DOCUMENT RESUME

ED 418 141 TM 028 236

AUTHOR	Wang, Wen-Chung
TITLE	An ANOVA-Like Rasch Analysis of Differential Item Functioning.
PUB DATE	1998-00-00
NOTE	22p.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Analysis of Variance; Foreign Countries; *Item Bias; Item Response Theory; Simulation; *Test Items
IDENTIFIERS	*Rasch Model

ABSTRACT

The conventional two-group differential item functioning (DIF) analysis is extended to an analysis of variance-like (ANOVA-like) DIF analysis where multiple factors with multiple groups are compared simultaneously. Moreover, DIF is treated as a parameter to be estimated rather than simply a sign to be detected. This proposed approach allows the investigation of the effects of DIF on items more thoroughly. Results of simulation studies show that the parameters of the proposed models were recovered very well. A real data set with 10 dichotomous items was analyzed. Implications and applications are discussed. (Contains 4 tables, 2 figures, and 10 references.) (Author/SLD)

An ANOVA-Like Rasch Analysis of Differential Item Functioning

WEN-CHUNG WANG

National Chung Cheng University

Correspondence:

Department of Psychology

National Chung Cheng University

Chia-Yi, Taiwan

Telephone: 886-5-2720411 ext. 6430

Fax: 886-5-2720857

E-mail: psywcw@ccunix.ccu.edu.tw

An ANOVA-Like Rasch Analysis of Differential Item Functioning

Abstract

The conventional two-group DIF analysis is extended to an ANOVA-like DIF analysis where multiple factors with multiple groups are compared simultaneously. Moreover, DIF is treated as a parameter to be estimated rather than simply a sign to be detected. This proposed approach allows us to investigate the effects of DIF on items more thoroughly. Results of simulation studies show that the parameters of the proposed models were recovered very well. A real data set with ten dichotomous items was analyzed. Implications and applications are addressed.

Keywords: differential item functioning, Rasch model, ANOVA, factorial design, marginal maximum likelihood estimation.

Item response theory (IRT) has been widely used to detect item differential functioning (DIF). Lord (1980) has pointed out that item characteristic curves are ideally suited to defining DIF. Since item parameters as well as person parameters determine the curves, the detection of DIF could be made by comparing item parameters between some focal group and some reference group. More specifically, within the framework of the Rasch model (Rasch, 1960), we can estimate the item difficulties separately for each group and then test their differences as follows:

$$Z_i = \frac{\hat{b}_{iF} - \hat{b}_{iR}}{\sqrt{Var(\hat{b}_{iF}) + Var(\hat{b}_{iR})}},$$

where $\hat{b}_{iF}$ and $\hat{b}_{iR}$ are maximum likelihood estimates of item i's difficulty for the focal group and the reference group, respectively; $Var(\hat{b}_{iF})$ and $Var(\hat{b}_{iR})$ are their estimated error variances, respectively. Z_i follows approximately the standard normal distribution.

Thissen, Steinberg and Gerrard (1986) and Thissen, Steinberg and Wainer (1988) have adopted a marginal maximum likelihood estimation (MML) to investigate DIF. They used the usual likelihood ratio test to compare a full model, where different item difficulty parameters are used for different groups, with a reduced model, where different groups yield the same item difficulty parameters.

The above conventional approaches have two shortcomings. First, the differences of the item difficulty parameters between focal and reference groups are to be tested rather than parameterized within the models. Therefore, the influences of DIF are not well investigated. Second, these item difficulty parameters are group-dependent. None of them can be treated as "item difficulty".

These shortcomings can be overcome by reparameterization of these item difficulty parameters. For example, the two item difficulty parameters (one for the focal group and the other for the reference group) can be reparameterized as one "grand item difficulty" and one DIF parameter. The grand item difficulty is in fact a weighted average of the two item difficulty parameters. The DIF parameter is the deviance of the item difficulty of the reference group to the grand item difficulty. It depicts how DIF influences the item characteristic curves. In this paper, the reparameterization is addressed.

Conventionally, DIF analysis focuses on two groups: one focal group and one reference group. This is analogous to the t-test of two means. As the t-test is extended to simple or factorial analysis of variance (ANOVA), DIF analysis can be extended to multiple factors with multiple groups. In this paper, an ANOVA-like Rasch DIF analysis is proposed. The proposed factorial DIF analysis has two major advantages. First, as ANOVA is statistically more powerful than the t-test, the ANOVA-like DIF analysis is more powerful than the conventional DIF analysis. Second, as main effects and interaction effects can be partitioned and investigated in ANOVA, they can also be done in the ANOVA-like DIF analysis, which in turn makes DIF analysis more thorough.

In the following, I give detailed description of the proposed modeling. Results of simulation studies for parameter recovery are shown. Finally, a real data set was analyzed to illustrate implications and applications of the proposed modeling.

Reparameterization of Item Parameters

Let there be one focal group and one reference group. In the terminology of ANOVA, this is a one-factor design. We can estimate the item difficulty parameters for each group. Within the Rasch model, it follows:

$$log\,(p\,/\,q)_1 = \theta_n - \delta_{i1}, \qquad (1a)$$

$$log\ (p\ /\ q)_2 = \theta_n - \delta_{i2}, \tag{1b}$$

where p is the probability of a correct answer of person n to item i; q is that of an incorrect answer of person n to item i; θ_n is person n's ability; δ_{i1} is item i's difficulty for the reference group (subscript 1); δ_{i2} is item i's difficulty for the focal group (subscript 2).

These item parameters can be reparameterized as

$$\delta_{ij} = \delta_i + \alpha_{ij}, \tag{2}$$

subject to

$$\textstyle\sum_j \alpha_{ij} = 0.$$

Equations (1a) and (1b) become

$$log\ (p\ /\ q)_1 = \theta_n - (\delta_i + \alpha_{i1}), \tag{3a}$$

$$log\ (p\ /\ q)_2 = \theta_n - (\delta_i + \alpha_{i2}), \tag{3b}$$

respectively. In the case of two groups, $\alpha_{i1} = -\alpha_{i2}$. δ_i can be viewed as the grand item difficulty of item i. α_{ij} represents the effect of group j on item i's difficulty and is referred to as a DIF parameter. If α_{ij} is significantly different from zero, the item expresses DIF. To test this hypothesis, on one hand, we can compare the ratio of α_{ij} over its estimated standard error to the standard normal distribution. On the other hand, we can adopt the likelihood ratio test to compare two nested models: a full model with DIF parameters and a reduced model without DIF parameters.

Equation (2) is analogous to one-way ANOVA. It can be extended to factorial ANOVA. For example, let there be two factors: Factor A, indexed $j = 1, \ldots, J$ (e.g., race) and Factor B, indexed $k = 1, \ldots, K$ (e.g., gender). More specifically, let there be four groups: White Male ($j = 1, k = 1$), Color Male ($j = 2, k = 1$), White Female, ($j = 1, k = 2$)

and Color Female ($j = 2$, $k = 2$). We could estimate the item difficulty parameters for each group as follows:

$$log\,(p / q)_{11} = \theta_n - \delta_{i11}, \qquad (4a)$$

$$log\,(p / q)_{21} = \theta_n - \delta_{i21}, \qquad (4b)$$

$$log\,(p / q)_{12} = \theta_n - \delta_{i12}, \qquad (4c)$$

$$log\,(p / q)_{22} = \theta_n - \delta_{i22}, \qquad (4d)$$

where δ_{i11} is item i's difficulty parameter for White Male; δ_{i21} is that for Color Male; δ_{i12} is that for White Female; δ_{i22} is that for Color Female; the others are defined as above.

Like the reparameterization of Equation (2), these item difficulty parameters can be reparameterized as

$$\delta_{ijk} = \delta_i + \alpha_{ij} + \beta_{ik} + (\alpha\beta)_{ijk}, \qquad (5)$$

subject to

$$\sum_j \alpha_{ij} = 0,$$

$$\sum_k \beta_{ik} = 0,$$

and

$$\sum_j (\alpha\beta)_{ijk} = \sum_k (\alpha\beta)_{ijk} = 0.$$

Equations (4a) to (4d) become

$$log\,(p / q)_{11} = \theta_n - (\delta_i + \alpha_{i1} + \beta_{i1} + (\alpha\beta)_{i11}), \qquad (6a)$$

$$log\,(p / q)_{21} = \theta_n - (\delta_i + \alpha_{i2} + \beta_{i1} + (\alpha\beta)_{i21}), \qquad (6b)$$

$$log\,(p / q)_{12} = \theta_n - (\delta_i + \alpha_{i1} + \beta_{i2} + (\alpha\beta)_{i12}), \qquad (6c)$$

$$log\,(p / q)_{22} = \theta_n - (\delta_i + \alpha_{i2} + \beta_{i2} + (\alpha\beta)_{i22}), \qquad (6d)$$

:tively. Consequently, δ_i can be viewed as the grand difficulty of item i, α_{ij} as the of Factor A_j, β_{ik} as the effect of Factor B_k, and $(\alpha\beta)_{ijk}$ as the interaction effect of A_j and Factor B_k, on item i. In the case of two levels in each factor,

$\alpha_{i1} = -\alpha_{i2}$,

$\beta_{i1} = -\beta_{i2}$,

$(\alpha\beta)_{i11} = -(\alpha\beta)_{i12} = -(\alpha\beta)_{i21} = (\alpha\beta)_{i22}.$

We can test if these DIF parameters are significantly different from zero by using the standard normal distribution or the likelihood ratio test. We may find some yield all kinds of DIF effects, some items yield only the interaction effect, some yield Factor A's main effect, and some items yielding Factor B's main effect, ; others.

Consider the influences of DIF parameters. At the Educational Testing Service, are classified into thee categories on the basis of Mantel Haenszel delta difference)-DIF). Category A contains the items with negligible or nonsignificant MH D- Category B contains the items with slight to moderate values of MH D-DIF. ›ry C contains the items with moderate to large values of MH D-DIF. Basically, ıbsolute value of MH D-DIF is less than 1.0, the item goes to Category A. If it is t 1.0 but less than 1.5, the item belongs to Category B. Finally, if it is 1.5 or more, m is classified as Category C.

Except for hard or easy items, a difference of 1.0 MH D-DIF is very roughly equal lifference of .1 in probability of correct answer between groups. Likewise, a nce of 1.5 MH D-DIF is very roughly equal to a difference of .15 in probability of correct answer between groups. In terms of the logit scale, 1.0 and 1.5 deltas ·rrespond to roughly .43 and .64 logits, respectively. Applying these criteria, three :egories could also be formed. If the difference of item difficulties of two groups (see uation (5)) is not significantly different from zero or it is less than .43, the item ongs to Category A. If it is between .43 and .64, the item belongs to Category B. ıally, if it is .64 or above, the item belongs to Category C.

To make the proposed models possible, a multidimensional random coefficients ıltinomial logit model is used and addressed in the following.

The Multidimensional Random Coefficients Multinomial Logit Model

The multidimensional random coefficients multinomial logit model (MRCML, ams, Wilson, & Wang, 1997) is a multidimensional extension of the random fficients multinomial logit model (Adams and Wilson, 1996). The MRCML model two levels. At the second level, a population model $f_\theta(\theta; \alpha)$ is formed, where θ is ector of latent variables and α is a set of parameters that characterize the distribution θ. In the case of multivariate normal distribution, α becomes a mean vector and a iance-covariance matrix. At the first level, a conditional item response model $(\mathbf{x}; \xi | \theta)$ is formed, where $\mathbf{x}$ is a vector of observation on items, ξ is a vector of rameters that describe those items, and θ is a vector of latent variables. The ıditional item response model describes the probability of observing a set of item sponses conditioned on the level of an individual on the set of latent variables.

BEST COPY AVAILABLE

The Conditional Item Response Model

Suppose a set of D latent traits underlies the examinees' test performances and examinees' positions are denoted $\theta = (\theta_1, \ldots, \theta_D)'$. Let there be I items indexed $i = 1, \ldots, I$, and K_i response categories in item i indexed $k = 1, \ldots, K_i$. A response in category k of item i is scored b_{ikd} on dimension d (the scoring schema is know a priori). The scores across D dimensions can be collected into a column vector $\mathbf{b}_{ik} = (b_{ik1}, \ldots, b_{ikD})'$, then into a scoring sub-matrix for item i, $\mathbf{B}_i = (\mathbf{b}_{i1}, \ldots, \mathbf{b}_{iK_i})'$, and then into a scoring matrix $\mathbf{B} = (\mathbf{B}_1', \ldots \mathbf{B}_I')'$ for the whole test.

Let $\xi = (\xi_1, \ldots, \xi_p)'$ denote a vector of p free item parameters. Let a design vector $\mathbf{a}_{ik}'$ denote a linear combinations of ξ corresponding to response category k of item i. They are denoted by a design matrix $\mathbf{A} = (\mathbf{a}_{11}', \mathbf{a}_{12}', \ldots, \mathbf{a}_{1k_1}', \mathbf{a}_{21}', \ldots, \mathbf{a}_{2k_2}', \ldots, \mathbf{a}_{Ik_I}')'$ for the whole test. Let an indicator variable X_{ink} denote as

$$X_{ink} = \begin{cases} 1 & \text{if response of person } n \text{ to item } i \text{ is in category } k, \\ 0 & \text{otherwise.} \end{cases}$$

Under the MRCML model, the probability of a response in category k of item i for person n is expressed as

$$f(X_{ink} = 1, \mathbf{A}, \mathbf{B}, \xi \mid \theta_n) = \frac{\exp(\mathbf{b}_{ik}' \theta_n + \mathbf{a}_{ik}' \xi)}{\sum_{u=1}^{K_i} \exp(\mathbf{b}_{iu}' \theta_n + \mathbf{a}_{iu}' \xi)}.$$

A marginal maximum likelihood estimation with EM algorithm (Bock & Aitkin, 1981) is developed. The proposed models in this study are all derived from the

MRCML by manipulating the design matrices **A** and **B**, although they are actually unidimensional. For example, suppose two items are administrated to four groups: White Male, Color Male, White Female, and Color Female. These two items are rearranged to eight virtual group-items. The left panel of Table 1 shows memberships and item responses of four persons. The right panel of Table 1 shows the rearranged responses of the eight virtual items, where V.1 indicates virtual item 1, V.2 indicates virtual item 2, and so on. V.1 to V. 4 belong to the first original item, and V.5 to V.8 to the second. In addition, V.1 and V. 5 go with White Male; V.2 and V.6 with Color Male; V.3 and V.7 with White Female; V.4 and V.8 with Color Female. The other cells are blank and treated as missing.

Table 1
Original item responses and rearranged item responses

	Item 1	Item 2	Eight Virtual Items							
			V.1	V.2	V.3	V.4	V.5	V.6	V.7	V.8
White Male	1	0	1				0			
Color Male	0	1		0				1		
White Female	0	1			0				0	
Color Female	1	1				1				1

Note: values are hypothetical scores.

In DIF analysis, the means of the distributions for various groups are usually quite different. Since in MML estimation with the normal case only a grand mean is assumed, we have to parameterize the differences among means for the groups. For four groups, there could be three parameters: one for the main effect of Factor A (i.e., difference between the race groups), one for the main effect of Factor B (i.e., difference between the gender groups), and the other for the interaction effect of Factors A and B (i.e., difference between the race by gender groups). In addition to these three "mean-difference" parameters, there is one grand item difficulty for Item 1 (the other grand item difficulty for Item 2 is constrained for model identification). Moreover, there are one Race DIF parameter (the other Race DIF parameter for Item 2 is constrained for model identification), one Gender DIF parameter (the other Gender DIF parameter for Item 2 is constrained for model identification), one Race by Gender DIF parameter (the other Race by Gender DIF parameter for Item 2 is constrained for model identification). Altogether seven item parameters are formed.

Figure 1 shows the corresponding scoring matrix **B** and the design matrix **A**. In Figure 1, ξ_1 indicates the mean difference between the race groups (main effect of Factor A); ξ_2 indicates the mean difference between the gender groups (main effect of Factor B); ξ_3 indicates the mean difference among the race by gender groups (interaction effect of Factors A and B). The mean of each group can be found from

$$\mu_{jk} = \mu + \xi_j + \xi_k + \xi_{jk},$$

where μ_{jk} stands for the mean for group jk; μ stands for the grand mean: Moreover, given two factors with two levels on each,

$\xi_j = \xi_1$, if $j = 1$; $\xi_j = -\xi_1$, if $j = 2$.

$\xi_k = \xi_2$, if $k = 1$; $\xi_k = -\xi_2$, if $k = 2$.

$\xi_{jk} = \xi_3$, if ($j = 1$, $k = 1$) or ($j = 2$, $k = 2$); $\xi_{jk} = -\xi_3$, otherwise.

For example, White Male ($j = 1$ and $k = 1$) has a mean of $\mu_{11} = \mu + \xi_1 + \xi_2 + \xi_3$; Color Male ($j = 2$ and $k = 1$) has a mean of $\mu_{21} = \mu - \xi_1 + \xi_2 - \xi_3$; White Female ($j = 1$ and $k = 2$) has a mean of $\mu_{12} = \mu + \xi_1 - \xi_2 - \xi_3$; Color Female ($j = 2$ and $k = 2$) has a mean of $\mu_{22} = \mu - \xi_1 - \xi_2 +$

BEST COPY AVAILABLE

ξ_3. In practice, we are not interested in testing the differences of the means among groups because they are usually expected to be different.

Regarding other item parameters in Figure 1, ξ_4 is the grand difficulty of the first item and $-\xi_4$ is the grand difficulty of the second item, because the mean of all the grand item difficulties is constrained to zero for model identification. ξ_5 is a Factor A DIF parameter, the main effect of Factor A on item difficulties. ξ_6 is a Factor B DIF parameter, the main effect of Factor B on item difficulties. ξ_7 is a Factors A by B DIF parameter, the interaction effect of Factors A and B on item difficulties. Item i's difficulty for group jk can be found from

$$\delta_{ijk} = \delta_i + \alpha_{ij} + \beta_{ik} + (\alpha\beta)_{ijk},$$

where

$\delta_i = \xi_4$, if $i = 1$; $\delta_i = -\xi_4$, if $i = 2$;

$\alpha_{ij} = \xi_5$, if ($i = 1, j = 1$) or ($i = 2, j = 2$); $\alpha_{ij} = -\xi_5$, otherwise;

$\beta_{ik} = \xi_6$, if ($i = 1, k = 1$) or ($i = 2, k = 2$); $\beta_{ik} = -\xi_6$, otherwise;

$(\alpha\beta)_{ijk} = \xi_7$, if ($j = 1, k = 1$) or ($j = 2, k = 2$), when $i = 1$; if ($j = 2, k = 1$) or ($j = 1, k = 2$), when $i = 2$; $(\alpha\beta)_{ijk} = -\xi_7$, otherwise

In fact, the computation of both the mean and the item difficulty of each group is similar to that of sample means in a factorial ANOVA design.

$$
\mathbf{B} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}_{\theta}
\qquad
\mathbf{A} = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & -1 & -1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 1 & -1 & -1 & 1 & -1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & -1 & -1 & -1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & -1 & 1 & -1 & 1 & 1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 1 & -1 & 1 & -1 & 1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & -1 & 1 & 1 & -1 & -1 & 1
\end{bmatrix}
\quad (\xi: 1\ 2\ 3\ 4\ 5\ 6\ 7)
$$

Figure 1. Scoring matrix **B** and design matrix **A** for two items administrated to four groups

The model in Figure 1 is a full model because all possible item parameters are estimated. We can form some reduced models by discarding some of the DIF parameters, for example, a model without interaction DIF. In addition, we can form models where all items, a subset of items, or no items express some kinds of DIF. Through model comparison, we can test if these DIF parameters are statistically significant. Although Figure 1 is an example of dichotomous items, the MRCML can be directly applied to polytomous items. Interested readers are referred to Adams, Wilson, and Wang (1997), Wang, Wilson, and Adams (1997) for details of how the two matrices were manipulated to form various models. The computer software ConQuest (Wu, Adams, & Wilson, 1997) could be used to estimate the parameters.

Simulation Studies

The design of the simulation studies is based on the real data analyses in the following section. Two-way factorial design was adopted with two levels on each, which leads to four groups. The sample sizes of these four groups are 214, 294, 83, and 182. There are ten dichotomous items in the test. Two conditions were conducted: one is a full model with all possible DIF parameters (see Figure 1) and the other is a reduced model with a few DIF parameters. Fifty replications were made under each condition.

In the full model, altogether 41 parameters were estimated, including two person distribution parameters (one grand mean parameter and one variance parameter), three mean-difference parameters, nine grand item difficulty parameters, nine Factor A DIF parameters, nine Factor B DIF parameters, and nine Factors A by B DIF parameters. Table 2 shows the generating values, the bias values (mean of fifty replications minus generating value), the standard errors, and the Z statistics (bias values divided with standard errors). According to the Z statistics, no parameters are statistically biased at the .05 level. In addition, all the parameters are recovered very well with the bias values between -.064 and .056.

Under the reduced model, altogether 17 parameters were estimated, including two person distribution parameters, and three mean-difference parameters, nine grand item difficulty parameters, and three Factor A DIF parameters (only three items show Factor A DIF). The results are summarized in Table 3. No parameters are statistically biased. All the parameters are recovered very well with the bias values between -.021 and .020.

Table 2
Generating values, bias values, standard errors, and *Z* statistics of various parameters in the full model (Model 1)

Parameter	Generating	Bias	SE	*Z*
Mean-difference				
1	.46	.030	.074	.41
2	-.06	-.007	.079	-.09
3	-.16	-.017	.093	-.19
Grand Item Difficulty				
1	-3.84	-.064	.196	-.33
2	-3.42	-.014	.139	-.10
3	-.70	.002	.117	.02
4	-.42	-.015	.107	-.14
5	-1.03	-.035	.125	-.28
6	1.06	-.018	.119	-.15
7	.34	.019	.126	.15
8	1.83	.021	.132	.16
9	2.65	.056	.181	.31
Factor A DIF				
1	-.16	-.004	.184	-.02
2	.32	-.019	.129	-.15
3	-.08	.044	.113	.39
4	.18	.024	.106	.23
5	-.28	.022	.113	.20
6	.02	.016	.128	.13
7	-.19	.023	.114	.20
8	-.04	-.009	.107	-.08
9	.04	-.013	.203	-.06
Factor B DIF				
1	.03	-.002	.183	-.01
2	-.21	.014	.144	.10
3	.01	-.010	.106	-.10
4	-.05	.006	.102	.05
5	.15	.004	.106	.03
6	-.01	.004	.123	.03
7	.23	-.018	.113	-.16
8	.09	-.042	.173	-.24
9	.05	.008	.159	.05
Factors A by B DIF				
1	.36	.014	.193	.07
2	.10	-.017	.141	-.12
3	.02	-.023	.095	-.24
4	.08	-.015	.092	-.16
5	.08	.007	.102	.07
6	.15	.005	.102	.05
7	.10	-.024	.132	-.18
8	-.14	.009	.111	.08
9	-.14	-.011	.217	-.05
Grand Mean	-1.11	-.016	.078	-.21
Variance	2.70	-.028	.244	-.12

Table 3
Generating values, bias values, standard errors, and *Z* statistics of various parameters in the reduced model (Model 4)

Parameter	Generating	Bias	SE	*Z*
Mean-difference				
1	.49	-.001	.067	-.02
2	-.06	-.015	.077	-.20
3	-.23	.000	.081	.00
Grand Item Difficulty				
1	-3.74	.009	.141	.06
2	-3.48	.008	.119	.07
3	-.69	-.021	.093	-.22
4	-.44	.018	.081	.22
5	-.98	.002	.093	.02
6	1.05	-.008	.112	-.07
7	.43	-.006	.102	-.06
8	1.84	.003	.129	.02
9	2.65	.001	.191	.01
Factor A DIF				
1	.28	-.010	.084	-.11
2	.19	-.007	.083	-.08
3	-.24	.020	.090	.22
Grand Mean	2.71	-.017	.195	-.08
Variance	-1.11	.005	.082	.07

Real Data Analyses

A personality test with ten dichotomous items from Wang (1997) was analyzed. Subjects are 773 secondary school teachers and college students, including 214 female teachers, 294 female students, 83 male teachers, 182 male students. There are two factors: status (teacher and student) and gender (male and female). We are interested in if the items show Status DIF, Gender DIF, or Status by Gender DIF. To investigate this, several models were formed. Model 1 is a full model with 41 parameters. It has a deviance G^2 of 6122.02. The estimated parameters are shown in Table 2 as the generating values. To test if the items show Status by Gender DIF, all the nine corresponding parameters were constrained to zero. The resulting model, Model 2, has a deviance of 6135.88. These two models are not statistically significant based on the likelihood ratio test. Therefore, no items show Status by Gender DIF DIF.

Further, to investigate Gender DIF, all the nine corresponding DIF parameters were constrained to zero. The resulting model, Model 3, is a nested model of Model 2 and has a deviance of 6147.96. Again, the likelihood ratio test is adopted to compare Models 2 and 3. It is found that Model 3 is preferred. Thus, no items show the Gender DIF. I further constrained all the nine Status DIF parameters to zero to test if the items show Status DIF. The model, Model 0, is a model without any DIF and has a deviance of 6168.88. Comparing Models 0 and 3, we find they are statistically significant. That is, at least one item shows Status DIF.

According to the estimated standard errors of the parameters in Model 3, as shown in Table 4, Items 2, 4, and 5 might have significant Status DIF effects. To investigate this, only the three DIF parameters are estimated and the other six DIF parameters are constrained to zero. The resulting model, Model 4, has a deviance of 6151.43. This model is not statistically different from Model 3. Therefore, only the three items express the status DIF. The estimated parameters in Model 4 are listed as the generating values in Table 3. Figure 2 shows the likelihood ratio tests for these five models.

The Status DIF parameters of items 2, 4, and 5 are .28, .19, and -.24, respectively. Since the teachers are indexed in front of the students, item 2 is .56 (= $2 \times .28$) more difficult for the teachers than for the students. Likewise, item 4 is .38 (= $2 \times .19$) more difficult for the teachers than for the students. Item 5 is .48 (= 2 × .24) easier for the teachers than for the students. Based on the classification stated above, Items 2 and 5 belong to Category B (slight to moderate effect); Item 4 and the other items belong to Category A (negligible effect). With this information, test developers and test users can gain deeper understanding about items on various groups.

Table 4.
Parameter estimates and their standard errors in Model 3

Parameter	Estimate	SE
Mean-difference		
1	.48	.06
2	-.06	.07
3	-.23	.09
Grand Item Difficulty		
1	-3.75	.19
2	-3.47	.12
3	-.70	.10
4	-.43	.09
5	-.97	.11
6	1.06	.11
7	.42	.12
8	1.84	.13
9	2.65	.15
Factor A DIF		
1	-.02	.19
2	.32	.11
3	-.07	.11
4	.20	.09
5	-.23	.09
6	.07	.11
7	-.14	.10
8	-.07	.13
9	.00	.16
10	-.07	.23

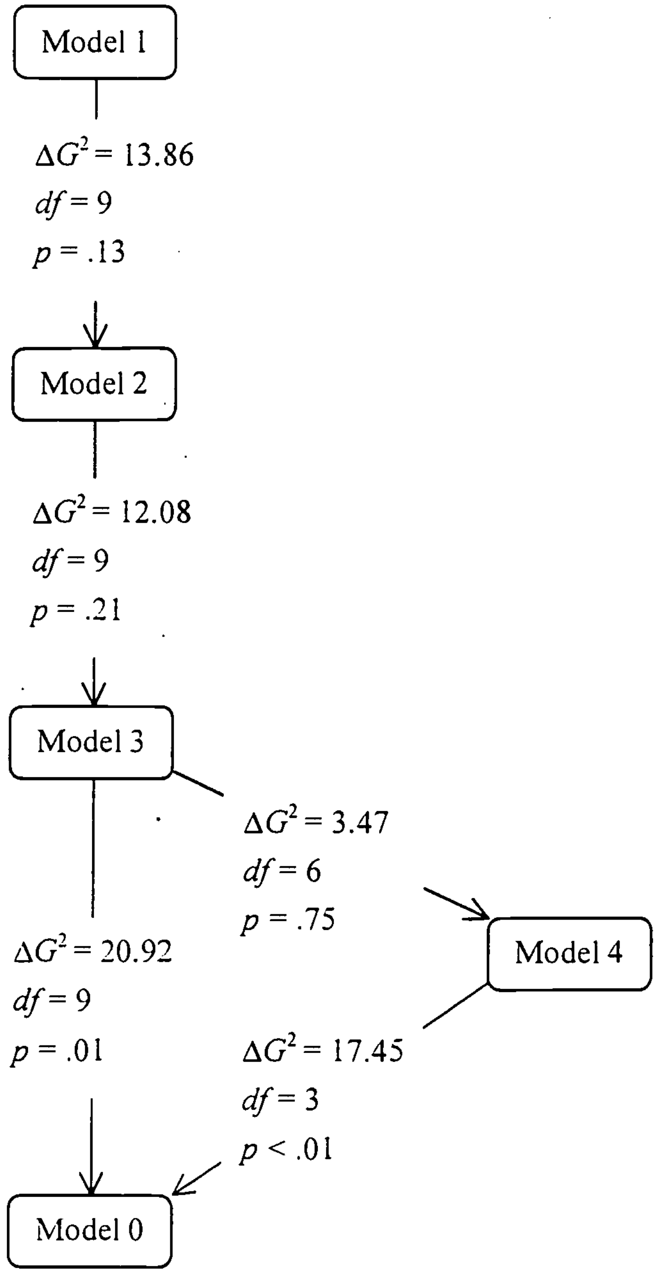

Figure 2. Likelihood ratio tests for the five nested models

Conclusion

Conventional DIF analysis is usually based on comparison of two groups, which is analogous to the *t*-test of two means. As the *t*-test is extended to ANOVA for multiple groups and multiple factors, the conventional DIF analysis is extended to the ANOVA-like DIF analysis in this study. Moreover, DIF is treated as a parameter to be estimated rather than simply a sign to be detected. In doing so, more thorough understanding of DIF can be acquired.

Results of the simulation studies show that all the parameters were recovered very well. A real data set with ten dichotomous items was analyzed. Various model were formed to test if the items show Status DIF, Gender DIF, or Status by Gender DIF. Neither Status by Gender DIF nor Gender DIF was found. However, three items show Status DIF. Although in this paper, a two-way factorial design with two levels on each factor was illustrated; this approach can be generalized to more than two ways with more than two levels on each. In addition, the approach is not limited to dichotomous items. It can be easily generalized to polytomous items.

References

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

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

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika, 46*, 443-459.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rasch, G. (1960). *Probabilistic Models for Some Intelligent and Attainment Tests.* Copenhagen: Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

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. Braum (Eds.), *Test validity* (pp. 147-169). Hillsdale, Lawrence Erlbaum Associates.

Wang, W. (1997). *The teacher personality inventory.* Unpublished manuscript.

Wang, W., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice* (pp. 139-155). Vol. 4, Norwood, NJ: Ablex.

Wu, M., Adams, R. J., & Wilson. M. R. (1997). *ConQuest.* Camberwell, Victoria: Australian Council for Educational Research.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: An ANOVA-Like Rasch Analysis of Differential Item Functioning

Author(s): Wen-Chung WANG

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 [X]	Level 2A [ ]	Level 2B [ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: [signature]
Printed Name/Position/Title: Wen-chung WANG Associate Prof.
Organization/Address: Dept. of Psychology, Nat'l Chung Cheng Univ. Chia-Yi, TAIWAN
Telephone: 886-5-2720411#6430
FAX: 886-5-2720857
E-Mail Address: psywcw@ccunix.ccu.edu.tw
Date: 2/24/98

(over)