Guilera, Georgina; Gómez-Benito, Juana; Hidalgo, M. Dolores
Scientific production on the Mantel-Haenszel procedure as a way of detecting DIF
Psicothema, vol. 21, núm. 3, 2009, pp. 492-498
Universidad de Oviedo
Oviedo, España

Available in: http://www.redalyc.org/articulo.oa?id=72711821025
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.

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.
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 it 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 (MHDF: Fidalgo, 1994; EZDIF: Wraller, 1998; DIFAS: Penfield, 2005). It has not high power for non-uniform DIF (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Utao & 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, Guiler, & Moreno, 2005; Guiler, 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:

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
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 (R² = .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>
<thead>
<tr>
<th>Zone (%)</th>
<th>Journals (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zone 1 (37%)</td>
<td>Journal of Educational Measurement (19)</td>
</tr>
<tr>
<td></td>
<td>Applied Psychological Measurement (18)</td>
</tr>
<tr>
<td>Zone 2 (31%)</td>
<td>Educational and Psychological Measurement (13)</td>
</tr>
<tr>
<td></td>
<td>Psicothema (10)</td>
</tr>
<tr>
<td></td>
<td>Applied Measurement in Education (8)</td>
</tr>
<tr>
<td>Zone 3 (32%)</td>
<td>17 journals</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Journal categories</th>
<th>Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychology, Mathematical</td>
<td>62</td>
</tr>
<tr>
<td>Psychology, Educational</td>
<td>46</td>
</tr>
<tr>
<td>Social Sciences, Mathematical Methods</td>
<td>30</td>
</tr>
<tr>
<td>Psychology, Applied</td>
<td>23</td>
</tr>
<tr>
<td>Education &amp; Educational Research</td>
<td>21</td>
</tr>
<tr>
<td>Mathematics, Interdisciplinary Applications</td>
<td>17</td>
</tr>
<tr>
<td>Psychology, Multidisciplinary</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th>Single-author</th>
<th>Collaborations</th>
<th>% collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>16</td>
<td>56</td>
<td>77.78</td>
</tr>
<tr>
<td>Spain</td>
<td>1</td>
<td>15</td>
<td>93.75</td>
</tr>
<tr>
<td>Taiwan</td>
<td>1</td>
<td>5</td>
<td>83.33</td>
</tr>
<tr>
<td>Canada</td>
<td>0</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>Germany</td>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>Argentina</td>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>Holland</td>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>Hungary</td>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>
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.
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 Donohue, 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 & Donohue, 1996; Donohue & 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; Elooua & 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**


