Letter to the Editor

The comparison of classification-system-based normalization procedures with source normalization alternatives in Waltman and Van Eck (2013)

Along with other co-authors, Ludo Waltman and Nees Jan van Eck have recently contributed to significantly increasing the fairness of bibliometric research assessments, in particular that of multidisciplinary assessments involving comparisons of citation impact between different fields of science (Van Eck, Waltman, Van Raan, Klautz, & Peul, 2013; Waltman & Van Eck, 2013a; Waltman, Van Eck, Van Leeuwen, & Visser, 2013). This note addresses some pending issues in their latest contribution – Waltman and Van Eck (2013b) for this journal, referred to as WVE hereafter – concerning a systematic large-scale empirical comparison of classification-system-based *versus* source normalization procedures.[1] Among the former, they focus on what we will call classification system *WoS*, namely, the system where publications are classified into fields based on the journal subject categories in the Web of Science bibliographic database. WVE study the normalization procedure based on this system that uses field mean citations as normalization factors.[2] To differentiate this procedure from others of the same type, we denote it by *NWoS* (rather than *NSC* as do WVE). On the other hand, according to WVE $SNCS^{(3)}$ exhibits the best performance among the source normalization procedures. Therefore, for our purposes, the issue put forward by WVE is the comparison of *NWoS versus* $SNCS^{(3)}$.

A key methodological feature of WVE's contribution is the distinction between the use of a classification system in the implementation and the evaluation of a normalization procedure. Sirtes (2012) first suggested that using a certain classification system for evaluation purposes would be generally biased in favor of normalization procedures based on that particular system. WVE concur with this idea, and provide further arguments about the possibility of this bias (see footnote 6 and Appendix C). Therefore, they recommend that the comparison between *NWoS* and $SNCS^{(3)}$ should be done using a second, independent classification system for evaluation purposes. Following this recommendation, WVE use three systems algorithmically constructed according to the methods in Waltman and Van Eck (2012), systems *A*, *B*, and *C*, consisting of 21, 161, and 1334 scientific fields at different granularity or aggregation levels.

Given a classification system, the degree to which differences in citation practices between fields have been corrected is indicated by the degree to which the field-normalized citation distributions coincide with each other. In particular, WVE use the measurement framework introduced in Crespo, Li, and Ruiz-Castillo (2013) where, given a classification system, the effect on citation inequality of differences in citation practices is captured by an *IDCP* between-group inequality term in a certain partition of the overall citation distribution by field and quantile, where *IDCP* stands for citation *I*nequality attributable to *D*ifferences in *C*itation *P*ractices. The evaluation of any set of normalization procedures in terms of a given classification system can take a graphical or a numerical form. Following the graphical approach, WVE reach the following conclusion:

The $SNCS^{(3)}$ procedure generally performs better than the NWoS procedure, specifically at higher levels of granularity.

Li and Ruiz-Castillo (2013) establish that the graphical and the numerical approaches are logically independent. Therefore, they can be used in a complementary fashion. To save space, in most of this note I will follow the numerical approach where, given a classification system, each normalization procedure is assessed in terms of the reduction it generates in the *IDCP* term. To understand the evaluation results obtained with this approach, some notation is needed. Recall that we have four classification systems, which will be indexed by *K* = *WoS*, *A*, *B*, and *C*. Given system *K*, denote by *IDCP(K)* the *IDCP* term that captures the effect on citation inequality of differences in citation practices across fields in *K*. Similarly, given system *K*, denote by *NK* the associated normalization procedure. Finally, given a classification system *G* for the evaluation of procedure

---

[1] As Li, Castellano, Radicchi, and Ruiz-Castillo (2013) and WVE indicate, the normalization need arises at the cardinal level, that is, in situations where the actual number of citation counts of individual publications – and not only their location in a percentile distribution (or a percentile class) – is needed. At the ordinal level, the percentile rank approach provides a sort of perfect normalization where, for any classification system, all citation distributions become equally distributed. For the percentile rank approach, see *inter alia* Bornmann and Marx (2013), Bornmann (in press), and Bornmann et al. (in press).

[2] This well-known, traditional, and inexpensive normalization procedure has been favorably evaluated from a number of different perspectives in recent contributions (Radicchi et al., 2008; Radicchi and Castellano, 2012; Crespo, Li, et al., 2013; Crespo, Herranz, Li, & Ruiz-Castillo (2013); Li, Castellano, Radicchi, & Ruiz-Castillo, 2013). Therefore, in this note we will always use it as a convenient representative of classification-system-based normalization procedures.

Table 1

The impact of normalization under the four classification systems.

Change in the value of the *IDCP* term after normalization by the different procedures, in %				
Normalization	Classification system used for evaluation purposes			
Procedures	*A*	*B*	*C*	*WoS*
WoS	87.0	82.5	72.2	86.8
SNCS[3]	85.4	83.7	73.4	80.1
NA	88.8	75.3	61.0	68.7
NB	88.8	87.8	71.4	75.6
NC	89.2	88.6	86.9	80.2

NK, with *G* not necessarily equal to *K*, denote by $IDCP^{NK}(G)$ the *IDCP* term within system *G* after normalization with *NK*. To rank any pair of procedures *NK* and *NL* under classification system *G*, we compare $IDCP^{NK}(G)$ with $IDCP^{NL}(G)$. We find it more useful to express the result as the percentage that the differences in the *IDCP* terms before and after normalization, $[IDCP(G) - IDCP^{NK}(G)]$ and $[IDCP(G) - IDCP^{NL}(G)]$, represent relative to the initial situation, $IDCP(G)$. Fortunately, WVE provide the values for $IDCP(G)$ and $IDCP^{NK}(G)$ for *K* = *WoS*, *SNCS*[3] when *G* = *A*, *B*, *C* in Tables D1, D2, and D3 in Appendix D. The remaining values – for *K* = *A*, *B*, and *C*, and when *WoS* is used for evaluation purposes – have been kindly provided by Ludo Waltman (to save space, this information is available on request). Thus, we have constructed Table 1 presenting the change in the *IDCP* term before and after each of the five normalization procedures using the four classification systems for evaluation purposes.

Consider, for example, the case where normalization procedure *NWoS* is applied to the data organized according to system *A*. The consequences are captured by $IDCP^{NWoS}(A) = 0.0237$ (row 2 and column *IDCP* in Table D1 in WVE). In turn, $IDCP(A) = 0.1818$ (row 1 and column 3 in Table D1 in WVE). We are interested in the percentage change in the *IDCP* term before and after applying *NWoS* in *A*, that is, in the expression

$$\frac{100[IDCP(A) - IDCP^{NWoS}(A)]}{IDCP(A)} = \frac{100(0.1818 - 0.0237)}{0.1818} = 87.0.$$

The value of this expression appears in row *NWoS* and column *A* in Table 1, indicating that the effect of differences in citation practices across fields in system *A* has been reduced by 87% as a consequence of normalization by *NWoS*. This figure can be compared, for example, with the 88.8% reduction caused by normalization with *NA* using *A* itself for evaluation purposes (row *NA* and column *A* in Table 1). On the other hand, the figures in columns *B*, *C*, and *WoS* in row *NWoS*, for example, are the values in expression $100 [IDCP(K) - IDCP^{NWoS}(K)]/IDCP(K)$ when the evaluation system is *K* = *B*, *C*, and *WoS* rather than *A*.

Our first observation is that when we evaluate *NWoS* and *SNCS*[3] using the independent classification systems *A*, *B*, and *C* in Table 1, we obtain the same results as WVE with the graphical approach: *SNCS*[3] performs better than *NWoS* (albeit by a small margin) using systems *B* and *C*, but the opposite is the case when we use system *A*. Since system *A* is less discriminating than *B* and *C*, we conclude that *SNCS*[3] performs generally better than *NWoS* using the numerical approach.[3] Given the independence between the two approaches, this is an important result pointing toward a certain superiority of source over classification-system-based normalization procedures. However, the next two observations go against this provisional conclusion.

1. As WVE point out in their concluding Section 5, any classification system can be expected to introduce certain biases in normalization, simply because any organization of the scientific literature into a number of perfectly separated fields of science is artificial. The obvious advantage of source normalization approaches is that they are independent of any classification system. However, in most practical situations researchers are limited to working within a single classification system, and do not have access to the (active) references needed to implement any citing-side normalization procedure. In this situation, only normalization procedures of the cited-side variety are available. Assume, for example, that we are limited to working with only a single classification system *K*, where *K* could be equal to *K* = *WoS*, *A*, *B*, or *C*. According to Table 1, normalization by *NK* reduces the $IDCP(K)$ term in the range 87–89%.[4] Furthermore, one could use other cited-side normalization procedures that perform even better. For example, judging from the results obtained in Li, Castellano, Radicchi, and Ruiz-Castillo (2013), one could use the two-parameter scheme originally suggested by Radicchi and Castellano (2012). There might be better alternatives, but this large reduction in the effect on citation inequality of

---

[3] Li and Ruiz-Castillo (2013) also find that the evaluation using a less discriminating classification system that assigns publications to fields in a random manner, so as to make the differences between them as small as possible, may lead to results that contradict the conclusions obtained under other, preferable approaches.

[4] This reduction in the *IDCP* term is of the same order of magnitude as the reduction generated with two average-based normalization procedures in situations where there is only a single classification system available (Crespo, Li, et al., 2013; Crespo, Herranz, et al., 2013).