The Spanish Journal of Psychology
2011, Vol. 14, No. 2, 1018-1022
http://dx.doi.org/10.5209/rev_SJOP.2011.v14.n2.47

Copyright 2011 by The Spanish Journal of Psychology
ISSN 1138-7416

A New Approach for Differential Item Functioning Detection Using Mantel-Haenszel Methods. The GMHDIF Program

Ángel M. Fidalgo

Universidad de Oviedo (Spain)

To date, the statistical software designed for assessing differential item functioning (DIF) with Mantel-Haenszel procedures has employed the following statistics: the Mantel-Haenszel chi-square statistic, the generalized Mantel-Haenszel test and the Mantel test. These statistics permit detecting DIF in dichotomous and polytomous items, although they limit the analysis to two groups. On the contrary, this article describes a new approach (and the related software) that, using the generalized Mantel-Haenszel statistic proposed by Landis, Heyman, and Koch (1978), permits DIF assessment in multiple groups, both for dichotomous and polytomous items. The program is free of charge and is available in the following languages: Spanish, English and Portuguese.
Keywords: differential item functioning, DIF software, GMHDIF, Mantel-Haenszel methods.

Hasta la fecha, los programas informáticos diseñados específicamente para detectar el funcionamiento diferencial de los ítems (DIF) usando los métodos Mantel-Haenszel han empleado los siguientes estadísticos: el estadístico ji-cuadrado de Mantel-Haenszel, el test generalizado de Mantel-Haenszel y el test de Mantel. El presente artículo describe una nueva aproximación (y el software correspondiente) que permite evaluar el DIF en múltiples grupos, tanto en ítems dicotómicos como politómicos, empleando el estadístico generalizado de Mantel-Haenszel propuesto por Landis, Heyman, and Koch (1978). El programa es gratuito y está disponible en español, inglés y portugués.
Palabras clave: funcionamiento diferencial de los ítems, DIF, GMHDIF, métodos Mantel-Haenszel.

This work was supported by the Ministerio Español de Educación y Ciencia (grant numbers PR2006-0424, SEJ2006-07491, PCI2006-A7-0553).

Correspondence concerning this article should be addressed to Ángel M. Fidalgo. Departamento de Psicología, Universidad de Oviedo, Plaza de Feijoo s/n, 33003 Oviedo (Spain). E-mail: fidalgo@uniovi.es

https://doi.org/10.5209/rev_SJOP.2011.v14.n2.47 Published online by Cambridge University Press

One basic requirement of the items making up tests for assessing knowledge, or any psychological test or scale, is that the measures obtained with them depend solely on respondents' level in the variable measured and not on irrelevant characteristics. When this is not the case (for equal levels of the variable measured, the probability of answering an item correctly differs between subgroups of a given population) we speak of differential item functioning (DIF). For example, suppose we have a test intended to measure spatial ability but having items that, because they are written in a complicated manner, also assess linguistic ability. Moreover, suppose this test is applied in schools having children who are native speakers and also children who are immigrants and are still learning the language. It is easy to see how the latter will obtain worse scores on the aforementioned items even if they have the same spatial ability as the native children. Statistical techniques for evaluating DIF allow us to compare any groups defined by the evaluator (for example: native/non-native population, men/women, African American/white/Hispanic) in order to guarantee that the members of the compared groups have the same probability of answering the items correctly if they present the variable under study to the same degree (Fidalgo, 1996). As applied researchers, is important to ensure that our psychological tests are free of DIF, given that DIF threatens the validity of our test results and may be a clear source of bias in the social sense. The purpose of this article is twofold: to inform applied researchers of a new theoretical framework for detecting DIF using Mantel-Haenszel (MH) methods, and to present the statistical software that uses it. The simulation studies and the software development were supported by public Spanish funds. We believe researchers are responsible for disseminating their work not only to the readers of journals focused on psychological measurement and statistics, but to all researchers who may benefit from the work.

The Mantel-Haenszel methods constitute one of the most popular DIF detection procedures. The most common MH statistics used for DIF detection are: the MH chi-square statistic (χ^2_{MH}; Holland & Thayer, 1988; Mantel & Haenszel, 1959), the generalized Mantel-Haenszel test (GMH, Mantel & Haenszel, 1959; Zwick, Donoghue, & Grima, 1993) and the Mantel test (MT; Mantel, 1963; Zwick et al., 1993). These statistics permit detecting DIF in dichotomous (χ^2_{MH}) and polytomous items (GMH and MT), although they limit the analysis to two groups. Recently Fidalgo and Madeira (2008) have shown a unified framework for the analysis of DIF using the generalized Mantel-Haenszel statistic proposed by Landis, Heyman and Koch (1978). As is pointed out there, this statistic subsumes the GMH and the Mantel Test, as well as the χ^2_{MH} statistic. Therefore, we can apply this generalized statistic to DIF assessment in multiple groups, both for dichotomous items and polytomous items (Fidalgo & Scalon, 2010). This provides a very important advantage given that, regardless of the number of groups compared, with a single test we can determine if an item is free of DIF, thus avoiding the spiraling Type I error rate resulting from making multiple comparisons. Unlike other statistics for analyzing DIF in multiple groups based on item response theory (IRT; Kim, Cohen, & Park, 1995), the proposed statistics may be applied both within IRT (Fidalgo, 2005b) and also in tests developed and analyzed from the perspective of classical test theory.

Going beyond the abstract of the GMHDIF program that has been published in *Applied Psychological Measurement* (Fidalgo, 2011), the chief objective of this work is to present a new framework for detecting DIF using Mantel-Haenszel methods, and to describe in a detailed way that program to applied researchers. First of all we shall describe the generalized Mantel-Haenszel statistics.

Generalized Mantel Haenszel Statistics

In 1978 Landis et al. proposed a generalized MH statistic for the analysis of Q: R x C contingency tables. Where Q are the number of strata or levels of the covariable (in DIF analysis the covariable, or matching variable, is usually the total test score), R corresponds to the levels of the factor (the number of groups) and C corresponds to the levels of the response variable (the number of item categories). The standard generalized Mantel-Haenszel tests the null hypothesis (H_0) of non-association between the factor and the response variable, controlling the effect of the covariable. The H_0 of no-association will be tested against different H_1 that will be a function of the scale on which factor and response are measured. Thus, we shall have a variety of statistics that will serve for detecting the general association (both variables are nominal; the *Generalized Nominal MH statistic* or $Q_{GMH(1)}$), mean score differences (factor is nominal and response ordinal; *the Generalized Ordinal MH statistic or* $Q_{GMH(2)}$), and linear correlation (both variables are ordinal; *the Generalized Correlation MH statistic or* $Q_{GMH(3)}$).

In the case of $Q_{GMH(1)}$ the alternative hypothesis (H_1) specifies that the distribution of the response variable differs in nonspecific patterns across levels of the row factor (groups). On the other hand, $Q_{GMH(2)}$, on considering the ordinal nature of the response variable, specifies in H_1 that the mean responses differ across the grouping variable. This requires assigning numbers, called scores, to the response categories that reflect their ordinal nature. One of the most common options is to assign to the response categories equally-spaced scores, applying, for example, successive integers, although other score systems are possible (Fidalgo & Bartram, 2010).

It should be noted the following equivalences between the MH statistics usually employed in the DIF literature and the $Q_{GMH(1)}$ and $Q_{GMH(2)}$ statistics: (a) Generalized Mantel-Haenszel test: for the special case of 2 groups,

https://doi.org/10.5209/rev_SJOP.2011.v14.n2.47 Published online by Cambridge University Press

$Q_{GMH(1)}$ is identical to the GMH proposed by Mantel and Haenszel (1959); (b) Mantel test: for the special case of 2 groups, $Q_{GMH(2)}$ is identical to the extended MH test proposed by Mantel (1963); and, (c) Mantel-Haenszel chi-square: when we have a dichotomously scored item and 2 groups, $Q_{GMH(1)} = Q_{GMH(2)} = \chi^2_{MH}$ (for this equivalence to be fulfilled must be calculated without the continuity correction it normally includes).

The interested reader can find more comprehensive information on these statistics in Fidalgo (2005a) and Fidalgo and Madeira (2008). Moreover, a detailed description of how the generalized *MH* statistics could be used for assessing DIF simultaneously in multiple groups and in polytomous items can be found in Fidalgo and Scalon (2010) and Fidalgo, Quintanilla, Fernández, Pons, and Aguerri (2010), respectively.

GMHDIF

GMHDIF has been developed to provide an easy-to-use program for conducting DIF analyses and runs in any Windows operating system. The main characteristics of the program are:

1. It can be used for DIF detection in dichotomously scored items, nominal-polytomous items and ordinal-polytomous items. To this end the program implements the Generalized Nominal MH statistic ($Q_{GMH(1)}$) and the Generalized Ordinal MH statistic ($Q_{GMH(2)}$). To compute $Q_{GMH(2)}$ sucesive integer scores are used to reflect the ordinal nature of the item categories.

2. It performs DIF analyses in multiple groups simultaneously. The generalized MH statistics computed by the program test the omnibus null hypothesis of no difference among all the groups. This means that, regardless of the number of groups compared, with a single test we can determine whether an item is free of DIF.

3. It performs pairwaise comparisons using a Bonferroni-adjusted alpha level (alpha level/no. of pairwise comparisons), when in a multiple groups DIF analyses the omnibus null hypothesis of no difference among all the groups is rejected. Moreover, it is possible to compare focal groups with a reference group or to compare all the groups with one another.

4. It performs two-stage DIF analyses. You can opt to perform a two-stage analysis. If you choose this option, the selected MH statistic will be applied in two stages. The items whose values were statistically significant at the significance level used in the first stage will be removed for calculating the matching criteria in the second stage. As suggested by Zwick et al. (1993), when the studied item is used to compute the matching variable, it will be always included in the matching score. Moreover, GMHDIF program is designed to automatically exclude those levels of the matching variable in which there is only one examinee.

Conducting DIF analysis

Using GMHDIF for DIF analyses is as simple as following these steps:

1. Import the data to be analyzed. Imported files must be delimited by space, tab, comma, or semi-colon delimiters. (The program can work properly with data files that have up to 201 variables and 10,000 cases).
2. Provide information about the following variables: (a) items to be studied for DIF, (b) items to be used as the matching variable, and (c) grouping variable.
3. Select the desired generalized MH statistic: $Q_{GMH(1)}$ or $Q_{GMH(2)}$.
4. Examine the results of the DIF analyses. The results can be saved as a text or rtf file upon request.

The program creates two output. The first contains descriptive information about the data file, the grouping and matching variables. The second output file contains the DIF analyses results: the $Q_{GMH(1)}$ or $Q_{GMH(2)}$ value for each analyzed item, its degrees of freedom (*df*), and the corresponding *p* value. Figure 1 shows the results of a two-stage DIF analysis involving four groups. As seen in the figure, the second item (Variable 3) did not show DIF at the .05 alpha level in either stage. The first item (Variable 2) showed DIF in both stages, making it necessary to determine between which particular groups the DIF existed. In this case, the decision was made to compare the focal groups (groups 2, 3, and 4) with the reference group (group 1) instead of comparing all the groups with each other. The pairwise comparisons show that the differential item functioning existed only between groups 1 and 2.

GMHDIF

GMH statistic = *QMH1*

Alpha level = 0.05

Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE 2

Stage 1: *QMH* = 8.5703 *df* = 3 *p* = 0.0356*

Stage 2: *QMH* = 9.0551 *df* = 3 *p* = 0.0286*

PAIRWISE COMPARISONS (PWC)

The pairwise comparisons statistically significant using a Bonferroni-adjusted alpha level (.05/ 3= 0.0167) are marked with an asterisk.

PWC between group 1 and group 2 : *QMH* = 5.9364 *df* = 1 *p* = 0.0148*

PWC between group 1 and group 3 : *QMH* = 0.7462 *df* = 1 *p* = 0.3877

PWC between group 1 and group 4 : *QMH* = 1.1769 *df* = 1 *p* = 0.2780

VARIABLE 3

Stage 1: *QMH* = 1.2350 *df* = 3 *p* = 0.7446

Stage 2: *QMH* = 0.6121 *df* = 3 *p* = 0.8937

Figure 1. Output file with the results of a two-stage DIF analysis involving four groups.

https://doi.org/10.5209/rev_SJOP.2011.v14.n2.47 Published online by Cambridge University Press

Which statistic should we use?

Although GMHDIF was specifically designed to make DIF analysis as simple as possible, the program necessarily asks the user, on some occasions, to choose from different options. The most important of these instances, no doubt, involves the choice of which statistic to use for the analysis. In the case of dichotomous items, it would be the same to use $Q_{GMH(1)}$ as $Q_{GMH(2)}$, since they are equivalent. In the case of polytomous items, when the response variable (the items) is nominal, the $Q_{GMH(1)}$ statistics should be used. For ordinal items, the choice between $Q_{GMH(1)}$ and $Q_{GMH(2)}$ will basically depend on the pattern of DIF we are interested in detecting. As described in Fidalgo and Madeira (2008), and Zwick et al. (1993) referring to the GHM and the Mantel test, $Q_{GMH(2)}$ increases the statistical power with respect to $Q_{GMH(1)}$ for detecting its particular pattern of association (the mean responses differ across the factor levels), so that it is much more effective for detecting constant DIF (the magnitude of DIF is constant across score categories). Stated more simply, in all response categories the item is consistently more difficult for one of the groups. On the other hand, $Q_{GMH(1)}$, on permitting detection of more complex patterns of association than $Q_{GMH(2)}$, has much more power than $Q_{GMH(2)}$ for detecting balanced DIF (the magnitude of DIF is balanced across score categories, so that the magnitudes of DIF are cancelled out within the item). In this case, the item does not systematically hurt any particular group but rather it might be more difficult for one group in some response categories but for another group in other response categories. In a recent study with large numbers of DIF patterns simulated, Fidalgo and Bartram (2010) found that depending on the DIF pattern the power difference between the two statistics can reach 86%. In this situation, they recommended $Q_{GMH(1)}$ if a single method is to be used given that $Q_{GMH(1)}$ is capable of detecting more complex patterns of association than $Q_{GMH(2)}$.

We have integrated all the information concerning MH statistics in a decision tree, displayed in Figure 2, that shows which MH statistics to use as a function of item characteristics and the nature and goals of the DIF analysis.

Software for DIF detection

A number of statistical software packages exist that permit a calculation of the MH statistics to a greater or lesser degree. The most complete package undoubtedly is the SAS software (2002, SAS Institute, Version 9.0), which includes an excellent module to perform categorical data analysis that uses the equation proposed by Landis et al. (1978) to compute all the generalized Mantel-Haenszel statistics presented here (Stokes, Davis, & Koch, 2000). Unfortunately, the program is not very easy to use for DIF detection. On the other hand, GMHDIF has been developed to provide an easy-to-use program for conducting DIF analyses. Some of the advantages of GMHDIF are that it automatically computes the total test score used as matching variable, it performs two-stage DIF analyses in multiple groups simultaneously and, if the H_0 is rejected, it performs pairwaise comparisons. Other programs also exist

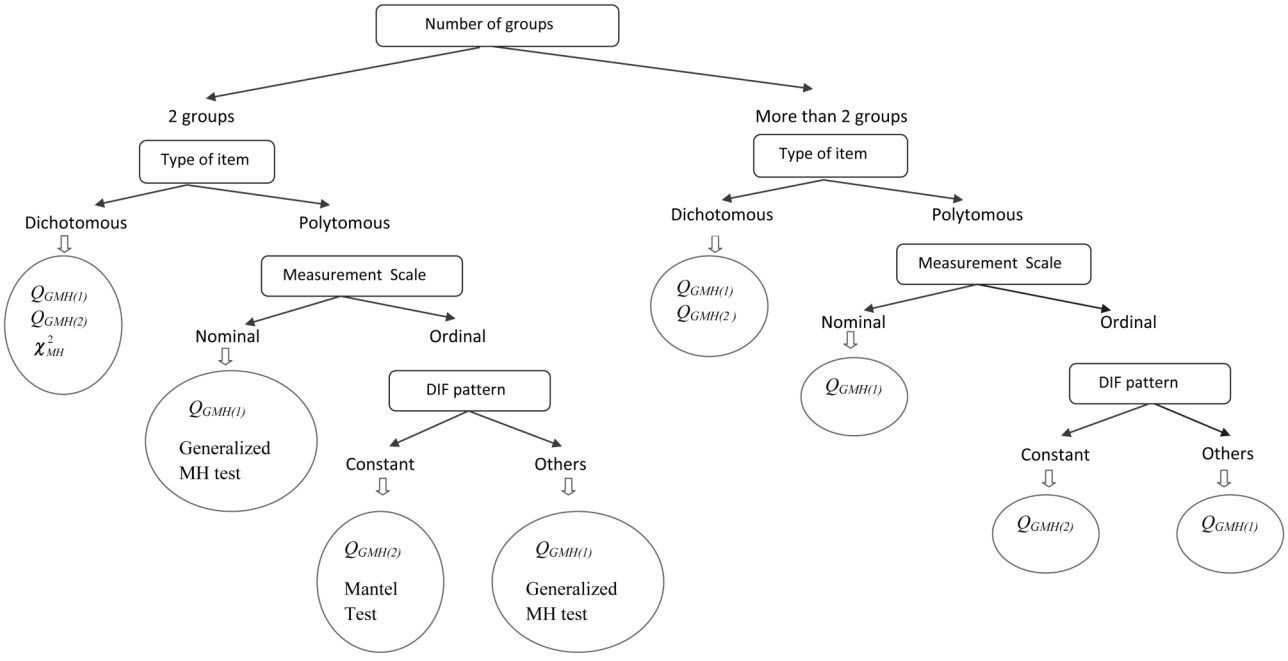


Figure 2. Types of MH statistics that may be used for detecting DIF as a function of item characteristics and objective of the analysis. If a circle contains two or more statistics then these statistics produce the same result (Fidalgo, 2010, Figure 1).

https://doi.org/10.5209/rev_SJOP.2011.v14.n2.47 Published online by Cambridge University Press

that are easier to use than SAS, or that are specifically designed for DIF detection, but these calculate only a small number of MH statistics. Among the commercially-available programs, SPSS (2007, SPSS for Windows, Rel. 16.0.0) only computes the χ^2_{MH} statistic (with the continuity correction) and the MH test of linear association. Among the free of charge software, we can mention MHDIF (Fidalgo, 1994), which calculates only the χ^2_{MH} statistic (with the continuity correction) and that can be used for detecting nonuniform DIF, and DIFAS (Penfield, 2005), which calculates the χ^2_{MH} statistic and the Mantel test. In conclusion, the GMHDIF program constitutes the most complete and simple option for assessing DIF in multiple groups using MH statistics.

Availability

The GMHDIF program, its user guide, and examples of input and output files can be obtained by e-mailing Angel M. Fidalgo at fidalgo@uniovi.es. The program and its related documentation are available in the following languages: Spanish, English, and Portuguese. Use is limited to academic or other nonprofit applications. Authors who use GMHDIF for their research are expected to cite this article.

References

Fidalgo, A. M. (1994). MHDIF: A computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure. *Applied Psychological Measurement, 18*, 300. doi:10.1177/014662169401800313

Fidalgo, A. M. (1996). Funcionamiento diferencial de los ítems [Differential item functioning]. In J. Muñiz (Ed.), *Psicometría* (pp. 371-455). Madrid, Spain: Universitas.

Fidalgo, A. M. (2005a). Mantel-Haenszel Methods. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol.3, pp. 1120-1126). Chichester, UK: John Wiley & Sons. doi:10.1002/0470013192.bsa364

Fidalgo, A. M. (2005b). Enfoque de la teoría de respuesta a los ítems [Item response theory framework]. In J. Muñiz, A. M. Fidalgo, M. A., García-Cueto, R. Martínez, & R. Moreno (Eds.), *Análisis de los ítems* (pp. 79-131). Madrid, Spain: La Muralla.

Fidalgo, A. M. (2010). GMHDIF: User's Manual. Oviedo, Spain: Universidad de Oviedo.

Fidalgo, A. M. (2011). GMHDIF: A computer program for detecting DIF in dichotomous and polytomous items using generalized Mantel-Haenszel Statistics. *Applied Psychological Measurement, 35,* 247-249. doi:10.1177/0146621610375691

Fidalgo, A. M., & Bartram, D. (2010). A comparison between some generalized Mantel-Haenszel statistics for detecting DIF in data simulated under the graded response model. *Applied Psychological Measurement, 34,* 600-606. doi:10.1177/0146621610378405

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for DIF detection. *Educational and Psychological Measurement, 68,* 940-958. doi:10.1177/0013164408315265

Fidalgo, A. M., & Scalon, J. D. (2010). Using Generalized Mantel-Haenszel Statistics to Assess DIF among Multiple Groups. *Journal of Psychoeducational Assessment, 28,* 60-69. doi:10.1177/0734282909337302

Fidalgo, A. M., Quintanilla, L., Fernández, R., Pons, F., & Aguerri, M. E. (2010). Detección del DIF en ítems politómicos mediante el uso de los métodos Mantel-Haenszel [Mantel-Haenszel methods for DIF detection in polytomous items]. *Revista Española de Metodología Aplicada, 15*, 12-18.

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

Kim, S. H, Cohen, A. H., & Park, T. -H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32,* 261-276. doi:10.1111/j.1745-3984.1995.tb00466.x

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

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

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. Computer Program Exchange. *Applied Psychological Measurement, 29,* 150-151. doi:10.1177/0146621603260686

SAS Institute (2002). SAS/STAT software: Version 9.0 (TS M0). Cary, NC: SAS Institute Inc.

SPSS for Windows, Rel. 16.0.0 (2007). Chicago, IL: SPSS Inc.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2000). *Categorical data analysis using the SAS system.* (2nd ed.). Cary, NC: SAS Institute.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233-251. doi:10.1111/j.1745-3984.1993.tb00425.x

Received June 1, 2010
Revision received September 23, 2010
Accepted November 22, 2010

https://doi.org/10.5209/rev_SJOP.2011.v14.n2.47 Published online by Cambridge University Press