Eindhoven University of Technology
Department of Industrial Engineering & Innovation Sciences
Human-Technology Interaction Research Group

Citation Counts as a Measure for Scientific Impact

Bachelor Thesis

J.H.P. Burgers

Supervisors:
D. Lakens
P. Isager

Final Version

Eindhoven, June 2019

Abstract

This thesis presents a look into citation counts as a measure for scientific impact which in turn is used to determine the replication value (RV). first, by comparing citation sources (WoS, Crossref, Scopus and Scite) from which citation counts can be retrieved. Secondly, by removing contradicting citations from the citation count, and comparing this new citation count without contradicting citations with the original total citation count. In both cases, based on the citation count, rank order lists are formed which are compared with the use of two tests. First, Kendalls tau is calculated to see how well the compared pairs of lists correlate. Second, the rank biased overlap (RBO) is calculated to see how well pairs of lists overlap. The RBO is different than Kendalls tau because it is able to give more weight to citation counts at the top of the list emphasizing the importance of high ranked articles as opposed to low ranked articles. Both measures indicate a significant correlation and overlap between ranked lists originating from Scopus and Crossref and WoS, and a lower correlation and overlap between Scite and all other sources. Based on the difference between Scite and all other sources, Scite is not yet the best choice as a citation source for determining scientific impact. Both measures also indicate a strong correlation and overlap between the ranked list formed from the total citation counts and the ranked list formed from the total citation count minus the contradicting citations. Based on this high correlation and overlap, taking out contradicting citations is not needed when determining scientific impact.

All used data and scripts can be found here:

`https://osf.io/b87de/?view_only=1ea26d1c7e6e485daf403531f812fffd`

Contents

Chapter 1

Introduction

In the past twenty years articles have appeared describing a replication crisis, especially in the field of psychology. Multiple events provide evidence for this replication crisis. First, there are documented cases of scientific fraud conducted by psychologists (e.g., Diederik Stapel). Secondly, there are articles by several authors (Ioannidis, 2005) criticizing research practices that produce bias in the published results. Thirdly the 100 replications done by the Open Science Collaboration resulting in only 36

This replication crisis has been answered with the creation of new methods, norms and an increase in replication studies. This shift is even being called the renaissance of psychology (Nelson, Simmons & Simonsohn, 2018). Calling it a renaissance may be a little over the top, but the message by Nelson (2018) carries some truth. In the article several unwanted research practices are addressed with the corresponding counter measures. To prevent selective reporting scientists should openly share their data, manipulations and sample size justification. An even better way of preventing selective reporting (or p-hacking) is pre-registration (a plan containing the hypothesis and proposed data analysis which is made and shared before the analysis is done). To address the problem of selective publishing in past research scientists should perform meta-studies and replication studies.

One of the approaches that aims at improving the way science is conducted within the field of psychology is the proposed quantitative approach by Isager, et al. (in prep) called the replication value (RV). This formula helps scientists to identify a study from a large set of similar studies that is expected to have had disproportionally large impact given the corroboration of that finding (i.e., it helps decide which study to replicate). The first targeted application, that will also be used to validate the RV, is choosing a study to replicate from a large set of studies (more than 3000) all in which fMRI plays a role.

For the purpose of selecting a study to replicate, a formula is proposed to quantify the reproducibility of publications based on the impact of a finding and the evidence underlying a finding. The way impact is defined here is similar to how impact is defined for the IF. Impact is based on the amount of citations within a certain time period. Research points out that using a methodical approach to select a replication study is suitable. Certain assumptions are made within the replication value project. These assumptions are similar to the assumptions underlying the IF, some of which center around the use of citations to quantify scientific impact. The assumption is made that citation count signals the theoretical impact of a study. It is assumed that highly theoretically important findings tend to be cited more than less theoretically important findings. As mentioned earlier the goal of the RV is to detect articles with a disproportionally large impact, citation counts are suitable for this purpose since they measure scientific impact (Radicchi, Weissman & Bollen, 2017).

Citation analysis - the act of assessing the impact of scientific work based on the number of times the work is being cited by others (A. F. J. van Raan, 1996) - can serve multiple purposes. Citations counts can and have been used to quantify the quality of scientists (Garfield, 1970), Journals (Garfield, 1972), and individual publications (A. F. J. van Raan, 1996), they are used

to determine where to allocate research funding (Garfield, 1962) or are used for recommending research articles (Vellino, 2015). In general citation analysis is assumed to be a suitable way to quantify impact towards the scientific world, and it is more time and resource efficient then peer reviews (Cronin & Overfelt, 1994).

Currently, performing citation analysis to assess the impact of publications is gaining popularity when being compared to the use of the informed judgement of peers. It is reasoned that the more an article is cited the greater its meaning and influence will be for science. 60 years ago citation counts were first used to measure scientific impact (Garfield, 1955). Eugene Garfield called this measure the Impact Factor (IF). The IF is simply the amount of citations within a certain time period, usually this period is two years. For a journal the IF is the amount of citations to all its articles published in the preceding two years divided by the number of published in that journal during the same time period. Calculating the IF is similar for scientists or complete research groups (Bloch & Walter, 2001).

With the increased popularity of the IF comes an increase in concerns and doubts regarding the use of citation analysis to quantify scientific impact It is argued that citations are affected by different behaviors that decrease the reliability of citation counts as a measure of scientific impact. Examples of unwanted behavior are citing without reading the paper, biased citations or self-citations, or distortion of the original text (Rodriguez-Ruiz, 2009). The IF relies on the fact that when a scientist cites a paper, he agrees with and confirms what is stated in the paper that he is citing. If the citations is made without properly reading the article, or if a scientist is making a citation to one of his older papers only to inflate his own IF, the citation loses this confirming power (Pandita & Singh, 2017).

Each individual citation can be positive or negative about the publication that is being cited, and therefore confirm or reject the findings (Nicolaisen, 2003). Beyond distinguishing between positive and negative, an even more specific categorization could be made. Research by (Bornmann & Daniel, 2008) suggests the following categorization of ways to look at the context of a citation:

- Affirmative citations; citations that agree with the findings of the cited work.
- Assumptive citations; citations that refer to assumed knowledge stated in the cited work.
- Conceptual citations; the use of theories or concepts of the cited work.
- Contrastive citations; citations contrasting the current work with the cited work or contrasting two different cited works.
- Methodological citations; citations that refer to methodological techniques or designs of the cited work.
- Negational citations; citations that corrects, questions or negates the cited work.
- Perfunctory citations; citations that are not immediately relevant or even redundant to the current work.
- Persuasive citations; citations made in a ceremonial fashion.

Especially the citations of the perfunctory (7), persuasive (8) and negational (6) type cause concerns within the scientific world when such citations are used to quantify the impact of scientific research. Multiple scientists address or conclude that the context of the citation should be taken into account when using citation data to predict scientific impact (Rodriguez-Ruiz, 2009) (Macroberts & Macroberts, 1989).

In the most ideal situation citations are sorted based on the eight categorizations defined in Bormann Daniel (2008). The higher the level of detail within the categorization, the better we can judge why and article is cited. A simplified and more general categorization would be to distinguish between confirming, mentioning or contradicting citations. When compared to the Bormann Daniel categorization confirming relates to affirmative (1), refuting to perfunctory (6)

and possibly contrastive (4) and mentioning to assumptive (2), methodological (5), perfunctory (7) and persuasive (8).

Besides the context of the citation, self-citations can also influence Scientific impact. Self-citations are seen as a technical problem that needs to be corrected for VanRaan1996. Prior research has been done to see what happens to ways of determining scientific impact when self-citations are filtered out (A. F. van Raan, 2008). This research concludes that the effect of leaving out self-citations are minimal and only influence impact assessment for lower performance groups. It is common for scientists to make a citations of ones own report, but it is only justified as long as the citation is made for scientific reasons, not when the citation is made with the reason to inflate ones own IF. For this reason, other research argues that self-citations should be left out, because they do influence the assessment of scientific impact, especially for scientists at the start of their career, when self-citations can form a bigger share of all citations (Schreiber, 2007).

The source from which the citations are being retrieved is also of great importance towards the predictive power citations have in measuring scientific impact. The citation count data retrieved from a specific source should give a realistic representation of the real world. There are four scientific databases from which bibliometric data is mostly retrieved; Web of Science (WoS), Elseviers Scopus, Crossref and Google Scholar(GS). GS contains significantly higher citation counts than WoS and Scopus across all subject areas (Martín-Martín, Orduna-Malea, Thelwall & Delgado López-Cózar, 2018). It cannot be assumed that this higher citation count of GS, which also includes non-scientific publications like blog posts, reflects scholarly impact. All in all, there is evidence that citation data retrieved from Google Scholar surpasses data retrieved from WoS and Scopus (Martín-Martín et al., 2018) (Levine-Clark & Gil, 2009) based on its completeness. Crossref is the result of an initiative that encourages publishers to openly share citation data, it now holds about half a billion open references (van Eck, Waltman, Lariviére & Sugimoto, 2018). Besides Crossref there are two other new kids on the block, from which the first is actually a revived one, Microsoft Academic and Dimensions. Microsoft Academic Search seemed to have find its final resting place when in 2016 a new updated version was released called Microsoft Academic (MA) (Harzing & Alakangas, 2017). While being relatively new this research suggests that it has reached the level of Scopus and WoS based on its coverage, citation counts are even actually higher within the MA database. Dimensions was launched in 2018 and tries to re-imagine the way we look at citations (Dimensions, n.d.). With the use of data science links between grants, publications trials and patents are made and presented to the user. Research points out that, while being even younger than MA, Dimensions already reach the level of Scopus and WoS based on their coverage and citations counts (Harzing, 2019).

Instead of evaluating citations based on the database from which they are retrieved, an evaluation based on the type of research output (e.g., books, article, dissertation) can be made. In 2005 WoS introduced the BKCI, the book citation index, and Google introduced Google Books, which also keeps track of citations made between books. While the BKCI database contain less citations than databases with journal citation counts, research suggests that the books within the BKCI are more heterogeneous information sources (Glänzel, Thijs & Chi, 2016). Other research (Kousha, Thelwall & Rezaie, 2011) found a moderate but significant correlation between Google Books citation counts and RAE scores (The Research Assessment Exercise conducted within the UK with the goal to produce quality profiles of Research institutions).

The different sources of bibliometric data all have their own metric which is used for assessing scientific impact. Scopus has SciteScore which is used to assess journals (Scopus, 2019). Google Scholar uses the h5-index (Silver & LeSauter, 2008) and WoS uses the IF. On all websites you are able to view a rank order of all journals based on their score. The RV will also be used to rank order candidate replication studies for the purpose of picking the best study to replicate. The citation counts used here can also be used to rank order the publications and the effect of leaving in or out negative citations or self-citations can be tested. Often only the top 50 of articles is presented on their sites, this is not surprise because the top of the list is only part scientists care about when selecting the best currently available articles or journals.

The goal of this thesis is to look into the assumption that citation counts measure scientific impact, were impact is used to select which study to replicate. By empirically testing if the

removal of negative citations or self-citations changes the perceived impact of a publication this assumption is verified. Besides looking at the citation data itself, the source from which the data is retrieved should not have an influence on how articles are ranked, ideally citation counts should be constant across different sources. Based on the citation counts retrieved from WoS, Scopus, Scite and Crossref ranked ordered lists are made of the same set of articles. By correlating and comparing the rank ordered lists

The main question this research tries to answers is:

(1) Is the reliability of scientific impact in the RV significantly affected when the type or context of citation is taken into account?

We hypothesize that negative citations form a fixed percentage of the total citations for each article, meaning that articles with many citations will also have more negative citations. When the percentage of negative citations is constant the rank ordered lists will have a high correlation. The second research question is:

(2) Does the source from which citation counts are received affect the rank ordered list based on the citation count which reflects the scientific impact in the RV of publications?

We hypothesize that there is a significant high correlation between all sources.

Chapter 2

Methods

2.1 Data

The dataset that was used for this project was the original unfiltered list extracted from WoS containing the bibliometric records for 8341 articles. This dataset contained basic bibliometric data (Authors, title, field, language, type, keywords, citation count, year, etc.) and the hand coded sample size for a part of the dataset. The articles, probable to be related to social psychological fMRI research, were extracted from WoS in two different ways, based on a search term and based on the journal they were published in. The used search term (fmri AND social) resulted in 5635 articles released within the time span of 2009-2019. The search term was refined by filtering for articles only. The other 2706 articles, also released within the time span of 2009 and 2019, were extracted from four manually picked journals that covered social neuroscience within the WoS core collection. Both extractions were done on 2019-02-21.To filter all retrieved records from WoS an R script was used to exclude articles based on keywords that appear within the article. Some studies cannot be replicated based on their subject or method. Studies that require, for example, animals are not feasible to replicate and are therefore filtered out. By checking if a DOI is mentioned twice in the dataset duplicate articles have been filtered out.

The purpose of the dataset within the RV project is to help the researcher browse options for replication. Therefore the dataset contains the year, citation count, sample size and altmetric score. For the purpose of this thesis the data is brought down to containing citation classification counts for citations retrieved with Scite and total citation counts for all citation sources.

2.2 Citation classification procedure

The classification of the citation count has been done automatically with the use of a machine learned artificial intelligent algorithm called Scite. Scite combines natural language processing to evaluate the veracity of scientific work. It does so by analyzing the language around the citation based on natural language processing knowledge. Experts are then used to train the algorithm by manually determining if a citation is confirming, mentioning, contradicting or unclassified. The output of Scite consists of a list of all the citations, the text on which the ai based his classification, and the number of citations per classification. The web interface used by Scite does not allow the easy extraction of large quantities of data. Therefore, Scite has been provided with a complete list containing the DOI for all articles within the raw compete database of 8341 articles. A dataset containing the classified citation count for 5587 articles was returned due to the fact that not all articles are already in their database. For this dataset all citation counts are gathered from all other sources.

Scite is a machine learned algorithm, this means that it is a black box model therefore manually inspecting if the correct classification has been made for the citations is needed. The classifications for 20 randomly selected articles within the larger database have been checked manually with the

use of the Scite web interface (scite.ai). The web-interface displays the piece of text on which the classification is based for each citation within the article, this allowed for efficiently checking the classifications. For each citation the piece of text around the citation is checked manually and classified by the researcher, the number of citations are also compared with the number of citations in the provided dataset to check for possible inconsistencies. The Scite website gives some extra information that is not used in this dataset, for each classification a percentage is given. This percentage represents the chance of the citation belonging to a specific classification (e.g., supporting (34%), mentioning (10%), contradicting (56%)). For this example the citation would be classified as contradicting. The percentages help the researcher to spot cases that need extra attention if the percentages are almost evenly spread across two or more classifications.

The data retrieved from WoS has already been curated but has been retrieved on 21-02-2019. Therefore the citation counts are manually checked to see how much they have changed over time. The citations retrieved from Crossref and Scopus have been retrieved more recently but are still checked in order to see if the retrieval process was correct. For all sources a sample of 20 DOIs is manually checked with the use of the corresponding website of all sources.

2.3 Retrieving citation counts form different sources

The dataset we started with already contained the WoS citation count. This dataset has been merged with the dataset retrieved from Scite. This resulted in a list of 5586 DOIs with the total citation count from Scite, the classified citation counts from Scite and the citation count for WoS. For all the DOIs in this remaining list citation counts have been retrieved from Crossref and Scopus. Retrieval from both sources was done automatically by interacting with APIs for both sources in Rstudio (Version 1.2.1335). The retrieval of citation counts resulted in 55 missing values for Scopus and none for Crossref. The Scite data was retrieved on 19-06-2019, Crossef data on 08-06-2019 and the Scopus data on 09-06-2019. The WoS citations counts that were already provided within the dataset were retrieved on 21-02-2019. Afterwards all retrieved citation counts were written into the original dataset resulting in the final dataset containing the classified citation count from Scite and the total citation counts for all sources. An effort has been made to extract self-citation counts from Scopus, which is the only source that lists the number of self-citations within the metadata stated on the site. Extracting the number of self-citations was not possible because there is no API that allows the user to extract the number of self-citations.

2.4 Analysis

The goal of the analysis is to compare rank ordered lists and statistically test if there is a significant difference between how elements within the list are ranked (Webber, Moffat & Zobel, 2010). A rank correlation based approach such as Kendall Tau fits this goal, but not without some challenges. First the dis-jointness problem, which occurs when elements are apparent in list A, but not in list B. The Kendall Tau approach compares the pairs, requiring that each element in A is also apparent in B. Therefore, when an article is in the WoS ranked list, but not in the Scite ranked list the Kendall Tau approach cannot be used. The second problem of top-weightiness is more relevant towards this thesis. With the Kendall Tau approach a difference within a pair at the bottom of the rank weights as much towards the score as a difference within a pair at the top of the list. An article that moves from the second place to the seventh place should have a much larger influence on the correlation score then an article that moves from the 70th place to the 75th place. Articles at the top of the list have a much higher citation count and therefore need a lot more extra citations for it to move up a place while articles at the bottom of the list only need a few citations to move up. there is only one article with 515 citations on WoS within the dataset used in this thesis while there are 125 articles with 17 citations on WoS. This means that an article with 17 citations can be placed 2303 or 2427 while having the same amount of citations. When the article with 17 citations gets two additional citations it will move up 200 places in the

ranking while the article with 515 citations would not move up at all.

A weighted overlap measure, Rank Biased Overlap (RBO) solves the problem of dis-jointness and top-weightiness. This measure is based on a model which compares the overlap of the two rankings at incrementally increasing depths (Webber et al., 2010). The RBO is calculated as the expected average overlap that is observed when comparing two ranks, the higher an element is in the rank, the higher it is weighted within the formula (1), the weights given to the elements in the list decrease geometrically, but never reach zero. The RBO falls in the range [0,1] where 1 means the lists are identical and 0 means that the lists are disjoint (do not overlap at all).

$$\mathrm{RBO}(S,T,p) = (1-p)\sum_{d=1}^{\infty} p^{d-1}\cdot A_d\ . \tag{1}$$

In this formula S and T are the two rankings, d is the position or depth within the ranking of an element A and p represents the initial weight given to the element A, which is dependent on the position of the element within the ranking. The value for p is decided by the experimenter, the higher the value for p, the steeper the curve that represents the weight given to elements within the ranking. With a low p the power decreases slowly with depth while with a high p the power decreases fast with depth, with a P of 0 only the number one ranked articles are compared resulting in a RBO that is either 0 or 1, with a P of 1 all articles are taken into account and all articles are given an equal weight. The second parameter is the depth (D), with a depth of 100 only the first 100 cases are taken into account, when the depth is set to infinity, all articles within the list are taken into account. To illustrate the effect of P, P = 0.9 means that the first 10 ranked articles have 86% of the weight while P=0.98 means that the first 50 ranked articles have 86% of the weight. To determine which P to use formula 2 can be used.

$$W_{\mathrm{RBO}}(1:d) = 1 - p^{d-1} + \frac{1-p}{p}\cdot d\cdot\left(\ln\frac{1}{1-p} - \sum_{i=1}^{d-1}\frac{p^i}{i}\right)\ . \tag{2}$$

First Kendalls Tau was calculated for all possible pairs of sources followed by the correlation between the total Scite citation counts and the Scite citation counts without the contradicting citations. Correlations were calculated in Rstudio, the input consisted of the list of raw citation counts for two different sources. The raw citation counts are then automatically ranked and compared resulting in the overall correlation between the two ranked lists. We aim to achieve a significance level below 0.01 and the sample size should be high enough for detecting a small correlation (τ =0.1), the width of the confidence interval is 0.1. The minimal sample size should be 1139 (Bonett & Wright, 2000), therefore N = 5586 will suffice.

The RBO presumes a strong relationship between the rankings therefore making it hard to test for statistical significance (Webber et al., 2010). If the desired statistical significance cannot be determined, determining the sample size a priori also becomes problematic. By comparing the sample size used in this analysis with the sample size used in other articles that apply the RBO a justification of the sample size in this thesis can be made. In the original article that proposes the RBO by Webber (2010) search queries for different search engines are compared. At least the first 10 results for 113 different queries are compared. Search results relate to citation counts in our dataset and queries to the source of the citation counts. The sample size of 5586 exceeds the minimal sample size of 10 used by Webber et al. (2010) significantly.

To calculate the RBO for each pair of sources the data had to be prepared. For the RBO algorithm to work properly two lists both containing only unique values is required. The algorithm compares the position of one unique value within the first list with the position of the same unique value in the second list. To turn the raw citation counts into this format all DOIs were numbered

and then sorted based on citation counts. The lists of numbers linked to the DOIs served as the input for the RBO. The RBO algorithm does not require both lists to be of equal lengths, therefore the cases with missing data were left in the dataset.

Chapter 3

Results

For the citations within the Scite citation data the median has a depth of 9 (IQR=17), the depths of the median for mentioning, supporting and contradicting citations are 8 (IQR=16), 1 (IQR=2) and 0 (IQR=0) respectively. The percentages of the classified Scite citations is given in figure 3.1. For the citations within the WoS citation data the median has a depth of 14 (IQR=23), 16 (IQR=23) for Crossref and 17 (IQR=25) for Scopus. A density plot for each different classification or citation source can be seen in figure 3.2 until **??**, all density plots show that citation counts are highly skewed towards zero, all citation counts are not normally distributed.

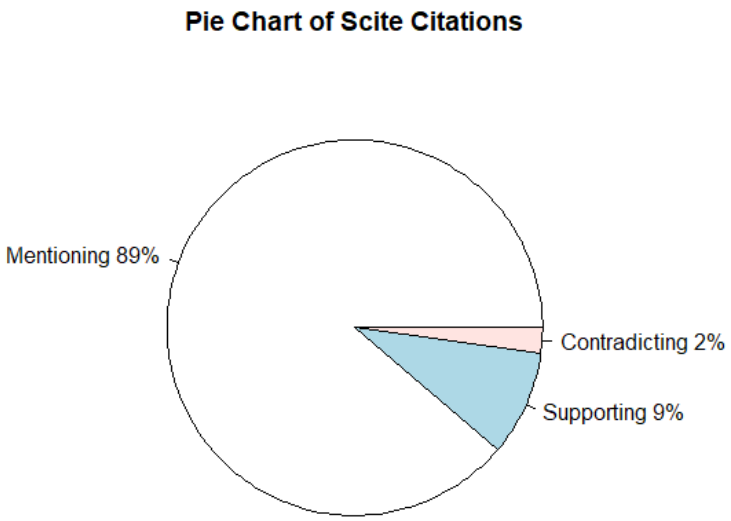

Figure 3.1: Pie chart of the classified citation counts retrieved from Scite

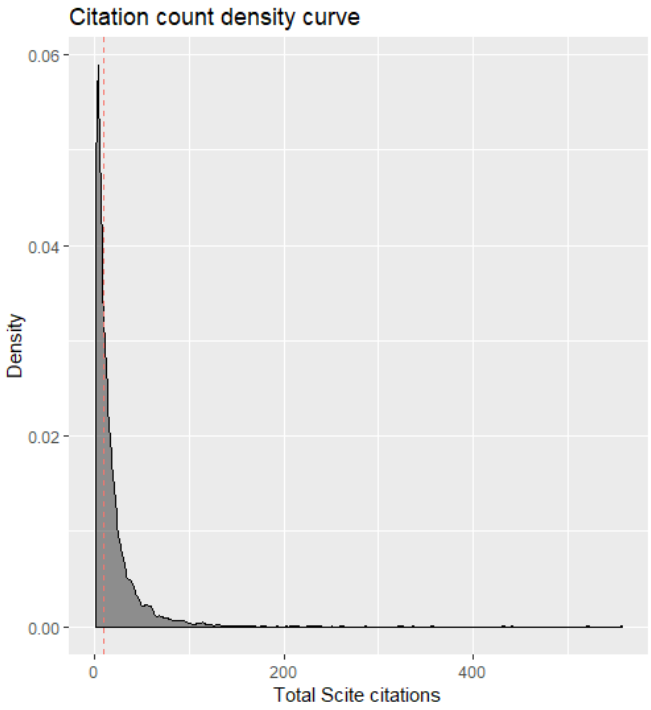

Figure 3.2: Citation count density curve for total Scite citations

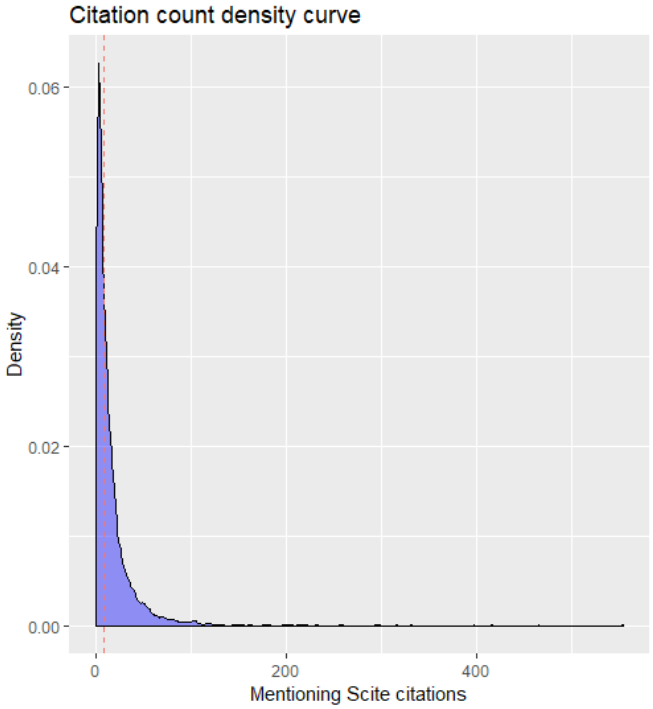

Figure 3.3: Citation count density curve for mentioning Scite citations

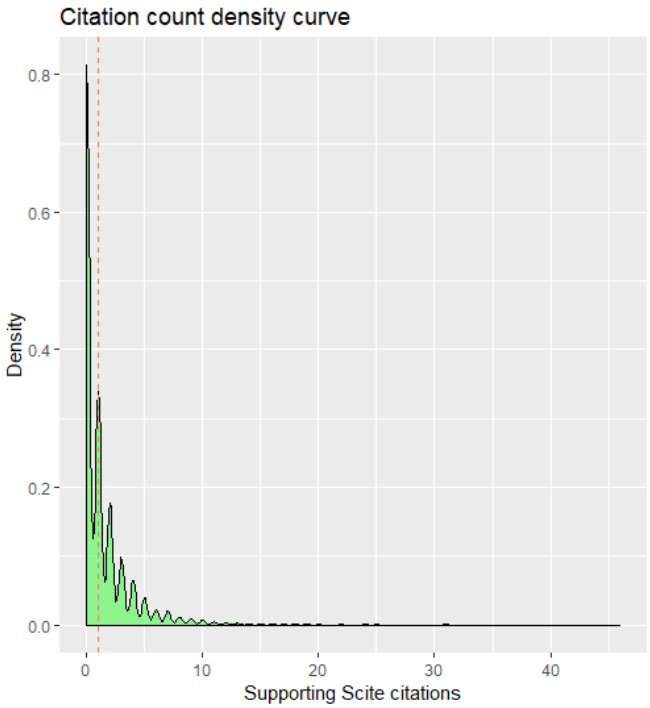

Figure 3.4: Citation count density curve for supporting Scite citations

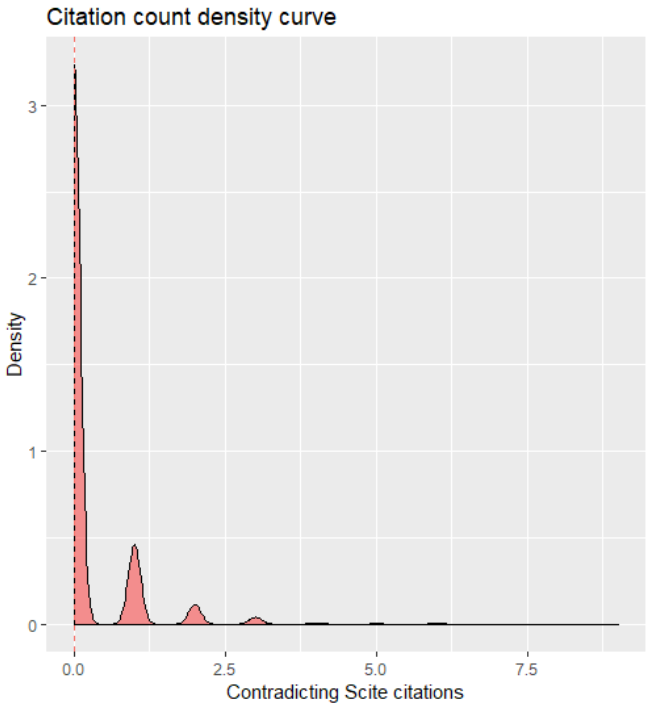

Figure 3.5: Citation count density curve for contradicting Scite citations

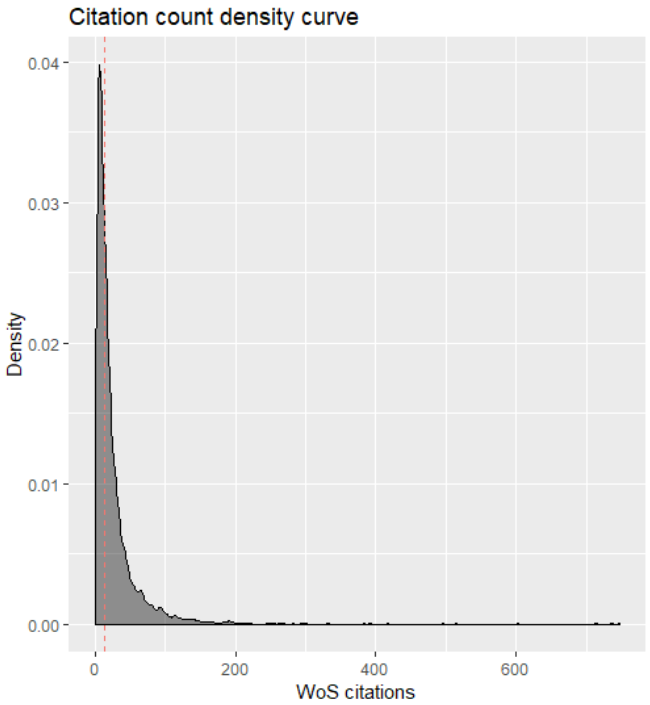

Figure 3.6: Citation count density curve for WoS citations

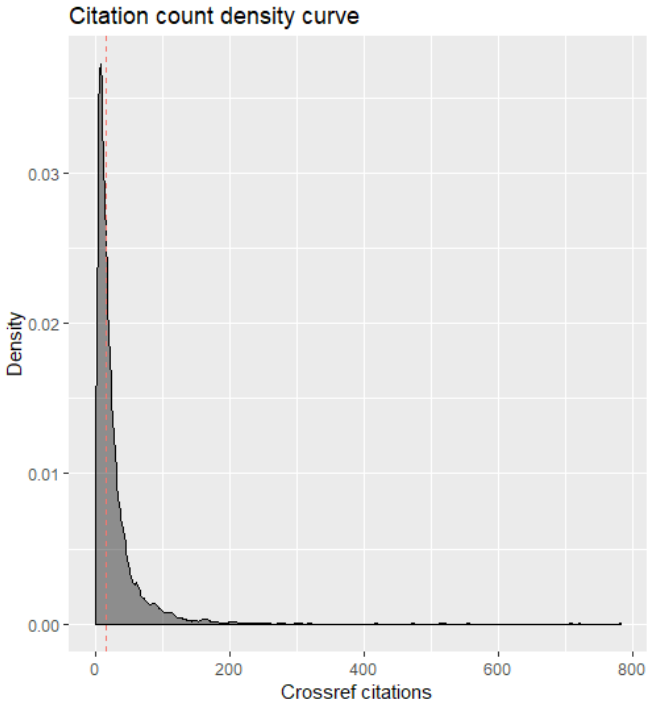

Figure 3.7: Citation count density curve for Crossref citations

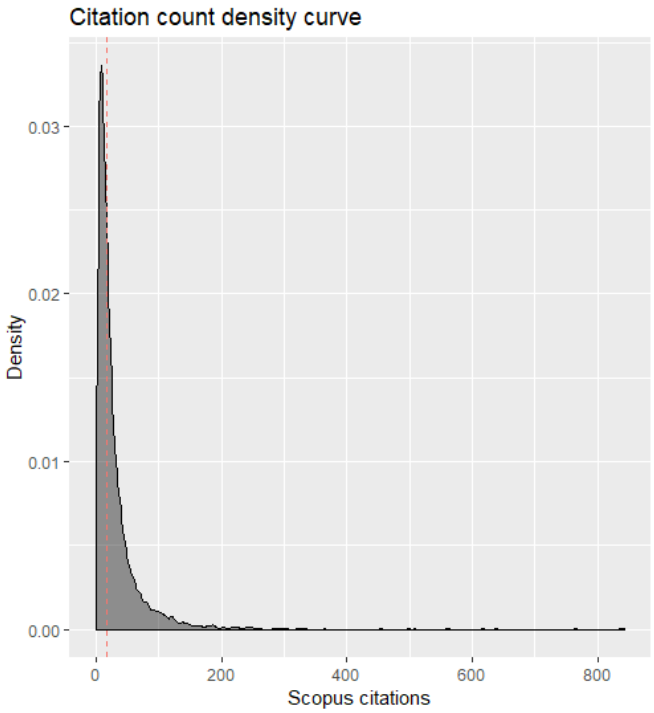

Figure 3.8: Citation count density curve for Scopus citations

Kendalls Tau was calculated for all possible pairs of citation count sources and for the Scite citation count with and without contradicting citations, the results can be seen in figure 3.9. The assumptions of Linearity and homoscedasticity, and normally distributed variables are all violated by the data as can be seen in figure 3.2 until 3.8. therefore Pearsons or Spearmans correlation could not be calculated. Kendalls Tau does not rely on these assumptions, all correlations (tau-b, $p < 0.001$) are represented in figure 3.9. The only assumptions that underly Kendalls Tau are that the data is ordinal or continuous and the data should follow a monotonic relationship. Both assumption are met by the data. There is a fairly strong correlation between WoS, Crossref and Scopus and a moderate correlation between Scite and all other sources. This means that when a publication is ranked high based on its citation count in one source, changes are strong that article is also ranked high in the other source.

The RBO requires, besides two lists of unique values, a depth (D) and a weight (P) as input. Depth D is set to infinity, meaning that the RBO focuses on all the articles within the ranking. While the depth is kept at a constant value the BSO is calculated for different weights (P=0.90, 0.98, 1.00). The measure with P=1.00 is done with the purpose to see what the average overlap is between the sources when all articles are given an equal weight independent of their position within the ranking. This shows if giving weight to the top articles is having an effect on the RBO. The results for the RBO are given in figure 3.10 until 3.12.

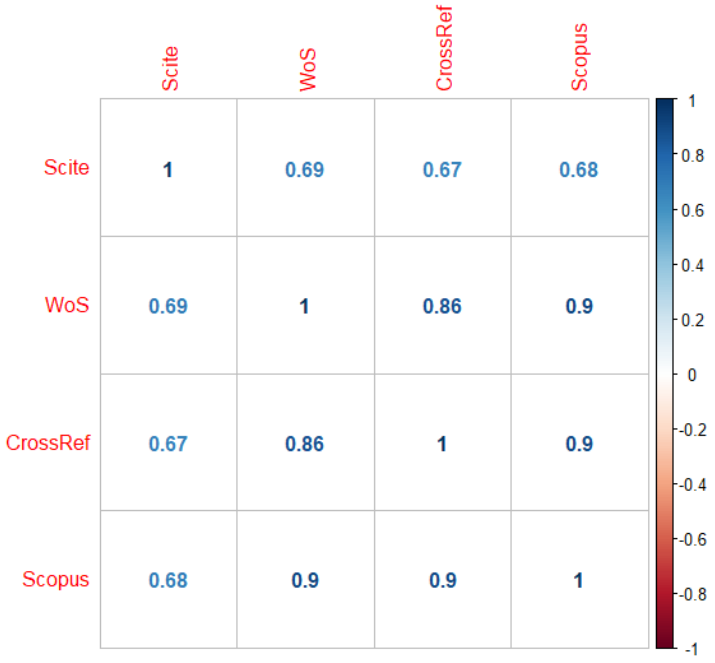

Figure 3.9: Matrix with all Kendalls tau correlation coefficients

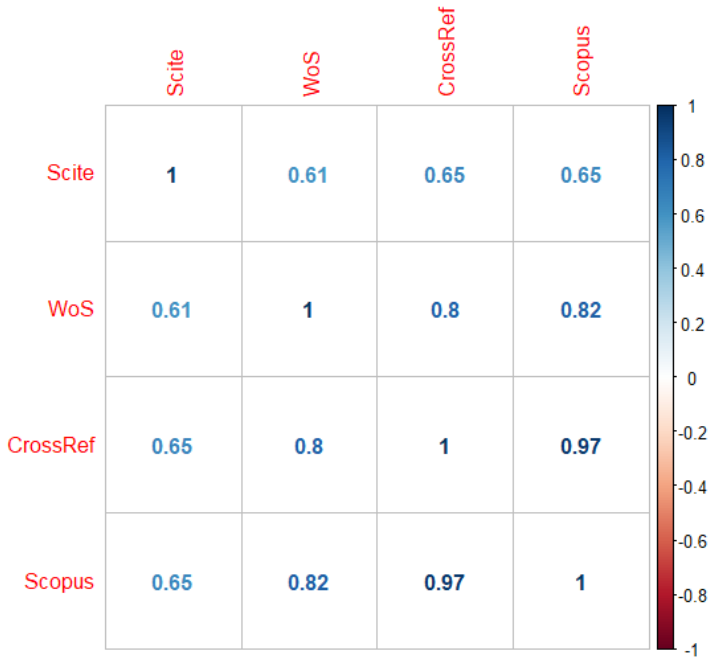

Figure 3.10: Matrix with all RBO overlap coefficients (P=0.90, D=∞)

Figure 3.11: Matrix with all RBO overlap coefficients (P=0.98, D=∞)

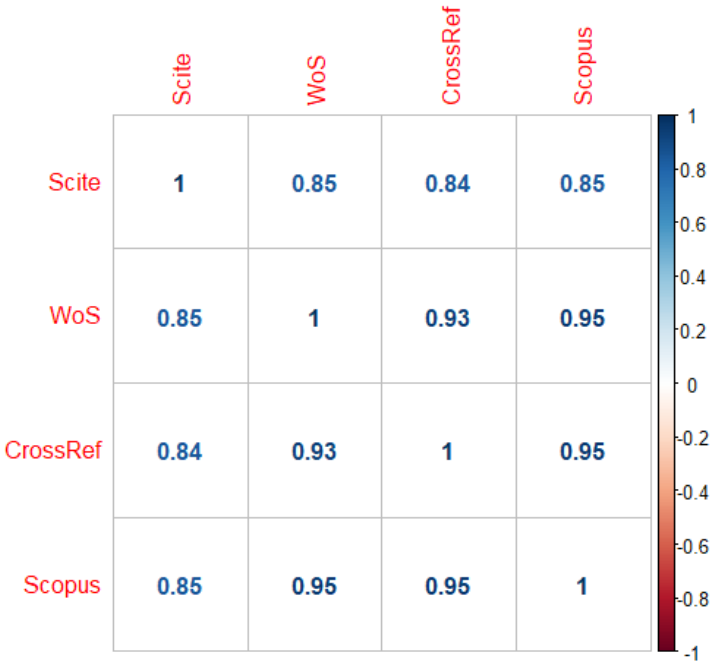

Figure 3.12: Matrix with all RBO overlap coefficients (P=1.00, D=∞)

Looking at how all RBO coefficients change when less weight is given to the top of the rank, we see immediately that the correlations get stronger except for the correlation between Crossref and Scopus, the correlation even gets weaker. When we compare Figure 3.9 with Figure 3.12 (RBO (P=1.0, D=inf.)) we can see how Kendalls tau relates to the RBO. Values for the RBO are generally closer to 1 but overall high correlations relate to high RBO values as we would expect. To get a better feeling of how the weight (P) given to top elements has an effect on the RBO, the RBO is plotted against different values for the weight (P), this can be seen in figure 3.13.

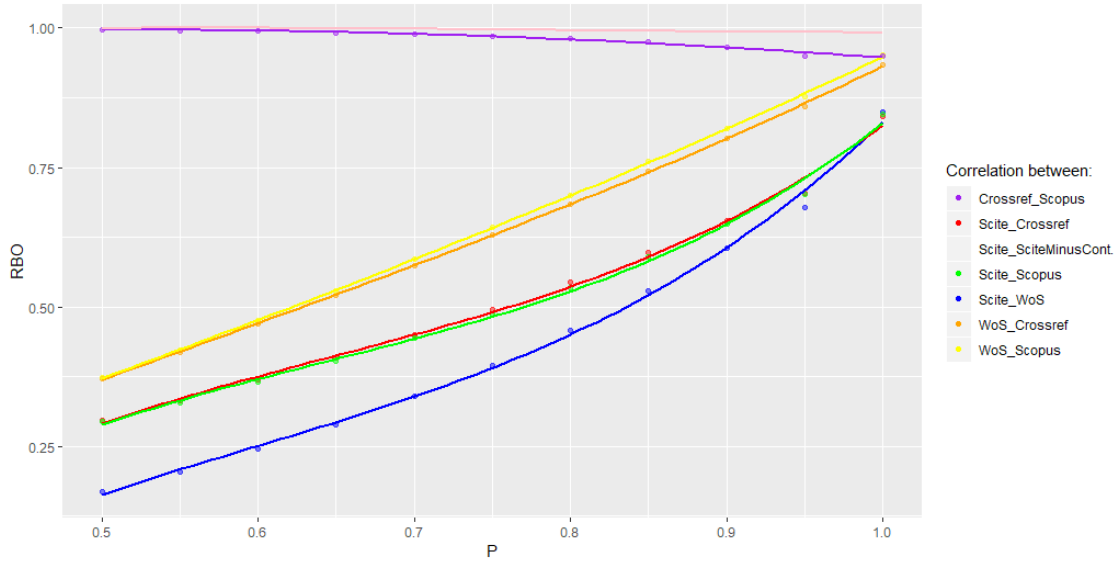

Figure 3.13: The RBO for different sources for different weights (P), (D=∞)

Figure 3.13 shows that the overlap between the pairs of ranked lists becomes stronger when the value for the weight (P) increases. When the value for the weight (P) increases the weight given to elements at the top of the list decreases, for weight (P=1.0) all articles get evenly weighted, and no extra weight is given to items at the top of the list. In figure 3.13 the depth is set to infinity, this is done to show which effect the weight (P) has on the overlap independent of the depth (D). The overlap between the total Scite citations (Scite) and the Scite citations minus the contradicting citations (SciteMinusCont.) does not seem to be affected by the weight (P). Again it is illustrated that the overlap between the rank ordered Crossref citations and Scopus citations increases when P increases. The overlap between Scite and all other sources are the lowest across all values for P, with a higher P the rate to which they overlap does get stronger. In other words, the top of ranked lists differ more than that they differ overall, the opposite is true for the top of the Crossref and Scopus ranked lists. The overlap between WoS and Crossref and Wos and Scopus follow a straight line for a change in weight (P) while the Scite overlap follows a curved line.

Kendall's Tau	0.9937
RBO (P=0.90, D=∞)	0.9929
RBO (P=0.98, D=∞)	0.9900
RBO (P=1.00, D=∞)	0.9924

Table 3.1: Correlations between all Scite citations and Scite citations minus the contradicting citations

Table 3.1 shows the Kendalls Tau correlation coefficients and the RBO scores for the ranked lists with all Scite citations (Scite) and the Scite citations minus the contradicting citations

(SciteMinusCont.). Independent of which test or which weight (P) is given the lists seem to be very similar and close to 1. With only 2% (fig. 1) of the citations being refuting this is to be expected. When taking a closer look into the contradicting citations there are only 18 articles that have 50% or more of their citations being contradicting from which none appear in the top 100. The article with the most negative citations (9 out of 521) (Lamm, Decety & Singer, 2011) appears on the third place within the ranking based on Scite citation counts. When the negative citations are left out, this article still appears on the third place.

Kendalls Tau for the correlation between the supporting and contradicting citations within the Scite data is 0.30 (P 0.001). This means that there is a small correlation between the supporting and contradicting citations while these values are expected to be discordant.

3.1 Data Validation and Qualification

A sample of 20 articles was randomly selected from the complete list of DOIs for manual classification. The citations counts that were generated by the Scite program and the citation counts resulting from manual classification are given in table 3.2. The manual check was performed on 19-06-2019, no differences were spotted between the data in the provided database and the data retrieved from the site.

	Provided Data				Manually checked				Retrieved online				error
DOI	T	M	S	C	T	M	S	C	T	M	S	C	
10.1093/scan/nst098	17	11	5	1	17	11	6	0	17	11	5	1	m
10.1016/j.ijpsycho.2012.02.011	2	2	0	0	2	2	0	0	2	2	0	0	
10.1093/scan/nsw031	8	8	0	0	4	4	0	0	4	4	0	0	
10.1093/ijnp/pyy068	3	1	0	2	3	1	0	2	3	1	0	2	
10.1093/scan/nsx042	2	2	0	0	2	2	0	0	2	2	0	0	
10.1093/cercor/bhp149	59	52	6	1	59	47	10	2	59	52	6	1	m
10.1016/j.cub.2015.06.043	19	18	1	0	19	17	2	0	19	18	1	0	m
10.1016/j.bbr.2015.04.047	2	1	1	0	2	2	0	0	2	1	1	0	m
10.1093/scan/nsu003	11	8	2	1	11	7	3	1	11	8	2	1	m
10.1093/scan/nsp006	16	15	1	0	16	15	1	0	16	15	1	0	
10.1037/a0024325	8	7	1	0	8	7	1	0	8	7	1	0	
10.1098/rstb.2017.0136	2	2	0	0	2	2	0	0	2	2	0	0	
10.1037/a0014121	11	10	1	0	11	7	4	0	11	10	1	0	m
10.1371/journal.pone.0054400	4	4	0	0	4	4	0	0	4	4	0	0	
10.1162/jocn.2010.21551	36	34	2	0	36	31	4	1	36	34	2	0	m
10.1093/scan/nst005	15	12	2	1	15	12	3	0	15	12	2	1	m
10.1080/17470919.2014.925503	5	4	1	0	5	5	0	0	5	4	1	0	m
10.1111/tops.12132	3	3	0	0	3	3	0	0	3	3	0	0	
10.1037/a0034154	6	6	0	0	6	5	1	0	6	6	0	0	m
10.1093/scan/nsq040	5	5	0	0	5	5	0	0	5	5	0	0	

Table 3.2: Citation counts provided in the data, checked manually and retrieved from scite.ai (T = total citations, M = mentioning citations, S = supporting citations and C = contradicting citations)

There is only one kind of error found by manually checking a sample the provided data. By manually reading the piece of text used by the algorithm to classify the citation 19 citations (out of a total of 230) are found that are misclassified. A misclassified case is a type 1 error when a mentioning citation is wrongly classified as a supporting or contradicting citation, in this case the algorithm spotted words that are indicators for supporting or contradicting while this is not the case. A type 2 error is made by the algorithm when a supporting or contradicting citation is classified as mentioning, in this case the algorithm missed the words that signal the citation to

be supporting or contradicting. A third error can also be made, in this case a supporting citation is seen as a contradicting citation or the other way around. For this type of error the algorithm wrongly interpreted a signaling word and missed a signaling word. Some examples of misclassified citations are given below, in table 3.2 these cases are indicated with an m.

Example of supporting citation that is classified as mentioning (Tune, Schlesewsky, Nagels, Small & Bornkessel-Schlesewsky, 2016):

The posterior MTG is connected to the angular gyrus via short fibers of the indirect segment of the AF and to anterior parts of the superior temporal lobes via the MLF (Buckner et al 2009; Turken and Dronkers 2011). The functional relevance of these connections as part of a neural circuit underlying conceptual processing is supported by the results of recent studies examining functional connectivity during task and rest, as well as by detailed meta-analyses (Simmons et al 2010; de Zubicaray et al 2011; Visser et al 2012; Noonan et al 2013; Hurley et al 2015). The fact that the discussed regions in the temporal and inferior parietal lobule appear to be functionally connected and show similar response profiles to semantic manipulations, however, does not entail that their specific contributions to the processing of semantic information are the same.

Example of a correctly classified supporting citation (Lamm et al., 2011):

Damage to the left temporal pole was associated with impaired naming of Faces specifically, consistent with studies using a variety of approaches (Damasio et al, 1996; Damasio et al, 2004; Tranel et al, 2009; Simmons et al, 2010) and stimulus materials (Belfi et al, 2014; Waldron et al, 2014). The requirement of additional anatomic substrate(s) for normal retrieval of proper names of faces is consistent with proposals that their retrieval entails additional process(es) (cf.

Example of a supporting citation that is classified as contradicting (Zhong, Chark, Hsu & Chew, 2016):

The vmPFC is anatomically and functionally well suited to play this role, as it projects to several brain areas that are heavily involved in reward valuation, preference generation, and decision-making (Behrens et al, 2009; Hare et al, 2010; Hare et al, 2009; Hare et al, 2008; Rangel et al, 2008). Our findings regarding the vmPFC also echo those of previous studies in which investigators, using different paradigms, reported data suggesting that activations in a neural network including the vmPFC positively reinforce social rewards (Hare et al, 2008; Li et al, 2009; ODoherty et al, 2004; Padoa-Schioppa, 2007; Tabibnia et al, 2008; Tricomi et al, 2010; Zaki et al, 2013; Zaki and Mitchell, 2011), and more generally social cognition (Gusnard et al, 2001; Miller and Cohen, 2001; Saxe, 2006).

To verify how much the data retrieved from WoS has changed over time and to check the reliability of the retrieval methods via automated scripts using APIs, data for 20 DOIs was retrieved manually on 18-06-2019 and compared with the data retrieved earlier. Table 3.3 shows the citation counts for all three sources retrieved two different dates, all number that have changed are marked. The citation counts that have changed all increased with a few citations, especially the WoS citations due to the fact that they have been retrieved almost half a year ago. This data does not indicate that the retrieval methods used for all sources was faulty.

	Automated retrieval			Manual Retrieval		
DOI	WoS	Crossref	Scopus	WoS	Crossref	Scopus
10.1038/s41598-017-01460-6	3	4	3	4	4	3
10.3389/fnagi.2016.00021	3	6	7	5	6	7
10.1080/17470919.2015.1040556	9	9	10	9	9	10
10.1037/a0015396	7	5	7	7	5	7
10.1016/j.dcn.2012.11.001	55	59	62	59	59	62
10.7554/eLife.38293	0	0	2	2	0	2
10.1073/pnas.1009164107	147	164	178	155	164	178
10.1038/s41598-017-10310-4	1	2	3	2	2	3
10.1080/17470919.2015.1037463	1	2	2	2	2	2
10.1016/j.biopsych.2012.03.027	83	91	97	87	92	99
10.1037/bne0000218	0	3	1	0	3	1
10.1037/xlm0000446	0	0	1	1	0	1
10.1111/nyas.12204	18	18	17	18	19	17
10.1007/s12038-015-9509-5	8	9	8	10	9	9
10.1038/nn.3781	33	38	36	34	38	36
10.1371/journal.pone.0206351	0	1	1	1	1	1
10.1007/s00221-010-2277-4	23	28	26	23	28	26
10.1037/bul0000073	22	29	31	25	29	31
10.7358/neur-2016-019-balc	0	0	0	0	0	0
10.1037/a0024403	17	15	17	17	15	17

Table 3.3: Data retrieved automatically from WoS (21-02-2019), Crossref (08-06-2019), Scopus (09-06-2019) and retrieved manually on 18-06-2019

Chapter 4

Discussion

The main goal of this thesis was to look into citation counts as a means to quantify scientific impact. This measure of impact is used to determine a replication value (RV) which therefore relies on these citation counts. Two questions were asked that underlie the goal of the thesis. The first question, Is the reliability of scientific impact in the RV significantly affected when the type or context of citation is taken into account?. The second question "Does the source from which citation counts are received affect the rank ordered list based on the citation count which reflects the scientific impact in the RV of publications?". Using the results an answer will be provided to these questions. At the start of this thesis a third question has also been addressed, whether leaving in or out self-citations has a significant effect on the RV. Using the literature an answers to this third question will also be provided here.

Before answering these question it is relevant to start with discussing the data validation, because the answers to the substantive questions rely heavily on the reliability of the data. As mentioned there is one main issue with the data provided by Scite. by manually checking the text on which the classification is based it was discovered that the algorithm is not always making the right classification. Most commonly confirming or contradicting citations were classified as mentioning which is a type II error, the algorithm just missed the signaling words classifying it as mentioning. Type I errors were spotted less frequently but were still there, meaning that there are cases for which a mentioning citation is classified as either confirming or contradicting. There were even cases found during manual checking which were classified as contradicting even though they were actually were confirming. 19 out of the 230 citations were wrongly classified based on the judgement of the researcher. Ideally the whole dataset is checked manually by a researcher common to the field to which the papers belong. Manually coding the citations for 20 cases took approximately three hours, for the whole dataset this would therefore take approximately 1100 hours (138 working days of 8 hours). Manually coding is extremely time consuming and therefore not feasible. Even when an experienced researcher would manually code the data mistakes will be made due to human error.

All in all, the concept of Scite can bring much value towards the scientific world, a tool that categorizes citations in a basic way distinguishing between mentioning, contradicting and supporting can help a scientist better interpret citation counts. There are other deep learning models (Hassan, Imran, Iqbal, Aljohani & Nawaz, 2018) which analyze citations using language processing reaching a similar accuracy (91% is the best result stated in the article) as the Scite model but those models only give a level of importance, which is less valuable than a classification. Based on our data Scite reaches an accuracy of 92% percent. The models mentioned in Hassan et al. (2018) measure something different than the model used by Scite, therefore relating the accuracy percentages is illogical. When Scite is able to match the number of articles held within their database to WoS or Scopus, and increase accuracy it has the capability to outvalue other sources of citations.

The issue concerning the misclassification of citations revolved around the Scite citation data, but there other issues with the data that should be looked into. Primarily the fact that not all

data was retrieved within a short time period. Citation counts increase over time, in an ideal situation all data is gathered on exactly the same moment such that time can have no effect on the data. A difference of a week is reasonable, but half a year is too much. This is illustrated by table 4 as one can easily see that the citation counts of WoS have increased relatively more than the citation counts of Crossref and Scopus.

Based on the data, with the discussion of the data validity in mind, an attempt can be made to answer the questions stated in the introduction. The correlation between the rank ordered lists based on the total number of Scite citations and the number of total Scite citations minus the contradicting citations shows how similar both lists are. A Kendalls tau correlation of 1 would have indicated that the lists are completely identical thus taking out contradicting citations would have had no effect. A Kendalls tau of -1 would indicate that the lists are completely different, meaning that taking out contradicting citations had completely turned the ranked list around. The value for Kendalls tau that was found in our data is 0.99 ($P < 0.001$). This correlation indicates that taking out contradicting citations has very little effect on the rank order meaning that there are not many articles that would change from position within the rank order.

A second analysis (RBO) was used to measure the overlap between pairs of rank ordered lists. This measure is able to give more weight to articles at the top of the list compared to articles further down the list. The RBO for the total citations with and without the contradicting citations is close to independent of the weight used in the calculation of the RBO. The RBO stays within the bounds of 0.98 and 0.99 ($0.5 < \text{weight} < 1.0$) with depth set to infinity. Again this is an indication that taking out refuting citations would not affect the rank order of the articles based on their citation count.

The second question concerns the comparison of the different citation count sources. Kendalls tau has been determined for all possible correlations between WoS, Crossref, Scopus and Scite. The correlations give an indication of how the citation sources compare to each other. In an ideal world all correlations would be 1 indicating a perfect match between the rank ordered lists of all sources. By comparing different sources this thesis tried to answer the question whether it matters which source is used for the retrieval of citation counts that will be used to assess the scientific impact used in the RV. If all correlations would have been 1 it would not have mattered which sources is being used while a correlation of -1 indicates that two sources could not be more different. Based on the results we can only determine how well the citation sources relate to each other, the citation source that has the highest overall correlation and overlap with the other citation sources

Figure 2 gives Kendalls tau for each possible correlation. Scite has a relatively weak correlation(τ = +-0.69) with all other sources. The errors mentioned earlier here might have caused the low correlations but the most likely cause is the amount of articles within the Scite database compared to the amount of articles within other databases. While Scite only contains around 200 million citations (Scite, n.d.), Scopus has around 1.4 billion citations (Scopus, n.d.) within their database, for WoS a similar number of 1.4 billion (Matthews, n.d.) and Crossref contains a little over 900 million citations (Crossref, n.d.). While almost all articles are within the databases of all sources, Scite is possibly missing a lot of articles that are citing articles that are within their database. The depth of the median for the citation count for articles in Scite (9) is lower than the depth of the median for WoS (14), Crossref (16 and Scopus (17). While most other papers report percentages of overlap between different sources and not average citation counts, these numbers are in line with the research already done in this field (Martín-Martín et al., 2018) (Harzing, 2019). There is no other research done that compares Scite as a citation source with the other mentioned sources, therefore a validation by comparison for these results cannot be made. The RBO is also determined for the overlap between Scite and the other sources which sketches a similar view as Kendalls tau. The RBO analysis does give some extra insights into how well citation counts from different sources overlap. As can be seen in figure 4 all the correlations with Scite are affected most by giving more weight to the top of the list, the part of the list we are interested in because difference at the top should account more towards the RBO than differences half-way down the list. Giving more weight to articles listed at the top of the list drastically decreases the overlap between the measured sources. Based on the substantial difference between Scite and all other

citation sources, both in terms of average citation scores, coverage, and rank-ordering of included records (based on the correlation and the overlap), Scite is not yet the best choice as a citation source for determining the scientific impact used in the RV.

The values for the correlation between WoS and Crossref (τ=0.86) and WoS and Scopus (τ=0.90) already indicate a better overlap between the ranked lists based on the citation counts retrieved from both sources. The correlation between Scopus and WoS is of equal strength (τ=0.90). Overall Scopus therefore has the highest correlation coefficients with all other sources except with Scite (the WoS / Scite correlation is 0.69). Independent of the weight Scopus also has the highest RBO between itself and the other sources. Especially the overlap score between Crossref and Scopus is interesting, it is the only RBO that increases when more weight is given to the top of the list. This indicates that there is more overlap between the articles at the top than throughout the whole ranked list. The overall high values for Kendalls tau and the RBO, especially the RBO between Crossref and Scopus, indicate that Scopus is the most suitable source of citation counts which can be used in determining the RV.

The effect of removing self-citations from the total citation count could not be tested. All sources evaluated in this thesis have no easy way of extracting the number of self-citations. Scopus does allow the user to filter out self-citations on their website but it is not possible to extract the number with automated extraction. This means that the only way of extracting self-citations is performing manual extraction. While this would be possible for a small data set (100), it is not feasible for large datasets as used in this thesis. To goal of the RV is to efficiently bring down a large dataset of articles to a smaller number of articles that can be checked manually. Manually extracting the self-citations for large datasets of around 1000 articles can be done within a reasonable amount of time but it would not be efficient.

Research suggests that self-citations can have a significant impact on career perspective for starting scientists (Ken, 2003) and form a significant part off the overall citations in some cases (Seeber, Cattaneo, Meoli & Malighetti, 2019) (Aksnes, 2003). According to the article by Aksnes (2003) the percentage of self-citations reaches as high as 31% in some fields. Spread over all different fields percentages lie around 20%. When at least one author (first author or co-author) is also an author (first author or co-author) of the citing paper, the citation is defined as a self-citation by Aksnes (2003). The article by Seeber et al. (2019) states that scientists that are more likely to benefit from self-citations also have an increased number of self-citations. When a scientist inflates his own citation count, his impact factor (IF) will also rise, meaning that the article is perceived to have a higher impact within the scientific world. When a self-citation is made truthfully to ones own article it can contribute towards the scientific field, but this correlation between the opportunity to benefit, and the increase of self-citation signals something else. Therefore it would be very convenient for citation databases to make it easy to exclude self-citations and allow scientists to extract the numbers via an API or other ways. Research has been conducted on the effect of self-citations but the effect of removing self-citations and measuring the effect it has on how articles rank has not been done. The question about the effect of removing self-citations from total citation counts and the effect the removal has on rank ordered lists will remain unanswered.

Another subject that is worth discussing is the added value of the BSO towards the analysis of the data. The top of the ranked lists that are compared here are of much greater value then the rest of the list for multiple reasons. Databases like WoS use citation counts to determine the IF for articles, scientists and journals. On their site they display the top of ranked order lists to show which journal or scientist is currently having the most impact within the scientific world. Clearly the most important cases reside in the top the list while the less important cases are somewhere down the list. There is a great difference between rising from the 5th to the 3rd and rising from the 152nd place to the 150th place. These differences would have an equal impact on Kendalls tau but not on the RBO. The same mechanism and reasoning apply to this thesis, and towards the goal of the RV which tries to quantify how much value replicating a specific study can bring towards the scientific field. Looking at the data, and especially at figure 4, it becomes clear that weighing top articles can significantly change the level of overlap between two ranked lists. The RBO between WoS and Scite increases from 0.16, a very weak overlap, to 0.83, a strong overlap, when more weight is given to the top. In other words the more emphasis is put on the articles at

the top of the list the lower the overlap between WoS and Scite becomes. There is also a downside to the RBO, the fact that it is not commonly used within the bibliometric field (there is no case found to the best of our knowledge). Kendalls tau is commonly used and can be used in null hypothesis significance testing (NHST), the most common way of rejecting a hypothesis within the field of statistics. It is important to be able to check beforehand which sample size is needed to reach a certain power with a certain level of significance and to check afterwards if the test is statistically significant, This is not yet possible for the RBO.

GS, MA and Dimensions are mentioned in the introduction as other sources from which citation data can be retrieved, but are not evaluated within this Thesis. GS is not used due to the fact that it does not allow for easy data extraction, unlike Scopus, Crossref or WoS they do not offer APIs or other mechanisms that allow the extraction of citation counts and metadata for a large batch of data. The only way to extract data from GS is by interacting with the website (Martín-Martín et al., 2018). The same goes for Dimensions, extracting data is possible but not yet publicly available (Harzing, 2019). MA does allow the extraction of metadata and citation counts via their AK API (Hug, Ochsner & Brändle, 2017) and therefore should have been included in this research. This thesis does not cover MA simply because it did not pop up during the literature research beforehand, and was only discovered after the data extraction form all other sources.

All in all, this study tried to take a closer look into the quantification of scientific impact with the use of citations. These citation counts can make or brake the career of a scientist and can determine where funding gos and where not, therefore being of great importance. Based on the presented data we cannot state that the context of a citation has an effect on the scientific impact. The development and enrichment of databases like Scite can possible prove otherwise in the future, until then they can bring value by encouraging scientists to take a closer look at citation data. The data does show that the citation databases that have the most data are more likely to give a good representation of reality. By enriching our citation databases, broadening the available information about citations, and making all information openly available, citation counts can be shown to be a robust measure, or actually no good measure for scientific impact at all.

References

Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics*, *56*(2), 235–246. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-0742310636{&}doi=10.1023{%}2FA{%}3A1021919228368{&}partnerID=40{&}md5=3f23dff5ba871de7ba6d33ee06108561 doi: 10.1023/A:1021919228368

Bloch, S. & Walter, G. (2001, oct). The Impact Factor: time for change. *Australian and New Zealand Journal of Psychiatry*, *35*(5), 563–568. Retrieved from http://www.embase.com/search/results?subaction=viewrecord{&}from=export{&}id=L32982603{%}0Ahttp://dx.doi.org/10.1046/j.1440-1614.2001.00918.xhttp://www.blackwell-synergy.com/links/doi/10.1046{%}2Fj.1440-1614.2001.00918.x doi: 10.1046/j.1440-1614.2001.00918.x

Bonett, D. G. & Wright, T. A. (2000, mar). Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, *65*(1), 23–28. Retrieved from http://link.springer.com/10.1007/BF02294183 doi: 10.1007/BF02294183

Bornmann, L. & Daniel, H.-D. (2008, jan). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80. Retrieved from https://www.emeraldinsight.com/doi/10.1108/00220410810844150 doi: 10.1108/00220410810844150

Cronin, B. & Overfelt, K. (1994). Citationbased auditing of academic performance. *Journal of the American Society for Information Science*, *45*(2), 61–72. doi: 10.1002/(SICI)1097-4571(199403)45:2⟨61::AID-ASI1⟩3.0.CO;2-F

Crossref. (n.d.). *crossref.org : : status.* Retrieved 2019-06-19, from https://data.crossref.org/reports/statusReport.html

Dimensions. (n.d.). *Dimensions.* Retrieved 2019-06-20, from https://www.dimensions.ai/

Garfield, E. (1955). Citation Indexes for Science. *Science*, *122*(3159), 108–111. Retrieved from http://www.jstor.org/stable/1749965

Garfield, E. (1962). Can Citation Indexing Be Automated? *Essays of an Information Scientist*, *1*, 84–90.

Garfield, E. (1970). Citation Indexing for Studying Science. , *227*, 669–671.

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation Author (s): Eugene Garfield Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/1735096 REFERENCES Linked references are available on JSTOR for thi. *Science*, *178*(3), 471–479.

Glänzel, W., Thijs, B. & Chi, P. S. (2016). The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: the book citation index. *Scientometrics*, *109*(3), 2165–2179. doi: 10.1007/s11192-016-2046-7

Harzing, A. W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, *120*(1), 341–349. Retrieved from https://doi.org/10.1007/s11192-019-03114-y doi: 10.1007/s11192-019-03114-y

Harzing, A.-W. & Alakangas, S. (2017, jan). Microsoft Academic: is the phoenix getting wings? *Scientometrics*, *110*(1), 371–383. Retrieved from http://link.springer.com/10.1007/s11192-016-2185-x doi: 10.1007/s11192-016-2185-x

Hassan, S.-U., Imran, M., Iqbal, S., Aljohani, N. R. & Nawaz, R. (2018, dec). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics*, *117*(3), 1645–1662. Retrieved from https://doi.org/10.1007/s11192-018-2944-yhttp://link.springer.com/10.1007/s11192-018-2944-y doi: 10.1007/s11192-018-2944-y

Hug, S. E., Ochsner, M. & Brändle, M. P. (2017, apr). Citation analysis with microsoft academic. *Scientometrics*, *111*(1), 371–378. Retrieved from http://link.springer.com/10.1007/s11192-017-2247-8 doi: 10.1007/s11192-017-2247-8

Ioannidis, J. P. A. (2005, aug). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. Retrieved from https://dx.plos.org/10.1371/journal.pmed.0020124 doi: 10.1371/journal.pmed.0020124

Ken, H. (2003). Self-citation and self-reference: Credibility and promotion in academic publication. *Journal of the American Society for Information Science and Technology*, *54*(3), 251–259. Retrieved from http://dx.doi.org/10.1002/asi.10204

Kousha, K., Thelwall, M. & Rezaie, S. (2011). Assessing the Citation Impact of Books:The Role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, *62*(11), 2147–2164. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/asi.21169/full doi: 10.1002/asi

Lamm, C., Decety, J. & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502. Retrieved from http://dx.doi.org/10.1016/j.neuroimage.2010.10.014 doi: 10.1016/j.neuroimage.2010.10.014

Levine-Clark, M. & Gil, E. L. (2009). A comparative citation analysis of web of science, scopus, and google scholar. *Journal of Business and Finance Librarianship*, *14*(1), 32–46. doi: 10.1080/08963560802176348

Macroberts, M. H. & Macroberts, B. F. (1989). Problems of Citation Analysis : A Critical Review.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M. & Delgado López-Cózar, E. (2018, nov). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. Retrieved from http://arxiv.org/abs/1808.05053https://linkinghub.elsevier.com/retrieve/pii/S1751157718303249 doi: 10.1016/j.joi.2018.09.002

Matthews, T. (n.d.). LibGuides: Web of Science platform: Web of Science: Summary of Coverage. Retrieved from https://clarivate.libguides.com/webofscienceplatform/coverage

Nelson, L. D., Simmons, J. & Simonsohn, U. (2018, jan). Psychology's Renaissance. *Annual Review of Psychology*, *69*(1), 511–534. Retrieved from http://www.annualreviews.org/doi/10.1146/annurev-psych-122216-011836 doi: 10.1146/annurev-psych-122216-011836

Nicolaisen, J. (2003). The Social Act of Citing : Towards New Horizons in Citation Theory. , 12–20.

Pandita, R. & Singh, S. (2017, jul). Self-citations, a trend prevalent across subject disciplines at the global level: an overview. *Collection Building*, *36*(3), 115–126. Retrieved from http://www.emeraldinsight.com/doi/10.1108/CB-03-2017-0008 doi: 10.1108/CB-03-2017-0008

Radicchi, F., Weissman, A. & Bollen, J. (2017, aug). Quantifying perceived impact of scientific publications. *Journal of Informetrics*, *11*(3), 704–712. Retrieved from http://dx.doi.org/10.1016/j.joi.2017.05.010https://linkinghub.elsevier.com/retrieve/pii/S1751157717300846 doi: 10.1016/j.joi.2017.05.010

Rodriguez-Ruiz, O. (2009). The citation indexes and the quantification of knowledge. *Journal of educational Administration*, *47*(2), 250–266. doi: https://doi.org/10.1108/09578230910941075

Schreiber, M. (2007, may). Self-citation corrections for the Hirsch index. *Europhysics Letters (EPL)*, *78*(3), 30002. Retrieved from http://stacks.iop.org/0295-5075/78/i=3/a=30002?key=crossref.999675785537c0d68077f0fb4947f4d0 doi: 10.1209/0295-5075/78/30002

Scite. (n.d.). *scite_*. Retrieved 2019-06-19, from https://scite.ai/{#}learn-more

Scopus. (n.d.). *The largest database of peer-reviewed literature - Scopus — Elsevier Solutions.* Retrieved 2019-06-19, from https://www.elsevier.com/solutions/scopus

Scopus. (2019). *How are CiteScore metrics used in Scopus? - Scopus: Access and use Support Center.* Retrieved 2019-05-28, from https://service.elsevier.com/app/answers/detail/a{_}id/14880/supporthub/scopus/

Seeber, M., Cattaneo, M., Meoli, M. & Malighetti, P. (2019, mar). Self-citations as strategic response to the use of metrics for career decisions. *Research Policy*, *48*(2), 478–491. Retrieved from https://doi.org/10.1016/j.respol.2017.12.004https://linkinghub.elsevier.com/retrieve/pii/S004873331730210X doi: 10.1016/j.respol.2017.12.004

Silver, R. & LeSauter, J. (2008, may). Circadian and Homeostatic Factors in Arousal. *Annals of the New York Academy of Sciences*, *1129*(1), 263–274. Retrieved from http://doi.wiley.com/10.1196/annals.1417.032 doi: 10.1196/annals.1417.032

Tune, S., Schlesewsky, M., Nagels, A., Small, S. L. & Bornkessel-Schlesewsky, I. (2016, aug). Sentence understanding depends on contextual use of semantic and real world knowledge. *NeuroImage*, *136*, 10–25. Retrieved from http://dx.doi.org/10.1016/j.neuroimage.2016.05.020https://linkinghub.elsevier.com/retrieve/pii/S1053811916301392 doi: 10.1016/j.neuroimage.2016.05.020

van Eck, N. J., Waltman, L., Lariviére, V. & Sugimoto, C. (2018). *Crossref as a new source of citation data: A comparison with Web of Science and Scopus.* Retrieved 2019-06-14, from https://www.cwts.nl/blog?article=n-r2s234

van Raan, A. F. (2008, aug). Selfcitation as an impactreinforcing mechanism in the science system. *Journal of the American Society for Information Science and Technology*, *59*(10), 1631–1643. Retrieved from http://doi.wiley.com/10.1002/asi.20868 doi: 10.1002/asi.20868

van Raan, A. F. J. (1996, jul). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*(3), 397–420. Retrieved from http://link.springer.com/10.1007/BF02129602 doi: 10.1007/BF02129602

Vellino, A. (2015, nov). Recommending research articles using citation data. *Library Hi Tech*, *33*(4), 597–609. Retrieved from https://www.emeraldinsight.com/doi/abs/10.1108/LHT-10-2014-0100?journalCode=lhthttp://www.emeraldinsight.com/doi/10.1108/LHT-06-2015-0063 doi: 10.1108/LHT-06-2015-0063

Webber, W., Moffat, A. & Zobel, J. (2010, nov). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, *28*(4), 1–38. Retrieved from http://portal.acm.org/citation.cfm?doid=1852102.1852106 doi: 10.1145/1852102.1852106

Zhong, S., Chark, R., Hsu, M. & Chew, S. H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *NeuroImage*, *129*, 95–104. Retrieved from http://dx.doi.org/10.1016/j.neuroimage.2016.01.040 doi: 10.1016/j.neuroimage.2016.01.040