 to 183. Additionally, the mean comparisons showed a statistically significant differences across the four latent groups for both forms of dataset (*p* < .01). The proportions for each latent group were in a range from .18 to .30. However, there still existed some slight differences. Specifically, the latent group 4 consisted of examinees with substantially higher score abilities than those in latent group 1 and group 2.

Having identified a potential four latent class membership, in the sequel we selectively chose the physicians’ demographic info as well as environment characteristic variables from the survey questionnaires to examine the relationship between the membership in the latent groups and the examinees’ characteristics. To build the CART model, the examinees who answered the questionnaires were matched with the specific latent groups identified by the MRM. The examinees who did not answer the surveys at all were deleted from the analysis. The final number of the examinees in the CART analysis was 5740 and the dependent variable was four latent classes in this step.

Figure 1 depicted both the left and the right branches of the Classification Tree (CT) since the outcome variable latent classes was categorical. From the root node, it clearly showed that latent class 1 (40.7%) and latent class 4 (38.6%) had a comparatively higher proportions than that for latent class 2 (15.0%) and latent class 3 (5.7%). The first variable that split the four latent groups was *race*. Under the splitting of first child node, in the left branch of the CT tree, they were American Indian or Alaska Native (1) and White (5); and in the right branch, they were Asian (2), Black or African American (3), Native Hawaiian or Other Pacific Islander (4) and Other (6). There was a big difference in the race population between the left branch and the right (41.2% =70.6% - 29.4%). This partly reflected the current distribution of population in the medical certification board’s examination for physicians. Additionally, the total number of latent class 1 and latent class 4 occupied more than 3500 (approximately 90%) of all in the left branch.

Among the American Indian or Alaska Native and White physicians who occupied 70.6% of the population in the left branch, a percentage of45.5% to see the patients for Age 65 + became an important cut-off number in splitting the groups. 59.5% of the physicians indicated that the percentage of patients they see who are older than 65 years old is ≤ 45.5%. Only 11% of the physicians expressed that the percentage of patients they see who are older than 65 years old is over 45.5%. Under the larger percentage of American Indian, Alaska Native and White physicians’ population who see the patients for Age 65+ equal with 45.5% or less, a child node in splitting this group was the variable *site size*. The site that was solo practice was only 19.4% and the site size which was at least 2 providers was about 40.1%.

For the right branch, the division was physicians of the Asian, Black or African American, Native Hawaiian or Other Pacific Islander and Other (29.4% of all). *Site owner* became an important variable in splitting the groups. For this group, the physicians who had no official ownership stake (100% employed), partial owner or shareholder, self-employed as a contractor (including locums) and other ownership types were the majority (sample size *N* = 1117; 19.5%) and the site of physicians who are the sole owners was much smaller (sample size *N* = 571; 9.9%). The local interactions between variables *satisfaction with the site partners* and *satisfaction with the principal practice locations* were also detected in the group dynamics. *Satisfaction with the site partners* was an important local variable, however the percentage difference for these two sub-groups was not obvious (9.3% versus 10.2%). Latent group 1, 2 and 4 all had much bigger percentages for both branches rather than that for latent group 3. Under the sub-group of people who had satisfaction rate 4.5 (between somewhat satisfied and very satisfied) or less (very dissatisfied, somewhat dissatisfied, neither satisfied nor dissatisfied), variable *satisfaction with the location* split the division of the final left branch. The percentage for this group of people who were very satisfied with the location (sample size *N* = 95; 1.7%) were much less than the percentage of the group of people who were less satisfied with the location (sample size *N* = 439; 7.6%).

On a basis of the CT analysis, it was reasonable to infer the important physicians’ characteristics that cause the latent DIF according to our sample. Variable *race* was shown up on the top of the tree, which indicated its role of first important predictor in predicting the outcome. Followed up with the parallel role of second important predictors: *percentage of patients they see across the age who are older than 65 years old* as well as *site owner*;and the third parallel important predictors were: *site size* and *satisfaction with partners*; and the fourth important predictor was *satisfaction with locations*. All of these were indicated as the important exploratory physicians’ characteristics that impacted the four latent groups from the root to the terminal nodes through the CT analysis. Among these, the traditional internal manifest variable was *race*, and the traditional external non-manifest variables were *percentage of patients they see across the age who are older than 65 years old, site owner, site size, satisfaction with partners* as well as *satisfaction with locations.*

In addition, to have a general understanding of what the physicians’ characteristics contributed to their scores in each latent class, specific RT technique was conducted for each latent class since the outcome variables were continuous (Figure 2) and the common and specific characteristics contributing to each latent class were provided (Table 4). Some obvious differences were detected, for instance, *race* was an obvious common variable impacting the splitting of scores for all latent classes. *Site size* impacted latent class 1, 2 and 4 but not latent class 3; *site owner* influenced latent class 1, 2 but not latent 3 and 4. Additionally, *panel* (Approximately, what is the size of your patient panel), *pts\_19 to 64 (*At your principal practice, what is the percentage of patients you see across the following age range?), and *mainsite* (Which of the following describes your principal practice site?) were some new variables that were uncovered in one or more specific trees. Based on the mean comparison of the four latent classes, latent class 4 had a comparatively higher mean than that for latent class 1 and latent class 2. Combined with the result in Table 4, we noticed that latent class 1 and latent class 2 both had several more factors such as *site owner*, *panel,* and *mainsite* than that for latent class 4 which only had two factors impacting the outcome *site size* and *ethnicity*. Through this comparison, we gained insight into a clearer explanation as to the comparison of the similar and different characteristics that cause the four latent classes.

Furthermore, the chi-square significance tests of association were performed to confirm the result of variables that have showed the important roles in the CART analysis that caused the latent DIF. There were statistically significant differences for the explanatory variable of *practice site size, site ownership, satisfaction towards partners* and *satisfaction towards location* across the four latent groups (*p* < .01). Table 5 was a demonstration of the chi-square test on the variable *site size*.

**Discussion**

With a combination of statistical techniques MRM and CART, this study identified the classes of latent DIF and assessed the important variables that caused the DIF classes overall as well as for each latent class. Using the dataset of physicians’ examination from a medical certification board, an MRM was conducted to classify the examinees into the four latent groups for both form A and form B of the exam, which guarantees the consistency and reliability of the result. From a data-driven approach, the hidden four latent classes were uncovered rather than the DIF flagged based on a common single manifest grouping variable, such as gender or social economic status. Compared with traditional manifest variable based DIF methods, the latent DIF strategy provided a method of identification of the advantaged and disadvantaged DIF subgroups in the examinee population from a data-driven perspective with a higher reliability and validity.

The follow-up CART analyses sought to determine whether the latent dimensions were associated in a meaningful way with the examinees’ characteristics. It allowed for the detection of the sources of DIF, but also for a more intuitive and direct interpretation of the interactions in the tree that cause DIF. which was pointed out as a limitation in Chen & Jiao’s (2014) study. In the current analyses, a few demographic and environment variables were identified as important examinee characteristics that caused DIF in the latent classes. The physician’s *race* was a top important characteristic causing DIF in the latent classes. Besides, *site owner*, *percentage of patients who are 65 + at the principal practice, site size, satisfaction with partners* and *satisfaction with the site location* were also assessed as levels of important variables impacting the latent DIF. The CT analyses demonstrated that multiple sources can altogether facilitate the classification of latent DIF. Additionally, four separate RT analyses were conducted in examining the relationship between the independent variables and the physicians’ performance in each latent class. RT helps the researchers with a better understanding of the specific characteristics that contribute to the differences in scores for each latent group. Compared to Chen & Jiao’s (2014) study, though they verified that a single manifest variable was not the proxy of latent class membership, they cannot provide the different interactions between several manifest variables and the outcomes. They tried to use the overlap between the latent classes and manifest groups in a DIF item to determine the consistency, which was not a realistic way to conduct in the complicated DIF analysis. What’s more, latent DIF items were not necessarily matched with the traditional or non-traditional manifest DIF items. Manifest DIF items might not be exhibited in the latent DIF whereas latent DIF items might not be indicated in the manifest DIF. It was hoped this combined statistical technique proposed in this study could made up the limitation in identifying the multiple sources interactively act together in predicting latent class membership and facilitate a more detailed explanation of latent DIF. More advanced technique such as Random Forest can be applied as well to dealing with the similar research questions since CART is a simplified derivative of Random Forest.

It was anticipated that the finding of this study could help test developers identify the sources that may cause latent DIF in real-world situations. After identifying the dimensions and sources of latent DIF, items can be extensively scrutinized by content specialists for the joint impact of multiple traditional and other nontraditional potential sources. During the item writing and test construction phases, these variables that flagged latent DIF should be taken into consideration. Or a bank review could be conducted to identify any items that may be impacted by these new variables. Once the gaps arising from the latent DIF resources are addressed, DIF in testing could decrease with a huge degree. In our study, latent DIF signals the qualitative heterogeneity among the physicians who took the tests in medical field. The exploration can also be extended to other areas of study, such as education and science.

**Limitations**

The latent DIF strategy also potentially suffers from shortcomings. The focus of this study was on the modeling and explanation of DIF, rather than on detection of items that display DIF. For the future studies, researchers can explore further on detecting the specific items that display DIF from the large pool with the identification of latent classes. Additionally, violation of unidimensionality assumption in the applications of the Rasch model could result in an inflated extraction of latent classes in the MRM. In this case, the overextraction of latent classes could result from model misspecification in the MRM but not from the true examinee heterogeneity.

**References**

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Education Research Association.

Alexeev, N., Templin, J., & Cohen, A. S. (2011). Spurious Latent Classes in the Mixture Rasch Model. *Journal of Educational Measurement, 48*(3), 313–332. https://doi.org/10.1111/j.1745-3984.2011.00146.x

Akaike, H. (1974). A new look at the statistical model identification. *Selected papers of Hirotugu Akaike*, 215-222. http://doi:10.1007/978-1-4612-1694-0\_16

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology*. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum.

Bozdogan, H. (1987). Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3). 345-370. http://doi:10.1007/bf02294361

Breiman, L., Friedman, J. H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth.

Chen, Y. F., & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 reading assessment. *Educational Assessment*, *19*(2), 77-96. http://doi:10.1080/10627197.2014.903650

Cho, S. J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35*, 336-370. <http://doi:10.3102/1076998609353111>

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *83*, 278–306.

Cho, S. J., Suh, Y., & Lee, W. Y. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, *35*(1), 48-61.http://doi:10.1111/emip.12093

Cohen, A. S., & Bolt, D. M. (2005). A Mixture Model Analysis of Differential Item Functioning. *Journal of Educational Measurement*, 42(2), 133-148.

http://doi:10.1111/j.1745-3984.2005.00007

de Ayala, R., Kim, S., Stapleton, L., & Dayton, C. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2,* 243-276.

http://10.1207/S15327574IJT023&4\_4

De Boeck, P., Cho, S. J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement, 35*(8), 583-603.

<http://doi:10.1177/0146621611428446>

French, A. W., & Miller, T. R. (1996). Logistic regresion and its use in detecting differntial item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.

http://doi:10.1111/j.1745-3984.1996.tb00495.x

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

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In A. Berk (Ed.), *Handbook of methods for detecting item bias* (pp. 117-155). Baltimore, MD: Johns Hopkins University Press.

Linn, R., & Drasgow, F. (1987). Implications of the golden rule settlement for test construction. *Educational Measurement: Issues & Practice, 6*(2), 13-17.

http://doi:10.1111/j.1745-3992.1987.tb00405.x

Linn, R., Levine, M., Hastings, C., & Wardrop, J. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*(2), 159-173.

<http://doi:10.1177/014662168100500202>

Ma, X. (2018). Using classification and regression trees: A practical primer. IAP.

McLachlan, G., & Peel, D. (2000). Finite mixture models. New York, NY: John Wiley.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*, 107-122. <http://doi:10.1111/j.1745-3984.1993.tb01069.x>

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Routledge.

Peterson L.E., Fang B., Phillips R. L., Avant R., & Puffer, J. C. (2019). The American Board of Family Medicine's Data Collection Method for Tracking Their Specialty. *The Journal of the American Board of Family Medicine, 32*(1), 89-95. <http://doi:10.3122/jabfm.2019.01.180138>

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institute, 1960. (Reprinted by University of Chicago Press, 1980)

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

Roussos, L., & Stout, W. (1996). A multidimensonal-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective* (Unpublished doctoral dissertation). University of Maryland, College Oark, MD.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*, 143-152.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461-464.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*(4), 317-375. <http://doi:10.2307/1164616>

Spray, J. & Miller, T. (1994). Identifying nonuniform DIF in polytomously scored test items (American College Testing Research Report Series 94-1). Iowa City, IA: American College Testing Program.

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.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differ ential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Willse, J. (2009). *mixRasch: Mixture Rasch Models with JMLE.* Retrived from [http://cran.r-project.org/web/packages/mixRasch/mixRasch .pdf](http://cran.r-project.org/web/packages/mixRasch/mixRasch%20.pdf)

Zieky, M. J. (2016). Developing fair tests. In S. Lane, M. R. Raymond & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 81-99). New York, NY: Routledge.