Comparing DIF Detection Accuracy with Observed Score and Latent Score Ordinal Logistic Model: a simulation study

Xiaoyun Wang

Vahab Khademi

Abstract

We studied and compared the performance of two approaches in the detection of differential item functioning: observed score and hybrid (latent-observed) methods. Twelve polytomous items were simulated using sample sizes of 100, 200, 500, 1000. The results show superior performance of the observed score approach at smaller sample sizes, and equal performance at larger samples. Under such parameter and test length conditions, the observed score approach seems a better candidate for different contexts, given large sample sizes may not be always available.

Citation: Khademi, Abdolvahab & Wang, Xiaoyun (2019). Comparing DIF detection accuracy with observed score and latent score ordinal logistic model: a simulation study. Poster presented at Northeastern Educational Research Association, Trumbull, CT, USA (October 21-23, 2019).

Introduction

In educational assessment, it is often important to ensure that the intended construct has the same meaning for individuals from different groups or subpopulations (e.g., defined by gender or testing mode). As a matter of fact, equal functioning of a construct and the related items is one required source of validity in establishing the internal structure of a test. A desirable property of an educational test is that individuals with the same observed score have the same standing on the construct underlying the instrument (Millsap, 2011; Schmitt & Kuljanin, 2008). In other words, if the respondents have the same standing on the construct continuum, the test must show equal or invariant measurement across groups. If we denote the observed score as X and the latent variable as W, then measurement invariance (MI), "expresses the idea that the

measurement properties of X in relation to the target latent trait W are the same across the populations" (Millsap, 2011, p. 46).

One rationale for investigating measurement invariance can be attributed to the increased group membership diversity (Schmitt & Kuljanin, 2008). In fact, comparison of subpopulations on a test may be meaningless unless measurement invariance is reasonably established (Cheung & Rensvold, 2002; Schmitt & Kuljanin, 2008).

As for the importance of meeting measurement invariance in a test, Standard 3.17 of the *Standards for Educational and Psychological Testing* (AERA/NCME/APA, 2014) requires,

> When aggregate scores are publicly reported for relevant subgroups--for example, males and females […]--test users are responsible for providing evidence of comparability and for including cautionary statements whenever credible research or theory indicates that test scores may not have comparable meaning across these subgroups (p. 71).

Measurement noninvariance is a type of systematic error introduced in the relationship between the latent factor and the manifest indicator (Vandenberg & Lance, 2000). Measurement noninvariance can occur both at item/question level and at the test level. Measurement noninvariance at the item level is commonly referred to differential item functioning (DIF) because the item is functioning differently on groups of similar ability (Hambleton, Swaminathan, & Rogers, 1991). Differential Item Functioning (DIF) occurs when individuals

having the same ability, but from different groups, do not have the same probability of getting the item right.

There is a plethora of statistical methods available for assessing measurement invariance, such as logistic regression (Swaminathan & Rogers, 1990), likelihood-ratio test (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993), multi-group confirmatory factor analysis (MG-CFA) (Muthen, 1989), and the item response theory (IRT) framework.

Item Response Theory

Classical measurement theory, which is based on sum observed scores, has some limitations, including the inter-dependence of item characteristics (such as item difficulty and discrimination) and examinees' ability. That is, the measurement of the ability is a function of the easiness or hardness of the test items. If the test is easy, the examinee will have high ability score; if the test is hard, the same examinee can have low ability score. Another limitation of CTT is that item discrimination and test reliability and validity are only specific to the group on which the test was piloted (they are group-dependent). Therefore, comparing examinees taking the same test will be difficult. In addition, comparing items whose characteristics are obtained from different groups is difficult. In practice, it is difficult to field test items for different populations of examinees.

Item response theory (IRT) is a more powerful measurement framework which overcomes the foregoing limitations. IRT is based on two postulates: (1) the performance of an examinee is the manifestation of an underlying latent trait; and (2) a function can be constructed to formalize the performance of the examinee as a manifestation of the underlying trait. The formal function characterizing an item is called the item characteristic function or the item characteristic curve (ICC). The function specifies that as the level of the latent trait increases, the probability of a

desired response to an item increases as well. The function has one set of parameters for the items (1, 2, and 3 parameters) and one set of parameters for the examinee (usually one parameter, the ability of the examinees). The item parameters include:

1. The b_i parameter is the point on the ability scale where the probability of a correct response is 0.5 (in the 1PL and 2PL models; in the 3PL model it is the point at which the probability of providing a correct response is (1+c)/2, where is *c* is the guessing parameter).The greater the *b*, the harder the item and the higher the ability needed to answer the item correctly.
2. The *a* parameter is the discrimination parameter and is manifested as the slope of the ICC: the steeper the ICC at a given *b* point, the more discriminating the item is.
3. The pseudoguessing parameter *c* shows the probability of low ability examinees answering the item correctly

The logistic function for a three-parameter IRT model is:

$$P(Y_{ij} = 1|a_j.\, b_j.\, c_j.\, \theta_i) = c_j + (1 - c_j)\frac{e^{1.7a_j(\theta_i - b_j)}}{1 + e^{1.7a_j(\theta_i - b_j)}} \quad (1)$$

When an IRT model fits the data well, several desirable features are achieved: (1) examinee ability estimates are not test-dependent, and (2) item parameters are not group-dependent. So except for measurement errors, ability estimates on different set of items will be the same and item indices over different groups will be the same (unless DIF is detected). This invariance

property is achieved through exchanging item information with ability estimation and ability information with item parameter estimation (traditionally through joint MLE).

When an IRT model fits the data, the same ICC is obtained for an item regardless of the distribution of the ability in the group of examinees used to estimate the item parameters. The property of invariance comes from regression analysis. When a regression model holds, the line produced has the same slope and intercept in different subpopulations within a restricted range of the X variable. IRT is based on nonlinear regression (logistic regression) and thus has this invariance property. The present study relies on the parameter invariance property of the IRT approach to trait measurement.

STATEMENT OF PROBLEM

One potential problem threatening the validity of test scores in educational and psychological tests is when the probability of answering correctly an item is different among examinees of different group membership but with similar ability or trait level. For example, test takers with similar ability may respond differently to an item if administered in two different ways, such as paper-based or online. For instance, in a recent PARCC test (Partnership for Assessment of Readiness for College and Careers), test takers had lower scores on the online version of the same test compared with the paper-based, as shown in Figure 1 below.

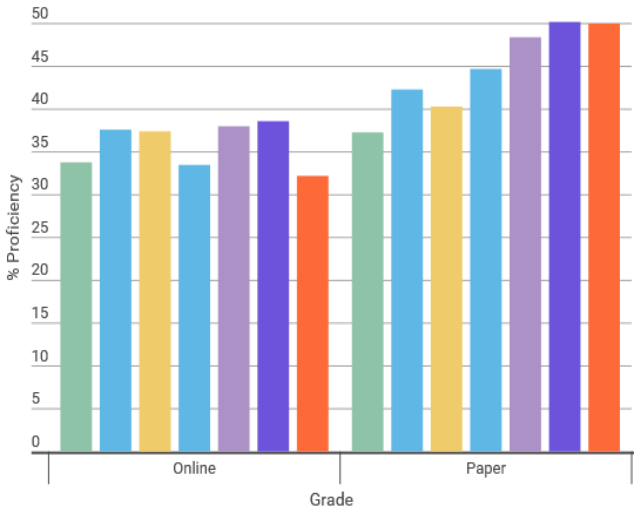


Figure 1: Test takers conditioned on their ability show different score on online and paper-based modes of test administration. Source: http://www.edweek.org/ew/articles/2016/02/03/parcc-scores-lower-on-computer.html

Therefore, for a fair interpretation of the test scores (i.e. the scores reflect the trait or the ability of the test takers, not an extraneous factor), we need to make sure that the influence of factors other than the target trait are minimized or eliminated. This quality is also very important when comparing groups, because we want to make sure that any difference between the groups (e.g. in the mean scores) is purely or mostly due to the trait they are being measured on.

An item displays DIF if people from different groups *with the same ability* have a different probability of giving a certain response. The following figures show two main types of DIF. Uniform DIF (Figure 2) occurs when an item is biased across the entire continuum of the trait or ability. Nonuniform DIF occurs when and item is flagged DIF at some interval of the trait continuum, but not at other intervals (Figure 3).

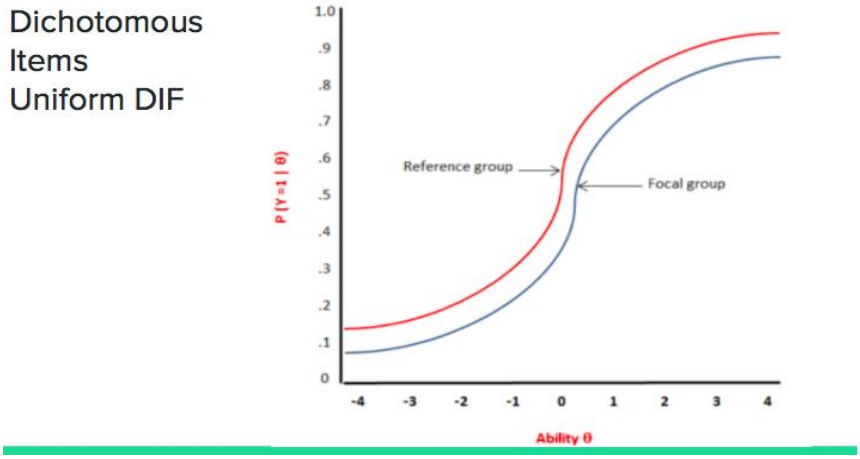


Figure 2: Uniform DIF

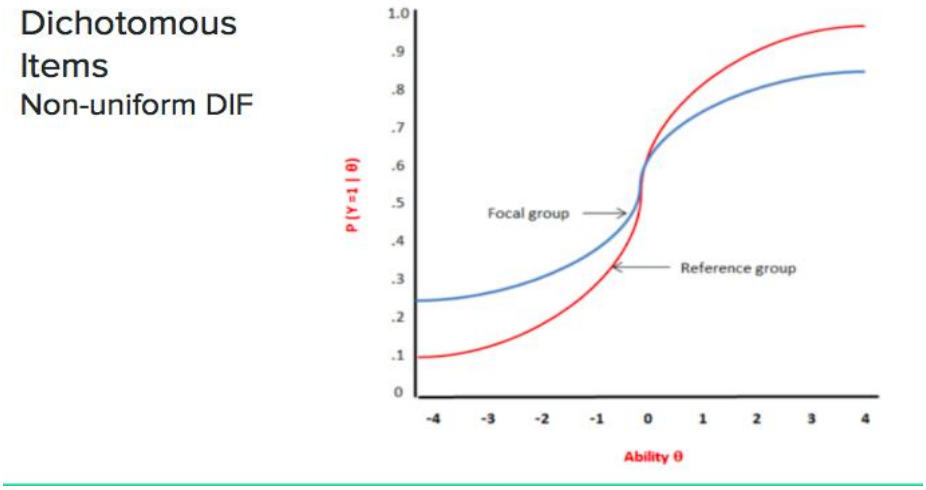


Figure 3: Non-uniform DIF

PROCEDURE

In this study, we simulated 12 ordinal response items, with five possible scores of 1,2,3,4,5. We simulated the data for different sample sizes to see how the two models perform across sample sizes. The polytomous IRT model to score such items (in probability scale) and estimate parameters (slope and category thresholds), is as follows:

$$P_1^*(\theta) \equiv P(u_i \geq 1|\theta) = \{1 + exp[-a_i(\theta - b_{i_1})]\}^{-1}$$
$$P_2^*(\theta) \equiv P(u_i \geq 2|\theta) = \{1 + exp[-a_i(\theta - b_{i_2})]\}^{-1}$$
$$\dots$$
$$P_{(m_i-1)}^*(\theta) \equiv P(u_i \geq m_i - 1|\theta) = \{1 + exp[-a_i(\theta - b_{i_{(m_i-1)}})]\}^{-1}$$

In the model above, θ represents the ability or the trait (e.g. math proficiency, attitude), a_i are the slopes of the curves (called item discrimination in psychometric terminology, which is the power of an item to distinguish poor performing and high performing test takers), and b_i are item difficulty thresholds (the ability it takes to transit from one category to another category, hence difficulty of the category).

To identify DIF items, we use the logistic model below. Ideally, the probability of responding correctly at a given category must be a function of the theta (ability), and characteristics of the items (e.g. difficulty), not any extraneous factor such as group membership of the test takers (e.g. gender ethnicity, language).

$$logitP(\mu \geq k) = a_k + b1 * \theta$$

$$logitP(\mu \geq k) = a_k + b1 * \theta + b2 * group$$

$$logitP(\mu \geq k) = a_k + b1 * \theta + b2 * \theta + b3 * \theta * group$$

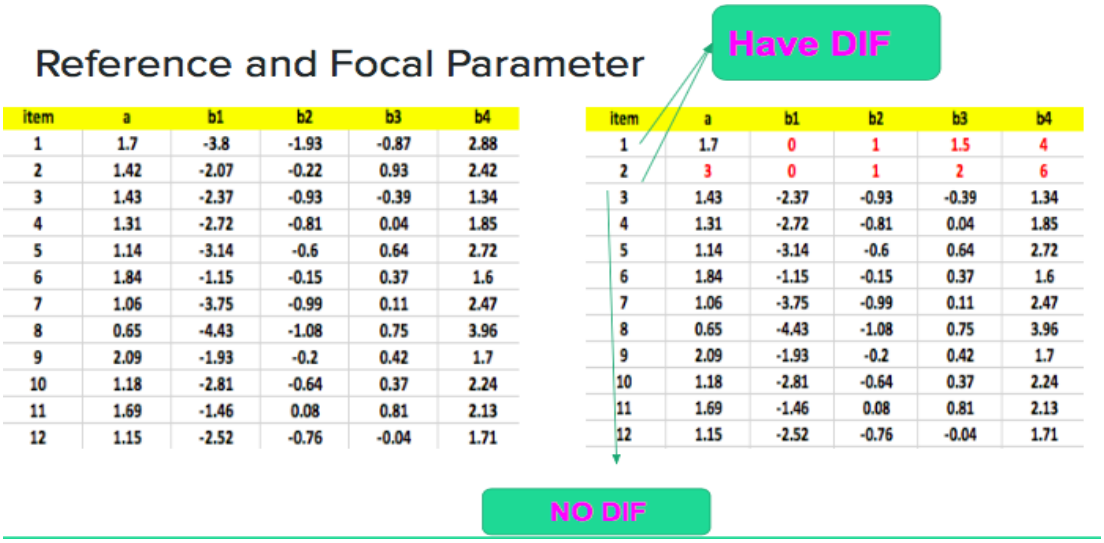
In the set of models above, the first one is the ideal model because the probability of correct response depends only on the ability (theta). In the second and third models, we introduce the dummy variable *group*. Group can be gender, ethnicity, language, etc. Ideally, we do not expect the group variable have a significant contribution to the probability of responding to an item. But if it does, it shows that given equal ability (i.e. conditioning on theta), test takers' group can affect the probability of responding. Hence that item is flagged as biased or showing DIF. In the third model, we interact the ability with group variable. If the interaction is significant, it shows that group interacts ability to affect the probability of responding correctly to an item, hence the item is flagged as DIF. However, this time the DIF is nonuniform because it is flagged for certain intervals of the ability continuum.

For conducting our simulation study, we took the following steps:

Step1

Parameter adjustment: we used actual item parameters to simulate data. We artificially manipulated the *a* and *b* parameters in one set of the item sets parameters to create DIF. We call the set of items with no DIF as the reference group, and the one with DIF items as the focal

group. The following table (Table 1) shows the item parameters in the reference (left) and focal (right) groups.



item	a	b1	b2	b3	b4
1	1.7	-3.8	-1.93	-0.87	2.88
2	1.42	-2.07	-0.22	0.93	2.42
3	1.43	-2.37	-0.93	-0.39	1.34
4	1.31	-2.72	-0.81	0.04	1.85
5	1.14	-3.14	-0.6	0.64	2.72
6	1.84	-1.15	-0.15	0.37	1.6
7	1.06	-3.75	-0.99	0.11	2.47
8	0.65	-4.43	-1.08	0.75	3.96
9	2.09	-1.93	-0.2	0.42	1.7
10	1.18	-2.81	-0.64	0.37	2.24
11	1.69	-1.46	0.08	0.81	2.13
12	1.15	-2.52	-0.76	-0.04	1.71

item	a	b1	b2	b3	b4
1	1.7	0	1	1.5	4
2	3	0	1	2	6
3	1.43	-2.37	-0.93	-0.39	1.34
4	1.31	-2.72	-0.81	0.04	1.85
5	1.14	-3.14	-0.6	0.64	2.72
6	1.84	-1.15	-0.15	0.37	1.6
7	1.06	-3.75	-0.99	0.11	2.47
8	0.65	-4.43	-1.08	0.75	3.96
9	2.09	-1.93	-0.2	0.42	1.7
10	1.18	-2.81	-0.64	0.37	2.24
11	1.69	-1.46	0.08	0.81	2.13
12	1.15	-2.52	-0.76	-0.04	1.71

Table 1: Parameters used to simulate response for 12 polytomous (ordinal multinomial) items.

Step2

Based on the IRT method, we used "simdata" function from the *lordif* package (Choi, S.W. 2016) in R to simulate the item response with different sample size from 100 to 1000. There are 5 categories from 0 to 4. To simulate the response, the item parametes in Step 1 were used. So, two sets of responses were produced: reference, and focal groups. We designated the two groups by the gender variable.

Step 3

For the latent model, we used the "lordif" package (Choi, S.W. 2016) to flag the DIF items. For example, there are total 12 items, with 2 DIF items. We bootstrapped the simulated 100 times

and calculated true positive/negative ratios in a confusion table. The criterion for flagging an item DIF is the change in chi-square from the null model.

Step 4.

For the observed model, we used the "lrm" function in R to test if the group parameter is significant or not the by p-value. If the parameters for group and ability is significant, we can say there are nonuniform DIF. Then we extracted the number of significant DIF items in a bootstrap of 100 times.

RESULT

Table 2 below shows the accuracy of detecting DIF by method and sample size

Monte Carlo simulation	Sample size	IRT model	Observed Score Logistic Model (Non-uniform)
100	100	35%	44%
100	200	79%	81%
100	500	98%	100%
100	1000	100%	100%

Table 2: Accuracy of the two approaches to DIF detection.

Table 3 below shows the Type-I error rate in the two models.

Type I error rate- 500 sample (IRT and LR)

	T	F
T	100% (2/2)	0% (Type 1-error)
F	30%(3/10)	70%

Table 3: Confusion table for the detection of DIF

Figures 4 and 5 below show the accuracy of the two models across different sample sizes in detecting DIF for Item 1 and Item 2. Item 2 has DIF both in threshold and in slope.

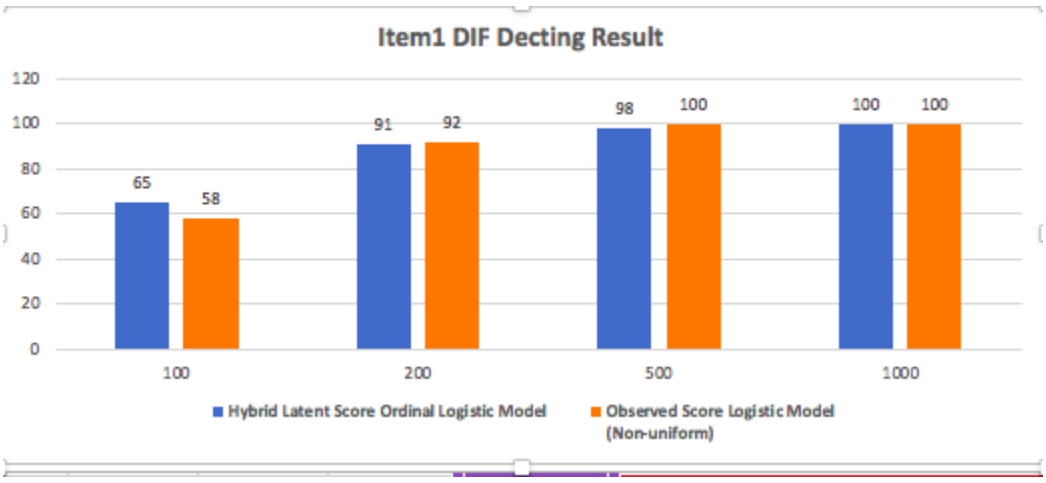


Figure 4: Accuracy of the two approaches in detecting DIF item 1.

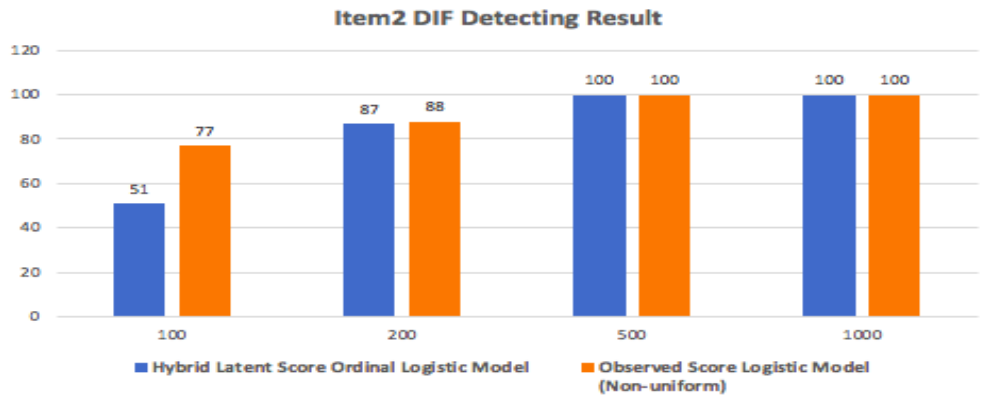


Figure 5: Accuracy of the two approaches in detecting DIF item 2.

Figure 6 below shows the DIF impact on overall ability estimate in the two groups. On the left in Figure 6 we can see that average total ability estimates for the reference and focal groups (F and M) does not differ much if there are two DIF items among the 12 items. So the impact is trivial, and hence the results are valid and comparable. In the right panel, the ability estimates based on only two items is shows (i.e. a test with length two items, both of which show DIF). We can see that those two items by themselves create biased estimates (nonuniform) of the ability of the two groups across the trait continuum.

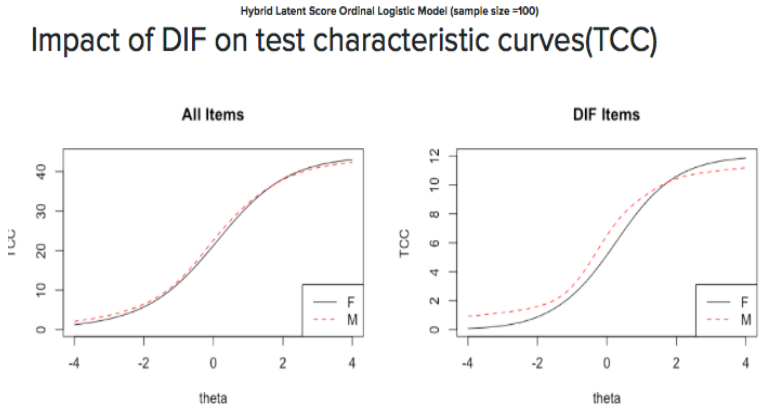


Figure 6: Impact of two DIF Items on Total Ability Estimates (Test Characteristic Curves).

Figure 7 below shows the impact of the DIF items on individual scores of test takes. This can be interpreted as residuals, and we expect scores to have minimum distance to the mean (the solid line). Individual scores are scattered in red and black colors around the mean true estimate of ability. A dotted line shows the average estimated ability (aggregate) of the population. The closer this dotted line is to the solid line, the less impact from the DIF items. In our study, the dotted line almost lies over the solid line, showing minimum impact from the DIF items.

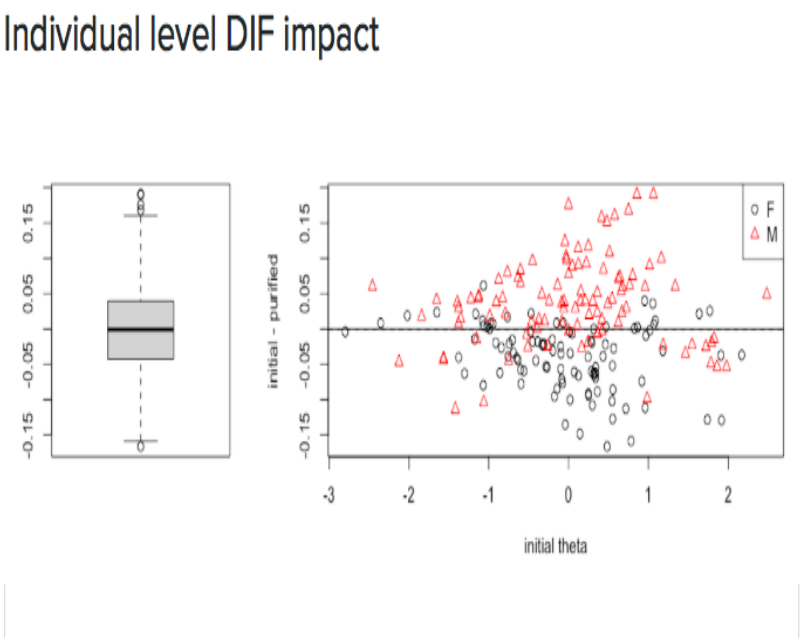


Figure 7: Impact of DIF items on Individual Scores

CONCLUSION

Valid inference must be based on the premise that measurement is purely or predominantly a function of the target trait and minimally affected by extraneous factors, such as ethnicity, language, gender, or age, among others. Several statistical models and approaches exist to help

identify items in instruments (like tests or surveys) that are potentially biased against a certain group. We used an observed score model (ordinal logistic model) and a model that combines latent variable (IRT) modeling with multinomial regression to detect DIF items in a simulated scenario across sample sizes of 100, 200, 500, 1000. The results of our simulation show that observed score ordinal logistic model performs better than the hybrid latent model in smaller sample sizes, and equally well in larger sample sizes. Given that flexibility of a model and its robustness is preferable to its sophistication, we prefer the observed ordinal logistic model to the more sophisticated and computational costly hybrid model. However, if the entire process of instrument construction is performed using latent modeling, the hybrid model should integrate into the framework if the sample size is larger than 1000.

References

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, D.C: AERA.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255.

Choi, S. W (2016). Logistic ordinal regression differential item functioning using IRT. *R Journal*

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

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

Muthen, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–85.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing (dsc 2003), Vienna, Austria*. ISSN 1609-395X.

Plummer, M. (2012). *JAGS version 3.3.0 user manual* [Computer software manual]. (October 1, 2012).

Schmitt, N. & Kuljanin, G. (2008). Measurement invariance: review of practice and implications. *Human Resources Management Review, 18,* 210-222.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression rocedures. *Journal of Educational Measurement, 27,* 361-370.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: the concept of item bias. *Psychological Bulletin, 99,* 118-128.

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. 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. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3(1),* 4-70.

Appendix (R code)

```
library(lordif)
library(mirt)
library(difR)
rm(list=ls())
set.seed(697)
ref.params <- read.csv("/Users/clarawang/Dropbox/697D-IDF-Project/ref_param.csv")
foc.params <- read.csv("/Users/clarawang/Dropbox/697D-IDF-Project/foc_param.csv")

# data for reference group
a.ref <- as.matrix(ref.params$a)
d.ref <- as.matrix(ref.params[,3:6])

#d.ref <-(d.ref+2)

d.ref <- -(d.ref - rowMeans(d.ref))
#d.ref <- d.ref + rnorm(12)


# data for the focal group
a.foc <- as.matrix(foc.params$a)
d.foc <- as.matrix(foc.params[,3:6])

d.foc <- -(d.foc - rowMeans(d.foc))
#d.foc <- d.foc + rnorm(12)


# do DIF analysis
# lordif
# Monte Carlo simulation
B=1
k=0
m=0
c=0
DIF.result <- matrix(data = NA, nrow = B, ncol = 12)
for (i in 1:B){
  ref.data <- simdata(a.ref, d.ref, 500,itemtype = 'graded')
  foc.data <- simdata(a.foc, d.foc, 500,itemtype = 'graded')
  all.data <- rbind(ref.data, foc.data)# rbind the two data drom focal and reference
     groups
  all.data <- as.data.frame(all.data)
  group.var <- rep(c(0,1), each=nrow(all.data)/2)
  sum.score <- rowSums(all.data[,1:12])
  all.data$sum.score <- sum.score
  all.data$mm.group <- group.var

  gender <- all.data$mm.group
  resp <- all.data[,1:12]
  genderDIF <- lordif(resp, gender, criterion="Chisqr", alpha=0.05, minCell = 5)
  flag = genderDIF$"flag"
  for (j in 1:12){
      DIF.result[i,j]=flag[j]
  }


  if (DIF.result[i,1]==TRUE)
      k=k+1

  if (DIF.result[i,2]==TRUE)
```

```
     m=m+1


  if((DIF.result[i,1]==TRUE) & (DIF.result[i,2]==TRUE) )
     c=c+1

  #print(genderDIF)
  #summary(genderDIF)
  #plot(genderDIF)
}
DIF.result
summary(genderDIF)
plot (genderDIF,labels=c("F", "M"))
k
m
c
```