Psicológica (2012), 33, 135-156.

Comparison of three software programs for evaluating DIF by means of the Mantel-Haenszel procedure: EASY-DIF, DIFAS and EZDIF

José Luis Padilla[*1], Mª Dolores Hidalgo[2], Isabel Benítez[1] & Juana Gómez-Benito[3]

[1]*University of Granada;* [2]*University of Murcia;* [3]*University of Barcelona*

The analysis of differential item functioning (DIF) examines whether item responses differ according to characteristics such as language and ethnicity, when people with matching ability levels respond differently to the items. This analysis can be performed by calculating various statistics, one of the most important being the Mantel-Haenszel, which can be carried out with software programs such as EZDIF, DIFAS and, more recently, EASY-DIF. In this context, the aim of the present study is to compare these three software programs by using simulated and real data. The procedural characteristics and the results obtained from the same dataset were thus compared by the three programs. DIFAS and EASY-DIF always provide equivalent results, while EZDIF is less accurate when using the thin matching strategy. The results also showed that DIFAS and EASY-DIF were the easiest to run, especially for testing practitioners, with the second offering a broader range of results for key characteristics for detecting DIF.

The items of a test or questionnaire show differential item functioning (DIF) when subjects with the same ability level for the characteristics or attributes being measured, but who belong to different groups (demographic, linguistic, or cultural), have a different probability of giving a specific item response (Millsap & Everson, 1993). DIF is usually studied by comparing two groups of subjects: the reference group (generally the majority), and the focal group (generally a minority group). Between these two groups DIF may appear as uniform or non-uniform. In the former, there

* This study was partially funded by the Spanish Ministry of Science and Innovation under the European Regional Development Fund (Project nº PSI2009-07280); and the Andalusia Regional Government under the Excellent Research Fund (Project nº SEJ-5188). Correspondence: José Luis Padilla, Department of Social Psychology and Methodology, University of Granada, Campus de Cartuja. 18071 Granada, Spain. E-mail: jpadilla@ugr.es

is no interaction between the score level of the attribute being measured and membership of a given group, i.e. the probability of giving a certain response to the item is uniformly higher for one group than the other across all score levels of the attribute. However, in the case of non-uniform DIF there is an interaction, i.e. the probability of giving a certain response to the item in the two groups is not the same for all score levels of the attribute (Mellenbergh, 1982).

There is a wide variety of statistical techniques for evaluating DIF in both dichotomous and polytomous items (Hidalgo & Gómez-Benito, 2010; Millsap & Everson, 1993; Potenza & Dorans, 1995). Among these, the Mantel-Haenszel (MH) statistic is regarded as a reference technique due to its ease of use and the fact that it can be applied to small samples. These characteristics have meant that numerous studies in the applied field, such as those conducted by the *Educational Testing Service* (ETS), have used the MH statistic to detect DIF. Its utility in this field has also been the focus of much research (Guilera, Gómez-Benito & Hidalgo, 2009).

A number of specific software programs aimed at detecting DIF by means of the MH procedure are now available, specifically, EZDIF (Waller, 1998), DIFAS (Penfield, 2005) and EASY-DIF (González, Padilla, Hidalgo, Gómez-Benito & Benítez, 2011). In this context, the aim of the present study was to analyse the characteristics of each one of these programs, as well as their advantages and disadvantages. This was done by conducting a comparative analysis of a simulated dataset using the three programs: EASY-DIF, DIFAS and EZDIF. This comparison was based not only on the instrumental and procedural characteristics of each software package, but also on the results they provided following the analysis of common simulated and real datasets.

Mantel-Haenszel statistic

The MH statistical procedure (Mantel & Haenszel, 1959) consists of comparing the item performance of two groups (reference and focal), whose members were previously matched on the ability scale. The matching is done using the observed total test score as a criterion or matching variable (Holland & Thayer, 1988). The Mantel-Haenszel statistic is based on a contingency table analysis. For dichotomous items, K contingency tables (2×2) are constructed for each item, where K is the number of test score levels into which the matching variable has been divided. Table 1 shows the 2 x 2 table for calculating the MH statistic for item *i* on a *j* score level in the test.

Table 1. Score on i^{th} item in *j* score

Group	1	0	Total
Reference	A_j	B_j	$N_{R.j}$
Focal	C_j	D_j	$N_{F.j}$
Total	$N_{1.j}$	$N_{0.j}$	$N_{..j}$

In typical applications of the MH procedure an item shows uniform DIF if the odds of correctly answering the analysed item at a given score level *j* is different for the two groups at some level *j* of the matching variable. The odds ratio (α) is given by:

$$\alpha = (p_{Rj}/1- p_{Rj})/ (p_{Fj}/1- p_{Fj})$$

in which p_{Rj} and p_{Fj} are the correct item response probabilities for the reference group and focal group, respectively. The test score level *j* is calculated as follows:

$$p_{Fj} = \frac{C_j}{N_{Fj}} \text{ and } p_{Rj} = \frac{A_j}{N_{Rj}}.$$

The MH statistic for detecting DIF in an item is expressed as:

$$MH = \frac{\left[\left| \sum_{j=1}^{K} A_j - \sum_{j=1}^{K} E(A_j) \right| - 0.5 \right]^2}{\sum_{j=1}^{K} Var(A_j)}$$

in which $E(A_j) = (N_{Rj}N_{1.j})/N_{..j}$ *and* $Var(A_j) = (N_{Rj}N_{Fj}N_{1.j}N_{0.j})/(N_{..j})2(N_{..j} - 1)$. The MH statistic, under the null hypothesis, is distributed as a χ^2 distribution with

one degree of freedom. Under the MH procedure an effect size estimate based on the common odds ratio α is expressed as

$$\alpha_{MH} = \frac{\sum_{j=1}^{K} A_j D_j / N_{..j}}{\sum_{j=1}^{K} B_j C_j / N_{..j}}$$

Holland and Thayer (1988) proposed a logarithmic transformation of α for interpretive purposes, with the aim of obtaining a symmetrical scale in which a zero value indicates an absence of DIF, a negative value indicates that the item favours the reference group over the focal group, and a positive value indicates DIF in the opposite direction. This transformation is expressed as

$$\Delta\alpha_{MH} = -2.35 \ln (\alpha_{MH})$$

Based on this transformation, Zwick and Ercikan (1989) proposed the following interpretation guidelines to evaluate the DIF effect size:

- Type A items—negligible DIF: items with $\Delta\alpha_{MH} < |1|$.
- Type B items—moderate DIF: items with $|1| \leq \Delta\alpha_{MH} \leq |1.5|$, and the MH test statistically significant.
- Type C items—large DIF: items with $\Delta\alpha_{MH} > |1.5|$, and the MH test statistically significant.

Zwick and Ercikan (1989) pointed out that Type B items could be used in the test if there are no others to replace them, and that Type C items will be selected only if they are necessary to meet test specifications.

In polytomous items the data is organised in K two-dimensional 2 x *c* tables, where *c* is the number of response categories in the item. Table 2 shows the contingency table for item *i* with level *j*.

Mantel (1963) proposed a statistic which is an extension of the standard Mantel-Haenszel procedure. The Mantel statistic is computed by means of the following expression:

$$MANTEL = \frac{\left[\sum_{j=1}^{K} F_j - \sum_{j=1}^{K} E(F_j)\right]^2}{\sum_{j=1}^{K} Var(F_j)}$$

in which F_j is expressed as:

$$F_j = \sum_{c=1}^{C} R_c N_{Fcj}$$

or the total score for the focal group at K ability level.

Table 2. Scores on i^{th} item

Group	R_1	R_2	R_3		R_c	**Total**
Reference	N_{R1j}	N_{R2j}	N_{R3j}	…	N_{Rcj}	$N_{R.j}$
Focal	N_{F1j}	N_{F2j}	N_{F3j}	…	N_{Fcj}	$N_{F.j}$
Total	$N_{.1j}$	$N_{.2j}$	$N_{.3j}$	…	$N_{.cj}$	$N_{..j}$

Based on the general characteristics of the MH procedure, new statistics have also been developed, for example, the Breslow-Day chi-square (Breslow & Day, 1980) and new procedures for DIF detection such as the combined decision rule (Penfield, 2003).

Description of the datasets

The comparative study was firstly based on simulated data for 1000 participants' responses, 500 participants in the reference group and 500 participants in the focal group. In the simulation, a normal ability distribution with a mean equal to zero and standard deviation equal to one was performed. Narayanan y Swaminathan (1994) parameters were considered including 40 dichotomous items with 2-p (difficulty and discrimination parameters). DIF was manipulated in the first four items, always favouring the reference group. Items 1 and 2 were flagged with uniform DIF (both differing in the difficulty parameter); and items 3 and 4 were flagged with non-uniform DIF (differing in difficulty and discrimination parameters). The rest of the items were manipulated to be free of DIF.

Secondly, a real dataset was used to compare the three software programs. The dataset comes from the responses to the Spanish National Health Survey (Spanish Ministry of Health and Social Policies, 2006). A short version of the General Health Questionnaire (GHQ; Goldberg, 1972) was included in the survey questionnaire. The GHQ is intended to measure non-psychotic psychiatric disorders in community and occupational contexts. The short version of GHQ used in the present study consists of 12 multiple-choice items with four response categories that can be coded according to either a dichotomous or polytomous system.

For the purposes of the study a sample of 290 respondents from Morocco and a random sample of 300 Spanish respondents, were extracted from the survey database. In the DIF analysis the group from Spain was designated as the reference group and the Moroccan participants as the focal group. The participants' responses were coded dichotomously.

Comparison of the software programs

The analysis of characteristics and the comparison of the EZDIF, DIFAS and EASY-DIF programs took into account a number of aspects. The first of these concerned procedural parameters such as how the programs could be obtained (availability, material, etc.), data handling and the analyses possible in each case. Subsequently the results they provided were compared by analysing a common dataset.

Availability

EZDIF, DIFAS and EASY-DIF are currently the most used free software programs for evaluating DIF. EZDIF uses the MS-DOS operating system, while the other two programs are run in Windows. All three can be obtained by contacting the authors. The program installer for each package comes with a user manual.

Data input

Firstly, procedural aspects related with the preparation and input of datasets for each of the three programs are considered. The steps required prior to analysis are described in each case.

EZDIF: In order to run the program the user has to create an input file in text format (*.in). This file must include the following data:

- Title of project

- Model of analysis: Model 1 calculates the Mantel-Haenszel statistic, while model 2 calculates logistic regression statistics.

- Reference group: The user must specify the name and location of the dataset corresponding to this group, the number of subjects and the number of items, entering this description in FORTRAN format statement.

- Focal group: The user must specify the name and location of the data file and the number of subjects. This description is also entered in FORTRAN format statement.

- Output: Name and location of the output data file.

- Levels: Number of levels of the matching variable and specification of their limits.

- Labels: labels used to identify the items.

Once this document has been prepared the user must specify its location on the start screen, which then enables the program to run the analyses. Figure 1 shows an example of an input document in *.in.

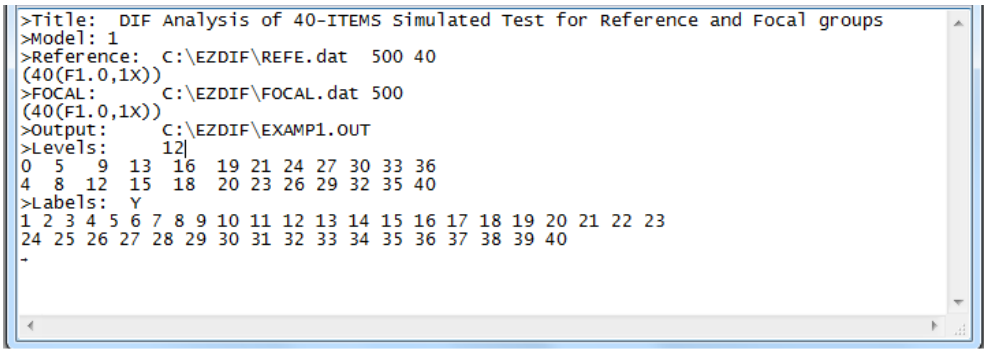


Figure 1. Example of the input document used in EZDIF

Once this file has been entered the program runs the analyses and creates an output file.

DIFAS: On the main window of the DIFAS program the user must specify a number of parameters. Firstly, it is necessary to determine the type of delimiter, i.e. whether the values for each variable should be separated by commas, a space or by tabs. The next step is to indicate the location of the data file containing the data of the two groups, i.e. the reference and focal groups. The file location is specified by means of a drop-down window that

enables the user to search the different folders of the computer system being used. The data file must be in text format (*.dat). Figure 2 shows the Main window of the DIFAS program.

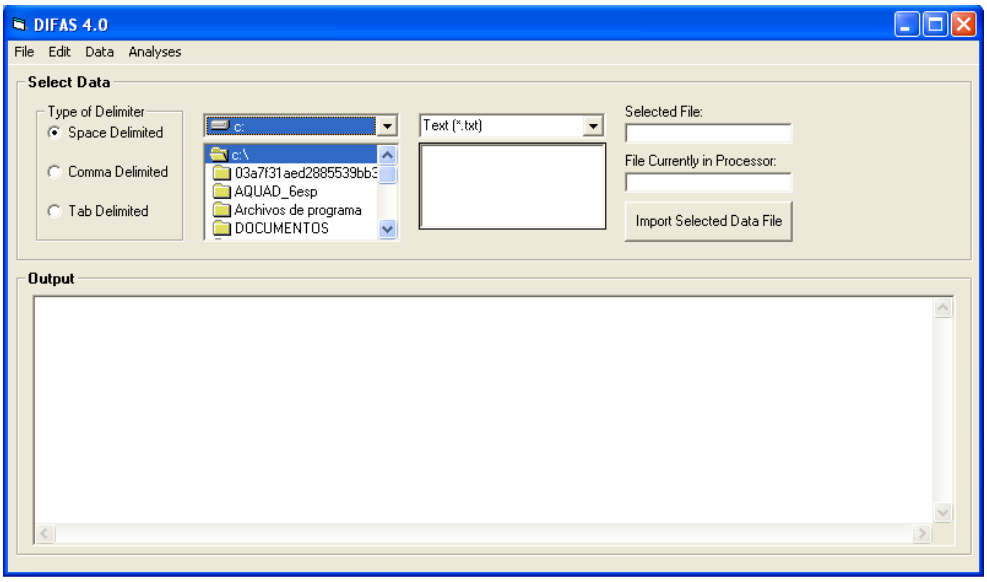


Figure 2. Main window of the DIFAS program

Once the file has been located the Output window shows the name of the file imported by the program, the number of subjects and the number of items, thus enabling the user to check that the data has been correctly interpreted by the program. Once the data has been entered the user must specify the type of analysis to be performed, which is done using the *Analyses* menu on the tool bar.

EASY-DIF: The Main window of the EASY-DIF program provides the user with direct access to the various options without having to use commands. However, it is first necessary to specify the characteristics of the data file (which must also be in text format *.dat but without any delimiter between the variables), i.e. from the outset the user must indicate the number of items and their location, the location of the grouping variable and the code used for each of the groups. Figure 3 shows the Main window of the EASY-DIF program.

In this window the user must also specify the format of the items, i.e. dichotomous or polytomous (in the latter case it is also necessary to indicate the number of response categories). Once the data has been entered the

program takes the user to another window where the analysis can be performed after selecting the item to be analysed.

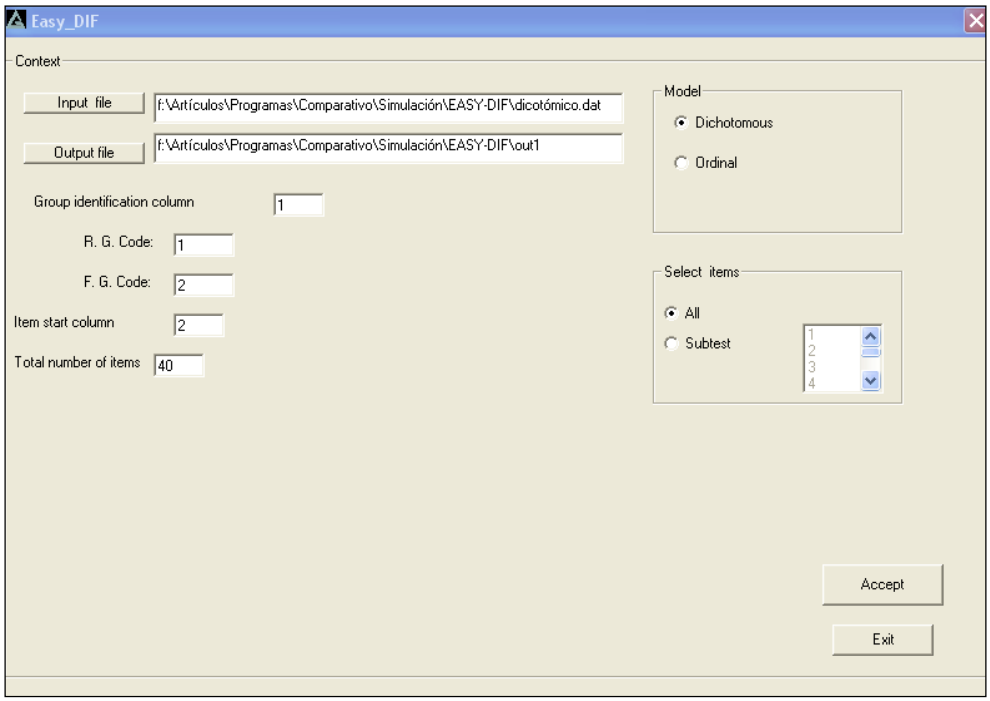


Figure 3. Main window of the EASY-DIF program

Specifications for the analysis

This section discusses those aspects that the user must know and determine in order to be able to execute the desired analyses. Some specific aspects of the sequence of actions are indicated, along with the decisions that the user must make to obtain the results.

EZDIF: This program requires users to have already determined the levels of the matching variable, which means that they must know the characteristics and distribution of the data. For example, if the aim is to establish two intervals based on the mean or the median, the user must first calculate this value using a tool outside the program. Should the user wish to establish intervals in the matching variable it is necessary to specify the number of intervals and the exact limits of each one of them.

DIFAS: This program requires a single dataset and, therefore, the user must specify the column containing the group code and the codes assigned

to each of the groups. After indicating the type of analysis to be performed (the next section discusses the types of analyses that can be done with each of the three programs) the user must select, in addition to the location and coding of the grouping variable, the items that will be included in the analysis, as well as what is known as the "stratifying" or matching variable. In this case the total test score or an external variable may be used as the matching variable. If the latter is chosen this external variable must be included in the dataset and its location must be specified at this point of the analysis. The user can also carry out a "thick matching" procedure, although the desired strata cannot be specified as these are set by the program, which divides the data into ten equal intervals.

EASY-DIF: After entering the dataset and specifying the characteristics, this program displays an analysis window, which shows the spaces that will contain the results once the user has specified the item to be analysed. In addition to the item the user must also specify the matching strategy to be used to calculate the Mantel-Haenszel statistic. The program allows the following data to be used as the matching variable: the total test score (thin matching), equal intervals, specific percentages of the total sample, the focal group or the reference group, outliers or a specific number of observations. Should the user require the matching strategy to be based on a percentage or a frequency, it is necessary to stipulate the specific percentage or frequency in the corresponding box. The program is also able to calculate the standardization statistic (Dorans & Holland, 1993). In this process the user must specify the standardization parameter. The program offers the possibility of using the complete sample, the total of the focal group or the total of the reference group.

Analysis

EZDIF: This program analyses DIF by means of the Mantel-Haenszel statistic and logistic regression (Clauser, Nungester, Mazor & Ripkey, 1996), and provides various statistical indices for each. EZDIF carries out a two-step purification of the matching variable, i.e. it analyses DIF in two steps; in the second it eliminates those items that showed DIF in the first. It also analyses non-uniform DIF and provides data that enables a visual inspection by means of empirical item characteristic curves.

DIFAS: This program analyses item characteristics and provides information about descriptive statistics (mean, standard deviation, minimum and maximum) and the frequencies of choice for each category of each item. It can also analyse DIF and differential test functioning for both

dichotomous and polytomous items. Non-uniform DIF is analysed according to the description of the empirical item characteristic curves.

EASY-DIF: This program analyses uniform DIF by means of the Mantel-Haenszel statistic and non-uniform DIF using the modified Mantel-Haenszel procedure (Clauser, Nungester, Mazor & Ripkey, 1996). It also includes the statistics used for standardization procedures (Dorans & Holland, 1993).

Statistical indices

EZDIF: This program displays the results of the DIF analysis and the empirical item characteristic curves. The first part includes two tables corresponding to the two steps of purification. The first table shows the results of the DIF analysis for all the items, while the second presents the same analyses but performed after eliminating the items that were labelled as showing DIF in the previous table. The tables include the following statistics: alpha, chi-square, probability of chi-square, MH D-DIF, and standard error of MH D-DIF. In the second part the program displays, for each item, the probability of a correct response by members of the reference and focal groups on each of the levels established for the matching variable.

DIFAS: The results offered by this program are displayed in two tables. The first of these shows the DIF statistics, while the second presents the conditional differences in the mean item scores between the reference and focal groups at ten intervals across the matching variable continuum. In the DIF analysis the program includes the following statistics: the Mantel-Haenszel chi-square statistic, the Mantel-Haenszel common log-odds ratio, the standard error of the Mantel-Haenszel common log-odds ratio, the Mantel-Haenszel log-odds ratio divided by the estimated standard error, the Breslow-Day chi-square test of trend in odds ratio heterogeneity, the combined decision rule and the ETS categorisation scheme.

EASY-DIF: This program offers different kinds of results. Firstly, it displays the frequency of choice of each category for both the reference and focal groups on each of the matching levels established for each item. It then gives the test mean and the Mantel-Haenszel results for the whole group and for the low- and high-performance groups. In each case it displays the following statistics: chi-square and its probability, the alpha value, the delta value and the delta error. As regards standardization the program provides the following indices: SPD, delta, the standard error, the mean of the reference and focal groups, and the total mean.

Output

All the programs produce a text file containing the results of the analysis, and in order to obtain this the user must specify the name of the output data file. Below we show the results screen obtained when analysing the simulated data with each of the three programs. Firstly, Figure 4 shows the output screen produced by the EZDIF program.

Mantel-Haenszel and Logistic Regression Analysis of
DIFFERENTIAL ITEM FUNCTIONING

Programmed by Niels G. Waller

	ITEM	Alpha	X^2	P-Value	MH D-DIF	SE (MH D-DIF)
A	1**	1.471	7.114	0.008	-0.906	0.332
A	2	1.151	1.001	0.317	-0.330	0.311
B	3**	1.607	6.906	0.009	-1.115	0.413
A	4	0.849	1.145	0.285	0.385	0.337
A	5	1.228	1.416	0.234	-0.483	0.381
A	6	1.015	0.001	0.976	-0.034	0.337
A	7	1.157	0.926	0.336	-0.342	0.332
A	8*	0.676	5.849	0.016	0.919	0.369
A	9	0.788	1.884	0.170	0.560	0.387
A	10	0.990	0.000	0.989	0.025	0.310
A	11	0.932	0.213	0.645	0.165	0.313
A	12	1.075	0.227	0.633	-0.170	0.312
A	13	1.055	0.096	0.757	-0.125	0.329
A	14	0.965	0.043	0.836	0.084	0.309
A	15	1.108	0.460	0.498	-0.241	0.324
A	16	1.063	0.107	0.744	-0.144	0.358
A	17	0.865	0.845	0.358	0.341	0.345
A	18	0.968	0.019	0.891	0.076	0.356
A	19	1.226	1.995	0.158	-0.479	0.323
A	20	0.799	1.055	0.304	0.527	0.469
A	21	0.973	0.010	0.922	0.063	0.361
A	22	1.097	0.313	0.576	-0.217	0.342
A	23	1.031	0.017	0.895	-0.072	0.351
A	24	1.079	0.237	0.627	-0.179	0.322
A	25	1.018	0.003	0.959	-0.041	0.336
A	26	0.834	1.261	0.261	0.426	0.356
A	27	1.076	0.170	0.680	-0.172	0.351
A	28	1.052	0.069	0.793	-0.119	0.353
A	29	0.915	0.351	0.554	0.210	0.318
A	30*	0.736	3.978	0.046	0.722	0.349
A	31	1.177	0.908	0.341	-0.382	0.370
A	32	0.996	0.002	0.965	0.008	0.327
A	33	1.048	0.071	0.790	-0.111	0.329
A	34	0.890	0.638	0.424	0.275	0.317
A	35	0.847	0.944	0.331	0.390	0.371
A	36	0.858	0.932	0.334	0.361	0.347
A	37	0.813	1.637	0.201	0.486	0.359
A	38	1.199	0.609	0.435	-0.426	0.482
A	39	0.873	0.734	0.391	0.320	0.344
A	40	0.995	0.001	0.970	0.011	0.334

Figure 4. Output screen of EZDIF

The output produced by EZDIF is displayed in two tables. As shown in Figure 4 the first of these includes the statistics obtained for each of the items. The next table shows the second step of purification, which includes the same statistics but calculated after eliminating those items for which DIF was detected in step one. The next figure (Figure 5) shows the output of the DIFAS program.

DIF analysis: Nonparametric tests for dichotomous items

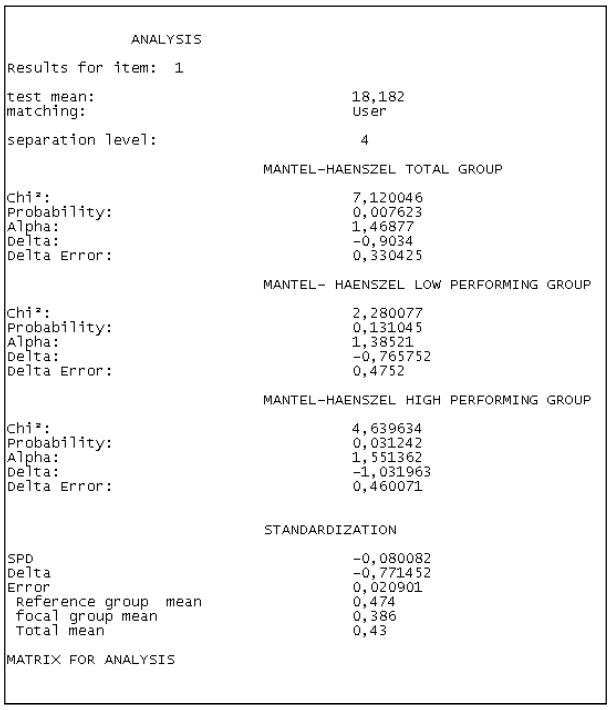
DIF STATISTICS: DICHOTOMOUS ITEMS

Name	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
Var 1	6,8814	0,38	0,1414	2,6874	1,288	Flag	A
Var 2	1,1609	0,1494	0,1311	1,1396	0,157	OK	A
Var 3	6,6009	0,459	0,1733	2,6486	0,234	Flag	A
Var 4	1,1731	-0,1663	0,1441	-1,1541	4,499	OK	A
Var 5	1,1626	0,1875	0,1618	1,1588	0,618	OK	A
Var 6	0,0009	-0,0144	0,1427	-0,1009	1,231	OK	A
Var 7	0,9441	0,1449	0,1398	1,0365	1,659	OK	A
Var 8	6,8178	-0,4194	0,156	-2,6885	0,006	Flag	A
Var 9	2,4344	-0,2691	0,1647	-1,6339	1,919	OK	A
Var 10	0,0005	0,0058	0,1312	0,0442	0,001	OK	A
Var 11	0,2032	-0,0685	0,1324	-0,5174	2,174	OK	A
Var 12	0,2603	0,076	0,1317	0,5771	1,436	OK	A
Var 13	0,0809	0,0497	0,1402	0,3545	1,264	OK	A
Var 14	0,0255	-0,0295	0,1312	-0,2248	1,041	OK	A
Var 15	0,42	0,0985	0,1376	0,7158	0,166	OK	A
Var 16	0,1715	0,0737	0,1508	0,4887	0,378	OK	A
Var 17	1,0298	-0,1589	0,1465	-1,0846	0,005	OK	A
Var 18	0,0566	-0,0475	0,1513	-0,3139	0,046	OK	A
Var 19	1,6899	0,1876	0,1367	1,3723	0,019	OK	A
Var 20	1,7565	-0,2845	0,1996	-1,4254	0,599	OK	A
Var 21	0,0112	-0,028	0,1531	-0,1829	0,011	OK	A
Var 22	0,2427	0,0828	0,1459	0,5675	0,763	OK	A
Var 23	0,0006	-0,0148	0,1492	-0,0992	0,321	OK	A
Var 24	0,3128	0,0857	0,1364	0,6283	0,001	OK	A
Var 25	0,0036	-0,0187	0,1423	-0,1314	1,815	OK	A
Var 26	1,8145	-0,2163	0,1516	-1,4268	0,55	OK	A
Var 27	0,0325	0,0378	0,1482	0,2551	0,348	OK	A
Var 28	0,0844	0,0549	0,1501	0,3658	0,182	OK	A
Var 39	0,4315	-0,098	0,1352	-0,7249	0,737	OK	A
Var 30	3,6635	-0,293	0,1475	-1,9864	0,001	OK	A
Var 31	0,743	0,1478	0,1571	0,9408	0,212	OK	A
Var 32	0,0019	-0,0036	0,1384	-0,026	0,011	OK	A
Var 33	0,1272	0,0589	0,1387	0,4247	0,265	OK	A
Var 34	1,0346	-0,1459	0,1347	-1,0831	0,651	OK	A
Var 35	1,3032	-0,192	0,1575	-1,219	0,634	OK	A
Var 36	1,1991	-0,1732	0,1478	-1,1719	0	OK	A
Var 37	1,6255	-0,206	0,1525	-1,3508	0,151	OK	A
Var 38	0,4085	0,1501	0,202	0,7431	2,224	OK	A
Var 39	1,2	-0,1721	0,1469	-1,1715	0,579	OK	A
Var 40	0,0035	-0,0017	0,1424	-0,0119	0,25	OK	A

Figure 5. Output screen of DIFAS

Here the output screen again displays two tables. The first (figure 5) includes the statistics calculated for the items. The MH CHI statistic refers to the Mantel-Haenszel chi-square, on the basis of which the remaining statistics are calculated. MH LOR is equivalent to the natural logarithm of alpha, LOR SE is the square root of the variance and LOR Z is the result of dividing the MH LOR by the LOR SE. The program also gives the Breslow-Day statistic and two combined decision rules which serve as criteria for classifying the DIF: the first is based on the significance of the Mantel-Haenszel chi-square and the Breslow-Day chi-square statistics (using a confidence level of 0.025), while the second uses the traditional criteria applied by the ETS (Educational Testing Service) to classify DIF (A= small, B=moderate and C= large).

The second table includes the conditional differences or the difference between the difficulty indices found for each of the groups, thus enabling the type of DIF present to be identified. The DIF is uniform if the same direction (positive or negative) is maintained across all the intervals, whereas the DIF is non-uniform when the direction changes, as this indicates that the benefit switches between the reference and focal groups. Finally, Figure 6 shows the output data file of the EASY-DIF program.



```
              ANALYSIS

Results for item:  1

test mean:                              18,182
matching:                               User

separation level:                        4

                                MANTEL-HAENSZEL TOTAL GROUP

Chi²:                                   7,120046
Probability:                            0,007623
Alpha:                                  1,46877
Delta:                                  -0,9034
Delta Error:                            0,330425

                                MANTEL- HAENSZEL LOW PERFORMING GROUP

Chi²:                                   2,280077
Probability:                            0,131045
Alpha:                                  1,38521
Delta:                                  -0,765752
Delta Error:                            0,4752

                                MANTEL-HAENSZEL HIGH PERFORMING GROUP

Chi²:                                   4,639634
Probability:                            0,031242
Alpha:                                  1,551362
Delta:                                  -1,031963
Delta Error:                            0,460071


                                STANDARDIZATION

SPD                                     -0,080082
Delta                                   -0,771452
Error                                   0,020901
 Reference group  mean                  0,474
 focal group mean                       0,386
 Total mean                             0,43

MATRIX FOR ANALYSIS
```

Figure 6. Output of EASY-DIF

As can be seen in Figure 6, EASY-DIF displays the results obtained for each of the items separately. In this case it includes the statistics for the whole group and for the low- and high-performance groups. The chi-square statistics obtained in the low- and high-performance groups enable the user to determine whether the DIF is uniform or non-uniform. The DIF is uniform when the probability of chi-square is statistically significant in both cases. If this is not the case, i.e. the probability of chi-square is only significant for one of the groups, the DIF detected is non-uniform. The results obtained by means of standardization procedures are also included.

The program enables the user to view the results simultaneously, as they are displayed in a small window within the analysis window. Figure 7 shows the analysis window with the results obtained. The graph shows the probability of giving a right answer in each level in which participants from both groups have been matched.

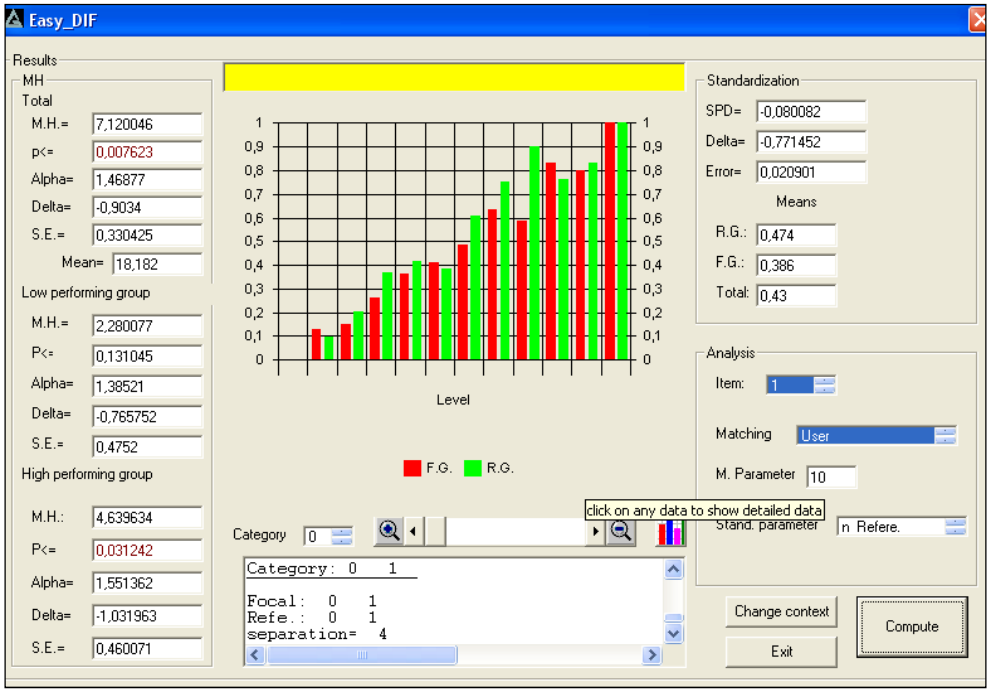


Figure 7. Analysis window of EASY-DIF

EASY-DIF also has a graph option in which the user can choose between a line and a bar graph. This enables detailed observation by enlarging the graphs and their content.

Comparison of results with simulated data

This section compares the results obtained with the three programs for the same set of simulated data. This comparison was made on two levels depending on the type of matching strategy used. The first approach involved a thin matching strategy in which the participants were divided across forty-one intervals, each one of which corresponded to one of the

possible scores on the test (0-40). Secondly, a thick matching strategy was applied, in which participants were grouped into twelve equals intervals (0-4, 5-8, 9-12, etc.).

In the first level of comparison the same results were obtained with DIFAS and EASY-DIF, whereas EZDIF yielded different results. Table 3 shows the values of the statistics that are common to the three programs for item 1 of the simulated data.

Table 3. Results obtained for item 1 with the thin matching strategy.

EASY-DIF		DIFAS		EZDIF	
Chi-square	Alpha	Chi-square	MH LOR	Chi-square	Alpha
7.7042	1.498	7.7042	0.4045	5.892	1.452

It can be seen that although the chi-square values were the same in EASY-DIF and DIFAS, it was necessary to calculate the MH LOR value in order to match the alpha values provided by the two programs. The MH LOR is equivalent to the Naperian logarithm of alpha, and by applying it to the value given by EASY-DIF (1.498) it was possible to obtain the value obtained with DIFAS (0.40). The different results obtained with EZDIF are probably due to a difficulty which arose with the program, whereby it stated "insufficient data found" for 22 of the 41 levels.

On the second level of comparison twelve equal intervals (0-4, 5-8, 9-12, etc.) were established in order to match the participants. In EASY-DIF this was done by selecting the option "User" and specifying the limits of the intervals; in EZDIF the interval limits were entered in the "Input" document, while in DIFAS an external variable was entered that divided the subjects into twelve groups. Table 4 shows the results obtained with the three programs for item 1.

Table 4. Results obtained for item 1 with the thick matching strategy.

EASY-DIF		DIFAS		EZDIF	
Chi-square	Alpha	Chi-square	MH LOR	Chi-square	Alpha
7.1200	1.469	6.8814	0.38	7.114	1.471

It can be seen that similar results were obtained by each of the programs. Once again, the value of MH LOR coincides with the natural logarithm of alpha given by EASY-DIF and EZDIF.

The comparison of results showed an equivalent performance between the three programs in the case of thick matching. However, with the thin matching strategy only DIFAS and EASY-DIF produced equivalent results, despite this being the most recommended option.

Comparison of results with real data

This section presents the results of the comparison between the three programs. The comparison was twofold using different matching strategies. The first approach involved a thin matching strategy in which the participants were divided across thirteen intervals, each one of which corresponded to one of the possible scores on the GHQ (0-12). Secondly, a thick matching strategy was applied, in which participants were grouped into two intervals (0-5 and 6-12), this being the default thick matching strategy implemented in EZDIF.

For the thin matching strategy, the same results were obtained with DIFAS and EASY-DIF, whereas EZDIF yielded different results. To illustrate the results obtained, Table 5 shows the values of the statistics that are common to the three programs for item 1.

Table 5. Results obtained for item 1 with the thin matching strategy.

EASY-DIF		DIFAS		EZDIF	
Chi-square	Alpha	Chi-square	MH LOR	Chi-square	Alpha
3.240204	1.833448	3.2402	0.6007	4.839	2.315

When computing MH LOR for the value given by EASY-DIF (1.83) it is possible to obtain the value obtained with DIFAS (0.60). The differences in the results obtained with EZDIF are probably due to the different criteria the program uses for including a matching category in the DIF analyse.

For the thick matching strategy, two equal intervals (0-5 and 6-12) were established in order to match the respondents. Table 6 shows the results obtained with the three programs for item 1.

Table 6. Results obtained for item 1 with the thick matching strategy.

EASY-DIF		DIFAS		EZDIF	
Chi-square	Alpha	Chi-square	MH LOR	Chi-square	Alpha
1.191292	1.381274	1.1913	0.323	1.191	1.381

It can be seen that the same results were obtained by each of the programs, with EZDIF and EASY-DIF being the least and most accurate, respectively. Once again, the value of MH LOR coincides with the natural logarithm of alpha given by EASY-DIF and EZDIF.

The comparison of results with real data also showed an equivalent performance between the three programs in the case of thick matching. However, with the thin matching strategy only DIFAS and EASY-DIF produced equivalent results.

Conclusions

The aim of this study was, firstly, to examine the characteristics of three available software programs (EASY-DIF, DIFAS and EZDIF) for analysing DIF by means of the Mantel-Haenszel statistic. A second objective was to compare the procedural aspects and the results obtained with each program for a common set of data so as to be able to make recommendations to potential users.

As regards data entry, that the DIFAS and EASY-DIF are running in Windows insures user friendliness, whereas EZDIF requires the MS-DOS operating system, thus of limited use for many people. More specifically, with DIFAS and EASY-DIF the user only has to locate the data file and specify the characteristics of the data, whereas in order to run EZDIF it is necessary to create a command file through which the program accesses the data.

With respect to the specifications for the analysis, the most relevant aspect of the comparison was the determination of the matching variable. The most advantageous program in this regard was EASY-DIF, since it enables up to six different matching strategies when the total score is taken as the variable. The program can also perform thick or thin matching depending on the user's needs, i.e. it is possible to establish a specific percentage or frequency of participants for each of the variable intervals. DIFAS enables an external variable or the total test score to be used as the matching variable, but users cannot establish specific intervals since the program automatically divides participants into ten equal intervals on the

basis of total scores. Finally, EZDIF is the program that possesses the most difficulties in this aspect since it requires specification of the number of intervals and their limits, which means that the user must have detailed knowledge of the data characteristics.

In terms of analysis the programs differ in a number of aspects. The first of these is the type of item that can be analysed: DIFAS and EASY-DIF can analyse DIF in both dichotomous and polytomous items, whereas EZDIF can only be applied to dichotomous items. However, one advantage of EZDIF is that it applies the purification procedure directly, i.e. it establishes two steps, eliminating in the second step those items that were found to show DIF in the first step. With both DIFAS and EASY-DIF the user must repeat the analysis and manually eliminate those items identified as showing DIF. As regards the detection of non-uniform DIF the clearest results are provided by EASY-DIF, since it yields results for the modified Mantel-Haenszel statistic as well as presenting the output data in the form of a graph. In contrast, DIFAS and EZDIF only provide the numerical data required to draw the empirical item characteristic curves. Finally, it should be noted that, among other characteristics, both DIFAS and EASY-DIF provide descriptive statistics for the data, which enables the user to observe the frequency of each category.

With respect to the statistical indices provided by each program, EZDIF and EASY-DIF give the results for chi-square, probability and alpha, as well as other statistics. DIFAS gives the chi-square values but includes other statistics based on transformations of alpha and the standard error. This could make interpretation more difficult for users with limited knowledge of these indices.

As regards output, DIFAS and EZDIF show a single table including all the results for all the items, whereas EASY-DIF presents a separate table for each of the items. Whether or not this is an advantage will depend on the user's objectives. One of the most important advantages of EASY-DIF over the other two programs is the possibility of obtaining a graphical representation of the results, as well as being able to view them instantly as the analysis progresses. This means that an applied researcher with little training would be able to interpret the results easily.

Finally, the comparison of the results from the three programs showed more similarities when the results were obtained via thick matching for simulated and real data, but more differences when a thin matching strategy was used. In the latter case, EZDIF had problems with calculating the statistics due to insufficient data on some of the levels. In the case of DIFAS the greatest difficulty appeared when conducting the thick matching

strategy, as the program requires the user to include a variable in the dataset that divides the subjects into intervals according to their scores and then use the option "stratify by external". To this end it is necessary to calculate the total scores of the subjects, determine the intervals to be established and assign the subjects to different groups on the basis of their total score. Therefore, it is important to note the ease with which both types of matching can be carried out with EASY-DIF, whereas both the other programs present certain difficulties. In relation to items detected when using simulated data, when thick matching was applied, items 2 and 4 were not flagged by any of the software, while other items, such as 8 and 30 (free of DIF), were flagged with DIF by EZDIF. Item 8 was also flagged by DIFAS. Finally, the DIFAS and ESAY-DIF programs provided quite similar results for the real dataset. The different results provided by EZDIF can be attributed to how the program implemented the thin matching strategy.

The characteristics of each software program have been shown in order to help the researcher choose depending on their interests.

RESUMEN

Comparación de tres programas para la evaluación del DIF mediante el procedimiento Mantel-Haenszel: EASY-DIF, DIFAS y EZDIF. El análisis del DIF (Funcionamiento Diferencial de los Ítems) examina si las respuestas a un ítem difieren en función de características como el idioma o el grupo étnico, entre personas igualadas respecto de la habilidad medida por el test. El análisis del DIF puede realizarse a partir de diferentes técnicas estadística, siendo el estadístico Mantel Haenszel uno de los más relevantes. El estadístico Mantel-Haenszel puede calcularse mediante programas como el EZDIF, el DIFAS y recientemente el EASY-DIF. En este contexto, el objetivo de este trabajo es comparar estos tres programas informáticos usando datos simulados y reales. Se analizaron las características procedimentales e instruccionales de los tres programas y se compararon los resultados proporcionados por cada uno de ellos para el mismo conjunto de datos. DIFAS y EASY-DIF siempre proporcionan resultados equivalentes, mientras que el EZDIF es menos preciso cuando se utiliza una estrategia de igualación delgada. Los resultados también mostraron una mayor facilidad en la ejecución de los programas DIFAS y EASY-DIF, especialmente para profesionales de la evaluación con tests en contextos aplicados, ofreciendo el segundo un mayor rango de resultados para características claves en la detección del DIF.

REFERENCES

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: Volume 1- The analysis of case-control studies*. Lyon: International Agency for Research on Cancer.

Clauser, B.E., Nungester, R.J., Mazor, K. & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33,* 202-214.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and Standarization. En P.W.Holland y H.Wainer (Eds.) *Differential Item Functioning* (pp. 35-66) Hillsdale, NJ: Erlbaum.

Goldberg, D. (1972). *The Detection of Psychiatric Illnes by Questionnaire*. Windsor. National Foundation for Educational Research.

González, A.; Padilla, J.L.; Hidalgo, M.D. Gómez-Benito, J. & Benítez, I. (2011). EASY-DIF: Software for analysing differential item functioning using the Mantel-Haenszel and standardization procedures. *Applied Psychological Measurement, 35,* 483-484.

Guilera, G.; Gómez-Benito, J. & Hidalgo, M.D. (2009). Scientific production on the Mantel-Hanszel procedure as a way of detecting DIF. *Psicothema, 21 (3),* 492-498.

Hidalgo, M. D., & Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (3rd edition)*. USA: Elsevier - Science & Technology.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Spanish Ministry of Health and Social Policies. National Health Survey 2006: <**http://www.msps.es/estadEstudios/estadisticas/encuestaNacional/encuesta2006.htm**> [Check: March 3, 2011].

Mantel, N. (1963). Chi-square tests with one degree of freedom, extension of the Mantel-Haenszel procedure. *American Statistical Association Journal, 58,* 690-700.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics 7***,** 105-118.

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement 17***,** 297-334.

Narayanan, P. y Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18(3)* 15-328.

Penfield, R. D. (2003). Application of the Breslow-Day test of trend in odds ratio heterogeneity to the detection of nonuniform DIF. *Alberta Journal of Educational Research*, *49*, 231-243.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*, 150-151.

Potenza, M. T. & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23-37.

Waller, N. G. (1998). EZDIF: Detection of Uniform and Nonuniform Differential Item Functioning With the Mantel-Haenszel and Logistic Regression Procedures. *Applied Psychological Measurement, 22,* 391.

Zwick, R. and Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26,* 55-66.

(Manuscript received: 24 January 2011; accepted: 23 March 2011)