Psicothema 2009. Vol. 21, nº 3, pp. 492-498
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Scientific production on the Mantel-Haenszel procedure as a way of detecting DIF

Georgina Guilera, Juana Gómez-Benito and M. Dolores Hidalgo*
Universidad de Barcelona and * Universidad de Murcia

For more than two decades, the Mantel-Haenszel (MH) procedure has been used to detect differential item functioning (DIF). A bibliometric study of this use of the MH procedure was conducted in order to better understand the current state of the research activity in the area (in terms of quantitative indicators and impact). Initially, we drew up a map of scientific output about this research area, and we subsequently conducted a detailed analysis of citations of authors and studies concerning the MH procedure. Main results suggest that the study of MH reached its peak in 1995; the most productive journal is the Journal of Educational Measurement, followed by Applied Psychological Measurement; the country with the greatest research output is the USA; the institutions that contribute to the research are mostly universities; the data fit Lotka's law of frequency of publication and do not confirm the exponential fit proposed by Price; and finally, a very high concentration of citations can be observed during the 1990s. In this context, the MH procedure is still being utilized and studied, thus several developments and applications may appear in the future, representing new theoretical, empirical, and simulation publications.

Producción científica sobre el procedimiento Mantel-Haenszel como método de detección del DIF. Tras más de dos décadas de la aplicación del procedimiento Mantel-Haenszel (MH) como técnica de detección del Funcionamiento Diferencial del Ítem (FDI), se realizó un estudio bibliométrico sobre el procedimiento MH para conocer en más profundidad el estado actual de la actividad investigadora en esta área (en términos de indicadores cuantitativos e impacto). Inicialmente, se define un mapa de la productividad científica en el tema de investigación que nos ocupa, y posteriormente se presenta un análisis detallado de las citas que reciben los autores y trabajos dedicados al estudio de MH. Los principales resultados sugieren que el estudio de MH alcanza su cenit en 1995, la revista más productiva es *Journal of Educational Measurement*, seguida de *Applied Psychological Measurement*, el país con una mayor productividad es EUA, las instituciones que contribuyen en la investigación son mayoritariamente universidades, los datos se ajustan a la ley de Lotka y no confirman la ley exponencial propuesta por Price, y finalmente se observa una elevada concentración de citas durante los años 90. En este contexto el procedimiento MH sigue siendo utilizado y estudiado, por lo que en un futuro pueden surgir ciertos desarrollos y aplicaciones, suponiendo nuevas publicaciones teóricas, empíricas y de simulación.

Bibliometric studies seek to describe the nature and development of a discipline or scientific field by means of counting and analysing the different aspects of written communication, and as such they are considered to be highly useful in the world of research and science in general. They provide valuable information about the research activity of a discipline, country, institution or journal in terms of quantitative indicators (i.e., number of publications, of authors, of journals) and impact (i.e., number of citations for a study, journal or author); they also enable comparative studies of these indicators to be carried out. It is not surprising, therefore, that governments use these data, among others, when making decisions about the awarding of grants or offering promotion to personnel; for example, in the USA, bibliometric indicators have been used since the 1970s in the process of distributing research funding, and in both Scandinavia and Switzerland researchers have drawn up detailed maps of scientific output in a wide range of knowledge areas (Ball & Tunger, 2006).

Bibliometry may also be used to suggest «future trends» in a given area of research, that is, to provide information about possible developments in the field in terms of productivity. As Ball and Tunger (2006) point out, the *past* of a discipline may be evaluated by counting the number of articles published during a given period (which may be extensive), its *present* by calculating the number of citations for the various studies, and its *future* by considering fluctuations in the number of publications and the number of times the work is cited.

Fecha recepción: 7-10-08 • Fecha aceptación: 31-1-09
Correspondencia: Georgina Guilera
Facultad de Psicología
Universidad de Barcelona
08035 Barcelona (Spain)
E-mail: gguilera@ub.edu

In recent decades, one area of research that has attracted much attention has been the study of differential item functioning (DIF) in order to ensure the metric equivalence of measurement instruments, to pay attention in recommendations from professional standards and guidelines, to improve test and questionnaire validity, among others. An item is considered to exhibit DIF when examinees from different groups (e.g., ethnicity, culture or gender) have a different probability of endorsing an item, when these are matched on the attribute measured by the item.

One of the pioneering methods used to detect DIF is known as the Mantel-Haenszel procedure (MH; Mantel & Haenszel, 1959). This method is based in contingency tables analysis and was first used to detect DIF by Holland and Thayer (1988). The MH procedure compares the item performance of the reference and focal groups, which were previously matched on the trait measured by the test; the observed total test score is normally used as the matching criterion. In the standard MH procedure an item shows DIF if the odds of correctly answering the item are different for the two groups at a given level *j* of the matching variable.

The MH procedure has widely been used to detect DIF because is conceptually simple, relatively easy to apply, offers a test of statistical significance and provides an estimation of the effect size based on the common odds ratio. Furthermore, the MH statistic can be calculated using easily accessible statistical software, whether that for general use (SPSS, SYSTAT, SAS) or more specific packages (MHDIF: Fidalgo, 1994; EZDIF: Waller, 1998; DIFAS: Penfield, 2005). It has not high power for non-uniform DIF (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994), but Mazor, Clauser and Hambleton (1994) proposed a variation that reduces this limitation. Finally, large sample sizes are not required and Mazor, Clauser and Hambleton (1992) found high power and good control of the Type I Error in samples of 200 subjects per group. Probably, these are the reasons why the MH is nowadays the gold standard for detecting items with differential functioning.

Borgman and Furner (2002) pointed out the importance accorded the evaluation of research by means of bibliometric indicators in order to determine in greater detail the current state of research in different areas. Two recent bibliometric articles have been published about DIF (Gómez, Hidalgo, Guilera, & Moreno, 2005; Guilera, Gómez, & Hidalgo, 2006), but they were centred on DIF in general without including any specification about procedures. So, taken into account the abovementioned and the fact that the MH procedure is currently used by the Educational Testing Service as the standard procedure for detecting items with DIF, it seems useful to conduct a specific bibliometric study that offers a scientific production map about the use of MH for detecting DIF. Thus, initially we drew up a map of scientific output regarding this research area, while subsequently we analysed citations in terms of authors, studies and year of publication.

Methods

Search strategy

Documents to be included in this study were located by searching in the databases SCI-EXPANDED and SSCI (Web of Science) of the Institute for Scientific Information (ISI) in February 2007. The strategy used was as follows:

(Differential item functioning or DIF) and (Mantel-Haenszel or MH)

The search was restricted to the period up to and including the year 2005. Once the studies had been complied, their titles and abstracts were reviewed by one expert in DIF to ensure that they did in fact refer to use of the MH procedure for detecting DIF.

Data analysis

The bibliometric analyses presented here are based on frequencies and percentages of studies, but also on other widely used indicators for analysing the growth of scientific production (Price's law), the dispersion of scientific output across journals (Bradford's law), and the authors' productivity (Lotka's law). The first one, Price's law (Price, 1963), proposes that the growth of the scientific production over time follows an exponential function. The Bradford's law (Bradford, 1934) describes how the articles in a specific area are scattered across journals; it postulates a model of concentric productivity zones with a decreasing information density; generally journals are divided in three concentric zones, each containing a similar number of articles. Finally, Lotka's law (Lotka, 1926) seeks to calculate the number of expected authors for a given number of published studies. The law is expressed as $y = C \times x^{-n}$, where *x* is the number of publications of interest, *n* is an exponent that is constant for a given set of data, *y* is the expected percentage of authors with frequency *x* of publications, and *C* is a constant. This means that productivity corresponds not to the number of articles published by an author but to its logarithm.

In the present study, firstly we analysed scientific output according to the variables *year of publication*, *number of authors*, *journal where published and corresponding subject area, type of study, country*, *institution* and *author*. Secondly we recorded the number of citations received by the various authors and studies, and analysed the evolution of citations over time.

All the analyses were conducted using Excel and CiteSpace 1.2 software (Chen, 2005).

Results

By using the abovementioned search strategy we identified 100 studies concerning the use of the MH procedure to detect DIF. The main results obtained are presented below. It should be noted that in the scientific output analysis the absolute frequencies coincide with the participation percentages in all cases, as the number of studies included was 100.

Scientific output

Year of publication. Figure 1 shows the evolution in the frequency of published studies for the period 1990-2005. The peak

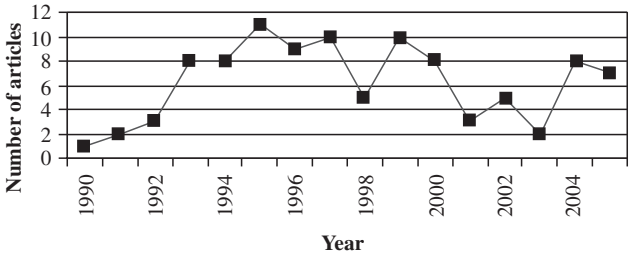

Figure 1. *Evolution over time of the number of publications*

output, with a total of 11 studies, corresponds to 1995, seven years after Holland and Thayer's original proposal regarding use of the MH procedure to detect DIF.

The visual examination of figure 1 may us think that the number of publications over time does not follow the exponential growth proposed by Price, an information that was corroborated after testing for the model's fit (R^2= .084; p= .276).

Number of authors. The overall mean number of authors per study was 2.27 (SD= 0.89; range 1-5). Studies with two or three authors account for well over half the total number of publications (73%); 41 studies were written by two authors and 32 by three.

As regards any changes in the number of authors per article over time it is worth to note that there is scarcely any variability over the years; practically all of them correspond to two or three authors.

Journals. A total of 23 journals have published articles about use of the MH procedure to detect DIF. Table 1 shows the most productive journals and the results of the Bradford's law scattering. Three concentric zones were defined, the zone 1 containing 37% of the articles, that were published in two journals, *Journal of Educational Measurement* and *Applied Psychological Measurement*, which contributed 19 and 18 documents, respectively. The second zone contained 31% of the articles published in three journals, *Educational and Psychological Measurement*, *Psicothema* and *Applied Measurement in Education*, which published 13, 10 and 8 documents, respectively. Finally, zone 3 comprised 17 journals accounting for a total of 32% of the output.

Subject area of journals. In order to assign journals to their subject areas we followed the classification system of the ISI *Journal Citation Reports*, which takes into account that the same journal may be assigned to more than one area. Table 2 shows the different areas of knowledge containing four or more studies, the total number of areas that have published reports about the use of the MH procedure to detect DIF being 17.

The central issue of the most widely represented areas refers to the mathematical aspect of both psychology and education. Specifically, the area making the largest contribution in terms of the MH procedure is *Psychology, Mathematical*, which alone accounts for over half the total number of studies. The areas *Psychology, Educational* and *Social Sciences, Mathematical Methods* also make a substantial contribution to this issue, with 46 and 30 published articles, respectively.

Type of study. In relation to type of study, it is worth to know that most studies are centred on the study of MH by means of simulated data (57%). The rest of the articles are empirical studies (32%; one of them also includes simulated data), theoretical reviews (7%) or are dedicated to develop new software for implementing MH procedure (5%).

Countries. In relation to scientific output according to the country of origin of authors, it should be noted that the same study may be counted more than once, as each report is counted for all the countries of the authors.

Among the total of 12 countries the greatest participation is that of the USA, which was involved in 72 studies; it is followed by Spain, Taiwan and Canada, which collaborated in 16, 6 and 5 articles, respectively.

As can be seen in Table 3, of those countries involved in some kind of joint work, that is, which publish collaborative studies, it is once again the USA that features most often, followed by Spain, Taiwan and Canada; however, in terms of the percentage of collaborations Canada comes top of the list.

Institutions. A total of 74 different institutions, mostly universities, participated in the collated studies. Figure 2 shows the contribution of those institutions involved in four or more studies; as in the case of output per country it should be borne in mind that the same article may be counted more than once as each report is counted for all the institutions of the authors.

The most productive organization is the *Educational Testing Service*, followed by three US universities (*University of Massachusetts*, *University of Illinois* and *University of California*) and one in Spain (*University of Oviedo*); together these account for almost 50% of the published studies.

Authors. Table 4 shows the output of the 135 authors included in the present analysis. It can be seen that a large number of authors (69.4%) participate sporadically in the study of the MH procedure and DIF, whereas only a few authors present several reports on the same issue.

Table 1
Most productive journals and scientific scattering

Zone (%)	Journals (n)
Zone 1 (37%)	*Journal of Educational Measurement* (19) *Applied Psychological Measurement* (18)
Zone 2 (31%)	*Educational and Psychological Measurement* (13) *Psicothema* (10) *Applied Measurement in Education* (8)
Zone 3 (32%)	17 journals

%: percentage of articles in the zone
n: number of published articles

Table 2
Number of publications in the most productive areas of knowledge

Journal categories	Articles
Psychology, Mathematical	62
Psychology, Educational	46
Social Sciences, Mathematical Methods	30
Psychology, Applied	23
Education & Educational Research	21
Mathematics, Interdisciplinary Applications	17
Psychology, Multidisciplinary	10

Table 3
Number of collaborative and single-author articles in the different countries

Country	Single-author	Collaborations	% collaboration
USA	16	56	77.78
Spain	1	15	93.75
Taiwan	1	5	83.33
Canada	0	5	100
Germany	0	1	100
Argentine	0	1	100
Holland	0	1	100
Hungary	0	1	100

After the analysis showed in Table 4, the value of *n* calculated by the least squares method was 1.97, giving a *C* value of 0.655. As the value of the maximum difference between the real and estimated accumulated frequencies was 0.052, that is, less than the critical value (c.v.= 0.138), the data obtained fit those estimated through application of Lotka's law.

Citation frequency

Citation of authors. Extreme caution must be exercised when interpreting the table showing the number of citations received by authors, as when counting citations it was only taken into account the first author of a study. Therefore, the data presented here could underestimate the number of citations received by authors depending on where their names appear in the list of contributors to a study.

Table 5 shows the top ten authors in terms of the number of times they have been cited by other authors (and journals) working on the use of the MH procedure to detect DIF. The names which appear in the table are of no surprise as, beginning with Mantel (and Haenszel) and Holland (and Thayer), they are all authors who have dedicated much effort to studying the use of the MH procedure for this purpose.

Citations of studies. As in the previous section, Table 6 shows the ten studies on the MH procedure and DIF which have received the highest number of citations.

Figure 3 illustrates the number of citations over the years. With the exception of 1959, the year in which the original study by Mantel and Haenszel was published, the years which correspond to the highest number of citations are grouped around the end of the 1980s and beginning of the 1990s; the studies published at this time have been the most important reference works to date as regards the application of the MH procedure to DIF.

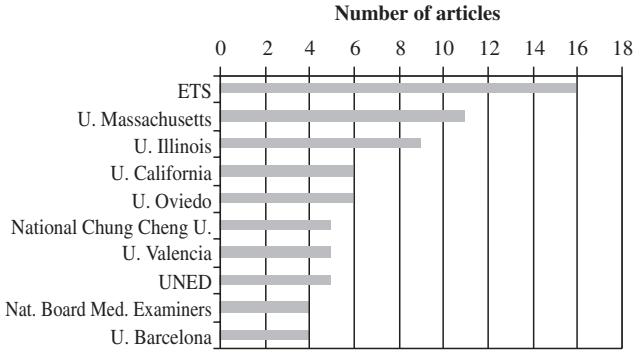

Figure 2. *Number of studies by the most productive institutions*

Table 4
Productivity of the authors and application of Lotka's law

x	y	X= lg x	Y= lg y	X^2	XY	$y_x/\Sigma y_x$	$\Sigma(y_x/\Sigma y_x)$	$f_e = C(1/x^a)$	Σf_e	D
1	95	0.000	1.978	0.000	0.000	0.704	0.704	0.656	0.655	0.048
2	23	0.301	1.362	0.091	0.410	0.170	0.874	0.167	0.822	0.052
3	4	0.477	0.602	0.228	0.287	0.030	0.904	0.075	0.897	0.007
4	4	0.602	0.602	0.362	0.362	0.030	0.933	0.042	0.939	-0.006
5	3	0.699	0.477	0.489	0.333	0.022	0.956	0.027	0.967	-0.011
6	3	0.778	0.477	0.606	0.371	0.022	0.978	0.019	0.986	-0.008
7	2	0.845	0.301	0.714	0.254	0.015	0.993	0.014	1.000	-0.007
11	1	1.041	0.000	1.084	0.000	–	–	–	–	–
Σ^a	135	3.702	5.799	2.489	2.019					

x: number of articles; y: number of authors
[a] Totals are presented excluding the data y= 1

Table 5
The top ten authors in terms of the number of citations received

Author	Cites
Holland, P. W.	86
Zwick, R.	85
Swaminathan, H.	68
Dorans, N. J.	66
Mantel, N.	63
Lord, F. M.	47
Shealy, R.	41
Raju, N. S.	40
Donoghue, J. R.	38
Camilli, G.	34

Table 6
The top ten studies in terms of the number of citations received

Author/s	Year	Source	Cites
Holland, P. W. and Thayer, D. T.	1988	Test validity	70
Swaminathan, H. and Rogers, H. J.	1990	J Educ Meas	50
Mantel, N. and Haenszel, W.	1959	J Natl Cancer I	41
Zwick, R.	1990	J Educ Stat	34
Shealy, R. and Stout, W. F.	1993	Psychometrika	31
Rogers, H. J. and Swaminathan, H.	1993	Appl Psych Meas	27
Dorans, N. J. and Kulick, E.	1986	J Educ Meas	25
Donoghue, J. R. Holland, P.W. and Thayer, D. T.	1993	Differential item functioning	25
Raju, N. S.	1988	Psychometrika	23
Camilli, G. and Shepard, L. A.	1994	Methods for identifying biased test items	21

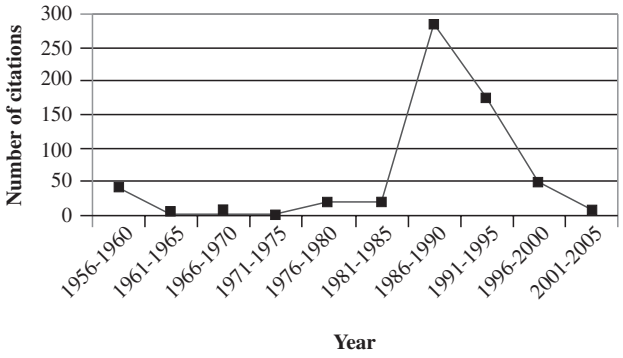

Figure 3. *Citation frequency over time according to the date of publication of the document*

Discussion

In general, greater interest in studying the use of the MH procedure was shown after publication of the paper by Holland and Thayer (1988) and this reached its high point in 1995. Although studies have continued to be published on the application of the MH procedure to DIF, output since 1995 has ranged from between 2 and 10 studies per year.

As regards authorship, most studies have been written by more than one author, a trend which, in the field of social science, seems to be more strongly associated with quantitative research than with theoretical or historical studies (Endersby, 1996). Moreover, the mean number of authors per study is 2-3, this figure being consistent with that reported by Beaver (2001) for various specialist areas such as natural sciences and mathematics.

The most productive journal is the *Journal of Educational Measurement*, followed by *Applied Psychological Measurement*, both located at Bradford's zone 1 and contributing to 37% of the articles. In this regard it is clear that the publication of articles on the use of the MH procedure to detect DIF is restricted to a small number of journals. The areas making the greatest contribution to this subject are those concerning the mathematical aspect of both psychology and education. Given that DIF studies refer to measurement instruments it is not surprising that it is these rather than other areas which appear in our results, as tests have traditionally been used in education and psychology (although their use has spread to other disciplines). However, many studies have also considered the use of the MH procedure to detect DIF by means of simulated data, and this would explain the predominance of areas with a mathematical aspect.

The country with the greatest research output is the USA, followed by Spain, Taiwan and Canada. In absolute terms the USA provides the highest number of author collaborations; however, among these four countries it is Canada which has the highest rate of collaboration as 100% of its studies are the work of more than one author, followed by Spain (93.75%), Taiwan (83.33%) and the USA (77.78%).

The institutions which contribute to research on the MH procedure and DIF are mostly universities. However, the most productive organization is the *Educational Testing Service*, most likely because, in addition to its great potential in terms of research in this area, it keenly supports the use of the MH procedure. The ranking also includes three US institutions with a long and varied tradition in the study of DIF, as well as four Spanish universities which, despite entering the field after their North American counterparts, have left their mark on the field of DIF research. The list also includes the National Chung Cheng University from Taiwan.

With respect to author productivity, the data fit Lotka's law of frequency of publication, indicating that most authors participate in the study of the MH procedure and DIF as an isolated activity (as showed by the *C* value, more than 65.5% of the authors have only published one article), and there are only a few who continue to work on the subject and illustrate their work through publications.

Finally, the results show the authors and studies that have been taken as the standard reference works in this field of research. Any researcher who has shown an interest in the study of the MH procedure and DIF will recognize the set of authors and studies listed here as being relevant for scientific progress in this area of knowledge, and they may well have read one or more of the authors or their works. Naturally, the list includes the original study by Mantel and Haenszel (1959) where they developed the test, as well as the report by Holland and Thayer (1988) in which its use for detecting DIF was first proposed. The book by Camilli and Shepard (1994) describes various techniques for detecting DIF, including the MH procedure, while the studies by Donoghue, Holland, and Thayer (1993) and that of Zwick (1990) have as their main objective the study of the MH procedure in various simulation conditions. The enormous impact of the studies by Swaminathan and Rogers (1990), Rogers and Swaminathan (1993) and Shealy and Stout (1993) is probably due less to the MH procedure as such than to the proposal of new techniques for detecting DIF (the first two use logistic regression while the latter employs the SIBTEST); however, the three articles did use the MH procedure for comparison purposes. Similarly, the study by Dorans and Kulick (1986) uses the standardization method, which has also been compared with the MH in subsequent studies. Finally, the article by Raju (1988) does not refer directly to the MH procedure, but does present the famous formulas for quantifying the amount of DIF in terms of the area between the item characteristic curves.

As regards the temporal evolution of the number of publications, the data do not confirm the exponential fit proposed by Price. It can be seen that the amount of research dedicated to the MH procedure is less at the start of the twenty-first century than during the 1990s, perhaps because a point has already been reached where the advantages and disadvantages of the method in different situations are well known (Allen & Donoghue, 1996; Donoghue & Allen, 1993; Ferreres Traver, Fidalgo Aliste, & Muñiz, 2000; Fidalgo, Ferreres, & Muñiz, 2004; Fidalgo, Mellenbergh, & Muñiz, 1998; Mazor et al., 1992, 1994; Narayanan & Swaminathan, 1994, 1996; Uttaro & Millsap, 1994; Wang & Su, 2004). Furthermore, lately the MH procedure and its extensions for polytomous items have been applied to several assessment settings (Dorans & Kulick, 2006; Elosua & López-Jáuregui, 2007; Kim, Cohen, Alagoz, & Kim, 2007; Ockey, 2007; Ponsoda, Abad, Francis, & Hills, 2008; Roever, 2007; Ross & Okabe, 2006), and most recent simulated studies have been focused on analyzing the functioning of these procedures under new simulation conditions or in comparison to other techniques (Bolt & Gierl, 2006; Fidalgo, Hashimoto, Bartram, & Muñiz, 2007; Williams, 2006), and suggesting new procedures for assessing DIF in the framework of the MH (Penfield, 2007; Sinharay, 2006). Subsequently, the MH procedure is still being utilized and studied, thus several developments and applications

could appear in the future, representing new theoretical, empirical, and simulation publications.

When considering the number of citations per study according to the year of publication a very high concentration of citations can be observed during the 1990s, this being the period when the key studies on the MH procedure were published. Given the above, the time was clearly ripe to conduct a bibliometric study of scientific output regarding the MH procedure and DIF in order to clarify and define the origin, development and current state of the scientific productivity on this subject.

Limitations and future research

Some of the main limitations of the present study refer to search strategy and data collection. It is well known that databases used here (SCI-EXPANDED and SSCI) are lack of non-English speaking journals, thus results should be read carefully keeping in mind this matter. Moreover, it is worth to know that article selection and data codification was carried out by only one reviewer; even though she is an expert on DIF and most indicators are easy-to-code (i.e., year of publication or journal), it could be seen as a study limitation.

Another weakness refers to time limitation in citation frequency studies. Reference counting is a dynamic process and results would be different depending on when the search was carried out; in this sense, results presented in citation frequency studies should be taken as temporary valid.

We have only presented some of the possible analyses which could have been conducted, but the present study provides an extensive bibliometric approach to research on the use of the MH procedure as a way of detecting DIF. In future research it would be interesting to carry out meta-analytic studies about the MH technique in order to know more in depth its functioning and to identify factors that may have an effect on Type I error rate and statistical power.

Acknowledgements

This study was partially supported by grants 2007FIC00736 and 2005SGR00365 from the «Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya», and SEJ2005-09144-C02-02 from Spain's «Ministerio de Ciencia y Tecnología» under European Regional Development Found (ERDF).

References

Allen, N.L., & Donoghue, J.R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*(2)*,* 231-251.

Ball, R., & Tunger, D. (2006). Bibliometric analysis - a new business area for information professionals in libraries? Support for scientific research by perception and trend analysis. *Scientometrics, 66*(3), 561-577.

Beaver, D.D. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics, 52*(3)*,* 365-377.

Bolt, D.M., & Gierl, M.J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement, 43*(4), 313-333.

Borgman, C.L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information and Technology, 36*(1)*,* 2-72.

Bradford, S.C. (1934). Sources of information on specific subjects. *Engineering, 23*(3)*,* 85-88.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Chen, C. (2005). CiteSpace: quick guide 1.2. http://cluster.cis.drexel.edu/~cchen/citespace/

Donoghue, J.R., & Allen, N.L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*(2)*,* 131-154.

Donoghue, J.R., Holland, P.W., & Thayer, D.T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In W.P. Holland & H. Wainer (Eds.): *Differential item functioning* (pp. 137-166). Hillsdale, NJ: LEA.

Dorans, N.J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4)*,* 355-368.

Dorans, N.J., & Kulick, E. (2006). Differential item functioning on the Mini-Mental State Examination: An application of the Mantel-Haenszel and standardization procedures. *Medical Care, 44*(11, Suppl. 3)*,* S107-S114.

Elosua, P., & López-Jáuregui, A. (2007). Application of four procedures for detecting differential item functioning in polytomous items [Aplicación de cuatro procedimientos de detección del funcionamiento diferencial sobre ítems politómicos]. *Psicothema, 19*(2)*,* 329-336.

Endersby, J.W. (1996). Collaborative research in the social sciences: Multiple authorship and publication credit. *Social Science Quarterly, 77*(2)*,* 375-392.

Ferreres Traver, D., Fidalgo Aliste, A.M., & Muñiz, J. (2000). Detection of non-uniform DIF: Mantel-Haenszel and logistic regression methods [Detección del funcionamiento diferencial de los ítems no uniforme: comparación de los métodos Mantel-Haenszel y regresión logística]. *Psicothema, 12*(Suppl. 2)*,* 220-225.

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*(3)*,* 300-300.

Fidalgo, A.M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement, 64*(6)*,* 925-936.

Fidalgo, A.M., Hashimoto, K., Bartram, D., & Muñiz, J. (2007). Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *Journal of Experimental Education, 75*(4)*,* 293-314.

Fidalgo, A.M., Mellenbergh, G., & Muñiz, J. (1998). Comparison of the Mantel-Haenszel procedure versus the loglinear models for detecting differential item functioning [Comparación del procedimiento Mantel-Haenszel frente a los modelos loglineales en la detección del funcionamiento diferencial de los ítems]. *Psicothema, 10*(1)*,* 209-218.

Gómez, J., Hidalgo, M.D., Guilera, G., & Moreno, M. (2005). A bibliometric study of differential item functioning. *Scientometrics, 64*(1), 3-16.

Guilera, G., Gómez, J., & Hidalgo, M.D. (2006). Differential item functioning: A bibliometric analysis of journals published in Spanish [Funcionamiento diferencial de los ítems: un análisis bibliométrico de las revistas editadas en español]. *Psicothema, 18*(4)*,* 841-847.

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

Kim, S.H., Cohen, A.S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*(2)*,* 93-116.

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.

Mazor, K.M, Clauser, B.E., & Hambleton, R.K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52,* 443-451.

Mazor, K.M, Clauser, B.E., & Hambleton, R.K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*(2)*,* 284-291.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(12)*,* 317-323.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4)*,* 315-328.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*(3)*,* 257-274.

Ockey, G.J. (2007). Investigating the validity of math word problems for English language learners with DIF. *Language Assessment Quarterly, 4*(2)*,* 149-164.

Penfield, R.D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*(2)*,* 150-151.

Penfield, R.D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*(3), 187-210.

Ponsoda, V., Abad, F.J., Francis, L.J., & Hills, P.R. (2008). Gender differences in the Coopersmith Self-Esteem Inventory: The incidence of differential item functioning. *Journal of Individual Differences, 29*(4)*,* 217-222.

Price, D.J.S. (1963). *Little science, big science*. New York: Columbia University Press.

Raju, N.S. (1988). The area between two item characteristics curves. *Psychometrika, 53*(4), 495-502.

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly, 4*(2), 165-189.

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2)*,* 105-116.

Ross, S.J., & Okabe, J. (2006). The subjective and objective interface of bias detection on language tests. *International Journal of Testing, 6*(3)*,* 229-253.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates tru bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics, 31*(1), 1-33.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4)*, 3*61-370.

Uttaro, T., & Millsap, R.E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*(1)*,* 15-25.

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*(4)*,* 391-391.

Wang, W.C., & Su, Y.H. (2004). Factors influencing the Mantel and the Generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*(6)*,* 450-480.

Williams, N.J. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement, 30*(1)*,* 22-42.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*(3)*,* 185-197.