Paper 2900-2015

Multiple Ways to Detect Differential Item Functioning in SAS

Yan Zhang, Educational Testing Service

ABSTRACT

Differential item functioning (DIF), as an assessment tool, has been widely used in quantitative psychology, educational measurement, business management, and insurance and healthcare industries. The purpose of DIF analyses is to detect response differences of items in questionnaires, rating scales, or tests across different subgroups (e.g., gender), while controlling for ability level. There are three general procedures to examine DIF: generalized Mantel-Haenszel (MH), logistic regression, and item response theory (IRT). The SAS system provides flexible and efficient procedures for all these approaches. The goal of this paper is to demonstrate multiple ways to conduct DIF analyses by using different SAS procedures (PROC FREQ, LOGISTIC, GENMOD, GLIMMIX, and NLMIXED).

INTRODUCTION

Differential Item Functioning (DIF) has been widely used in healthcare, business management, and educational measurement. Assessment developers design and construct questionnaires or tests including sets of items that measure, for example, cognition, personality traits, or political views. DIF occurs if responders to a questionnaire or test item from different groups with the same overall scores have different probabilities of giving a correct or positive response to the item. Specifically, DIF has been recognized as a standard tool to measure significant item function differences across groups (e.g., gender or race) while controlling the overall scores on the trait being measured. The usual convention is to designate one group as the reference group (e.g., male, or white) and the others as focal groups (female, or African Americans, Asian Americans, etc.). The SAS system can be used to conduct multiple DIF analyses, conveniently and efficiently, by using current SAS procedures. It should also be pointed out that DIF analyses are just statistical procedures to identify items. A content review by committees of experts carefully examining the contents must be carried out for the identified items. The content experts need to assess whether there is any aspect of the item content that is irrelevant to the trait being measured but may be resulting in differential performance between groups.

In general, there are three main DIF analysis methods: the generalized Mantel-Haenszel (MH) test (Mantel & Haenszel, 1959), logistic regression, and item response theory (IRT). The Mantel-Haenszel DIF statistic was proposed as a method for detecting DIF by Holland & Thayer (1988). It has been widely used in educational measurement due to its easy implementation in testing programs. However, it is usually used to detect uniform DIF for dichotomous items. The logistic regression procedure for DIF was introduced by Swaminathan and Rogers (1990). It can detect both uniform and nonuniform DIFs and can also include exogenous variables (variables besides overall scores that are controlled for in the analyses, such as, age) in the models. Item response theory (IRT) DIF procedures have received increased attention because they can model differences in item difficulty and discrimination parameters. Differences in difficulty of items between groups reflect uniform DIF while differences in item discrimination parameters reflect nonuniform DIF. Logistic regression and IRT procedures are model-based. They both can identify uniform and nonuniform DIF for dichotomous and polytomous items. It is essential to understand the related SAS procedures and apply them appropriately to identify DIF items. First, I showed how to conduct DIF analyses by using the FREQ, LOGISTIC, GENMOD, GLIMMIX, and NLMIXED procedures. Second, I used SAS/IML to call an R package - difR to conduct DIF analyses and then compared the results between the SAS procedures and difR. All the analyses were conducted using SAS 9.3 and R 2.15.3.

DATA SOURCES

In this paper I used two datasets, collected from a questionnaire, called Verbal Aggression Assessment, for illustrative purposes. The datasets can be downloaded from http://bearcenter.berkeley.edu/page/materials-explanatory-item-response-models. The two datasets have the same variables and data structure but different coding categories. One dataset, referred to *verb_poly*, has polytomous items, coded as 0, 1, or 2, whereas the other, as *verb_dich*, has dichotomous items, coded as 0 or 1. The item response values 1 and 2 in *verb_poly* were recoded as 1 in *verb_dich*. The datasets include 26 variables (24 items in the questionnaire, Gender, and Anger) and 316 responders (243 females and 73 males). The responders filled out a behavioral questionnaire about how to describe their frustrating situation in verbal aggression responses (i.e. S1DoScold, S1WantScold). The male group was designated as the reference group (gender = 1) and the female group (gender = 0) as the focal group. In this paper, the 0.05 significance level was used as the significance threshold.

GENERALIZED MANTEL-HAENSZEL PROCEDURE

The MH DIF procedure compared dichotomous item performance between two groups after matching responders on overall scores. Responders in the focal and reference groups were matched on total test or questionnaire scores by dividing responders in both groups into defined strata on those scores. The total scores were generated by summing item scores across all items. Estimates of the odds ratio for a given item, across the strata of the matching variable, can be computed from a 2 x 2 x *K* contingency table with *k* denoting the *k*-th stratum, (*k* = 1,2,…*K*). Table 1 shows the 2 x 2 contingency table for the *k*-th stratum of an item. The a_k, b_k, c_k, and d_k represent the numbers of responders in the cells. N_k denotes the number of responders in the *k*-the stratum.

	Response (1)	Response (0)	Total
Reference group	a_k	b_k	
Focal group	c_k	d_k	
Total			N_k

Table 1. *k*-th stratum of 2 x 2 table

Some DIF occurs if the odd ratio for an item is greater than 1 or less than 1. The common odds-ratios formula is:

$$OR_{MH} = \frac{\sum a_k d_k / N_k}{\sum b_k c_k / N_k}$$

In the SAS system, the Cochran-Mantel-Haenszel statistic (Landis, Heyman, & Koch, 1978) can be generated using the FREQ procedure. The CMH option in the TABLE statement requests this statistic. Total scores were summed across all 24 items as the matching variable. Responders were stratified on total scores using PROC RANK. The CMH statistics were separately obtained for each item. For any given item, the null hypothesis is that there is no association between the defined groups (gender in the examples below) and the item responses across strata. The SAS statements for the example dataset are as follows.

```
DATA verbal_dich;
  SET verbal_dich;
  score = sum(of S1WantCurse -- S4DoShout);
RUN;
PROC RANK DATA=verbal_dich OUT=Verbal group=10;
  VAR score;
  RANKS stratum;
RUN;
PROC FREQ DATA=Verbal;
  TABLES stratum * Gender * S3DoCurse /CMH noprint;
RUN;
```

Table 2 shows the SAS output. The CMH χ^2 statistic is 8.002 with 1 degree of freedom with a probability level of 0.005. This means that the males performance on this item (S3Docurse) is significantly different from that of the females after controlling for total scores. The odds ratio is 2.125, which suggests that the odds of males responding positively to this item are 2.125 times higher than that of the females. The results indicate that uniform DIF occurs in this item. The Breslow-Day Test results imply that the odds ratios between the two groups vary significantly across strata, indicating nonuniform DIF.

	DF	Value	Prob	95% CI
Cochran-Mantel-Haenszel Statistics	1	8.002	0.005	
(Odds Ratio)		2.125		(1.112, 4.053)
Breslow-Day Test for Homogeneity of the Odds Ratios	9	23.614	0.005	

Table 2. Summarized Results of MH methods

LOGISTIC REGRESSION

The logistic regression DIF procedure can identify uniform and nonuniform DIF for both dichotomous and polytomous items. For each item, three models with increasing numbers of predictors are used. The probability of a positive response to an item is modeled as a function of total scores, group, and the interaction between total score and group. The item shows DIF if the model fit statistic (-2LogL) is improved when group and interaction are added to the model, in order. The group estimate coefficients (β_2) are associated with uniform DIF while the interaction estimates (β_3) are related to nonuniform DIF. Like the MH procedure, responders' values on the trait being measured are represented by their total scores. The models are:

$$Model\ 1:\ logit(P) = \beta_0 + \beta_1\theta$$
$$Model\ 2:\ logit(P) = \beta_0 + \beta_1\theta + \beta_2 Z$$
$$Model\ 3:\ logit(P) = \beta_0 + \beta_1\theta + \beta_2 Z + \beta_3\theta Z$$

where θ denotes the value of the responder on the trait, and Z denotes group membership (e.g. gender or race), and $logit(P)$ denotes the logit of the probability of responders answering positively or correctly. The likelihood ratio test (LRT) is used to compare the likelihood of two models. The model with the smaller -2logL has better fit to the data. The LRT statistic is calculated by:

$$G^2 = [-2lnL(model\_r)] - [-2lnL(model\_f)] \sim \chi^2_{(d)}$$

where *model_r* denotes reduced models and *model_f* denote full models. G^2 follows the chi square distribution and *d* is the difference in numbers of parameter between the reduced and full models. The null hypothesis is that item parameters between reference and focal group do not differ. Uniform DIF can be identified by comparing the LRT statistic between Models 1 and 2, with degree of freedom (*df*) = 1. Nonuniform DIF is tested by comparing Models 2 and 3, with *df* = 1. An overall test of DIF can be conducted by comparing Models 1 and 3, with *df* = 2.

Several SAS procedures can be used to carry out logistic regression analysis. For illustrative purposes, this paper only shows how to use the LOGISTIC and GENMOD procedures to detect DIF for dichotomous and polytomous items, respectively. Since the items in the polytomous data used in the example are ordinal, the LINK=CLOGIT option in PROC GENMOD was chosen because it fits the cumulative logit model when items have more than two response categories. The LOGISTIC procedure fits linear logistic regression models for dichotomous or polytomous response categories using Fisher's method to maximize the likelihood (ML) function. PROC GENMOD fits generalized linear logistic regression models using the Newton-Raphson method to maximize the likelihood.

The LOGISTIC and GENMOD procedures, by default, set the reference level to 0, which can be changed in the CLASS statement. The three logistic regression models are fit in order and the item called "S3DoCurse" was chosen as a response variable. The variable *score* is the total score. The SAS codes are as follows:

```
PROC LOGISTIC DATA=verbal_dich;  *Model 1;
 MODEL S3DoCurse=score /LINK=logit;
RUN;
PROC LOGISTIC DATA=verbal_dich; *Model 2;
  MODEL S3DoCurse=score gender / LINK=logit;
RUN;
PROC LOGISTIC DATA=verbal_dich;  *Model 3;
  MODEL S3DoCurse=score gender score*gender / LINK=logit;
RUN;
PROC GENMOD DATA=verbal_poly;  *Model 1;
  MODEL S3DoCurse=score /LINK=clogit DIST=mutl;
RUN;
PROC GENMOD DATA=verbal_poly;  *Model 2;
  MODEL S3DoCurse=score gender/LINK=clogit DIST=mutl;
RUN;
PROC GENMOD DATA=verbal_poly;  *Model 3;
  MODEL S3DoCurse=score gender score*gender/LINK=clogit DIST=mutl;
RUN;
```

A summary of results is shown in Outputs 1 and 2.

	-2logL	β_1		β_2		β_3	
		Estimate	*p*	Estimate	*P*	Estimate	*P*
Model 1	351.511	-0.217	<.0001				
Model 2	344.368			-0.822	0.008		
Model 3	344.273					0.021	0.757

Output 1. Summarized output from PROC LOGISTIC

	Full Log Likelihood	β_1		β_2		β_3	
		Estimate	*P*	Estimate	*P*	Estimate	*P*
Model 1	-175.7556	-0.125	<.0001				
Model 2	-172.1839			-0.745	0.006		
Model 3	-251.5534					0.032	0.305

Output 2. Summarized output from PROC GENMOD

In Output 1, the chi square statistic $\chi^2_{(1,0.05)}$ = 3.841 is smaller than the G^2(Model1 – Model3) = 7.143. Thus, the item may have uniform DIF. However, G^2(Model2 – Model3) = 0.095 is smaller than the chi square statistic. The item does not show nonuniform DIF. In Model 1, β_1 is significantly different from zero, which suggests that the log odds of answering 1 on this item decreases as total scores increase. In Model 2, β_2 indicates that the odds of performance on this item are significant different between the reference and focal groups. In Model 3, β_3 is not significantly different from zero, indicating that there is no nonuniform DIF between the groups. Each of the 24 items was separately analyzed with logistic regression models. PROC GENMOD provides the log likelihood statistic denoted "Full Log Likelihood", which is multiplied by -2 to get -2LogL. The results generated from PROC GENMOD are consistent results with those from PROC LOGISTIC. Six items have shown uniform DIF: 6, 14, 16, 17, 19, and 20. No items show nonuniform DIF.

ITEM RESPONSE THEORY

In this section, I illustrated using IRT mixed effects models to detect DIF by using two SAS procedures - PROC NLMIXED and PROC GLIMMIX. The general idea of mixed effect models is to treat item parameters as fixed effects and responders' parameters as random effects. PROC NLMIXED is modeling nonlinear mixed models and PROC GLIMMIX is modeling linear mixed models with various fitting methods. One advantage of the NLMIXED procedure is that mathematical formulations can be stated explicitly in this procedure, which is not available for the GLIMMIX procedure.

IRT DIF procedures have three model types: 1-parameter (1PL), 2-parameter (2PL), and 3-parameter (3PL) logistic regression. 1PL IRT can only detect uniform DIF while 2PL- and 3PL- IRT can detect both uniform and nonuniform DIF. The models for 1PL-, 2PL- and 3PL-IRT are:

$$1PL: Pr(x_{ij}=1) = \frac{1}{1+\exp\left[-(\theta_i - b_j)\right]}$$

$$2PL: Pr(x_{ij}=1) = \frac{1}{1+\exp\left[-\alpha_j(\theta_i - b_j)\right]}$$

$$3PL: Pr(x_{ij}=1) = c_j + (1-c_j)\frac{1}{1+\exp\left[-\alpha_j(\theta_i - b_j)\right]}$$

where b_j is the difficulty for item *j*, θ_i is the *i-th* responder's trait parameter, α_j is the discrimination parameter for item *j* that allows different items to differentially discriminate among responders, and c_j is the pseudoguessing parameter for item *j*. $Pr(x_{ij}=1)$ is the probability of *i-th* responder responding positively (with the response coded as 1 rather than that coded as 0) to item *j*.

The discrimination parameter (α) can't be specified in PROC GLIMMIX so that this procedure can only model the 1PL IRT and detect uniform DIF. Thus, this paper uses PROC GLIMMIX to fit the 1PL IRT model (also called the Rasch

model) and PROC NLMIXED is used to fit 1PL- and 2PL- IRT models for illustrative purposes. Since there are only 316 responders in the verbal aggression dataset and it may not work well for IRT models with polytomous items (Bond & Fox, 2007). Therefore, this paper only illustrates IRT models with dichotomous items. For polytomous items with sufficient numbers of responders, researchers can modify PROC NLMIXED statements to fit 1-PL or 2PL- partial credit and graded response models.

Both the GLIMMIX and NLMIXED procedures require the data with "long format". This means that a single response variable contains all of the item responses. Dummy variables are created for all items so that they can index and differentiate items. In the verbal aggression data, a response variable called *y* is created to include all the responses for 24 items, one variable called *Person* for each responder, one variable *Item* for each item, and the 24 dummy variables (called *item1, item2, …, item24*) for item indexing. Table 3 illustrates this data structure.

Person	Item	Gender	Item1	Item2	Item3	Item4	…
1	Item1	1	1	0	0	0	
1	Item2	1	0	1	0	0	
1	Item3	1	0	0	1	0	
1	Item4	1	0	0	0	1	
						…	
2	Item1	2	1	0	0	0	

Table 3. Data Structure for the GLIMMIX and NLMIXED procedures

IRT - PROC GLIMMX

In this paper I used an exploratory approach to detect DIF. First, a null model (intercept-only) was applied to the data, which specified that all item difficulty parameters (*b*) are the same. Second, all 24 items were added to the model and thus the difficulty parameters for items were allowed to vary. Third, the interaction effects between gender and items were added to the model to detect DIF. Since the models were nested, the likelihood ratio test was used to compare the likelihood of pairs of models. Items show DIF in overall tests if the model fit statistics are improved.

The following statements are used to build the three models in order: the null model with random responders, the item model with random responders, and item and the interaction effect model. In the MODEL statement, LINK and DIST options specify using logit link functions and the response variable *y* is binary. The RANDOM statement specifies random intercept effects for each responder and their G-side covariance structures. Item 24 was set as the reference item by default. The estimated item difficulty for each item needs to be adjusted accordingly.

```
/*Model 1: all items have the same difficulty b*/
PROC GLIMMIX DATA=verbal_dich METHOD=quadrature GRADIENT NOCLPRINT ORDER=data;
      CLASS person;
      MODEL y = /DIST=binary LINK=logit SOLUTION;
      RANDOM intercept/SUBJECT =person TYPE=UN;
RUN;
/*Model 2: random person and fixed item*/
PROC GLIMMIX DATA=verbal_dich METHOD =quadrature GRADIENT NOCLPRINT ORDER=data;
      CLASS person item;
      MODEL y (event='1')= item/SOLUTION LINK=LOGIT DIST=BINARY;
      RANDOM INTERCEPT / SUBJECT=person TYPE=UN G;
run;
/*Model 3: detect DIF for gender difference*/
PROC GLIMMIX DATA=verbal_dich METHOD =laplace GRADIENT NOCLPRINT NOITPRINT ORDER=data;
      CLASS gender person item;
      MODEL y (event='1')= item gender item*gender/SOLUTION LINK=CLOGIT DIST=BINARY;
      RANDOM INTERCEPT / SUBJECT=person TYPE=UN G;
RUN;
```

In Table 4, G^2 = 9506.36 – 8073.88 = 1432.48 between Model 1 and 2 is larger than χ^2_{23} = 35.172 at the .05 significance level. Thus, Model 2 is a better fit than Model 1. Similarly, Model 3 with the lowest -2LogL has the best fit to the data. The estimated coefficients are shown in SAS output "Solutions for Fixed Effects", see Output 3. Based on

estimated coefficients and significance levels, items 14, 16, 17, 19, and 20 show significant uniform DIF between the reference and focal groups.

	GLIMMIX (-2LogL)	NLMIXED (-2LogL)
Model 1: Intercept model with random responder	9506.36	8129.6
Model 2: Item model with random responder	8073.88	8101.2
Model 3: Item and interaction model	8005.06	9223.7

Table 4. Model fit statistics

Solutions for Fixed Effects

Effect	item	gender	Estimate	Standard Error	DF	t Value	Pr > |t|	Gradient
Intercept			-1.9783	0.2098	314	-9.43	<.0001	0.001866
⋮								
gender*item	item13	1	0.6923	0.5424	7222	1.28	0.2019	-0.00132
gender*item	item14	1	1.0729	0.5227	7222	2.05	0.0401	0.003251
gender*item	item15	1	0.1687	0.5180	7222	0.33	0.7446	-0.00118
gender*item	item16	1	1.3882	0.5486	7222	2.53	0.0114	-0.00356
gender*item	item17	1	1.3487	0.5179	7222	2.60	0.0092	0.006446
gender*item	item18	1	0.5374	0.5296	7222	1.01	0.3103	-0.00157
gender*item	item19	1	1.2547	0.5147	7222	2.44	0.0148	0.002797
gender*item	item20	1	1.1399	0.5241	7222	2.17	0.0297	-0.00433
gender*item	item21	1	0.7463	0.6216	7222	1.20	0.2299	-0.00072
gender*item	item22	1	0.8786	0.5277	7222	1.67	0.0959	0.002774
gender*item	item23	1	0.9036	0.5112	7222	1.77	0.0772	-0.00263
gender*item	item24	1	0	.	.	.	.	.

Output 3. Solution for Fixed effects (Model 3) from PROC GLIMMIX

IRT- PROC NLMIXED

Next I used PROC NLMIXED to build logistic mixed-effect models to detect DIF. This method differs from the GLIMMIX procedure by introducing DIF parameters into the model. The mathematical formulation and procedures were introduced in detail by Meulders and Xie (2004) and van den Noortgate and De Boeck (2005). In general, questionnaire items are regarded as fixed effects and responders' parameters as random effects. The main idea is that DIF occurs if item parameters, discrimination α and difficulty *b*, differ between groups. Therefore, two DIF parameters δ and ξ are added to the model as a linear function of predictors. The δ reflects uniform DIFs and ξ nonuniform DIF. The two parameters interact with group variable *Z*. The item shows DIF if the DIF parameters are significantly different from 0. The probability of a success or positive response is modeled by:

$$Pr(x_{ij}=1)=\frac{1}{1+\exp\left[-(\alpha_j+Z\zeta_j)(\theta_i-(b_j+Z\delta_j))\right]}$$

where *j* denotes item *j* and *i* denotes the *i-th* responder. For the 1PL model, two types of parameters, *b* and *δ*, need to be defined in the statements of PROC NLMIXED. For 2PL models, the four types of parameters, α, *b*, *δ*, and *ζ* are free to be estimated. The 1PL model statements are as follows:

```
   PROC NLMIXED DATA=verbal_dich NOAD METHOD=gauss QPOINT=22;
     PARMS b0-b24=0 d1-d23=0 sd1=1 sd0=0.8 mu1=0;
     b=b0+b1*item1+b2*item2+b3*item3+b4*item4+b5*item5+b6*item6+b7*item7+b8*item8+b9*it
em9+b10*item10+b11*item11+b12*item12+b13*item13+b12*item12+b15*item15+b16*item16+b17*i
tem17+b18*item18+b19*item19+b20*item20+b21*item21+b22*item22+b23*item23+b24*item24;
```

```
    delta=d1*item1+d2*item2+d3*item3+d4*item4+d5*item5+d6*item6+d7*item7+d8*item8+d9*i
tem9+d10*item10+d11*item11+d12*item12+d13*item13+d12*item12+d15*item15+d16*item16+d17*
item17+d18*item18+d19*item19+d20*item20+d21*item21+d22*item22+d23*item23;
    ex= exp(theta-b-delta*gender);
    prob = ex/(1+ex);
    MODEL y ~ BINARY(prob);
    RANDOM theta ~ NORMAL(gender*mu1,(1-gender)*sd0*sd0+gender*sd1*sd1)SUBJECT=person;
  RUN; QUIT;
```

The coefficients *b*1 – *b*24 are difficulty parameters of items, *d*1 – *d*23 are DIF parameters, and *d*24 is the reference parameter. The theta (θ) reflecting the responders' parameters are assumed to follows group-specific normal distributions. Thus, theta (θ) follows a normal distribution with mean = 0 and *SD* = 0.8 for females and a *N*(0, 1) distribution for males. The estimates of several DIF parameters are shown in Output 4.

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
⋮									
d16	-1.3302	0.5492	315	-2.42	0.0160	0.05	-2.4107	-0.2496	-0.01212
d17	-1.3003	0.5171	315	-2.51	0.0124	0.05	-2.3178	-0.2829	0.007943
d18	-0.5217	0.5272	315	-0.99	0.3232	0.05	-1.5590	0.5156	0.000569
d19	-1.2103	0.5137	315	-2.36	0.0191	0.05	-2.2211	-0.1996	0.000604
d20	-1.1203	0.5220	315	-2.15	0.0326	0.05	-2.1472	-0.09330	-0.00322
d21	-0.7633	0.6188	315	-1.23	0.2183	0.05	-1.9809	0.4543	-0.00165
d22	-0.8244	0.5277	315	-1.56	0.1192	0.05	-1.8626	0.2139	0.002227
d23	-0.8657	0.5098	315	-1.70	0.0905	0.05	-1.8688	0.1374	-0.00583

Output 4. DIF parameters for 1PL IRT from PROC NLMIXED

For the 2PL IRT model, three more statements are added to the NLMIXED procedure for detecting nonuniform DIF.

```
*discrimination parameters;
alpha=a1*item1+a2*item2+a3*item3+a4*item4+a5*item5+a6*item6+a7*item7+a8*item8+a9*it
   em9+a10*item10+a11*item11+a12*item12+a13*item13+a12*item12+a15*item15+a16*item1
   6+a17*item17+a18*item18+a19*item19+a20*item20+a21*item21+a22*item22+a23*item23+
   a24*item24;
*nonuniform DIF parameter;
zeta=e1*item1+e2*item2+e3*item3+e4*item4+e5*item5+e6*item6+e7*item7+e8*item8+e9*ite
   m9+e10*item10+e11*item11+e12*item12+e13*item13+e12*item12+e15*item15+e16*item16
   +e17*item17+e18*item18+e19*item19+e20*item20+e21*item21+e22*item22+e23*item23;
ex= exp((alpha+zeta*gender)*(theta-b-delta*gender));
```

Table 4 shows that Model 2 has the best model fit. Based on estimated coefficients and their *p*-values, items 16, 17, 19, and 20 have been identified as having uniform DIF. None of the 24 items show any nonuniform DIF. In Table 5, we can see that DIF in item 6 can be detected by both the MH and logistic regression procedures, but not the IRT procedures. It is possible that there are not sufficient data for IRT to detect DIF.

R PACKAGE – difR

In addition to using the SAS procedures to measure DIF, researchers can also link R packages to the SAS system by SAS/IML and then call R functions to conduct DIF analyses. There are several R packages available for different DIF analyses. One R package – difR (Magis et al., 2010) is used here for illustrative purposes. difR includes many DIF procedures, for example, MH, logistic regression, and IRT, but they requires dichotomous items. It is worth noting that, unlike the SAS system, R is a case-sensitive software. This paper uses R 2.15.3 for demonstration because the higher R versions don't work well with SAS 9.3. The package difR is also needed under the R 2.15.3 version. The –

RLANG option must be specified in the SAS system to support access to R language interfaces. The codes are as follows:

```
PROC OPTIONS OPTION=rlang; RUN;
PROC IML;
RUN ExportDataSetToR("Work.Verbal", "verbal" );
  submit / R;
    library(difR);
    verbal <- verbal[colnames(verbal)!="Anger"]
    difMH(verbal, group="Gender", focal.name=1);
    difLogistic(verbal[,1:24], group=verbal[,25], focal.name=1);
    difLord(verbal, group="Gender", focal.name=1, model="1PL");
  endsubmit;
QUIT;
```

The ExportDataSetToR() function is used here to transform data from SAS to R. In difR, the reference level = 1 is set by default. Thus, the coefficient estimates from R are opposite in sign to those from SAS outputs. The variable *Anger* needs to be removed from the data before DIF analyses otherwise it will be treated as an item. As shown in Table 5, item 16, 17, and 19 have significant DIF for all the DIF procedures while the other items need more careful investigation due to the possibility of other effects, for example, test or questionnaire length and effect size.

Item	MH		Logistic Regression			IRT		
	FREQ	difMH	LOGISTIC	difLogistic	GENMOD	GLIMMIX	NLMIXED	difLord
6	Y	Y	Y	Y	Y			Y
14	Y	Y	Y		Y	Y		
16	Y	Y	Y	Y	Y	Y	Y	Y
17	Y	Y	Y	Y	Y	Y	Y	Y
19	Y	Y	Y	Y	Y	Y	Y	Y
20	Y	Y	Y		Y	Y	Y	

Table 5. Summarized results from SAS and R

CONCLUSION

This paper demonstrates how to flexibly conduct multiple DIF analyses by using different SAS procedures. The Mantel-Haenszel procedure can be easily implemented in the FREQ procedure. It is generally used to detecting uniform DIFs for dichotomous items. For each item, DIF occurs if the odd ratios is either less than or greater than 1. The LOGISTIC and GENMOD procedures can be used to conduct logistic regression DIF for both dichotomous and polytomous items. For each item, three models are built in order, likelihood functions are compared between models, and estimates of coefficients are tested for significance. In order to identify DIF for all items conveniently, SAS macros need to be coded by researchers.

The GLIMMIX and NLMIXED procedures are used to build mixed effect models for IRT DIF procedures. They give researchers a comprehensive and parsimonious way to conduct DIF analyses since all items, group variables, and interaction variables are included into one model. However, it may take a long time to run PROC NLIMIXED when there are a large number of responders and items in a dataset. The GLIMMIX procedure can only implement the 1PL IRT model while the NLMIXED procedure can apply the 1PL-, 2PL-, and 3PL- IRT models, by adding corresponding mathematical formulas in its statements. Researchers may need to spend more time on an exploratory approach on data. All the procedures don't require programming expertise but they require basic statistical knowledge of DIF and these SAS procedures and cautious interpretation of SAS outputs.

REFERENCES

Bond, T. G., and Fox, C. M. 2007. *Applying the Rasch model: Fundamental measurement in the human sciences.* 2nd ed. Mahwah, NJ: Erlbaum.

Holland, P. W., and Thayer, D. T. 1988. *Differential item performance and the Mantel-Haenszel procedure.* In H. Wainer & H. Braun (Eds), *Test validity* (pp.129-125).Hillsdale, NJ: Erlbaum.

Landis, J. R., Heyman, E. R., and Koch, G. G. 1978. “Average partial association in three-way contingency tables: A review and discussion of alternative tests.” *International Statistical Review,* 46:237-254.

Magis, D., Beland, S., Tuerlinckx, F., and De Boeck, P. 2010. “A general framework and an R package for the detection of dichotomous differential item functioning.” *Behavior Research Methods,* 42: 847-862.

Mantel, N, and Haenszel, M. W. 1959. “Statistical aspects of thee analysis of data from retrospective studies of disease.” *Journal of the National Cancer Institute*, 22:719-48.

Meulders, M., and Xie, Y. 2004. *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 213–240). In P. De Boeck & M. Wilson (Eds.), *Person-by-item predictors* (pp. 213–240). New York, NY: Springer-Verlag.

Swaminathan, H., and Rogers, H. J. 1990. “Detecting differential item functioning using logistic regression procedures.” *Journal of Educational Measurement,* 27: 361-370.

van den Noortgate, W., and De Boeck, P. 2005. “Assessing and explaining differential item functioning using logistic mixed models.” *Journal of Educational and Behavioral Statistics,* 30: 443–464.

ACKNOWLEDGEMENTS

I would like to thank James Carlson, Jianbin Fu and Matthew Duchnowski for their valuable editing advice.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Yan Zhang, Ph.D.
Educational Testing Service
660 Rosedale Rd, MS T20
Princeton, NJ 08540
(609) 734-5889
yzhang002@ets.org