Mathématiques et sciences humaines
Mathematics and social sciences

199 | 2012
Psychologie et mathématiques

Differential item functioning detection with logistic regression

Détection du fonctionnement différentiel d'items par régression logistique

Martha Cuevas and Victor H. Cervantes

Electronic version
URL: http://journals.openedition.org/msh/12274
DOI: 10.4000/msh.12274
ISSN: 1950-6821

Publisher
Centre d'analyse et de mathématique sociales de l'EHESS

Printed version
Date of publication: 15 September 2012
Number of pages: 45-59
ISSN: 0987-6936

Electronic reference
Martha Cuevas and Victor H. Cervantes, « Differential item functioning detection with logistic regression », *Mathématiques et sciences humaines* [Online], 199 | 2012, Online since 04 December 2012, connection on 23 July 2020. URL : http://journals.openedition.org/msh/12274 ; DOI : https://doi.org/10.4000/msh.12274

Math. Sci. hum. / Mathematics and Social Sciences (50e année, n° 199, *2012*(3), p. 45–59)

DIFFERENTIAL ITEM FUNCTIONING DETECTION WITH LOGISTIC REGRESSION

Martha CUEVAS [1], Víctor H. CERVANTES [2]

RÉSUMÉ – Détection du fonctionnement différentiel d'items par régression logistique
La régression logistique *a été utilisée comme une méthode d'identification du DIF dans différents contextes. Certaines études ont montré que cette procédure peut être affectée par des variables comme le ratio des tailles entre groupes, la taille de l'échantillon, et qu'elle semble liée avec les gammes de difficulté et la discrimination des items [Herrera, 2005 ; Santana, 2009]. Nous avons fait une étude de simulation avec quatre variables indépendantes partiellement traversées qui ont abouti à 270 conditions et simulé 200 répliques pour chacune d'elles. La différence des McFadden R^2 ($R^2\Delta$) entre modèles a été utilisée comme une mesure de la taille de l'effet et comme variable dépendante afin de minimiser les taux de faux positifs et négatifs que le test statistique n'aurait pas été en mesure de contrôler. Nous avons utilisé des modèles linéaires pour définir les variables qui affectent les mesures de la taille de l'effet : $R^2\Delta$ pour la détection des items avec du DIF uniforme (DRU) et $R^2\Delta$ pour détecter les items avec du DIF non uniforme (DRN). Les résultats montrent que les variables manipulées et leurs interactions affectent de différentes manières le DRU et le DRN. Nous avons également obtenu des seuils pour les variables dépendantes, aussi bien pour DRU que pour DRN, pour plusieurs niveaux des variables en jeu.*

MOTS-CLÉS – Ampleur du DIF, Fonctionnement différentiel d'items, Longueur des tests, Ratio de la taille du groupe de l'échantillon, Régression logistique, Taille de l'échantillon

SUMMARY – *Logistic Regression has been used as a method to identify differential item functioning (DIF) in different contexts. Some studies have shown that DIF detection through this procedure may be affected by variables such as sample size ratio, and sample size. It also seems related with specific item parameters like certain ranges of difficulty and discrimination [Herrera, 2005 ; Santana, 2009]. We made a simulation study with four partially crossed independent variables which resulted in 270 conditions and simulated 200 replications for each experimental condition. The difference of McFadden's R^2 between models ($R^2\Delta$) was used as an effect size measure and as a dependent variable in order to minimize type I and II errors that the statistical test would not have been able to control. We used linear models to define which variables affected the effect size measures : $R^2\Delta$ for detecting items with uniform DIF (DRU) and for detecting items with non uniform DIF (DRN). The results show that manipulated variables and some of their interactions affect DRU and DRN differently. We also obtained cut-off points, both for DRU and DRN, for several levels of the variables that affect the $R^2\Delta$ measures.*

KEYWORDS – Differential item functioning, Length of test, Logistic Regression, Magnitude of DIF, Sample size, Sample size ratio

[1] Instituto Colombiano para la Evaluación de la Educación - ICFES - Calle 17 No. 3-40, Bogotá (Colombia), mcuevas@icfes.gov.co

[2] Instituto Colombiano para la Evaluación de la Educación - ICFES - Calle 17 No. 3-40, Bogotá (Colombia), vcervantes@icfes.gov.co

1. INTRODUCTION

Reliability and validity are two characteristics that all measurement instruments must have, including educational and psychological tests. The American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME) [1999] claim "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Thus, any test parameter that is different between two or more subpopulation groups, like item difficulty, may be a sign of a threat to test validity because the test results would need different interpretations for each group. In this context, differential item functioning (DIF) becomes an important validity and bias issue of test analysis.

Camilli and Shepard [1994], cited by Wu & Ercikan [2006], define DIF as a statistical procedure that checks whether examinees with comparable total test scores belonging to different groups answer similarly the individual items of the test. In a more general way, DIF refers to differences in psychometric properties of the items between groups [Fidalgo, 1996]. In conducting DIF analyses it is usual that there are, at least, two groups of interest : the focal group and the reference group. The former generally refers to a minority or traditionally considered disadvantaged group, while the latter is the majority or privileged group.

In the past, DIF and bias were interchangeable words, but in 1988 Holland and Thayer helped to precise the difference between these two concepts [Herrera, 2005]. Nowadays, bias is used to refer to an informed opinion [Holland & Wainer, 1993, p. XIV, cited by Herrera, 2005] that takes into account the aim of the test as well as contextual information about groups, which can explain DIF on a given item. In general, the DIF analysis is considered the first step, statistical step, in order to decide if an item could be biased against a particular group.

An additional concept that we should take into account is impact ; this refers to actual differences in attribute or ability distribution between groups. This is important because if an item is detected by a statistical procedure as having DIF, it does not necessarily mean the item is biased. In this case, it is crucial to determine if the reason for which the groups score differently in an item is relevant or not regarding the measurement object. In the first case, the DIF is due to actual differences, and in the second one it is due to bias [Gómez & Navas, 1998].

1.1. LOGISTIC REGRESSION FOR IDENTIFYING DIF

Different methods have been proposed to identify items with DIF : Mantel-Haenzel, difference of difficulties, Lord's χ^2, Non compensatory DIF (NCDIF), SIBTEST, logistic regression, etc. Logistic regression has been used as a method to identify DIF in different contexts like health, education and psychology (v. g. [Bennett *et al.*, 1987 ; Clauser *et al.* 1996 ; Pertersen *et al.* 2003]). The procedure for identifying DIF by using logistic regression consists of fitting the models shown in equations (1) to (3), which we are rewriting from [Camilli & Shepard, 1994].

$$\mathsf{logit}\,(P(U=1)) = \beta_0 + \beta_1\theta + \beta_2 g + \beta_3\theta g \tag{1}$$

$$\mathsf{logit}\,(P(U=1)) = \beta_0 + \beta_1\theta + \beta_2 g \tag{2}$$

$$\mathsf{logit}\left(P(U=1)\right) = \beta_0 + \beta_1\theta \tag{3}$$

where :

–$(P(U=1))$ is the probability of giving a correct response to a specific item,
– θ is examinee's ability in the test or her/his total score, and
– g is the group which the examinee belongs to (reference or focal).

The comparison between these models through the G^2 statistic (likelihood ratio test statistic with χ^2 distribution of degrees freedom equal to the difference between the number of parameters of the compared models) allows to identify if there is DIF for an item [Thissen *et al.*, 1993] as well as the type of DIF : 'Uniform', 'Non Uniform' or 'Mixed' DIF. Additionally, the null hypothesis statistical test may be complemented by the use of an effect size measure in order to better inform the test developers about the magnitude of the differences between the focal and reference groups. According to Kirk [1996, cited by Zumbo, 1999], small sample sizes cannot show interesting statistical effects and large sample sizes can lead to statistically significant results where the effect is very small and there is not a practical significance. In this context, the $R^2\Delta$, defined as difference of the R^2 between the models, has been proposed as the natural effect size measures [Zumbo, 1999].

Thus, according to equations (1) to (3) and Santana [2009] an item presents uniform DIF, in statistical terms, if the G^2 statistic is significant between models (2) and (3). Similarly, an item presents non uniform DIF if the G^2 statistic is significant between models (1) and (2). If an item is deemed to present both uniform and non uniform DIF, it is classified as presenting mixed DIF.

Within the Item Response Theory (IRT) framework, this classification of DIF may be interpreted as follows : When an item is classified as presenting uniform DIF, the difficulty parameter changes but discrimination is the same [Camilli & Shepard, 1994]. This may be seen as evidence that an irrelevant dimension is being tackled by the item and the groups differ in the distribution of this dimension [Herrera, 2005]. When an item is classified as presenting non uniform DIF, the difficulty parameter is the same but the discrimination is not. In this case, the interpretation of DIF would imply that the variance of the groups on irrelevant dimension is not the same or the correlation between both dimensions is different between the groups [Herrera, 2005]. Finally, when an item is classified as presenting mixed DIF, both difficulty and discrimination are different between focal and reference groups [Herrera, 2005].

In order to know which group is favored when an item has DIF, the coding should be known. For example, for uniform DIF, if 1 is the code for the focal group and 2 for the reference group and the sign of β_2 is negative, the favored group is the focal one. In the non uniform DIF case, if $\beta_3 > 0$ the item favors the reference group's persons who have high magnitude attribute and the focal group's persons with low ability. In the same way, $\beta_3 < 0$ shows that low ability people from the reference group and persons with high ability in the focal group are favored [Herrera *et al.*, 2005].

As stated previously, one or several $R^2\Delta$ measures can be used as effect size measures and according to a defined cut-off point establish some level of practical significance. Furthermore, this classification might be incorporated into the decision process about whether an item presents DIF or not as is often encountered in applied

analyses. The latter strategy has been identified as a "blended" statistical test by Zumbo [2008]. It is worth noting that a "blended" approach produces a different statistical test with different properties than the initial test [Zumbo, 2008]. Regarding the blended strategy, Zumbo [2008] considers it is conservative because the cut-offs counter-act the power of the statistical test, being this the first reason for including a size effect measure. In the same way, he inquires why if the effect size does not add information to the statistical decision then we use it.

Some studies have shown that DIF detection through the logistic regression procedure may be affected by variables such as sample size ratio, sample size, and that it seems related with specific item parameters like certain ranges of difficulty and discrimination. For instance, the work of Herrera [2005] showed that the size of the reference group does not significantly affect the Type I error of the logistic regression but that the size ratio between reference and focal groups had a significant effect for items with low discrimination. There was also a significant interaction between the two factors (size and size ratio) for items with high difficulty. Overall, Herrera [2005] found an adequate control of Type I error, with a maximum false positive rate mean of 6.8 %. Moreover, her results showed that larger sample sizes presented higher rates of correct detections but fewer correct detections as the sample size ratio grew between both groups. Anyhow, Herrera [2005] found low power for detecting uniform DIF using the logistic regression procedure for joint uniform and non uniform DIF detection given the sample sizes and sample size ratios she manipulated. Furthermore, Herrera *et al.* [2005] assert, citing Rogers (1990) and Rogers & Swaminathan (1993), that the degree freedom lost by including an extra parameter for detecting non uniform DIF could decrease the power of logistic regression for detecting uniform DIF.

Santana [2009] included more extreme sample size ratios than Herrera *et al.* [2005], impact, the percentage of DIF items and simulation model (one or three parameters logistic IRT model) and found that these factors affect both Type I error and power for this procedure. Regarding Type I error, Santana [2009] found that it is larger under conditions in which the samples size ratio is less extreme and when there is impact between the focal and reference groups. Furthermore, she found greater power in less extreme conditions of sample size ratio.

Given the above context and other research (such as Cromwell [2006] ; Swaminathan & Rogers [1990]), and taking into account some characteristics of tests developed and applied in Colombia, we set the following objectives for this study : (1) identify some variables that can affect an effect size measure for the logistic regression procedure in the detection of DIF, and (2) find the best cut-off points for an effect size measure of RL procedure in the detection of DIF that takes into account the variables that affect it.

2. METHOD

2.1. Simulation

We designed an experimental study with the following partially crossed independent variables :

Sample size : n = 7500, 8500, 26000 and 33000.

Group sample size ratio : 1, 3, 4, 6, 10, 20, 150 and 250. A group sample size ratio of 3 means there are three examinees in the reference group for each examinee in the focal group.

Test length : K = 18 and 26.

DIF magnitude : No DIF, two items with both small and big DIF and four items with both small and big DIF. The DIF magnitude was taken as the difference between the difficulty parameters (b) and the discrimination parameters (a) for the focal group and the reference group. For the "no DIF" condition the parameters were the same for both groups. There were two items with uniform DIF (items 10 and 15) when the condition had two DIF items, and there were two items with uniform DIF and two items with non uniform DIF (items 4, 7, 10, and 15) when the condition had four DIF items. The item parameters that were changed for the focal group appear in Table 1, they were the same regardless of test length.

TAB. 1. Item parameters for DIF items

	Reference group		Focal group (big DIF)		Focal group (small DIF)	
Item	a	b	a	b	a	b
4	0.64695	-0.41164	1.44695	-0.41164	1.04695	-0.41164
7	0.3103	0.27266	1.1103	0.27266	0.7103	0.27266
10	0.49396	-1.36204	0.49396	-0.56204	0.49396	-0.96204
15	0.31593	1.25932	0.31593	2.05932	0.31593	1.65932

The crossing of the above variables produced 270 experimental conditions that were replicated 200 times. Responses for all conditions were simulated using the two parameters logistic IRT model for dichotomous items. The number of replicates was chosen following Atar [2007] who showed that this number produces stable results and more replications do not improve stability. The levels of the other variables were chosen bearing in mind the characteristics of the Colombian tests that are presented by all students in 5th and 9th grades every three years. The data were simulated with R software (version 2.12.1) [R Development Core Team, 2010].

2.2. DATA ANALYSIS

We ran the logistic regression procedure in order to identify DIF items in each experimental condition adapting the script developed by Cervantes [2008] and used in Santana's [2009] study. The logistic regression procedure was run with R software (version 2.12.1) [R Development Core Team, 2010]. This procedure was run in two phases performing item purification as is recommended by Clauser *et al.* [1993] and Zenisky *et al.* [2003] for performing DIF analyses and controlling (matching) by examinees' number of correct responses in the model. In the first phase, we identified items with DIF if the G^2 statistic was significant at the 0.05 level ($p<0.05$) when comparing models (1) and (3) (simultaneous test of no uniform nor non uniform DIF hypothesis [Swaminathan & Rogers, 1990]). This strategy has shown an improvement of power and less computational cost than other like comparing models (1) and (2)

with (3) separately. However, this does not allow us to identify the type of DIF [Hidalgo *et al.*, 2005]. In the second phase, items identified in the first phase were not considered for the correct responses. Additionally, the procedure to identify DIF through model comparison was the depicted in section 1.1. in order to detect uniform and non uniform DIF.

McFadden's $R^2\Delta$ was used as an effect size measure and as a dependent variable. Specifically, we used the $R^2\Delta$ when the group effect was introduced for detecting items with uniform DIF (DRU), and the $R^2\Delta$ when the interaction effect between correct responses number and group was introduced for detecting items with non uniform DIF (DRN). We chose McFadden's R^2 since it may be : "interpreted as the ratio of the estimated information gain when using the current model M in comparison with the null model to the estimate of the information potentially recoverable by including all possible explanatory variables" (see Kent (1983) and Hastie (1987)) [Shtatland *et al.*, 2000, p. 2].

We ran linear models with the R software (version 2.12.1)[R Development Core Team, 2010] and the packages lme4 [Bates *et al.*, 2011] and car [Fox & Weisberg, 2011] to define which variables affected the effect size measure. Due to computational limitations, it was necessary to make random samples of replications to run these models. We obtained 30 samples of six replications each for DIF items and considered a significant effect if $p < 0.05$ in two or more samples.

After identifying DIF items, we took the DRU and DRN of all items in DIF conditions in order to define the best cut-off points through a effect size measure that would allow us to get a measure of DIF magnitude. When an item was not detected by statistical test ($p < 0.01$), its DRU or DRN was substituted for 0, depending on if the procedure was detecting uniform DIF or non uniform DIF respectively. We used the areas under the curve (AUC) of receiver - operating curves (ROC) between specificity (1 - Type I error rate) and sensitivity (power) to find the best cut-off points for each combination of variables that affected the DRU and the DRN. This last analysis was run with R software (version 2.12.1) [R Development Core Team, 2010] and the package ROCR [Sing *et al.*, 2009].

3. RESULTS

Given the amount of different experimental conditions, most results are presented in figures where labels for the independent variables are encoded in the following manner :

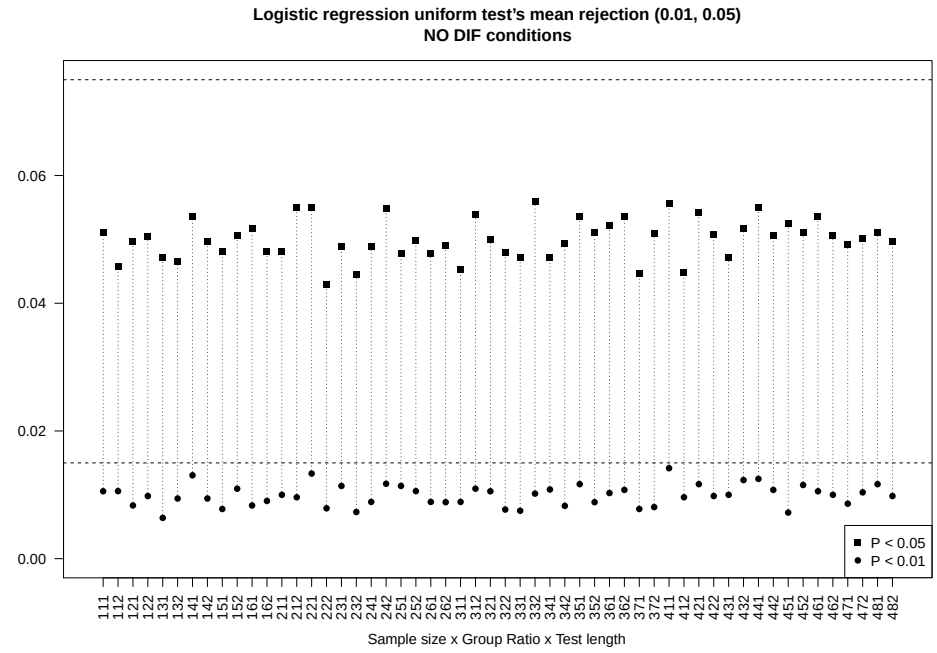
Sample size : 1 = 7500, 2 = 8500, 3 = 26000 and 4 = 33000.

Group sample size ratio : 1 = 1, 2 = 3, 3 = 4, 4 = 6, 5 = 10, 6 = 20, 7 = 150 and 8 = 250.

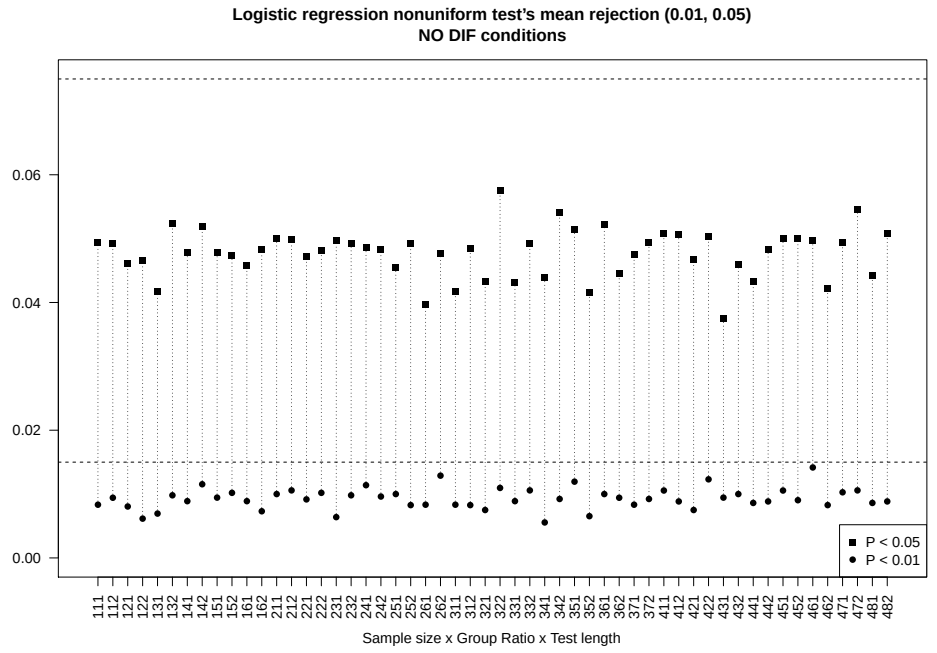
Test length : 1 = 18 and 2 = 26.

3.1. TYPE I AND TYPE II ERRORS

As shown in Figure 1, Type I error does not seem affected by some particular variables. Furthermore, according to Bradley's liberal criteria (1.5 * α, [Bradley, 1978]), shown by the horizontal pointed lines in the figure, the G^2 statistic testing for uniform and non uniform DIF properly controls Type I error under all conditions.



(a)

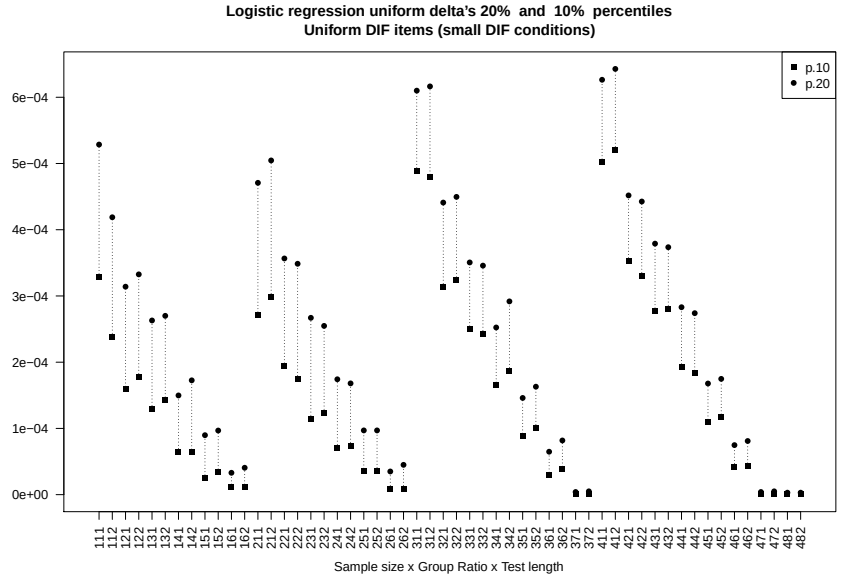


(b)

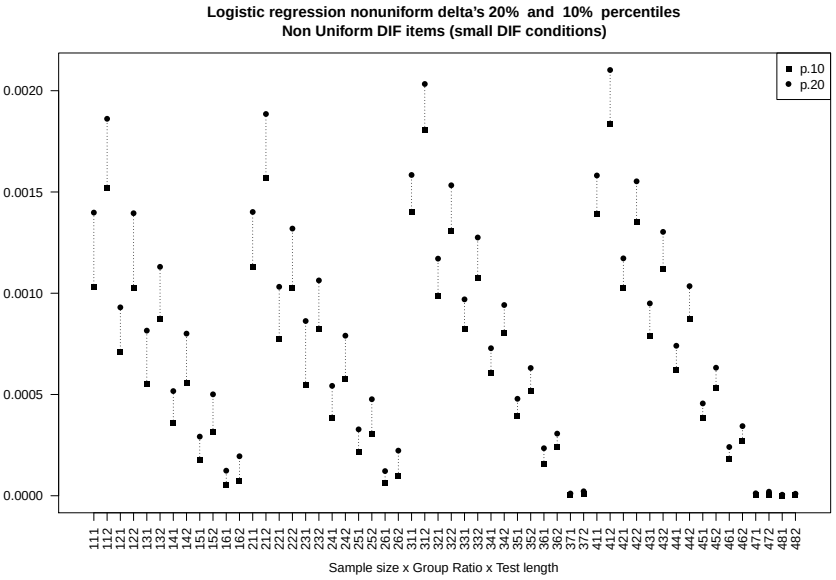
FIGURE 1. Type I error rates for DIF detection under conditions with no DIF items : (a) uniform and (b) non uniform

Figures 2 and 3 show that sample size affects power for detecting both uniform

and non uniform DIF in small DIF conditions, and big conditions when using the $R^2\Delta$ measure. Also, since $R^2\Delta$ gets bigger when sample size increases, it may be necessary to define lower cut-off points for conditions with small sample size in order to detect small DIF, both uniform and non uniform, and big uniform DIF instead of using constant cut-off points. However, for conditions with items presenting big non uniform DIF, the relation with sample size is not so clear.

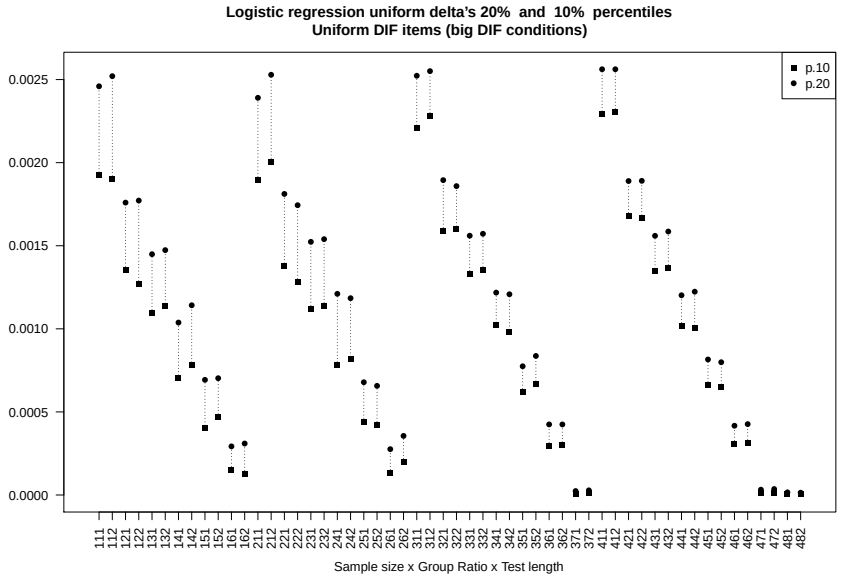


(a)

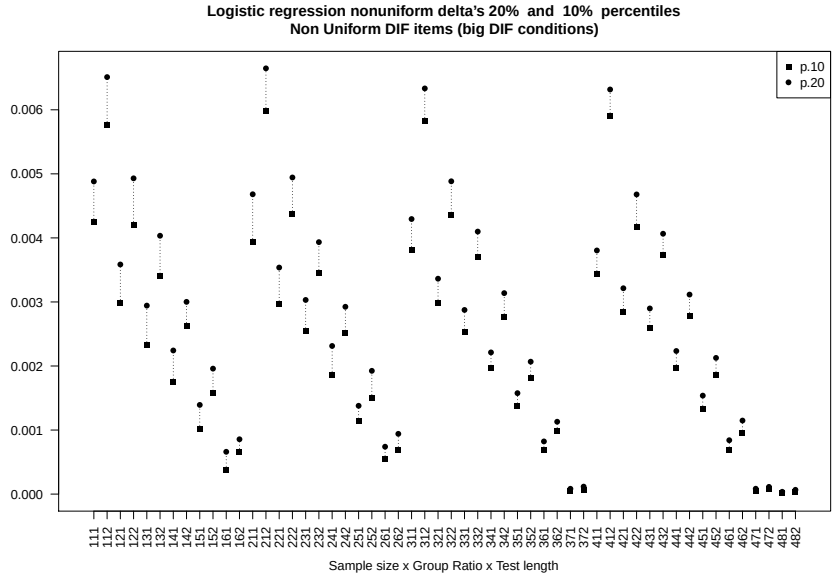


(b)

FIGURE 2. $R^2\Delta$'s 20 % and 10 % percentiles for DIF detection under conditions with small DIF items : (a) uniform and (b) non uniform.



(a)



(b)

FIGURE 3. $R^2\Delta$'s 20 % and 10 % percentiles for DIF detection under conditions with big DIF items : (a) uniform DIF and (b) non uniform.

On the other hand, there is a clear effect of sample size ratio on Type II error because the 20 and 10 percentiles are localized in lower $R^2\Delta$ values for more extreme conditions. In other words, $R^2\Delta$ becomes smaller as the ratio gets more extreme for conditions with either small or big DIF. Finally, there appears to be an effect of test length on power but it is clearer for non uniform DIF than it is for uniform DIF.

3.2. EFFECTS ON DRU AND DRN

Tables 2 and 3 show ANOVA tables of the linear models performed on DRU and DRN where effects with a significance higher than 0.05 have been omitted (except for main effects involved in significant interaction terms). Since the linear models were obtained for 30 random samples of six replications from each of the experimental conditions, the tables shown correspond to one of the 30 samples. These example tables were chosen among those which depicted all effects that were significant in at least two of the samples. These results support the observations on Figures 2(a) to 3(b) although the interaction of DIF magnitude and test length for DRN, which was not noted before, seems stronger than the effect due to test length alone.

TAB. 2. Effects on uniform $R^2\Delta$

Variable	F	p
Size	70.610	9.69E-44
Ratio	473.287	0
DIF Magnitude	892.215	0
Test length	3.983	0.046
Size × DIF Magnitude	3.778	9.64E-05
Ratio × DIF Magnitude	17.012	3.40E-58

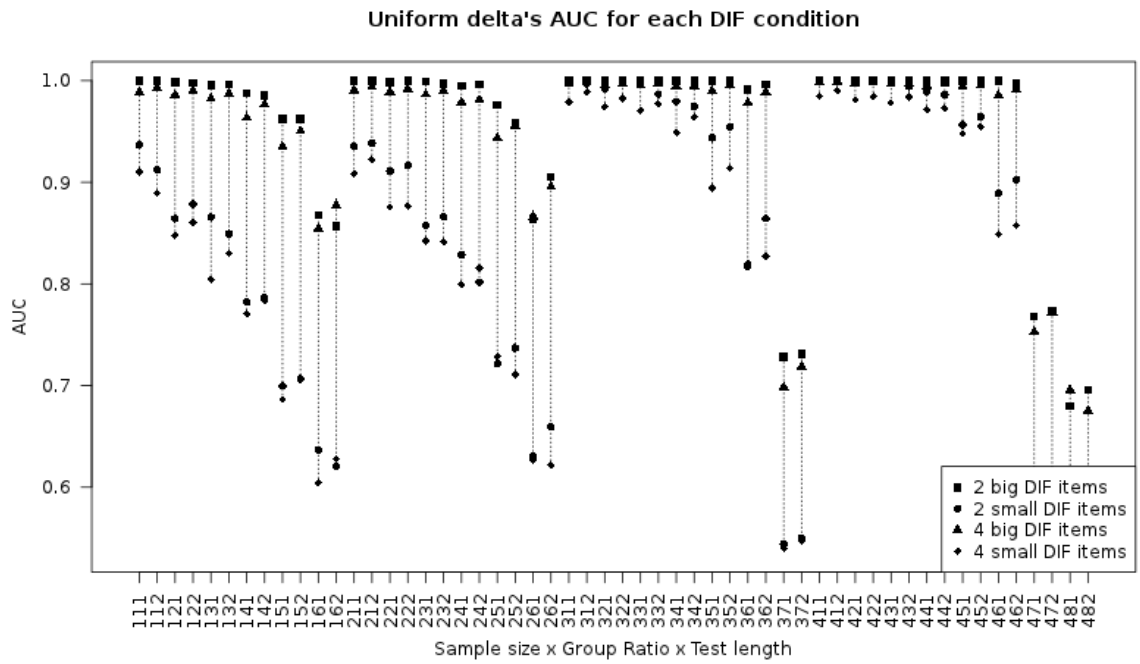
TAB. 3. Effects on non uniform $R^2\Delta$

Variable	F	p
Size	27.601	2.95E-17
Ratio	133.511	6.49E-145
DIF Magnitude	638.029	4.98E-113
Test length	0.039	0.844
Size × DIF Magnitude	4.081	0.007
Ratio × DIF Magnitude	30.403	7.42E-39
DIF Magnitude × Test lenght	4.122	0.043

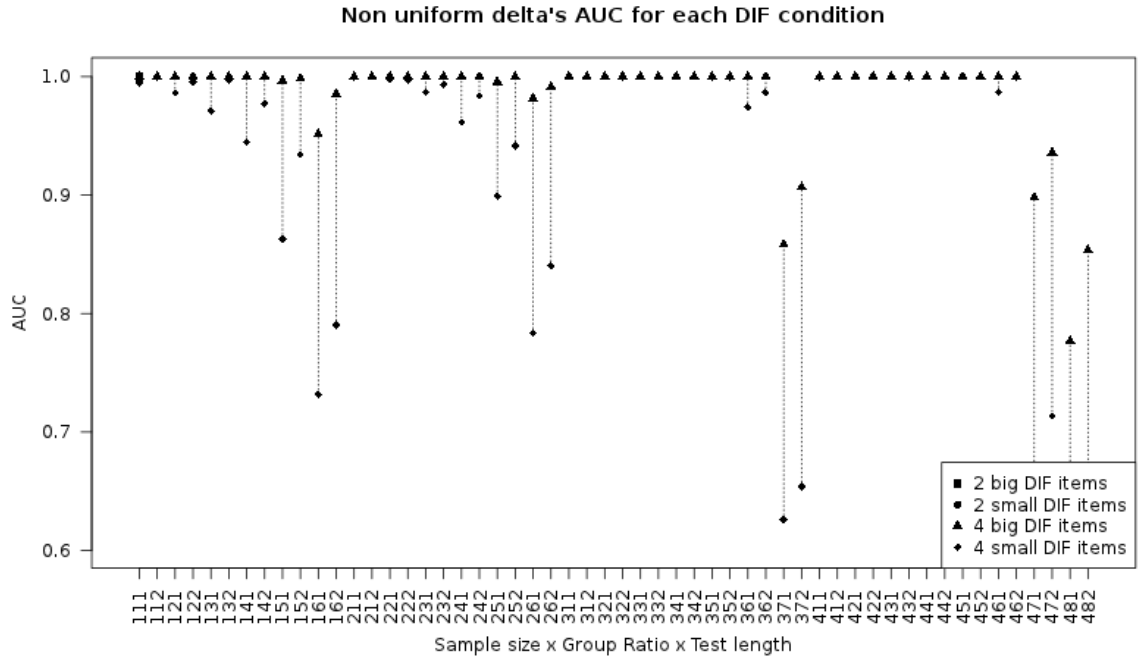
3.3. CUT-OFF POINTS

Figures 4(a) and 4(b) show AUC for each condition with DIF. AUC for uniform DIF increase with sample size and decrease for extreme sample size ratios although there is no clear trend for test length. AUC for non uniform DIF also increase with the sample size and decrease for extreme sample size ratios, and increase with test length. Furthermore, the areas were mostly stable (and nearly equal to the possible maximum) for big DIF conditions. As it is to be expected, AUC for conditions with big DIF were higher than those for conditions with small DIF.

We found that cut-off points were not too different for varying test lengths when holding other variables constant ; based on this, we obtained cut-off points that did



(a)



(b)

FIGURE 4. DRU's and DRN's areas under the curve (AUC) for each condition with DIF : (a) uniform DIF and (b) non uniform.

not took this variable into account. In addition, for setting cut-off points for small DIF condition, we considered the cut-off points obtained to control Type I error in

conditions with no DIF that were more similar, in terms of the specificity, with the best cut-off points in small DIF conditions. For this reason, the cut-off points for small DIF may be smaller for some conditions with large sample sizes when they are compared to similar conditions of small sample size.

The best cut-off points for each condition and type of DIF are presented in Tables 4 and 5. Cut-off points that controlled Type I error within Bradley's liberal criterion for $\alpha = 0.05$ [Bradley, 1978] and had power of at least 0.7 are shown. Although the cut-off point for the condition with sample size of 8500 and sample size ratio of 20 for uniform big DIF showed a specificity that did not meet Bradley's criterion, we decided to present this cut-off point since its specificity was close to the required value, and the similar condition of sample size 7500 did meet the criterion.

Tab. 4. Cut-off points for each condition without taking into account test length for DRU

Sample size	Sample size ratio	Big DIF			Small DIF (big DIF)		
		Cut-off	Specificity	Sensitivity	Cut-off	Specificity	Sensitivity
7500	1	0.001012	0.961	0.989	0.000363	0.928	0.859
7500	3	0.000696	0.955	0.98	0.000364	0.928	0.765
7500	4	0.000698	0.96	0.971	0.000356	0.929	0.722
7500	6	0.000507	0.944	0.949			
7500	10	0.000413	0.93	0.907			
7500	20	0.000356	0.927	0.764			
8500	1	0.001047	0.964	0.989	0.000338	0.93	0.874
8500	3	0.000697	0.96	0.989	0.000326	0.931	0.823
8500	4	0.000584	0.955	0.983	0.00032	0.93	0.747
8500	6	0.000512	0.953	0.961			
8500	10	0.000385	0.933	0.918			
8500	20	0.000319	0.922	0.796			
26000	1	0.001672	0.983	0.995	0.000218	0.965	0.983
26000	3	0.001048	0.975	0.993	0.000174	0.959	0.98
26000	4	0.000813	0.97	0.996	0.000155	0.953	0.965
26000	6	0.000565	0.967	0.996	0.000138	0.947	0.94
26000	10	0.000366	0.966	0.989	0.000113	0.932	0.871
26000	20	0.000182	0.956	0.966	0.000103	0.93	0.708
33000	1	0.001658	0.983	0.996	0.000208	0.967	0.995
33000	3	0.001146	0.979	0.996	0.000146	0.959	0.996
33000	4	0.000818	0.97	0.996	0.000144	0.96	0.974
33000	6	0.000627	0.972	0.993	0.000115	0.948	0.959
33000	10	0.000395	0.97	0.991	0.000096	0.937	0.923
33000	20	0.000166	0.958	0.982	0.000082	0.926	0.79

3.4. CONCLUSIONS

Type I error results agree with previous studies such as Santana [2009], if we take into account only conditions without impact in that research, and Herrera [2005]. This means that the G^2 statistic test for logistic regression has a good control on Type I error under conditions similar to those manipulated in this study. Results on power differ regarding uniform DIF of those of Herrera [2005]'s study since the maximum power found there was 0.4. This result may be due to the use of different sample sizes. Nevertheless, our results are similar to those of Santana [2009] where

Tab. 5. Cut-off points for each condition without taking into account test length for DRN

Sample size	Sample size ratio	Big DIF			Small DIF (big DIF)		
		Cut-off	Specificity	Sensitivity	Cut-off	Specificity	Sensitivity
7500	1	0.002708	1	0.995	0.00062	0.988	0.978
7500	3	0.001287	1	0.998	0.000499	0.978	0.97
7500	4	0.001542	1	0.995	0.000539	0.981	0.938
7500	6	0.00109	0.999	0.993	0.000408	0.956	0.904
7500	10	0.000621	0.987	0.985	0.00038	0.951	0.803
7500	20	0.000372	0.953	0.939			
8500	1	0.002618	1	0.999	0.000628	0.993	0.995
8500	3	0.00183	1	0.998	0.000534	0.988	0.978
8500	4	0.001637	1	0.994	0.000471	0.98	0.953
8500	6	0.000991	1	0.995	0.000359	0.959	0.938
8500	10	0.000658	0.996	0.983	0.000333	0.951	0.836
8500	20	0.000408	0.97	0.96			
26000	1	0.003626	1	0.968	0.000589	1	1
26000	3	0.002615	1	0.985	0.000458	1	1
26000	4	0.002029	1	0.993	0.00043	1	1
26000	6	0.00134	1	0.996	0.000303	0.999	0.996
26000	10	0.000957	1	0.995	0.000197	0.992	0.994
26000	20	0.000464	1	0.99	0.000138	0.973	0.946
26000	150	0.000099	0.942	0.795			
33000	1	0.002802	1	0.993	0.000588	1	1
33000	3	0.002362	1	0.99	0.000468	1	1
33000	4	0.002083	1	0.99	0.000324	1	0.999
33000	6	0.001748	1	0.979	0.000287	1	0.999
33000	10	0.000983	1	0.996	0.000204	0.998	0.995
33000	20	0.000318	1	0.999	0.000128	0.984	0.973
33000	150	0.000078	0.94	0.858			

the lowest power for uniform DIF found was 0.33 under extreme conditions such as a ratio sample size of 250 and a three parameters model. The three studies found similar results on the effect of sample size ratio on power – it is low when there are extreme differences between focal and reference groups sample sizes.

Moreover, this research showed that it is better to have specific cut-off points on an effect size measure, such as $R^2\Delta$, than using a single overall cut-off value. Although all manipulated variables affected the effect size measures, the cut-off points could be simplified without affecting power nor Type I error control. The use of ROC and AUC based cut-off points, as shown in this research, allows us to characterize the statistical properties of DIF detection and classification under the "blended" approach described by Zumbo [2008]. This also allows us to gain information on the sensitivity of DIF classification in comparison to the regular null hypothesis testing approach.

A limitation of this study is that we did not manipulate impact between the focal and reference groups. This variable was shown by Santana [2009] to have an effect on DIF detection with the logistic regression procedure. A second limitation is that the areas between item characteristic curves (ICC, [Raju, 1988]) of the focal and the reference groups varied greatly (between 1.89 and 0.4). Authors like Swaminathan & Rogers [1990] have said areas of 0.6 and 0.8 would be considered moderate to high

DIF and that they would represent the actual limits of discrimination item values. However, we have observed applied data where the area between ICC may be as high as the areas simulated in this study. These limitations suggest further research that should explore smaller areas between the ICC for non uniform DIF and the effect of impact.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA) & the National Council on Measurement in Education (NCME) (1999), *Standards for educational and psychological testing*, AERA Publications.

ATAR B. (2007), *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures*, Doctoral thesis, Florida State University. http ://CRAN.R-project.org/package=lme4

BATES D.M., MAECHLER M., BOLKER B. (2011), "lme4 : Linear mixed-effects models using S4 classes", R package version 0.999375-42. http ://CRAN.R-project.org/package=lme4.

BENNETT R.E., ROCK D.A., KAPLAN B.A. (1987), "SAT Differential Item Performance for Nine Handicapped Groups", *Journal of Educational Measurement* 24(1), pp. 41-55.

BRADLEY J.V. (1978), "Robustness ?", *British Journal of Mathematical and Statistical Psychology* 31, pp. 144-152.

CAMILLI G., SHEPARD L.A. (1994), *Methods for Identifying Biased Test Items*, SAGE Publications.

CERVANTES V.H. (2008), Detección de DIF-Dos etapas. *Purificación de medida de equiparación y reporte de ítems que presentan DIF. Procedimiento de R-L. Uso de pseudos R^2 como medidas de tamaño del efecto* [Script para R], Proyecto de identificación de sesgo cultural en el Examen de Estado ICFES, Grupo Métodos e instrumentos de investigación en Salud, Universidad Nacional de Colombia.

CLAUSER B., MAZOR K., HAMBLETON R. K. (1993), "The Effects of Purification of Matching Criterion on the Identification of DIF Using the Mantel-Haenszel Procedure", *Applied Measurement in Education* 6(4), pp. 269-279.

CLAUSER B., NUNGESTER R.J., SWAMINATHAN H. (1996), "Improving the Matching for DIF Analysis by Conditioning on Both Test Score and an Educational Background Variable", *Journal of Educational Measurement* 33(4), pp. 453-464.

CROMWELL S. (2006), *Improving the prediction of differential item functioning : a comparison of the use of an effect size for logistic regression DIF and Mantel-Haenszel DIF methods*, Doctoral thesis, Texas A & M University (Texas).

FIDALGO A.M. (1996), "Funcionamiento diferencial de los ítems", in J. Muniz (ed.), *Psicometria*, Madrid Universitas (S.A.), pp. 371-455.

FOX J., WEISBERG S. (2011), "An R Companion to Applied Regression", second edition, Thousand Oaks (CA), Sage publications.
URL : http ://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

GÓMEZ J., NAVAS M.J. (1998), "Impacto y Funcionamiento Diferencial de los Ítems Respecto al Género en una Prueba de Aptitud Numérica" *Psicothema* 10(3), pp. 686-696.

HERRERA A.N. (2005), *Efecto del tamaño de muestra y razón de tamaños de muestra en la detección de funcionamiento diferencial de los ítems*, Doctoral thesis, Universidad de Barcelona, Barcelona (Spain).

HERRERA A.N., GÓMEZ J., HIDALGO M.D. (2005), "Detección en los ítems mediante análisis de tablas de contingencia", *Avances en Medición* 3, pp. 29-52.

HIDALGO M.D., GÓMEZ J., PADILLA J.L. (2005), "Regresión Logística : alternativas de análisis en la detección del funcionamiento diferencial del ítem", *Psicothema* 17(3), pp. 509-515.

PETERSEN M.A., GROENVOLD M., BJORNER J.B., ARONSON N., CONROY T., CULL A., FAYERS P., HJERMSTAD M., SPRANGERS M., SULLIVAN M. (2003), "Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire", *Quality of Life Research* 12(4), pp. 373-385.

R DEVELOPMENT CORE TEAM (2010), *R : a language and environment for statistical computing*, Version 2.12.1, Vienna (Austria), avalaible from URL : http ://www.r-project.org.

RAJU N. (1988), "The Area Between Two Item Characteristic Curves", *Psychometrika* 53(4), pp. 495- 502.

SANTANA A.C. (2009), *Efecto de la razón de tamaños de muestra en la detección de funcionamiento diferencial de los ítems a través del procedimiento de regresión logística*, Master thesis, Universidad Nacional de Colombia, Bogotá (Colombia).

SHTATLAND E.S., MOORE S., BARTON M.B. (2002), "One more time about R^2 measures of fit in logistic regression", *NESUG : Statistics, Data Analysis & Econometrics* 15, Retrieved 28 December 2011, from http ://www.lrz.de/ wlm/ST004.pdf.

SING T., SANDER O., BEERENWINKEL N., LEGAUER T. (2009), "ROCR : Visualizing the performance of scoring classifiers", R package version 1.0-4.
http ://CRAN.R-project.org/package=ROCR.

SWAMINATHAN H., ROGERS H.J. (1990), "Detecting Differential Item Functioning Using Logistic Regression Procedures", *Journal of Educational Measurement* 27(4), pp. 361-370.

THISSEN D., STEINBERG L., WAINER H. (1993), "Detection of differential item functioning using the parameters of item response model", in P. W. Holland & H. Wainer (eds), *Differential item functioning*, Hillsdale (New Jersey), Lawrence Erlbaum Associates, Inc. Publishers, pp. 67-113.

WU A.D., ERCIKAN K. (2006), "Using Multiple-Variable Matching to Identify Cultural Sources of Differential Item Functioning", *International Journal of testing* 6(3), pp. 287-300.

ZENISKY A.L., HAMBLETON R.K., ROBIN F. (2003), "Detection of Differential Item Functioning in Large-Scale State Assessments : A Study evaluating a Two-Stage Approach", *Educational and Psychological Measurement* 63(1), pp. 51-64.

ZUMBO B.D. (1999), *A handbook on the theory and methods of differential item functioning (DIF) : Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*, Ottawa (Ontario), Directorate of Human Resources Research and Evaluation, Department of National Defense.

ZUMBO B.D. (2008), *Statistical Methods for Investigating Item Bias in Self-Report Measures*, Università degli Studi di Firenze E-prints Archive, Florence (Italy).