The Application of PageRank and HITS in Chinese and UK University Ranking

Huang Zhanmiao **Luo Yuanhui** **Xia Wencan**
Department of Mathematics
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

Abstract

PageRank is an algorithm first used by Google Search that ranks website pages based on their importance. In this report, we apply PageRank and Kleinberg's HITS algorithm (authority and hub) to the university website pages of China (mainland) and the United Kingdom to investigate the university rankings. Besides, we also explore the performances of PageRank with different hyperparameters α. Different ranking results are compared against the research ranking with Spearman's ρ, Kendall's τ and Page's trend test for their ranking consistency. In case study, we further analyze the ranking of some representative universities in the two countries.

1 Introduction

University rankings, which evaluate universities based on a variety of factors, have become an increasingly important aspect of academia. Ranking algorithms such as PageRank and Kleinberg's HITS provide a perspective to rank universities based on the importance of their website pages. In this report, we explore Chinese and British university rankings based on PageRank and Kleinberg's HITS algorithm. To evaluate different ranking approaches, we compare their performances in top universities and test their consistency with the research ranking with non-parametric statistical methods including Spearman's ρ, Kendall's τ and Page's trend test. In addition, more investigations are conducted on the PageRank results with different damping factors α's and rankings of representative universities.

2 Dataset

2.1 Data Background

PageRank is an algorithm used by Google Search to rank web pages in their search engine results. We believe there is a connection between research ranking and PageRank of universities. From that point of view, we investigate the PageRank of universities based on two datasets.

The first dataset contains Chinese (Mainland) University Weblink during 12/2001-01/2001.[1] The second data set contains United Kingdom university Weblinks in 2006.[2] The research ranking for UK universities are from 'Shanghai Jiao Tong Ranking' with top 42 university.[3]

[1] `https://github.com/yao-lab/yao-lab.github.io/blob/master/data/univ_cn.mat`

[2] `http://cybermetrics.wlv.ac.uk/database/`

[3] `https://www.universityrankings.ch/results?ranking=Shanghai&region=World&year=2006&q=uk`

2.2 Data Preprocessing

The first data set is already processed, containing univ_cn, rank_cn and W_cn. The univ_cn contains the webpages of universities, rank_cn is the research ranking of these universities in that year, and W_cn is the link matrix whose (i,j)-th element gives the number of links from university i to j. There are 76 universities recorded in total.

The second data set contains only domains and link information. We use the sum of the number of times the valid domains of all university-j which are referred by university-i as the (i,j)-th element of W_cn. Since there is no ranking information, we use Shanghai Jiao Tong Ranking as a research ranking for reference. There are 112 universities recorded in total.

3 Methodology

3.1 Ranking Methods

- **PageRank**

  PageRank is the ranking method used by Google and it works by counting the number and quality of links to a page to determine how important the website is. Intuitively, the more links a website receives from other websites, the more important it is.

  Suppose the visits to websites are assumed to be a Markov chain on a connected graph $G = \{V, E, W\}$ with transition probability $P_{ij} = \mathrm{Prob}\,(x_{t+1} = j \mid x_t = i) \geq 0$. Thus P is a row-Markov matrix, i.e. $P \cdot \mathbf{1} = \mathbf{1}$ where $\mathbf{1} \in R^V$ is the vector with all elements being 1. Let the weight matrix W decodes the webpage link structure, and define an out-degree vector $d_i^o = \sum_{j=1}^n w_{ij}$, which measures the number of out-links from i. With a diagonal matrix $D = diag(d_i^o)$ and a row Markov matrix $P_1 = D^{-1}W$, we define

  $$P_\alpha = \alpha P_1 + (1-\alpha)E \tag{1}$$

  where $E = \frac{1}{n}\mathbf{1} \cdot \mathbf{1}^T$. the Perron-Frobenius theory tells us that P_α determines a unique stationary distribution π which can be used as ranking scores.

- **HITS Authority**

  HITS Authority uses primary **right** singular vector of W as scores to give the ranking. To be specific, let $L_\alpha = W^T W$ with $L_\alpha(i,j) = \sum_k W_{ki} W_{kj}$. Then the primary right singular vector of W is just a primary eigenvector of a non-negative symmetric matrix L_α. The higher value of $L_\alpha(i,j)$ means the more references received on the pair of nodes. Therefore Perron vectors tend to rank the webpages according to authority.

- **HITS Hub**

  HITS Hub uses the primary **left** singular vector of W as scores to give the ranking. Let $L_h = WW^T$ with $L_h(i,j) = \sum_k W_{ik} W_{kj}$. The primary left singular vector of W is just a primary eigenvector of a non-negative symmetric matrix L_h. The higher value of $L_h(i,j)$ means they are hitting the same target, thus giving hub ranking.

3.2 Statistical Methods

- **Spearman's ρ**

  Spearman's rank correlation coefficient (Spearman's ρ) is a non-parametric measure of the statistical correlation between the rankings of two random variables. Intuitively, it measures the similarity between two rankings, which will be high when the observations have similar or even identical rankings. Mathematically, Spearman's ρ is the Pearson correlation coefficient between two ranking values $\rho = \frac{\mathrm{cov}(R_X, R_Y)}{\sigma_{R_X}\sigma_{R_Y}} \in [-1, 1]$ and can be simplified into $\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^n (R_{X_i} - R_{Y_i})^2$ for distinct integers, where R_X and R_Y are the ranking values of random variable X and Y. When two rankings are independent, $T = \rho\sqrt{\frac{n-2}{1-\rho^2}}$ follows a t-distribution with $n-2$ degrees of freedom, then the Spearman correlation test can be conducted.

- **Kendall's τ**

Kendall rank correlation coefficient (Kendall's τ) is a non-parametric measure of rank correlation between two random variables by the number of concordant pairs. Similarly to Spearman's rank correlation coefficient, Kendall's τ lies in $[-1, 1]$ and will be high when the rankings of the two observations are similar. Mathematically, $\tau = \frac{2}{n(n-1)} \sum_{1 \leq i \leq j \leq n} \operatorname{sign}\left[(X_j - X_i)(Y_j - Y_i)\right]$ and $3\sqrt{\frac{n(n-1)}{2(2n+5)}}\tau$ approximates standard normal distribution for large samples when two rankings are independent, then Kendall correlation test can be conducted.

- **Page's Trend Test**

  Page's trend test is a non-parametric test for multiple comparisons on whether repeated observations of rankings follow a pre-specified trend. Intuitively, it tests the overall consistency between multiple rankings and the particular trend, which will be significant when multiple rankings almost have the pre-specified trend. Mathematically, the Page's trend statistics is $P = \sum_{i=1}^{n} \sum_{j=1}^{k} iR_{ij}$, where R_{ij} is the rank of subject i in the j-th ranking. It approximates the normal distribution $N(kn(n+1)^2/4, kn^2(n+1)^2(n-1)/144)$ for large samples under the null hypothesis, then Page's trend test can be conducted.

4 Data Analysis and Visualization

4.1 China (Mainland)

In this section, We first apply the PageRank algorithm to the web linking data of Chinese universities and expect to obtain an explainable rankings. Comparisons will be made with other ranking approaches including HITS authority ranking and HITS hub ranking, as well as among the transitions of varying damping factors α in PageRank.

4.1.1 Implementations of different approaches and parameters

PageRank is well applicable to rank the universities based on the linking matrix W between their websites. Note that we always set the damping factor $\alpha = 0.85$ in Google PageRank as the default unless otherwise stated below. By applying the method in Section 3.1, we calculate the ranking scores of universities and sort them in descending order.

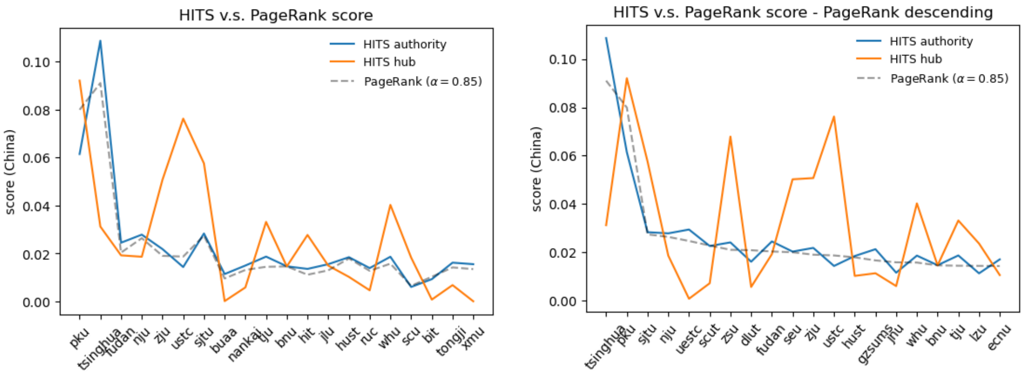
The ranking results are listed in Table 1, where the top-6 are Tsinghua University (Tsinghua), Peking University (PKU), Shanghai Jiao Tong University (SJTU), Nanjing University (NJU), University of electronic science and technology of China (UESTC), and South China University of Technology (SCUT). It excludes three of top-6 universities in research ranking: Fudan University (Fudan), Zhejiang University (ZJU), and the University of Science and Technology of China (USTC).

Besides, we choose HITS authority ranking and HITS hub ranking which measure the out-degree and in-degree, respectively. Different ranking orders are indicated in Table 1.

Table 1: Research ranking, PageRank ($\alpha = 0.85$), HITS authority and hub ranking (China)

	Research ranking	PageRank	HITS authority	HITS hub
1	PKU	Tsinghua	Tsinghua	PKU
2	Tsinghua	PKU	PKU	USTC
3	Fudan	SJTU	UESTC	ZSU
4	NJU	NJU	SJTU	SJTU
5	ZJU	UESTC	NJU	ZJU
6	USTC	SCUT	Fudan	SEU

Comparisons across ranking methods We next turn to the ranking scores of the above methods to see where the differences occur. The comparisons are made in the left Figure 1(a) and right 1(b). The left sub-figure follows the research ranking order in the x-axis, and the right one is rearranged as descending PageRank scores. Score curves of HITS authority and HITS hub are plotted, with the PageRank scores marked in the dashed line as a control group. Note that he names of universities in the figures are directly extracted from the URLs.



(a) x-axis in the order of Research rank

(b) x-axis in the order of PageRank

Figure 1: Compare PageRank ($\alpha = 0.85$), HITS authority and HITS hub ranking (China)

1. The two x-axes of Figure 1(a) and 1(b) are the top 20 universities in Research rank and PageRank, respectively. By comparing these two groups, we observe that certain universities rank highly in PageRank but not in the research ranking. Examples of such universities include zsu, seu, and uestc. Additionally, buaa does not feature among the top 20 universities according to PageRank.

2. In Figure 1, it can be noticed that PageRank and HITS authority ranking exhibit similar behavior, as they both consider the degree of being cited by other links. However, HITS hub ranking demonstrates an apparently different trend. Some universities receive high hub scores despite having relatively lower PageRank and authority scores, which might be accounted as high out-degree only. Notable examples include pku, ustc, and zsu, with ustc being particularly noteworthy, which will be discussed in Section 5).

Choosing varying α in PageRank The value of the damping factor α could influence the performance and reliability of the ranking algorithm. Therefore, we attempted a series of $\alpha \in (0, 1)$, and observed a powerful score trend in Figure 2(a) with varying α that smaller α generates lower variances among all the scores of universities. Then some institutions rank higher for lower α, while some are reversed. This is because smaller α (e.g. $\alpha = 0.15$, yellow line) strengthens the function of the uniform random walk matrix E in Equation (1). The link matrix is suppressed too much and thus renders the distinguishing ability to be worse and less arresting.

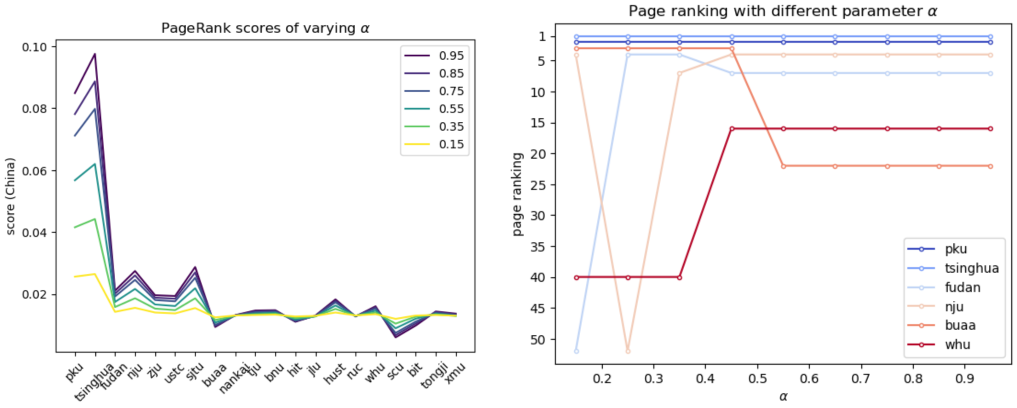
The variation and instability of small α are also revealed in Figure 2(b), except for the dominant top university PKU and Tsinghua. Therefore, α is needed for the stationary distributions of Markov chains, and sufficient high α is also necessary to guarantee the linking effect.

4.1.2 Rankings Consistency

In order to investigate the consistency of the above rankings, we adopt different statistical methods to compare these rankings against the research ranking. As shown in Table 2, the small Spearman's p-value and Kendall's p-value reveal that there is a significant correlation between these three rankings and the research ranking and the small Page's trend p-value validates the overall consistency between different rankings and the research ranking. From the values of Spearman's ρ and Kendall's τ we can see that PageRank and HITS authority show a stronger statistical correlation to the research ranking than that of hub ranking, which further verifies our conclusion in Section 4.1.1.

4.2 The United Kingdom

In this section, We conduct a similar analysis on the web linking data of UK universities. The three ranking approaches (PageRank, HITS authority and hub) and the effect of choosing distinct $\alpha \in (0, 1)$ are applied repeatedly. We discovered some interesting and unique outcomes in UK dataset when using HITS authority, and try to explain it from the theoretical and graph structure aspect. More details will be discussed and compared with the China case in Section 5.



(a) Score, univs in the order of research ranking

(b) Ranking of several example universities with different α's, see legend

Figure 2: Compare PageRank with different damping factor α (China)

Table 2: Comparison between different rankings and the research ranking of Chinese (Mainland) universities with Spearman's ρ, Kendall's τ and Page's test.

Rankings Statistics	PageRank	HITS authority	hub ranking
Spearman's ρ	0.706	0.751	0.540
Kendall's τ	0.520	0.572	0.378
Spearman's p-value	1.105×10^{-12}	5.944×10^{-15}	4.915×10^{-7}
Kendall's p-value	3.000×10^{-11}	2.665×10^{-13}	1.334×10^{-6}
Page's p-value		9.488×10^{-24}	

4.2.1 Implementations of different approaches and parameters

Just like in Section 4.1.1, our aim is to examine and compare the outcomes of PageRank with different α values, along with HITS authority and HITS hub rankings.

Referring to Table 3, we can observe the overall ranking of UK universities. Cambridge and Oxford consistently maintain their top positions in the ranking, reflecting their strong academic reputation and social influence. Besides, the top university rankings in the UK are more stable than those in China.

Table 3: Research ranking, PageRank ($\alpha = 0.85$), HITS authority and hub (UK)

	SJTU rank	PageRank	HITS authority	HITS hub
1	Cambridge	Cambridge	Cambridge	Dundee
2	Oxford	Oxford	UCL	Oxford
3	IC	Nottingham	Nottingham	UCL
4	UCL	UCL	Oxford	Bristol
5	Manchester	Soton	ED	ED
6	ED	ED	IC	Manchester

Comparisons across ranking methods and varying α Similarly to Chinese university rankings, we use different ranking criteria or distinct values of α as shown in Figure 3 and 4.

It is interesting that we have observed an extremely high ranking-score of Cambridge University when using HITS authority ranking (Figure 3, blue line), with almost 10 times of other universities including Oxford. It is also much higher than its PageRank or HITS hub score. Despite the one score deviation, the rank is not much affected (Table 3).

To explain this phenomenon, it might be helpful to investigate the specific characteristics of the web linking data, such as the distribution of incoming links, or the presence of influential nodes. PageRank takes into account the overall link structure of the web graph, which may assign higher importance to Cambridge University based on its extensive connections and incoming links. In contrast, HITS emphasizes the authority of a node based on how much it is cited and the quality of those citations. One possibility is that the structure of the web linking data might influence the ranking scores. If there are specific patterns or clusters in the web graph that disproportionately favor Cambridge University, it can contribute to the higher score for that particular institution. Therefore, the difference in scores between PageRank and HITS for Cambridge University could stem from the different perspectives and criteria employed by these algorithms in determining importance and relevance within the network.

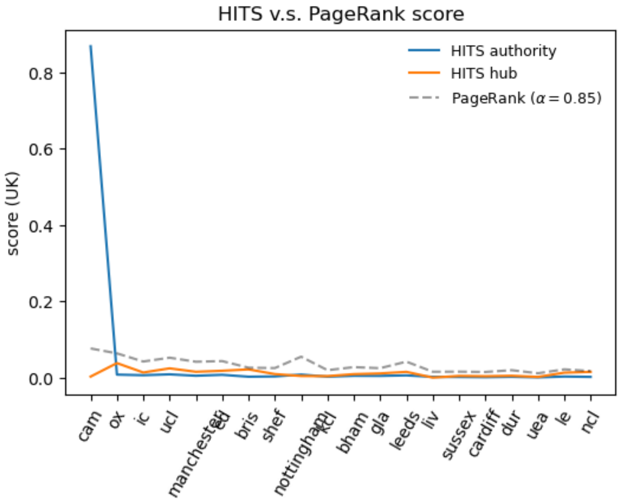
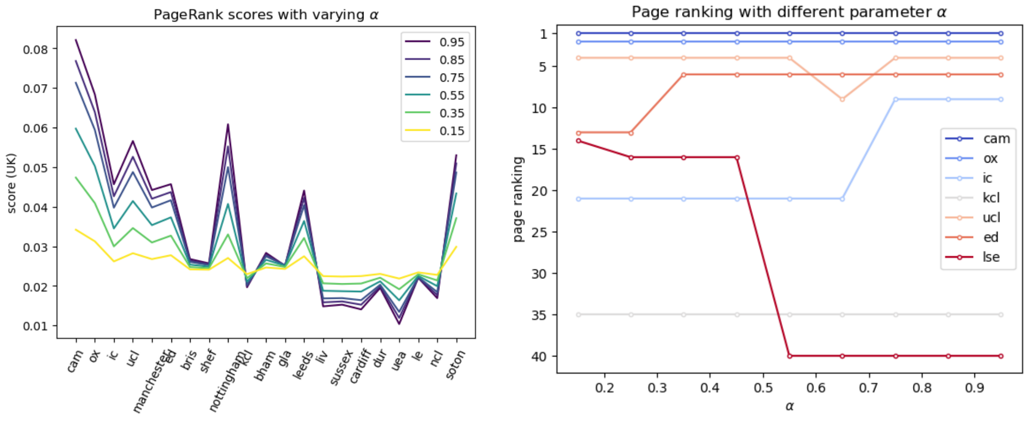


Figure 3: Compare PageRank ($\alpha = 0.85$), HITS authority and HITS hub rankings (UK)

We once more verify the trend of varying α values in Figure 4 with the Chinese range (Figure 2) that small α weakens the distinguishing and ranking. Here, top universities in UK are quite stable against different α, including Cambridge, Oxford, UCL (University College London), and ED (Edinburgh).



(a) Score, univs in the order of research ranking

(b) Ranking of several example universities with different α's, see legend

Figure 4: Compare PageRank with different α (UK)

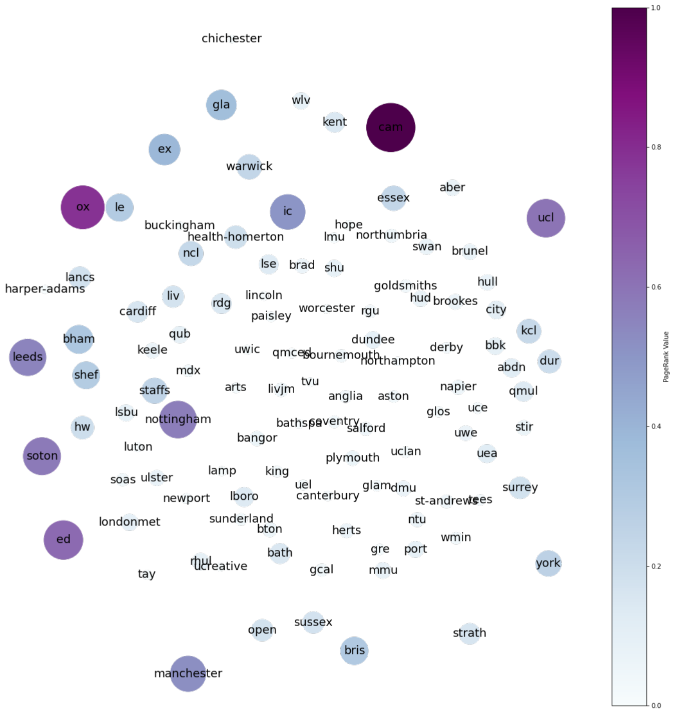


Figure 5: PageRank Visualization (To have a general view of UK univsersities)

4.2.2 Rankings Consistency

Similarly, we compare these rankings with different statistical methods to investigate their consistency with the research ranking. As shown in Table 4, Spearman's p-value and Kendall's p-value indicate that PageRank and HITS authority are weakly correlated with the research ranking while the statistical correlation between hub ranking and the research ranking is not significant. The weak correlation differs from our findings in the dataset about Chinese universities, which might result from the different ranking criteria adopted by the "Shanghai Jiao Tong Ranking". However, we can see that these rankings remain overall consistent since the Page's trend p-value is small. In addition, as shown by the values of Spearman's ρ and Kendall's τ, PageRank and HITS authority also show a stronger correlation to the research ranking than that of hub ranking.

Table 4: Comparison between different rankings and the research ranking of UK universities with Spearman's ρ, Kendall's τ and Page's test.

Rankings Statistics	PageRank	HITS authority	hub ranking
Spearman's ρ	0.260	0.264	0.105
Kendall's τ	0.180	0.185	0.085
Spearman's p-value	0.096	0.091	0.506
Kendall's p-value	0.093	0.085	0.429
Page's p-value		2.355×10^{-13}	

5 Case Study

We investigate the ranking results of "golden triangle universities" in UK, which consists of the University of Cambridge, the University of Oxford, Imperial College London, King's College London, the London School of Economics and University College London. LSE is a highly specialized school and lacks diversity from other domains, and thus has a low rank in the PageRank algorithm. A similar phenomenon shows up in PageRank of the University of Science and Technology of China in Chinese universities. In conclusion, the PageRank or HITS hub rankings give a lower rank to specialized schools while a higher rank to comprehensive universities.

On the other hand, we investigate the HITS hub algorithm with the top 5 universities in UK, which are Dundee University, Oxford University, University College London, University of Bristol, and the University of Edinburgh. However, it is not consistent with research ranking and Dundee University is over-ranked. We found that there are tons of links from the Dundee domain to the Cambridge domain, causing a relatively high rank of Dundee under HITS hub. This might be a misleading factor in ranking universities.

6 Conclusion

In this report, we explore the Chinese and British university rankings with PageRank and Kleinberg's HITS algorithm. By comparison, PageRank and HITS authority give similar ranking results, which might differ from that of HITS hub for specialized universities such as USTC and LSE. When varying the hyperparameter α, the variance of PageRank scores will decrease as α increases since large α can strengthen the influence of the Markov matrix. Besides, the non-parametric statistical results by Spearman's ρ and Kendall's τ reveal that there is a statistical correlation between the three rankings and the research ranking, where PageRank and HITS authority show a stronger correlation to the research ranking than that of hub ranking and Chinese universities show a stronger correlation to the chosen research ranking than that of British universities. Generally, the three rankings are consistent with the research ranking, which is supported by Page's trend test results.

7 Contribution

Huang Zhanmiao: Ranking algorithm, data visualization, report, and presentation.
Luo Yuanhui: Statistical analysis, methodology theory, report, and presentation.
Xia Wencan: Data processing, ranking algorithm, case study, and report.

Reference

[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: Bring order to the web. technical report, Stanford University.

[2] Ding, Chris, et al. "PageRank, HITS and a unified framework for link analysis." Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. 2002.

[3] Spearman, C. 1904. "The Proof and Measurement of Association Between Two Things,". American Journal of Psychology, 15 January: 72–101.

[4] Abdi, Hervé. "The Kendall rank correlation coefficient." Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA (2007): 508-510.