Field- and time-normalization of zero-inflated data: An empirical analysis using citation and Twitter data

Robin Haunschild¹  Lutz Bornmann²

¹ r.haunschild@fkf.mpg.de
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart (Germany)

² bornmann@gv.mpg.de
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich (Germany)

Abstract
Thelwall (2017a, 2017b) proposed a new family of field- and time-normalized indicators, which is intended for sparse data. These indicators are based on units of analysis (e.g., institutions) rather than on the paper level. They compare the proportion of mentioned papers (e.g., on Twitter) of a unit with the proportion of mentioned papers in the corresponding fields and publication years (the expected values). We propose a new indicator (Mantel-Haenszel quotient, MHq) for the indicator family. The MHq goes back to the MH analysis. This analysis is an established method, which can be used to pool the data from several 2x2 cross tables based on different subgroups. We investigate (using citations and assessments by peers, i.e., F100Prime recommendations) whether the indicator family (including the MHq) can distinguish between quality levels defined by the assessments of peers. Thus, we test the convergent validity. We find that the MHq is able to distinguish between quality levels (in most cases) while other indicators of the family are not. Since our study approves the MHq as a convergent valid indicator, we apply the MHq to four different Twitter groups as defined by the company Altmetric (e.g., science communicators). Our results show that there is a weak relationship between all four Twitter groups and scientific quality, much weaker than between citations and scientific quality. Therefore, our results discourage the use of Twitter counts in research evaluation.

Introduction
Alternative metrics (altmetrics) is a new fast-moving area in scientometrics (Galloway, Pease, & Rauh, 2013). Initially, altmetrics – a collection of many web-based indicators – have been proposed as a supplement to traditional bibliometric indicators. They measure attention related to research papers on internet platforms. The core of altmetrics is gathered from social media platforms, but mentions in mainstream media or in policy documents also belong to the umbrella term altmetrics (National Information Standards Organization, 2016; Work, Haustein, Bowman, & Larivière, 2015). According to Haustein (2016), sources of altmetrics can be grouped into (i) social networks, (ii) social bookmarks and online reference management, (iii) social data (e.g., data sets, software, presentations), (iv) blogs, (v) microblogs, (vi) wikis, and (vii) recommendations, ratings, and reviews. Recently, some indicators based on altmetrics have been proposed which are normalized with respect to the scientific field and publication year. These indicators were developed because studies have shown that altmetrics are – similar to bibliometric data – field- and time-dependent (see, e.g., Bornmann, 2014). Some fields are more relevant to the general public or a broader audience than other fields (Haustein, Larivière, Thelwall, Amyot, & Peters, 2014). The Mean Normalized Reader Score (MNRS) was introduced by Haunschild and Bornmann (2016) for normalization of data from social bookmarks and online reference management platforms (with a special emphasis on Mendeley readers) (see also Fairclough & Thelwall, 2015). The Mean Discipline Normalized Reader Score (MDNRS) was tailored specifically to Mendeley by Bornmann and Haunschild (2016b). The MDNRS uses Mendeley disciplines for

¹ This is a substantially extended version of a conference paper which has been presented at the 16th International Conference on Scientometrics & Informetrics (ISSI) 2017 (Haunschild & Bornmann, 2017).
field normalization. The employed normalization procedures rely on average value calculations across scientific fields and publication years as expected values. However, normalization procedures based on averages (and percentiles) of individual papers are problematic for zero-inflated data sets (Haunschild, Schier, & Bornmann, 2016). The overview of Work, et al. (2015) on studies investigating the coverage of papers on social media platforms show that many platforms have coverages of less than 5% (e.g., Blogs, or Wikipedia). Erdt, Nagarajan, Sin, and Theng (2016) reported similar findings in their meta-analysis. They found that former empirical studies dealing with the coverage of altmetrics show that about half of the platforms are at or below 5%; except for three (out of eleven) where the coverage is below 10%.

Bornmann and Haunschild (2016a) propose the Twitter Percentile (TP) – a field- and time-normalized indicator for Twitter data. Bornmann and Haunschild (2016a) circumvent the problem of zero-inflated Twitter data by including in the TP calculation only journals with at least 80% of the papers having at least 1 tweet each. However, this procedure leads to the exclusion of many journals from the TP procedure. Very recently, Thelwall (2017a, 2017b) proposed a new family of field- and time-normalized indicators. These indicators are based on units of analysis (e.g., a researcher or institution) rather than on single papers. They compare the proportion of mentioned papers (e.g., on Twitter) of a unit with the proportion of mentioned papers in the corresponding fields and publication years (the expected values). The family consists of the Equalized Mean-based Normalized Proportion Cited (EMNPC) and the Mean-based Normalized Proportion Cited (MNPC). Hitherto, this new family of indicators has only been studied on rather small samples.

In this study, we investigate the new indicator family empirically on a large scale (multiple complete publication years) and add another member to this family. In statistics, the Mantel-Haenszel (MH) analysis is frequently used for pooling the data from multiple 2×2 cross tables based on different subgroups. In this study, we have mentioned and not-mentioned papers of a unit, which have been published in different subject categories and publication years and are compared with the corresponding reference sets. We name the new indicator Mantel-Haenszel quotient (MHq). In the empirical part of this study, we compare the indicator scores with assessments by peers. We are interested whether the indicators can discriminate between different quality levels, which peers assigned to publications. In other words, we investigate the convergent validity of the indicators. The convergent validity can only be tested by using citations, since we can assume that they are related to quality (Diekmann, Naf, & Schubiger, 2012). Thus, the first empirical part is based on citations. In the second part (after confirmation of convergent validity), MHq values are exemplarily presented for Twitter data.

**Indicators for zero-inflated count data**

The next sections focus on the formulas not only for the calculation of the EMNPC, MNPC, and MHq, but also for the corresponding 95% confidence intervals (CIs). The CI shows the range of possible indicator values: We can be 95% confident that the interval includes the “true” indicator value in the population. Thus, we assume to analyze sample data and infer by using CIs to a larger, inaccessible population (Williams & Bornmann, 2016). Claveau (2016) argue for using inferential statistics with scientometric data as follows: “these observations are realizations of an underlying data generating process … The goal is to learn properties of the data generating process. The set of observations to which we have access, although they are all the actual realizations of the process, do not constitute the set of all possible realizations. In consequence, we face the standard situation of having to infer from an accessible set of observations – what is normally called the sample – to a larger, inaccessible one – the population. Inferential statistics are thus pertinent“ (p. 1233).
Equalized Mean-based Normalized Proportion Cited (EMNPC)

Thelwall (2017a, 2017b) proposed the EMNPC as an alternative indicator for zero-inflated count data. Here, the proportion of mentioned publications is calculated: suppose that the publication set of a group $g$ has $n_{gf}$ papers in the publication year and subject category combination $f$. $s_{gf}$ denotes the number of mentioned papers (e.g., on Twitter). $F$ denotes all publication year and subject category combinations of the publications in the set. The overall proportion of $g$’s mentioned publications is the number of mentioned publications ($s_{gf}$) divided by the total number of publications ($n_{gf}$):

$$p_g = \frac{\sum_{f \in F} s_{gf}}{\sum_{f \in F} n_{gf}} \quad (1)$$

However, $p_g$ could lead to misleading results, if the publication set $g$ includes many publications which appeared in fields with many mentioned papers. Thus, Thelwall (2017a, 2017b) proposes to artificially treat $g$ as having the same number of publications in each year and subject category combination. Thelwall (2017a, 2017b) fixes it to the arithmetic mean of numbers in each combination. However, he recommends not to include combinations of $g$ with only a few publications in the analysis. Thus, the equalized sample proportion $\hat{p}_g$ is the average of the proportions in each combination:

$$\hat{p}_g = \frac{\sum_{f \in F} s_{gf}}{\sum_{f \in F} n_{gf}} \quad (2)$$

The corresponding equalized sample proportion of the world ($w$) is:

$$\hat{p}_w = \frac{\sum_{f \in F} s_{wf}}{\sum_{f \in F} n_{wf}} \quad (3)$$

In Eqs. (2) and (3), $[F]$ is the number of subject category and publication year combinations in which the group (in case of Eq. (2)) and the world (in case of Eq. (3)) publish. The equalized group sample proportion has the following undesirable property: it treats $g$ as if the average number of mentioned publications does not vary between the subject categories. The ratio of both equalized sample proportions (group and world) is the EMNPC:

$$\text{EMNPC} = \frac{\hat{p}_g}{\hat{p}_w} \quad (4)$$

According to Thelwall (2017a), CIs for the EMNPC can be calculated:

$$\text{EMNPC}_L = \exp \left( \ln \left( \frac{\hat{p}_g}{\hat{p}_w} \right) - 1.96 \sqrt{\frac{\left( n_g - \hat{p}_g n_g \right) / \hat{p}_g n_g + \left( n_w - \hat{p}_w n_w \right) / \hat{p}_w n_w}{n_g} + \frac{\left( n_w - \hat{p}_w n_w \right) / \hat{p}_w n_w}{n_w}} \right) \quad (5)$$

$$\text{EMNPC}_U = \exp \left( \ln \left( \frac{\hat{p}_g}{\hat{p}_w} \right) + 1.96 \sqrt{\frac{\left( n_g - \hat{p}_g n_g \right) / \hat{p}_g n_g + \left( n_w - \hat{p}_w n_w \right) / \hat{p}_w n_w}{n_g} + \frac{\left( n_w - \hat{p}_w n_w \right) / \hat{p}_w n_w}{n_w}} \right) \quad (6)$$

In Eqs. (5) and (6), $n_w$ is the total sample size of the world, and $n_g$ is the total sample size of group $g$. 
Mean-based Normalized Proportion Cited (MNPC)

The other indicator proposed by Thelwall (2017a) is the Mean-based Normalized Proportion Cited (MNPC) which is calculated as follows: For each publication which is mentioned at least once (e.g., mentioned on Twitter), the reciprocal of the world proportion mentioned for the corresponding subject category and publication year replaces the number of mentions. All unmentioned publications remain at zero. If \( p_{gf} = s_{gf}/n_{gf} \) is the proportion of mentioned publications in set \( g \) in the corresponding subject category and publication year combination \( f \) and \( p_{gw} = s_{gw}/n_{gw} \) the proportion of world’s mentioned publications in the same year and subject category combination \( f \), then:

\[
   r_i = \begin{cases} 
   0, & \text{if } c_i = 0 \\
   1/p_{wf}, & \text{if } c_i > 0 
   \end{cases}
\]

where \( p_{wf} \) and \( c_i \) are the proportion of world’s mentioned publications and the number of mentions for the corresponding subject category and publication year combination \( f \), then:

(7)

The MNPC calculation follows the calculation of the MNCS (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011) and is defined as:

\[
   MNPC = \frac{\left( r_1 + r_2 + \cdots r_{ng} \right)}{ng}
\]

Thelwall (2016, 2017a) proposes an approximate CI for the MNPC. The lower limit \( L \) (MNPC_{gl}) and upper limit \( U \) (MNPC_{gu}) for group \( g \) in year and subject category combination \( f \) is calculated in the first step:

\[
   MNPC_{gL} = \exp\left( \ln\left( \frac{p_{gf}}{p_{wf}} \right) - 1.96 \sqrt{\frac{(n_{gf} - p_{gf}n_{gf})/(\hat{p}_{gf}n_{gf}) + (n_{gw} - p_{gw}n_{gw})/(\hat{p}_{gw}n_{gw})}{n_{gf}}} \right)
\]

(9)

\[
   MNPC_{gU} = \exp\left( \ln\left( \frac{p_{gf}}{p_{wf}} \right) + 1.96 \sqrt{\frac{(n_{gf} - p_{gf}n_{gf})/(\hat{p}_{gf}n_{gf}) + (n_{gw} - p_{gw}n_{gw})/(\hat{p}_{gw}n_{gw})}{n_{gf}}} \right)
\]

(10)

The group-specific lower and upper limits are used to calculate the MNPC CIs in a second step:

\[
   MNPC_{gL} = MNPC - \sum_{f \in F} \frac{n_{gf}}{ng} \left( \frac{p_{gf}}{p_{wf}} - MNPC_{gL} \right)
\]

(11)

\[
   MNPC_{gU} = MNPC + \sum_{f \in F} \frac{n_{gf}}{ng} \left( MNPC_{gU} - \frac{p_{gf}}{p_{wf}} \right)
\]

(12)

If any of the world proportions are equal to zero, the MNPC cannot be calculated. If any of the group proportions are equal to zero, CIs cannot be calculated. As solutions, either the corresponding subject category and publication year combination can be removed from the data or a continuity correction of 0.5 can be added to the number of mentioned and not mentioned publications (Thelwall, 2017a). We prefer to use the continuity correction. Plackett (1974) recommends this approach for the calculation of odds ratios.

Mantel-Haenszel quotient (MHq)

The recommended method for pooling the data from multiple 2×2 cross tables – based on different subgroups (which are part of a larger population) – is the Mantel Haenszel (MH) analysis (Hollander & Wolfe, 1999; Mantel & Haenszel, 1959; Sheskin, 2007). According to Fleiss, Levin, and Paik (2003) the method “permits one to estimate the assumed common odds ratio and to test whether the overall degree of association is significant. Curiously, it is
not the odds ratio itself but another measure of association that directly underlies the test for overall association … The fact that the methods use simple, closed-form formulas has much to recommend it” (p. 250). The results of Radhakrishna (1965) point out that the MH approach is empirically and formally valid against the background of clinical trials.

The MH analysis yields a summary odds ratio for multiple 2×2 cross tables. We call this summary odds ratio MHq. If the impact of units in science is compared with reference sets (the world), the 2×2 cross tables (which are pooled) consist of the number of publications mentioned and not mentioned in subject category and publication year combinations. The results of Radhakrishna (1965) point out that the MH approach is empirically and formally valid against the background of clinical trials.

The MHq is defined as:

\[ MHq = \frac{R}{S} \]  

(16)

The MHq CI is calculated as recommended by Fleiss, et al. (2003). The variance of \( \ln(MHq) \) is estimated by:

\[ \tilde{\text{Var}}[\ln(MHq)] = \frac{1}{2} \left( \frac{\sum_{f=1}^{F} P_f R_f}{R^2} + \frac{\sum_{f=1}^{F} (P_f S_f + Q_f R_f)}{RS} + \frac{\sum_{f=1}^{F} Q_f S_f}{S^2} \right) \]  

(17)

Calculation of the MHq CI is performed as follows:

\[ MHq_L = \exp \left[ \ln(MHq) - 1.96\sqrt{\tilde{\text{Var}}[\ln(MHq)]} \right] \]  

(18)

\[ MHq_U = \exp \left[ \ln(MHq) + 1.96\sqrt{\tilde{\text{Var}}[\ln(MHq)]} \right] \]  

(19)

It is an advantage of the MHq that the world average has a value of 1. This is similar to the EMNPC and MNPC and simplifies the interpretation. A further advantage of the MHq is that

Table 1. 2×2 subject-specific cross table

|          | Number of mentioned publications | Number of unmentioned publications |
|----------|----------------------------------|-----------------------------------|
| Group g  | \( a_f \)                        | \( b_f \)                          |
| World    | \( c_f \)                        | \( d_f \)                          |

The MHq calculation starts by defining some auxiliary variables:

\[ R_f = \frac{a_f d_f}{n_f} \text{ with } R = \sum_{f=1}^{F} R_f, \]  

(13)

\[ S_f = \frac{b_f c_f}{n_f} \text{ with } S = \sum_{f=1}^{F} S_f, \]  

(14)

\[ P_f = \frac{a_f + d_f}{n_f} \text{ with } Q_f = 1 - P_f \]  

(15)

Where: \( n_f = a_f + b_f + c_f + d_f \)

The MHq is defined as:

\[ MHq = \frac{R}{S} \]  

(16)

The MHq CI is calculated as recommended by Fleiss, et al. (2003). The variance of \( \ln(MHq) \) is estimated by:

\[ \tilde{\text{Var}}[\ln(MHq)] = \frac{1}{2} \left( \frac{\sum_{f=1}^{F} P_f R_f}{R^2} + \frac{\sum_{f=1}^{F} (P_f S_f + Q_f R_f)}{RS} + \frac{\sum_{f=1}^{F} Q_f S_f}{S^2} \right) \]  

(17)

Calculation of the MHq CI is performed as follows:

\[ MHq_L = \exp \left[ \ln(MHq) - 1.96\sqrt{\tilde{\text{Var}}[\ln(MHq)]} \right] \]  

(18)

\[ MHq_U = \exp \left[ \ln(MHq) + 1.96\sqrt{\tilde{\text{Var}}[\ln(MHq)]} \right] \]  

(19)

It is an advantage of the MHq that the world average has a value of 1. This is similar to the EMNPC and MNPC and simplifies the interpretation. A further advantage of the MHq is that
the result can be interpreted as a percentage relative to the world average. \( MHq = 1.20 \), e.g., means that the publication set under study has an impact which is 20% above average.

**Data sets used**

As data sets, we used publications of the Web of Science (WoS) from our in-house database – derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), and Arts and Humanities Citation Index (AHCI) provided by Clarivate Analytics. All papers of the document type “article” with DOI published between 2010 and 2013 were included in the study. The citation counts are restricted to citations with a three-year citation window – excluding the publication year (Glänzel & Schoepflin, 1995). The overlapping WoS subject categories are used for field classification (Rons, 2012, 2014). We include only subject categories in this study where (1) more than 9 papers are assigned to and (2) the number of cited and uncited publications is greater than zero. This should avoid statistical and numerical problems. These restrictions resulted in a dataset including 4,490,998 publications.

We matched the publication data with peers’ recommendations from F1000Prime via the DOI. F1000Prime is a post-publication recommendation system of publications which have been published mainly in medical and biological journals. A peer-nominated global “Faculty” of leading scientists and clinicians selects and rates the publications and explains their importance. Thus, only a restricted set from the publications in these disciplines covered is reviewed. Most of the publications are actually not. Faculty members can select and evaluate any publication of interest. Faculty members rate the publications as “Recommended,” “Must read”, or “Exceptional”. This is equivalent to recommendation scores (RSs) of 1, 2, or 3, respectively. Since publications can be recommended multiple times, we calculated an average RS (FFa):

\[
\text{FFa} = \frac{1}{i_{\text{max}}} \sum_{i_{\text{max}}}^{i_{\text{max}}} \text{RS}_i
\]  

(20)

The papers are categorized depending on their \( \text{FFa} \) value:
- \( \overline{\text{FFa}} = 0 \): publications which are not recommended (Q0)
- \( 0 < \overline{\text{FFa}} \leq 1.0 \): recommended publications with a rather low average score (Q1):
- \( \overline{\text{FFa}} > 1.0 \): recommended publications with a rather high average score (Q2):

Following Waltman and Costas (2014), we only considered subject categories where a publication with an F1000Prime recommendation is assigned to.

Whereas the first empirical part of this study is based on citation counts, the second empirical part focusses on Twitter data. The Twitter data were taken from a data set retrieved from the company Altmetric. With the number of publications and proportion of uncited and untweeted publications, Table 1 and Table 2 show overviews of the data included in this study. It is clearly visible in Table 2 that Twitter data are affected by zero-inflation, but citation data are not.

Table 1. Number of papers included in this study broken down by different sources (citations and Twitter groups), publication year, and \( \overline{\text{FFa}} \)

| Year | \( \overline{\text{FFa}} \) | Citations | Twitter groups |
|------|-----------------|-----------|----------------|
|      |                 | Researchers | Science communicators | Practitioners | Members of the public |
| 2010 | Q0              | 628,862    | 587,563           | 505,205       | 504,052       | 626,116       |
| Year | Q0 | Q1  | Q2  | Q0  | Q1  | Q2  | Q0   | Q1  | Q2  | Q0  | Q1  | Q2  | Q0  | Q1  | Q2  | Q0  | Q1  | Q2  | Q0  | Q1  | Q2  |
|------|----|-----|-----|-----|-----|-----|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|      |    |     |     | 6,576 | 6,286 | 5,941 | 6,189 | 6,622 | 6,576 | 6,286 | 5,941 | 6,189 | 6,622 | 6,576 | 6,286 | 5,941 | 6,189 | 6,622 | 6,576 | 6,286 | 5,941 | 6,189 | 6,622 |
|      | 681,749 | 669,966 | 614,564 | 598,772 | 683,815 | 681,749 | 669,966 | 614,564 | 598,772 | 683,815 | 681,749 | 669,966 | 614,564 | 598,772 | 683,815 | 681,749 | 669,966 | 614,564 | 598,772 | 683,815 |
|      | 733,813 | 730,917 | 704,025 | 706,521 | 736,992 | 733,813 | 730,917 | 704,025 | 706,521 | 736,992 | 733,813 | 730,917 | 704,025 | 706,521 | 736,992 | 733,813 | 730,917 | 704,025 | 706,521 | 736,992 |
|      | 785,961 | 786,486 | 772,364 | 770,555 | 788,758 | 785,961 | 786,486 | 772,364 | 770,555 | 788,758 | 785,961 | 786,486 | 772,364 | 770,555 | 788,758 | 785,961 | 786,486 | 772,364 | 770,555 | 788,758 |

Table 2. Proportion of uncited and untweeted papers broken down by different sources (citations and Twitter groups), publication year, and FFa.

| Year | FFa | Citations | Researchers | Science communicators | Practitioners | Members of the public |
|------|-----|-----------|-------------|-----------------------|--------------|-----------------------|
| 2010 | 10.36 | 98.88 | 99.43 | 99.08 | 96.44 |
| 2010 | 0.84 | 95.31 | 98.32 | 97.24 | 88.95 |
| 2010 | 0.43 | 89.89 | 96.50 | 94.49 | 80.04 |
| 2011 | 10.61 | 97.04 | 98.69 | 98.19 | 89.96 |
| 2011 | 0.61 | 89.11 | 95.50 | 94.49 | 89.96 |
| 2011 | 0.68 | 76.74 | 90.21 | 89.10 | 72.71 |
| 2012 | 10.41 | 92.03 | 96.16 | 95.68 | 77.73 |
| 2012 | 1.08 | 75.74 | 84.76 | 84.00 | 44.78 |
| 2012 | 0.46 | 56.11 | 72.10 | 75.94 | 29.42 |
| 2013 | 10.84 | 89.21 | 94.33 | 93.62 | 73.53 |
| 2013 | 1.39 | 70.35 | 81.87 | 77.79 | 37.53 |
| 2013 | 0.50 | 50.58 | 69.30 | 69.71 | 26.87 |

In most of the previous studies, which are based on Twitter data, counts of Twitter mentions are used. In this study, we differentiate the Twitter data further on and focus on different Twitter groups: researchers, science communicators, practitioners, and members of the public. These groups are defined by Altmetric based on keyword matching (Adie, 2016):

- **Researcher sample keywords**: "Post doc", "post-doc", "post-doctoral", "scientista", "researcher", "scientist", "postdoc", "postdoctoral", "principal investigator", "Profesora", "professor", "prof", "professoring", "Aspirante a PhD", "phd candidate", "Ph.D researcher", "doctoral student", "phd researcher", "grad student", "phd student", "postgrad", "postgraduate", "biologist", "molecular biology", "genomes", "genomics", "molecular evolution", "microbiology", "biomechanics", "bioinformatics", "geologist", "oceanographer", "geochemist", "anthropologist", "archeologist", "paleontologist", "astrophysicist", "astrophysics", "physics", "physicist", "mathematics", "mathematics", "maths", "mathematical", "chemist", "chemistry", ...
- **Science communicator sample keywords**: "news service", "Magazine", "editor-in-chief", "editors-in-chief", "correspondent", "Jornalista", "reporter", "journalist", "journo", "sci comms", "science communicator", "science communication", "editor", "editor"
Practitioner sample keywords: "medic", "medical student", "Pediatrician", 'M.D.', "doctor", "nurse", "surgeon", "emt", "physician", "radiologist", "radiology resident", "psychiatrist", "dentist", "medical", "psychologist", "doula", "Nutricionista", "dietician", "dietitian", "nutrition", "fitness", "public health", "health counselors", "health counselor", "Hospital", "dental", "NHS", ...

Member of the public is a Twitter user which does not fall into another of the previous three groups.

Results

Empirical analysis using citations

It is an established way of analyzing the convergent validity of indicators comparing the indicator values with peer evaluations (Garfield, 1979; Kreiman & Maunsell, 2011). Convergent validity is defined as the degree to which two measurements of a construct (here: two indicators of scientific quality) with a theoretical relationship are also empirically related. This approach has been justified by Thelwall (2017b) as follows: “if indicators tend to give scores that agree to a large extent with human judgements then it would be reasonable to replace human judgements with them when a decision is not important enough to justify the time necessary for experts to read the articles in question. Indicators can be useful when the value of an assessment is not great enough to justify the time needed by experts to make human judgements” (p. 4). Several publications studying the relationship between Research Excellence Framework (REF) outcomes and citations reveal considerable relationships in different fields, such as psychology and biological science (Butler & McAllister, 2011; Mahdi, d'Este, & Neely, 2008; McKay, 2012; Smith & Eysenck, 2002; Traag & Waltman, 2017; Wouters et al., 2015). Similar results have been reported for the Italian research assessment exercise: “The correlation strength between peer assessment and bibliometric indicators is statistically significant, although not perfect. Moreover, the strength of the association varies across disciplines, and it also depends on the discipline internal coverage of the used bibliometric database” (Franceschet & Costantini, 2011, p. 284). Bornmann (2011) shows in an overview of studies on journal peer review that better recommendations from peers are related to higher citation impact of the corresponding papers. The correlation between citation impact scores and RS from F1000Prime has already been investigated in other studies. The results of Bornmann (2015) reveal that about 40% of papers with RS=1 are highly cited papers; for publications with RS=2 and RS=3 the percentages are 60% and 73%, respectively. Waltman and Costas (2014) report “a clear correlation between F1000 recommendations and citations” (p. 433). The previous results on F1000Prime might point out, therefore, that citation-based indicators differentiate between the three quality levels. Looking at it the other way round, the validity of new indicators does not seem to be given if they do not differentiate. Against this backdrop, we analyze in this study the ability of MHq, EMNPC, and MNPC to differentiate between the F1000Prime quality groups. Figure 1 shows the MHqs with CIs for Q0, Q1, and Q2 across four publication years.
The results in Figure 1 point out that the MHq values are very different for Q0, Q1, and Q2. The average MHq for all years is close to (but below) 1 for Q0. The mean MHq for Q1 is about eight times and that for Q2 is about 15 times higher than the mean MHq for Q0. Thus, the MHq indicator seems to significantly separate between Q0, Q1, and Q2; the MHq values seem to be convergent valid with respect to F1000Prime scores.

Figure 2. MNPC with CIs for Q0, Q1, and Q2 across four publications years. The horizontal line with MNPC=1 is the worldwide average.
Figure 3. EMNPC with CIs for Q0, Q1, and Q2 across four publications years. The horizontal line with EMNPC=1 is the worldwide average.

The results for MNPC and EMNPC – the two indicators proposed by Thelwall (2017a) – are shown in Figure 2 and Figure 3. It is clearly visible in both figures that the indicator values for all groups are very close to 1. This result does not accord with Figure 1 where the MHq values for Q1 and Q2 are significantly greater than 1. Also, as visible in Figure 2 and Figure 3, the indicator values do not show the expected ordering (analogous to the ordering in Figure 1: MHq(Q0)<MHq(Q1)<MHq(Q2)) for 2010 in the case of EMNPC and 2010 and 2011 in the case of MNPC. These differences in the results can be interpreted as a first indication that MNPC and EMNPC do not differentiate properly between quality groups as defined by FFa values.

**Empirical analysis using Twitter data**

In the previous section, we demonstrated that the MHq is convergent valid using citation data compared with post-publication recommendation scores from F1000Prime. The MHq is able to distinguish between different scientific quality levels as defined by F1000Prime scores. In this section, we determine whether and to which extent different Twitter groups (as defined by the company Altmetric) can distinguish between the same quality levels. Figure 4 shows the MHq results for researchers, science communicators, practitioners, and members of the public.
All four Twitter groups show the expected ordering with $\text{MHq}(Q0) < \text{MHq}(Q1) < \text{MHq}(Q2)$. For all four Twitter groups, the quality group Q0 is close to but below 1. The MHq values for the quality groups Q1 and Q2 are between 2 and 4 and between 4 and 6, respectively. Compared to citation data, all four Twitter groups show a much weaker association to scientific quality than citations, by a factor of about three. In Figure 1, which shows the results for citation data, $\text{MHq}(Q1)$ is between 7 and 9, and $\text{MHq}(Q2)$ is between 11 and 18. The results in Figure 4 further reveal that the association to scientific quality is on a similar level for all four Twitter groups. Since researchers should assess scientific quality better than the other groups in the figure, the association to quality is somewhat stronger for researchers in the figure than for the other groups.

**Discussion**

Much of the altmetrics data is sparse (Neylon, 2014). A metric inflated by zero values is not informative for research evaluation purposes in the first place (Thelwall, Kousha, Dinsmore,
& Dolby, 2016). Thelwall (2017a, 2017b) introduced a new family of field- and time normalized indicators for sparse data including EMNPC and MNPC. Here, the proportion of mentioned publications of a unit (e.g., a researcher) is compared with the expected values (the proportion of mentioned publications in the corresponding publication years and fields).

EMNPC and MNPC differ from most of the indicators used in bibliometrics and altmetrics. Usually, an indicator value is calculated for each publication. The publication-based indicator values can then be aggregated by the user, for example, by averaging or summing. Instead, the indicators of the new family are based on the calculation of the indicator values for publication sets of groups (e.g., universities). This property implies that the new indicators cannot be used as versatilely as the usual bibliometric (and altmetric) indicators. However, they are able (by construction) to handle zero-inflated data properly which the usual indicators are not.

In this study, we added a further variant to the family – the MHq – and analyzed all three variants empirically. We started by analyzing the convergent validity of the indicators based on citation data. Citation data can be used to formulate predictions, which can be empirically validated. Thus, we studied whether EMNPC, MNPC, and MHq are able to validly differentiate between three different quality levels – as defined by RS from F1000 (FFa). By comparing the indicator values with peer recommendations, we were able to test whether the indicators discriminate between different quality levels.

The results point out that EMNPC and MNPC cannot validly discriminate between different quality levels as defined by peers. The EMNPC and MNPC values are close to the worldwide average – independent of the quality levels. The CIs substantially overlap in many comparisons. Thus, the convergent validity of the EMNPC and MNPC does not seem to be given. In contrast to these indicators, the MHq is able to discriminate between the quality levels.

Because of the positive results for the MHq, we applied the MHq to Twitter counts of four different groups as defined by the company Altmetric: researchers, science communicators, practitioners, and members of the public. If Twitter counts are intended to be used for research evaluation, a substantial relationship to scientific quality should be given. Otherwise, Twitter counts should not be employed in research evaluation. Our investigation of MHq values based on Twitter data reveals a weak relationship between Twitter counts and scientific quality. This relationship is much weaker than that between citation counts and scientific quality.

Our study of the relationship of different Twitter groups’ data to scientific quality is directed at specific societal groups. Earlier, we studied the directed societal impact measurement for different status groups of Mendeley data (Bormann & Haunschild, 2017). Researchers on Twitter show a slightly stronger relationship to scientific quality than other societal groups, but the differences are only minor.

This study follows the important initiative of Thelwall (2017a, 2017b) to develop new indicators for zero-inflated data. The current study is the first independent attempt to investigate the new indicator family empirically. This family is important for altmetrics data. Thus, we need further studies focusing on various sources with sparse data (in addition to Twitter). Since F1000 concentrates on biomedicine, future empirical studies should analyze the new family in other disciplines.

Acknowledgements

The bibliometric data used in this paper are from an in-house database developed and maintained by the Max Planck Digital Library (MPDL, Munich, Germany) and derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), and Arts and Humanities Citation Index (AHCI) prepared by Clarivate Analytics (formerly the IP
and Science business of Thomson Reuters). The Twitter data were taken from a data set retrieved from Altmetric on June 04, 2016 and stored in a local database and maintained by the Max Planck Institute for Solid State Research (Stuttgart, Germany). The F1000Prime recommendations were taken from a data set retrieved from F1000 in November, 2017. We would like to thank Mike Thelwall for helpful correspondence regarding calculation of the CIs for MNPC and EMNPC.

References

Adie, E. (2016). [Personal Communication].

Bornmann, L. (2011). Scientific peer review. Annual Review of Information Science and Technology, 45, 199-245.

Bornmann, L. (2014). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. Journal of Informetrics, 8(4), 935-950.

Bornmann, L. (2015). Inter-rater reliability and convergent validity of F1000Prime peer review. Journal of the Association for Information Science and Technology, 66(12), 2415-2426.

Bornmann, L., & Haunschild, R. (2016a). How to normalize Twitter counts? A first attempt based on journals in the Twitter Index. Scientometrics, 107(3), 1405-1422. doi: 10.1007/s11192-016-1893-6.

Bornmann, L., & Haunschild, R. (2016b). Normalization of Mendeley reader impact on the reader- and paper-side: A comparison of the mean discipline normalized reader score (MDNRS) with the mean normalized reader score (MNRS) and bare reader counts. Journal of Informetrics, 10(3), 776-788.

Bornmann, L., & Haunschild, R. (2017). Measuring field-normalized impact of papers on specific societal groups: An altmetrics study based on Mendeley Data. Research Evaluation, 26(3), 230-241. doi: 10.1093/reseval/rvx005.

Butler, L., & McAllister, I. (2011). Evaluating university research performance using metrics. European Political Science, 10(1), 44-58. doi: 10.1057/eps.2010.13.

Claveau, F. (2016). There should not be any mystery: A comment on sampling issues in bibliometrics. Journal of Informetrics, 10(4), 1233-1240. doi: http://dx.doi.org/10.1016/j.joi.2016.09.009.

Diekmann, A., Naf, M., & Schubiger, M. (2012). The impact of (Thyssen)-awarded articles in the scientific community. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 64(3), 563-581. doi: 10.1007/s11577-012-0175-4.

Erdt, M., Nagarajan, A., Sin, S.-C. J., & Theng, Y.-L. (2016). Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. Scientometrics, 1-50. doi: 10.1007/s11192-016-2077-0.

Fairclough, R., & Thelwall, M. (2015). National research impact indicators from Mendeley readers. Journal of Informetrics, 9(4), 845-859. doi: http://dx.doi.org/10.1016/j.joi.2015.08.003.

Fleiss, J., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions (3. ed.). Hoboken, NJ, USA: Wiley.

Franceschet, M., & Costantini, A. (2011). The first Italian research assessment exercise: a bibliometric perspective. Journal of Informetrics, 5(2), 275-291. doi: DOI: 10.1016/j.joi.2010.12.002.

Galloway, L. M., Pease, J. L., & Rauh, A. E. (2013). Introduction to Altmetrics for Science, Technology, Engineering, and Mathematics (STEM) Librarians. Science & Technology Libraries, 32(4), 335-345. doi: 10.1080/0194262X.2013.829762.

Garfield, E. (1979). Citation indexing - its theory and application in science, technology, and humanities. New York, NY, USA: John Wiley & Sons, Ltd.

Glänzel, W., & Schoepflin, U. (1995). A Bibliometric Study on Aging and Reception Processes of Scientific Literature. Journal of Information Science, 21(1), 37-53. doi: Doi 10.1177/016555159502100104.

Haunschild, R., & Bornmann, L. (2016). Normalization of Mendeley reader counts for impact assessment. Journal of Informetrics, 10(1), 62-73.

Haunschild, R., & Bornmann, L. (2017). Normalization of zero-inflated data: An empirical analysis of a new indicator family. Paper presented at the ISSI 2016 Wuhan: 16th International Society of Scientometrics and Informetrics Conference, Wuhan (China). https://arxiv.org/abs/1704.02211
Haunschild, R., Schier, H., & Bornmann, L. (2016). Proposal of a minimum constraint for indicators based on means or averages. *Journal of Informetrics, 10*(2), 485-486. doi: 10.1016/j.joi.2016.03.003.

Haustein, S. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics, 108*(1), 413-423. doi: 10.1007/s11192-016-1910-9.

Haustein, S., Larivière, V., Thelwall, M., Amiot, D., & Peters, I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ? *Information Technology 2014; 56*(5): 207-215.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. New York, NY, USA: Wiley.

Kreiman, G., & Maunsell, J. H. R. (2011). Nine criteria for a measure of scientific output. *Frontiers in Computational Neuroscience, 5*(48). doi: 10.3389/fncom.2011.00048.

Mahdi, S., d'Este, P., & Neely, A. D. (2008). *Citation counts: are they good predictors of RAE scores? A bibliometric analysis of RAE 2001*. London, UK: Advanced Institute of Management Research.

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute, 22*(4), 719-748.

McKay, S. (2012). Social policy excellence - peer review or metrics? Analyzing the 2008 Research Assessment Exercise in social work and social policy and administration. *Social Policy & Administration, 46*(5), 526-543. doi: 10.1111/j.1467-9515.2011.00824.x.

National Information Standards Organization. (2016). *Outputs of the NISO Alternative Assessment Metrics Project*. Baltimore, MD, USA: National Information Standards Organization (NISO).

Neylon, C. (2014). Altmetrics: What are they good for? Retrieved October 6, 2014, from http://blogs.plos.org/opens/2014/10/03/altmetrics-what-are-they-good-for/#.VC8WETI0JAM.twitter

Plackett, R. L. (1974). *The analysis of categorical data*. London, UK: Chapman.

Radhakrishna, S. (1965). Combination of results from several 2 x 2 contingency tables. *Biometrics, 21*, 86-98.

Rons, N. (2012). Partition-based Field Normalization: An approach to highly specialized publication records. *Journal of Informetrics, 6*(1), 1-10. doi: 10.1016/j.joi.2011.09.008.

Rons, N. (2014). Investigation of Partition Cells as a Structural Basis Suitable for Assessments of Individual Scientists. In P. Wouters (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden “Context Counts: Pathways to Master Big and Little Data”* (pp. 463-472). Leiden, the Netherlands: University of Leiden.

Sheskin, D. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL, USA: Chapman & Hall/CRC.

Smith, A., & Eysenck, M. (2002). *The correlation between RAE ratings and citation counts in psychology*. London: Department of Psychology, Royal Holloway, University of London, UK.

Thelwall, M. (2016). Three practical field normalised alternative indicator formulae for research evaluation. Retrieved from https://arxiv.org/abs/1612.01431

Thelwall, M. (2017a). Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics, 11*(1), 128-151. doi: 10.1016/j.joi.2016.12.002.

Thelwall, M. (2017b). *Web Indicators for Research Evaluation: A Practical Guide*. London, UK: Morgan & Claypool.

Thelwall, M., Kousha, K., Dinsmore, A., & Dolby, K. (2016). Alternative metric indicators for funding scheme evaluations. *Aslib Journal of Information Management, 68*(1), 2-18. doi: doi:10.1108/AJIM-09-2015-0146.

Traag, V. A., & Waltman, L. (2017). *Replacing peer review by metrics in the UK REF?* Paper presented at the ISSI 2016 Wuhan: 16th International Society of Scientometrics and Informetrics Conference, Wuhan (China).

Waltman, L., & Costas, R. (2014). F1000 Recommendations as a Potential New Data Source for Research Evaluation: A Comparison With Citations. *Journal of the Association for Information Science and Technology, 65*(3), 433-445. doi: 10.1002/asi.23040.
Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: an empirical analysis. *Scientometrics, 87*(3), 467-481. doi: 10.1007/s11192-011-0354-5.

Williams, R., & Bornmann, L. (2016). Sampling issues in bibliometric analysis. *Journal of Informetrics, 10*(4), 1253-1257.

Work, S., Haustein, S., Bowman, T. D., & Larivière, V. (2015). Social Media in Scholarly Communication. A Review of the Literature and Empirical Analysis of Twitter Use by SSHRC Doctoral Award Recipients. Montreal, Canada: Canada Research Chair on the Transformations of Scholarly Communication, University of Montreal.

Wouters, P., Thelwall, M., Kousha, K., Waltman, L., de Rijcke, S., Rushforth, A., & Franssen, T. (2015). *The Metric Tide: Correlation analysis of REF2014 scores and metrics (Supplementary Report II to the Independent Review of the Role of Metrics in Research Assessment and Management)*. London, UK: Higher Education Funding Council for England (HEFCE).