Scientometrics
DOI 10.1007/s11192-017-2559-8

And now for something completely different: the congruence of the Altmetric Attention Score's structure between different article groups

Bhaskar Mukherjee[1] · Siniša Subotić[2,3] · Ajay Kumar Chaubey[1]

Received: 2 May 2017

Abstract Altmetric Attention Score (AAS) is an increasingly popular composite altmetric measure, which is being criticized for an inappropriate and arbitrary aggregation of different altmetric sources into a single measure. We examined this issue empirically, by testing unidimensionality and the component structure congruence of the five 'key' AAS components: News, Blogs, Twitter, Facebook, and Google+. As a reference point, these tests were also done on different citation data: WoS, Scopus, and Google Scholar. All tests were done for groups of articles with: (1) high citations, but lower AAS (HCGs), and (2) high AAS, but lower citations (HAGs). Changes in component structures over time (from 2016 to 2017) were also considered. Citation data consistently formed congruent unidimensional structures for all groups and over time. Altmetric data formed congruent unidimensional structures only for the HCGs, with much inconsistency for the HAGs (including change over time). The relationship between Twitter and News counts was shown to be curvilinear. It was not possible to obtain a satisfactory congruent and reliable linear unidimensional altmetric structure between the groups for any variable combination, even after Mendeley and CiteULike altmetric counts were included. Correlations of altmetric aggregates and citations were fairly inconsistent between the groups. We advise against the usage of composite altmetric measures (including the AAS) for any group comparison purposes, until the measurement invariance issues are dealt with. The

✉ Siniša Subotić
sinisa.subotic@pmf.unibl.org

Bhaskar Mukherjee
mukherjee.bhaskar@gmail.com

Ajay Kumar Chaubey
ajaybhu21@gmail.com

[1] Department of Library and Information Science, Banaras Hindu University, Varanasi, India

[2] Laboratory of Experimental Psychology – LEP-BL, University of Banja Luka, Banja Luka, Bosnia and Herzegovina

[3] CEON/CEES, Belgrade, Republic of Serbia

Published online: 30 October 2017

underlying pattern of associations between individual altmetrics is likely too complex and inconsistent across conditions to justify them being simply aggregated into a single score.

Keywords Altmetrics · Altmetric Attention Score (AAS) · Citations · Measurement invariance · Measurement congruence

Introduction

Emergence of the Web 2.0 (O'Reilly 2005) brought upon an array of innovations, including a massive rise of social media, which also has had a major (and increasing) influence on scholarly communication practices. Scientific research can now be shared, (re)viewed, liked, blogged, tweeted, etc., in a very fast and global way via online platforms, reaching an audience that potentially expands well beyond the formal academic circles. This has opened a possibility of using such online activities as a proxy for measuring the broader impact of scientific research (Galligan and Dyas-Correia 2013). A term 'altmetrics' has been proposed as a collective name for the data related to the presence of scientific output in social web tools, such as Facebook, Twitter, blogs, news media, online reference management tools, etc. (Priem et al. 2010). Note, however, that some authors have argued against this term. For example, Ronald and Fred (2013) have instead proposed the terms 'influmetrics' and 'web-based social influmetrics'. Others did not object to the 'altmetrics' name specifically, but did point out to the necessity of distinguishing it from the slightly older concept of 'usage metrics' (Glänzel and Gorraiz 2015). Usage metrics, such as views and downloads (Gorraiz et al. 2014; Wang et al. 2016), and similarly, a count of times an article was mentioned on the internet (i.e., 'web citations') (Kousha and Thelwall 2007; Vaughan and Shaw 2005) were also considered as 'alternative' proxies for scientific impact before the 'actual' altmetrics. Regardless, 'altmetrics' currently seems to be the most widely accepted term when referring to scientific output in social web tools (e.g., Bornmann 2014, 2015; Erdt et al. 2016; Galligan and Dyas-Correia 2013; Glänzel and Gorraiz 2015; Haustein et al. 2014; Ortega 2015; Priem et al. 2010; Sud and Thelwall 2014; Thelwall et al. 2013; Wouters and Costas 2012).

It has been suggested that both usage metrics and altmetrics are useful frameworks for gaining "a much broader and more complete picture of scientific communication" (Glänzel and Gorraiz 2015, p. 2163). This, however, is arguably even more true for altmetrics than it is for usage metrics. There also appears to be a slight shift in research focus from web citations and usage statistics towards altmetrics (Bornmann 2014). Usage metrics have been around for a longer time than altmetrics and their relationship with the traditional bibliometric scientific impact metrics (i.e., citation counts) has been thoroughly studied, with correlations being generally high or moderate to high (Gorraiz et al. 2014, Kousha and Thelwall 2007; Vaughan and Shaw 2005; Wang et al. 2016). This implies that usage metrics and citation counts still measure a common impact core and point to the same or similar conclusions.

Conversely, correlations between altmetrics and citations appear to be lower, ranging from negligible values to about .50 on a high end. Specifically, Bornmann (2015) conducted a meta-analysis on the topic of association of altmetrics with traditional citations, obtaining pooled citations—microblogging (i.e., Twitter) correlation of .003, citations—blogging correlation of .12, with somewhat higher correlations for online reference managers, i.e., .23 for citations—CiteULike and .51 for citations—Mendeley. He concluded that "the more a social media community is dominated by people focussing on research, the higher the correlation between the corresponding altmetric and traditional citations is" (p. 1140), with the remark that "[l]ow correlations point to altmetrics which might be of special interest for the broad impact measurement of research, i.e. impact on other areas of society than science" (p. 1140).

While promising, research on altmetrics is still very preliminary and incomplete, and mostly focuses on associations with the traditional citation metrics, which is the first, necessary, but an insufficient research step (Bornmann 2015; Sud and Thelwall 2014). We are at the point where correlations of citations with the most popular altmetrics have already been fairly well established. This has led some prominent authors (e.g., Bornmann 2015) to argue that researchers should now focus on more in-depth studies which would examine who actually uses research/articles outside academia and for which purposes. This is especially important given the fact that it is not yet particularly clear which kind of a broader impact a given altmetric taps into the most (Sud and Thelwall 2014). For example, is it social, cultural, environmental, economic, etc. (Bornmann 2014), or do altmetrics have more to do with networking abilities and science popularization (Ortega 2015), or something else entirely?

Current research background

There are still several obstacles and technical issues related to altmetrics research (Bornmann 2014). This includes a commercial bias concern (i.e., social media providers have a financial interest in promoting as much communication through their portal/tool as possible) and a susceptibility towards manipulation (i.e., it is easy to generate high 'fake' altmetric counts). Furthermore, while altmetric data is arguably readily available, there are many data quality related issues, such as obtaining sufficient information about user groups, deciding on proper normalizations, ensuring stable replicability, etc. (Bornmann 2014; Erdt et al. 2016).

Luckily, all these issues are mostly technical in nature and it is currently 'an open season' for the development of the best altmetric framework, which would hopefully solve all the major problems. One of the arguably most popular emerging frameworks is the Altmetric.com project (Adie and Roe 2013), which is a form of the primary altmetric aggregator (Erdt et al. 2016). It is a commercial website and service that tracks, analyses, and collects an ongoing online activity around published research outputs from a large selection of online sources such as blogs, Twitter, Facebook, Google+, mainstream news outlets, media, and other sources. They show the data both individually and in a form of the automatically calculated 'Altmetric Attention Score' (AAS), which is intended to reflect both a quantity (higher attention means higher score) and a quality (weighting per different sources) of attention a research output has received. The AAS is a composite measure comprised of more than a dozen individual altmetric data sources, including Twitter, Facebook, Google+, and so on (see: https://goo.gl/E2M05n).

The AAS is based on three main factors (see: https://goo.gl/24o7fJ): (1) Volume—AAS raises as more people mention an article, but only one mention from each person per source is taken into consideration (which is intended as a form of built-in protection against manipulation). (2) Sources—categories of the mentions are weighted differently (see: https://goo.gl/E2M05n), e.g., a newspaper article mention has a higher weight than a blog post, which has a higher weight than a tweet. (3) Authors—who wrote the mention and to whom (i.e., who is the audience) is also taken into consideration, controlling for a potential bias towards a journal or a publisher. As the same source states, if a doctor shares a link with other doctors that is weighted more than an automated share from a journal account.

The AAS is becoming increasingly present in research communication channels (Gumpenberger et al. 2016), likely due to its very intuitive nature and a convenience factor. There are, however, several criticisms of the AAS (Gumpenberger et al. 2016). One of the major concerns is the mere fact that AAS is a single composite measure, which arguably simplifies the multidimensional nature of the data it contains (Gumpenberger et al. 2016). As Gumpenberger et al. (2016) point out, a single score might be convenient, but there is a danger of the AAS being misused, just like the journal impact factor, e.g., it might soon be used as a basis for ranking individuals, institutions, or even countries. Directly related to this is a criticism of the AAS' content being chosen somewhat arbitrarily. For example, AAS does not include reference managers such Mendeley or CiteULike, because, as Altmetric.com's explanation states (see: https://goo.gl/E2M05n), full details of who is making a mention cannot be shown for these sources, which conflicts with the stated idea behind the AAS—that everything included in the score must be fully transparent and visible. Gumpenberger et al. (2016), however, suggest that reference managers reflect captures and that it is a mistake to exclude them, as they are an important source of information regarding a level of online activity surrounding a given research output, which is, in fact, the intended goal of the AAS. They also point out that data on which the AAS is based on is not normalized or standardized in any way, that rounding AAS scores to integers is a mistake, and that the relative importance (i.e., statistical weights) of the parts from which the AAS is comprised of are determined using arbitrary criteria instead of valid scientific principles. Some concerns regarding a proprietary nature of the AAS and data validity were also raised (Gumpenberger et al. 2016).

On one hand, nobody is denying the fact that AAS has obviously been a huge commercial success (Gumpenberger et al. 2016). It is probably not going anywhere any time soon and it might very well become 'one altmetric to rule them all'. It seems very likely that AAS will also become increasingly popular for research purposes. It has already been established that the AAS is correlated around .30 with citations on an article level, and around .61 on a journal level (for Economics and Business Studies journals; Nuredini and Peters 2016). On the other hand, Gumpenberger et al. (2016) have raised several important criticisms of the AAS that should, in fact, be addressed if the AAS is to be considered as a valid overall representation of the 'altmetric impact,' and if it is to be used responsibly, productively, or even used at all.

Research problem

Gumpenberger et al. (2016) voiced their concerns about the AAS based on conceptual considerations. We are interested in approaching the issues from a data driven point of view, by testing if it is empirically justifiable to aggregate individual altmetrics into a single composite measure, as the AAS does, and by testing if such an aggregate/composite is suitable for group comparison purposes. Specifically, we intend to examine: (1) Can the AAS elements, i.e., altmetrics that the AAS is comprised from, be conceptualized as a meaningful unidimensional measure? (2) Are the AAS elements related to each other in a similar way, i.e., do they form a similar structure across different conditions? The reason why this is important is because if we are to compare any groups, experimental conditions, etc., using any kind of test or measure, that particular test or measure must have the same general meaning for every group or condition that is being compared. In other words, there is a necessary assumption that a measure on which comparisons are being made is on the same measurement level, i.e., that there is so called measurement invariance/equivalence across the compared groups (Drasgow 1984).

Many of the key concepts regarding measurement invariance originate from a field of psychology (e.g., Drasgow 1984; Meredith 1993; Reise et al. 1993), but the general idea holds true for any kind of comparative measurement and expands beyond psychology, as it basically addresses ‘comparing apples to oranges’ problem. One of the proposed definitions of measurement invariance/equivalence states that: “Equivalent measurement is obtained when the relations between observed test scores and the latent attribute measured by the test are identical across subpopulations.” (Drasgow 1984, p. 134). If a measurement invariance does not hold, “then differences between groups in mean levels or in the pattern of correlations of the test with external variables are potentially artifactual and may be substantively misleading” (Reise et al. 1993, p. 552). For group comparisons to be sound, structures of latent factors or components for a given composite measure should be at least partially equivalent between the groups. On the most basic level, this means that at least factor/component loadings of the measure/test/etc. should be similar between compared conditions (Meredith 1993). One simple way of testing this empirically is to determine a ‘congruence’ of loadings for a given measure’s factor or component between two (or more) conditions. A congruence is, essentially, a standardized measure of proportionality of elements (i.e., loadings) between vectors (i.e., factors or components) (Lorenzo-Seva and Ten Berge 2006).

To test if the AAS is indeed functionally unidimensional and minimally measurement invariant, i.e., congruent, we will compare its structure between groups of two different types of articles: (1) articles that are highly cited and have some, but not necessarily a very high altmetrics presence, and (2) articles that have high AAS and are cited, but not necessarily highly. Furthermore, we will also examine if the AAS is congruent across different times, i.e., does its structure remain functionally equivalent one year after an initial measurement. As an illustrative point of a direct comparison, we will also examine unidimensionality and congruence for a composite of ‘traditional’ citation measures derived from different sources, between the same groups that we will use to test the AAS on. We will also examine how altmetric and citation measures interrelate, and how these correlations vary between the groups. If we obtain unidimensional and congruent altmetric composite structures between different groups, that would not negate the criticisms of Gumpenberger et al. (2016), but it would at least provide a certain level of empirical support for the concept of AAS. Conversely, unfavorable results (i.e., low congruence, violated and/or not reliable unidimensionality) would point to the necessity for the AAS to be reconsidered.

Methods

Data and procedure

Data gathering was done in two phases. Data for phase one was collected in mid July 2016. In this phase, two groups of articles were selected, initially comprised of 100 articles each: (1) articles with generally high citation counts (i.e., ‘highly cited group’, HCG), and (2) articles with generally high ‘Altmetric Attention Score’ (i.e., ‘high AAS group’, HAG).

Articles for the HCG were selected from both Web of Science (WoS) and Scopus databases. When choosing articles, we opted to rely on a combination of WoS and Scopus in order to obtain a more diverse selection of highly cited articles and their citation counts, according to each of these databases, since their coverages are somewhat different

(Mongeon and Paul-Hus 2016). A selection process was conducted manually, using BRICS and G8 country names as arbitrary starting search terms. Only articles published in 2015 were considered, and first-appearing unique records from each database were included, until a desired number of articles was reached. We included 2015 articles with 2016 data in order to give selected articles enough time to receive citations.

Articles for the HAG were obtained from the pre-existing, publicly available Altmetric.com selection called 'The Altmetric Top 100'. We used a list from 2015 (see: https://goo.gl/dH2Rsl). Altmetric.com provides lists of the top 100 articles as per the AAS, for each year, starting from 2013.

There was a small overlap between the two article groups, as three articles from the HAG also appeared amongst the selected HCG. These articles were removed from both groups, reducing the number of articles to 97 per group. Furthermore, we removed seven articles from the HCG, since they had no altmetric records according to the available Altmetric.com data, i.e., they had no AAS at that time. Thus, the final number of articles in the HCG was 90, with 97 articles in the HAG.

We counted WoS, Scopus, and Google Scholar citations for each article, in both groups. Note that all the articles from the high AAS group had indexed citations according to at least one of the three citation sources. Furthermore, the AAS and all the individual altmetrics from which the score is comprised of (see: https://goo.gl/E2M05n) were gathered, relying upon publicly available Altmetric.com data. Due to very low or (almost) non-existent frequencies of several altmetric indices included in the AAS calculation (e.g., Wikipedia mentions, policy documents, etc.), especially in the HCG, only the following five altmetrics, with the most frequent counts (across these groups), were used in the analyses, in addition to the AAS itself: News, Blogs, Twitter, Facebook, and Google+. As most frequent, these altmetrics can arguably be seen as the 'key elements' of the AAS and should be sufficient enough to test a similarity of the AAS configuration for the two conditions (HCG vs. HAG). For the purpose of extended analyses, we also gathered Mendeley and CiteULike counts, which are not used in the calculation of the AAS (see: https://goo.gl/E2M05n), but their counts are nevertheless available from Altmetric.com.

In phase two, which was conducted in mid July 2017, we gathered updated altmetric and citation counts for the two article groups selected in phase one, using the same sources as previously described. Several HCG articles, which were previously removed due to the lack of altmetric counts, now had Altmetric.com entries, but we opted not to include them in the update, in order to be able to make direct comparisons between phase one and phase two data. In this phase, we also gathered altmetric and citation counts for an additional HAG and HCG, comprised of articles published in 2016. For the additional HAG, we used 'The Altmetric Top 100' articles for the year 2016 (see: https://goo.gl/8jtppD). For the additional HCG, we selected another collection of highly cited articles, using the same procedure as in phase one.

For 2015 articles, we refer to the data gathered in July 2016, i.e., phase one, as $\text{HCG}_{2015\_\text{ph1}}$ and $\text{HAG}_{2015\_\text{ph1}}$, for highly cited and high AAS groups, respectively. We refer to the data obtained for these groups in phase two, i.e., July 2017, as $\text{HCG}_{2015\_\text{ph2}}$ and $\text{HAG}_{2015\_\text{ph2}}$, respectively. We refer to the data obtained in July 2017 for 2016 high AAS group as HAG_{2016}, while 2016 highly cited group was named HCG_{2016}. This is summarized in Table 1.

Table 1 Article groups summary

	Article groups					
	HCG_{2015_ph1}	HCG_{2015_ph2}	HAG_{2015_ph1}	HAG_{2015_ph2}	HAG_{2016}	HCG_{2016}
Number of articles	90	90	97	97	100	100
Year of articles' publication	2015	2015	2015	2015	2016	2016
Date of data collection	Mid-July 2016	Mid-July 2016	Mid-July 2017	Mid-July 2017	Mid-July 2017	Mid-July 2017
Phases of data collection	One	Two	One	Two	Two	Two

HCG_{2015_ph1} = highly cited group from 2015, based on data from 2016. HCG_{2015_ph2} = highly cited group from 2015, based on data from 2017. HAG_{2015_ph1} = high AAS group from 2015, based on data from 2016. HAG_{2015_ph2} = high AAS group from 2015, based on data from 2017. HCG_{2015_ph1} is comprised of the same articles as HCG_{2015_ph2}. HAG_{2015_ph1} is comprised of the same articles as HAG_{2015_ph2}

Unidimensionality and measurement invariance testing procedure

There are many ways to examine if a composite measure is essentially/functionally unidimensional (e.g., Ferrando and Lorenzo-Seva 2017) or measurement invariant (e.g., Meredith 1993; Reise et al. 1993), but we opted to rely on the simplest and least restrictive tests and conditions available, all of which are based on a principal component analysis (PCA). PCA is a statistical technique used to summarize patterns of correlations among a set of observed variables, by reducing them to a smaller number of coherent subsets of variables, which are called (principal) components and represent linear combinations (i.e., vectors) formed from the observed variables (Tabachnick and Fidell 2013). PCA is similar to the exploratory factor analysis (EFA), but it is arguably less restrictive than the latter, as it uses all of the observed variables' variance, instead of only a shared variance as the EFA does (Tabachnick and Fidell 2013). If various altmetrics do, in fact, form a unidimensional structure, as the AAS implies, then this should be reproducible by PCA via extraction of a single principal component. Note that the AAS itself is functionally a component, as it is a linear combination of a handful of observed variables, i.e., individual altmetric indices. PCAs were based on log-transformed values of the variables and conducted in the FACTOR program (Lorenzo-Seva and Ferrando 2006, 2013). Other analyses were done in R (R Development Core Team 2005) or manually.

Given a proposed unidimensionality of the AAS, one component was extracted by default for each of the considered article groups. However, two procedures, namely Horn's parallel analysis (PA; Horn 1965) and Velicer's minimum average partial test (MAP; Velicer 1976) were also consulted in order to double check the justification of a single component extraction, as both of them are commonly used to assess the optimal number of components to retain in the PCA (e.g., Subotić 2013; Zwick and Velicer 1986). As a necessary and the most important condition when evaluating unidimensionality, we expect that tested altmetrics should have a minimal required loading in all conditions/groups, i.e., correlation of variables with their principal component should be at least .30, which is typically the lowest used conventional threshold (other common lower bound values are .32 or .40; e.g., Grice 2001; Tabachnick and Fidell 2013). Taking into consideration that the AAS is calculated by the addition of the altmetric values (after they are weighted, but not recoded), it is inherently assumed that the relationship between them is positive. Thus, we also expect a direction of all component loadings' signs to be positive. Finally, we took into consideration a degree to which the components reliably measure the same construct, i.e., how internally consistent they are. For this purpose, we relied on a McDonald's ω coefficient (McDonald 1999; Zinbarg et al. 2005). As a general rule of a thumb, internal consistency values should ideally be around .90 and should not fall below .70 (e.g., Kline 2010).

There are several ways and stages of assessing measurement invariance of a composite measure, and there are several levels of measurement invariance itself (e.g., Chen 2007; Meredith 1993; Lorenzo-Seva and Ten Berge 2006; Reise et al. 1993; Vandenberg and Lance 2000; Wu et al. 2007). Assuming that the unidimensionality holds, the most fundamental step in invariance testing is to establish an equivalence of loadings between the compared groups and one of the easiest (and arguably best) ways of doing so is to determine a level of factor/component congruence between groups. The most popular way of doing so, which we opted to use as well, is via the Tucker's congruence coefficient—ϕ (Tucker 1951), which is obtained by calculating a cosine of the angle between the factors/components (Lorenzo-Seva and Ten Berge 2006). If we assume two vectors (factors or components), the Tucker's congruence coefficient is calculated as

$$\phi(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

where x_i and y_i are factor/component loadings of variable i on factors x and y, respectively (Tucker 1951; Lorenzo-Seva and Ten Berge 2006). The value of ϕ in the range between .85 and .94 implies a fair similarity between two factors or components, a value higher than .95 implies that they can be considered equal, while a value of .84 and lower means that factors or components should be treated as structurally different (Lorenzo-Seva and Ten Berge 2006).

Results

Findings are presented in stages, beginning with the comparisons between article groups in mean values of the AAS and citations, followed by the unidimensionality tests, congruence tests, and finishing with the correlation analyses between the composite altmetric scores and citations. We also included several non-planned tests, to answer specific questions raised by the results themselves. One of these is a linearity test for two variables that showed inconsistent loadings between different conditions. The second one is a set of additional PCAs, conducted on subset of positively and negatively loaded variables. The third one is an ad hoc exploratory attempt to obtain a congruent component configuration, by including additional altmetric indices.

Difference between groups comparing the AAS values and citations

We first established that every HAG had significantly higher AAS values than every HCG (all $ps < .001$) and that every HCG had significantly higher mean citation counts[1] than

[1] Note that we expressed mean citations as a log-transformed average of WoS, Scopus, and Google Scholar citation counts. Using combined citations was justified given the high unidimensionality and congruence of

every HAG (all *p*s < .001). All these effects were of a high magnitude, with Cohen's *d*s ranging from 1.66 to 3.14 for the log-transformed mean AAS and from 1.53 to 3.27 for the log-transformed mean citations (Cohen 1992). Expressed using so called common language effect statistics—*CL* (Dunlap 1994; McGraw and Wong 1992), it can be stated that there is a probability between 88.02 and 98.69% that any randomly selected article from the HAGs will have a higher AAS than any randomly selected article from the HCGs. Conversely, there is a probability between 86.04% and 98.96% that any randomly selected article from the HCGs will have a higher mean citation score than any randomly selected article from the HAGs.

The changes in AAS and mean citations over time, i.e., from phase one to phase two, are shown in Table 2. While 2016 and 2017 counts were highly correlated, all the values showed a significant increase from mid-2016 (HCG_{2015_ph1} and HAG_{2015_ph1} counts) to mid-2017 (HCG_{2015_ph2} and HAG_{2015_ph2} counts), except the AAS score for the HCG_{2015_ph1} versus HCG_{2015_ph2} comparison, for which the increase over time was non-significant and trivial (Cohen 1992). Increases in mean citations were high in both groups, while the increase over time in the AAS from HAG_{2015_ph1} to HAG_{2015_ph2} was of a moderate intensity (Cohen 1992).

Unidimensionality tests

We conducted seven separate PCAs, for all six HCGs and HAGs and also for the combined data of four non-repeated groups (i.e., everything except HCG_{2015_ph2} and HAG_{2015_ph2}). The five most frequent altmetric indicators used in the calculation of the AAS were used as observed variables. A single principal component was extracted for every group by default, but in order to assess and verify unidimensionality, Horn's PA (1965) and Velicer's MAP (1976) test results were taken into consideration. Component loadings were expected to be positive and no less than .30 (Grice 2001). Internal consistency reliability (McDonald's ω; McDonald 1999; Zinbarg et al. 2005) of at least .70 and ideally around .90 was also expected. In addition, we also calculated the percentages of variance accounted for by the components (in order to assess how much information is lost due to aggregation into a single component) and the correlations of component scores with the corresponding 'raw' AAS values. The results are shown in Table 3.

For all three highly cited groups/conditions (i.e., HCG_{2015_ph1}, HCG_{2015_ph2}, and HCG_{2016}), as well as for the combined data, unidimensionality assumptions are met. Both MAP and PA suggest that it is optimal to retain one component. All loadings are positive and above the minimal .30 threshold, with internal consistencies being around the upper threshold of .90, meaning that all five altmetric variables are strongly and reliably represented on the principal components for these groups. Furthermore, components explain fairly large percentages of the variables' variance and all of the correlations of component scores with the corresponding 'raw' AAS values are high (Cohen 1992).

When it comes to the high AAS groups (i.e., HAG_{2015_ph1}, HAG_{2015_ph2}, and HAG_{2016}), unidimensionality is generally not supported, especially for the HAG_{2015_ph1}. With the exception of HAG_{2015_ph2}, several loadings in the other two HAGs are below the .30 threshold and/or are negative. MAP and PA do not conclusively suggest a single component to be the most optimal solution in these three groups. Internal consistencies are low

Footnote 1 continued
citation data from these sources, as shown in the Unidimensionality and congruence tests section of the Results. The log-transformation was used to account for highly non-normal distributions.

Table 2 Differences between phase one and phase two AAS and mean citations

Comparisons		*t*	*df*	p_t	*d*	*CL* (%)	*r*	ρ
AAS	$HCG_{2015\_ph1} = HCG_{2015\_ph2}$	− 1.70	89	.093	0.18	55.04	.97	.98
	$HAG_{2015\_ph1} < HAG_{2015\_ph2}$	− 5.86	96	< .001	0.59	66.30	.72	.83
Mean citations	$HCG_{2015\_ph1} < HCG_{2015\_ph2}$	− 24.00	89	< .001	2.52	96.32	.91	.94
	$HAG_{2015\_ph1} < HAG_{2015\_ph2}$	− 21.97	96	< .001	2.20	94.18	.97	.93

HCG_{2015_ph1} = highly cited group from 2015, based on data from 2016. HCG_{2015_ph2} = highly cited group from 2015, based on data from 2017. HAG_{2015_ph1} = high AAS group from 2015, based on data from 2016. HAG_{2015_ph2} = high AAS group from 2015, based on data from 2017. Mean citations are average of WoS, Scopus, and Google Scholar counts. *t*-statistics were calculated on log-transformed variables. *d* = Cohen's measure of effect size (Cohen 1992). *CL* = common language effect statistics (Dunlap 1994; McGraw and Wong 1992). *r* = Pearson's product-moment correlation of phase one and phase two log-transformed data. ρ = Spearman's rank correlation of phase one and phase two untransformed data. All *p*s for *r* and ρ are < .001

in two out of three groups. Percentages of explained variance were also low (below 50%) for all three HAGs, suggesting a substantial loss of information due to aggregation into a single component.

High negative loading of the Twitter variable in HAG_{2015_ph1} is mainly due to the inverse bivariate correlation of Twitter and News counts ($r_{\text{log-transformed}} = -.51$, $p < .001$; $\rho = -.45$, $p < .001$). Negative loading of News counts in HAG_{2016} (albeit below the .30 threshold) is also a result of an inverse correlation with Twitter ($r_{\text{log-transformed}}$ = - $\rho = -.34$, $p < .001$). Twitter—News correlation does seem to 'stabilize' to generally high (Cohen 1992) positive values on the combined sample ($r_{\text{log-transformed}} = .70$, $p < .001$; $\rho = .59$, $p < .001$), with both variables having positive component loadings (correlations in HCGs are also positive). However, a deeper investigation into the relationship between these variables (see Fig. 1) reveals that there is an evidence of their association being curvilinear, most likely cubic (increases in explained variance of quadratic model over linear model and cubic model over quadratic model are both statistically significant: $\Delta p = .004$, and $\Delta p < .001$, respectively. A curvilinear relationship between Twitter and News does explain why there is such an inconsistency in correlations and loading of these variables across the groups. Note that for individual groups there was no strong indication of the non-linear relationship between Twitter and News in HCGs, but with obvious signs of curvilinearity for individual HAGs.

All three HAGs' components explain much lower percentages of the observed variables' variance in comparison to the HCGs or the combined data. Their correlations with the corresponding AAS values are also lower than those of the HCGs and the combined data, i.e., correlation is low and nonsignificant for the HAG_{2015_ph1}, moderate for the HAG_{2016}, and high only for the HAG_{2015_ph2} (Cohen 1992). Note that reversing the sign of variables with negative loadings before the PCAs are conducted would not have changed the percentages of the explained variance or the correlations with the 'raw' AAS. However, given that internal consistency is dependent on the directions of the variables, we did explore what would happen if empirically observed signs were ignored, and all the values were instead treated as positive, as the AAS implicitly assumes. Doing so resulted in only a slight increase or no increase (within a rounding error) of the HAGs' internal consistencies.

Given that the HAG_{2015_ph1} is obviously displaying the highest departure from unidimensionality, we conducted two additional ad hoc PCAs, one on only positively loaded

Table 3 Principal component analysis of five AAS elements for the HCGs and the HAGs

Variables	Loadings						
	$\text{HCG}_{2015\_\text{ph1}}$	$\text{HCG}_{2015\_\text{ph2}}$	$\text{HAG}_{2015\_\text{ph1}}$	$\text{HAG}_{2015\_\text{ph2}}$	HCG_{2016}	HAG_{2016}	Comb.
News	.846	.880	.897	.661	.839	− .266	.848
Blogs	.875	.910	.702	.831	.895	.601	.921
Twitter	.887	.911	− .723	.488	.903	.885	.921
Facebook	.891	.893	− .154	.374	.911	.832	.895
Google+	.691	.749	− .187	.559	.796	.765	.834
Number of components suggested by MAP	1	1	1	1	1	1	1
Number of components suggested by PA	1	1	1 or 2	2	1	1 or 2	1
Explained variance	70.80%	75.78%	37.57%	36.35%	75.72%	49.82%	78.26%
Internal consistency (ω)	.90	.92	.49/.56	.51	.92	.76/.76	.93
Correlation with the AAS (ρ)	.95*** [.91, .98]	.95*** [.90, .98]	.10 [− .10, .29]	.77*** [.64, .87]	.94*** [.90, .96]	.45*** [.26, .61]	.89*** [.86, .92]

$\text{HCG}_{2015\_\text{ph1}}$ = highly cited group from 2015, based on data from 2016. $\text{HCG}_{2015\_\text{ph2}}$ = highly cited group from 2015, based on data from 2017. $\text{HAG}_{2015\_\text{ph1}}$ = high AAS group from 2015, based on data from 2016. $\text{HAG}_{2015\_\text{ph2}}$ = high AAS group from 2015, based on data from 2017. HCG_{2016} = highly cited group from 2016, based on data from 2017. HAG_{2016} = high AAS group from 2016, based on data from 2017. Comb. = Combined data ($N = 387$ articles) for the non-repeated groups: $\text{HCG}_{2015\_\text{ph1}} + \text{HAG}_{2015\_\text{ph1}} + \text{HCG}_{2016} + \text{HAG}_{2016}$. Reliability of internal consistency is measured by McDonald's ω coefficient (McDonald 1999; Zinbarg et al. 2005). For HAG_{2015_1} and HAG_{2016} ω was calculated for loadings as is (the first number) and with all loadings converted to positive values (the second number). ρ = Spearman's rank correlation. Correlations of the component scores with the 'raw' AAS values was calculated with the AAS' of the same year (values given in [] are 95% confidence intervals of ρ, based on 5000 bootstrap samples). $^{***}p < .001$

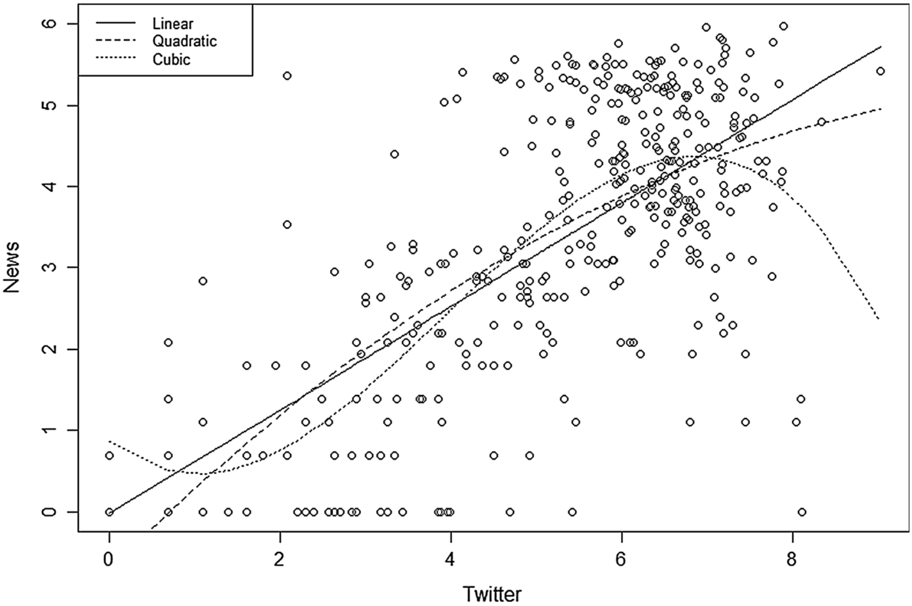

Fig. 1 Linear ($R^2 = .487$, $R^2_{\text{adjusted}} = .486$, $F(1, 385) = 365.49$, $p < .001$), quadratic ($R^2 = .498$, $R^2_{\text{adjusted}} = .495$, $F(2, 384) = 190.44$, $p < .001$), and cubic ($R^2 = .539$, $R^2_{\text{adjusted}} = .536$, $F(3, 383) = 149.41$, $p < .001$) fitting curve of the relationship between Twitter and News counts (based on log-transformed data)

items (News and Blogs), and the other on only negatively loaded items (Twitter, Facebook, and Google+). In the first case, loadings were .887 and .887, respectively, with 78.62% of variance explained. In the second case, loadings were .768, .380, and .746, respectively, with 43.05% of variance explained. These two (sub)components' scores correlated negatively with each other ($r = -.22$, $p = .03$; $\rho = -.20$, $p = .05$), but their individual correlations with the AAS: $\rho = .46$, $p < .001$, and $\rho = .48$, $p < .001$, respectively, were higher than the value of $\rho = .10$, which was originally obtained when they were combined into a single HAG$_{2015\_ph1}$ altmetric component.

As a point of comparison, we also conducted PCAs for three (log-transformed) citation counts: WoS, Scopus, and Google Scholar. PCAs were again conducted for the six article groups, plus the combined data. This is shown in Table 4. For all groups, including the combined data, principal components account for large percentages of variance, while also having generally good internal consistencies. All loadings are positive, and well above the .30 threshold, with MAP and PA both suggesting that one component is optimal for all the groups. Thus, it can be concluded that unidimensionality for citation data is fully supported in all examined groups.

Congruence tests

The similarity of the components is shown in Table 5. Reported are the values of Tucker's congruence coefficients—ϕ (Tucker 1951; Lorenzo-Seva and Ten Berge 2006) for every pair of the examined groups' components. For an easier comparison, we reported ϕ values of the components calculated on altmetric indicators (loadings from Table 3) and on citation counts (loadings from Table 4) in the same table. Congruences of the altmetric's

Table 4 Principal component analysis of citation counts for the HCGs and the HAGs

Variables	Loadings						
	HCG_{2015_ph1}	HCG_{2015_ph2}	HAG_{2015_ph1}	HAG_{2015_ph2}	HCG_{2016}	HAG_{2016}	Comb.
WoS citations	.828	.700	.949	.955	.930	.992	.974
Scopus citations	.795	.850	.962	.962	.895	.989	.970
Google Scholar citations	.899	.837	.831	.753	.786	.979	.946
Number of components suggested by MAP	1	1	1	1	1	1	1
Number of components suggested by PA	1	1	1	1	1	1	1
Explained variance	70.87%	63.73%	83.87%	80.12%	76.17%	97.37%	92.83%
Internal consistency (ω)	.81	.73	.91	.90	.86	.99	.96

HCG_{2015_ph1} = highly cited group from 2015, based on data from 2016. HCG_{2015_ph2} = highly cited group from 2015, based on data from 2017. HAG_{2015_ph1} = high AAS group from 2015, based on data from 2016. HAG_{2015_ph2} = high AAS group from 2015, based on data from 2017. HCG_{2016} = highly cited group from 2016, based on data from 2017. HAG_{2016} = high AAS group from 2016, based on data from 2017. Comb. = Combined data ($N = 387$ articles) for the non-repeated groups: $HCG_{2015\_ph1} + HAG_{2015\_ph1} + HCG_{2016} + HAG_{2016}$. ω = McDonald's coefficient of internal consistency reliability (McDonald 1999; Zinbarg et al. 2005)

Table 5 Congruence coefficients

Comparison pairs	HCG$_{2015\_\mathrm{ph1}}$	HCG$_{2015\_\mathrm{ph2}}$	HAG$_{2015\_\mathrm{ph1}}$	HAG$_{2015\_\mathrm{ph2}}$	HCG$_{2016}$	HAG$_{2016}$	Comb.
HCG$_{2015\_\mathrm{ph1}}$	–	1.00	.99	.99	.99	1.00	1.00
HCG$_{2015\_\mathrm{ph2}}$	1.00	–	.99	.99	.99	1.00	1.00
HAG$_{2015\_\mathrm{ph1}}$	.18/.88	.18/.89	–	1.00	1.00	1.00	1.00
HAG$_{2015\_\mathrm{ph2}}$	.96	.97	.36/.92	–	1.00	.99	1.00
HCG$_{2016}$	1.00	1.00	.16/.87	.96	–	1.00	1.00
HAG$_{2016}$	.79/.94	.79/.94	− .34/.73	.70/.87	.81/.95	–	1.00
Comb.	1.00	1.00	.17/.87	.97	1.00	.81/.95	–

HCG$_{2015\_\mathrm{ph1}}$ = highly cited group from 2015, based on data from 2016. HCG$_{2015\_\mathrm{ph2}}$ = highly cited group from 2015, based on data from 2017. HAG$_{2015\_\mathrm{ph1}}$ = high AAS group from 2015, based on data from 2016. HAG$_{2015\_\mathrm{ph2}}$ = high AAS group from 2015, based on data from 2017. HCG$_{2016}$ = highly cited group from 2016, based on data from 2017. HAG$_{2016}$ = high AAS group from 2016, based on data from 2017. Comb. = Combined data ($N = 387$ articles) for the non-repeated groups: HCG$_{2015\_\mathrm{ph1}}$ + HAG$_{2015\_\mathrm{ph1}}$ + HCG$_{2016}$ + HAG$_{2016}$. Values below diagonal are congruence coefficients for the altmetric variables PCAs (as shown in Table 3). Values above the diagonal are congruence coefficients for the citation variables PCAs (as shown in Table 4). When two values are reported, the first one is calculated for loadings as is (the first number) and with all loadings converted to positive values (the second number)

data components are shown below the diagonal and congruences of the citation's data components are shown above the diagonal. It is obvious that congruences for the altmetric's data components are very inconsistent, with 11 out of 21 values being below the critical value of .85, which means that in more than half of the pairs, compared altmetric components are, in fact, structurally different. Note, however, that components for altmetric indicators are fully congruent for the HCGs, as all the ϕ values are well above the upper threshold of .95. Thus, these components can be regarded as functionally equal. This includes the equivalence over time for the altmetric data gathered in 2016 and 2017, for the same group of highly cited articles from 2015 (i.e., HCG$_{2015\_\mathrm{ph1}}$ vs. HCG$_{2015\_\mathrm{ph2}}$ comparison). This also includes a full equivalence of every HCGs' components with the combined altmetric's data component. Conversely, none of the altmetric's data components are congruent between themselves, including the lack of congruence over time for the altmetric data gathered in 2016 and 2017, for the same group of high AAS articles from 2015 (i.e., HAG$_{2015\_\mathrm{ph1}}$ vs. HAG$_{2015\_\mathrm{ph2}}$ comparison). All four cases in which altmetric's data components do show congruence include HAG$_{2015\_\mathrm{ph2}}$, whose component is congruent with HCGs' components and with the combined data component. If we were to remove negative signs from all loadings, then 10 additional component pairs (which include HAGs' components) would become somewhat congruent (i.e., they would be in the .85 to .94 range, which implies fair similarity), leaving only HAG$_{2015\_\mathrm{ph1}}$ vs. HAG$_{2016}$ to remain incongruent.

Congruence values of all the citation's data components are well above the .95 cutoff, meaning that all the citation's components for either HCGs or HAGs are structurally equivalent, exhibiting much stronger similarity in comparison to the altmetric's data components.

Note that before the AAS calculation, Altmetric.com weights the altmetric variables according to arbitrarily predefined criteria (Gumpenberger et al. 2016). Thus, we also considered PCAs and corresponding congruence coefficients in which (prior to log-transformation) the variables were weighted according to a default Altmetric.com criteria (described at: https://goo.gl/E2M05n). While the obtained numerical values were not identical, they pointed to the same conclusion that was derived from the reported unweighted values, thus we do not report the results based on weighted values, due to their redundancy.

Ad hoc question: can a congruent altmetric combination be achieved?

Given the results of fairly incongruent structures of the altmetric's data components, we also wanted to explore if it is actually possible to obtain any combination of altmetric variables which would be congruent on our data, for all the groups. Since 'Blogs' was the only altmetric variable which had a positive sign and a loading above .30 in the initial analysis for all the groups (see Table 3), we decided to expand the variable selection by including Mendeley and CiteULike altmetric counts, even though they are not actually parts of the AAS (but there are arguments that they should be; see, e.g., Gumpenberger et al. 2016). After this, we repeated the analyses previously reported in Table 3. However, adding these variables did not result in much improvement, and most of the values actually worsened, with the most notable exception being HAG_{2015_ph1}—AAS correlation which improved from $\rho = .10$ to $\rho = .51$ ($p < .001$). Mendeley integrated mostly poorly into the components, with loadings below .30 in four cases, including the combined data. Specifically, the Mendeley loadings were: .274, .239, .687, .695, .260, .738, and .181, respectively. CiteULike integrated slightly better, with its loadings being: .348, .264, .535, .449, .439, .616, and .397, respectively. In an effort to improve the results, we did several iterations of the PCAs, each time removing a different 'problematic' variable, starting from the ones with negative loadings. After all the variables with negative loadings were removed, Mendeley and CiteULike loadings increased above the .30 threshold in all the groups. In this solution, however, only Blogs, Mendeley, and CiteULike variables remained. This is shown in Table 6.

Unidimensionality seems to hold for every group. Note that there is a high level of congruence between every group pair, with all the ϕ values being well above the .95 cutoff (e.g., the lowest ϕ value is .97, for the HCG_{2015_ph2} vs. HAG_{2015_ph2} comparison and all other coefficients are in .98–1.00 range). However, with the exception of HAG_{2016}, internal consistencies are well below the minimal cutoff (.70), and principal components explain only around 50% of the variables' variance. Correlations with the 'raw' AAS are mostly moderate (Cohen 1992).

Taken together, these ad hoc tests suggest that it is technically possible to obtain a congruent altmetric combination with the addition of Mendeley and CiteULike altmetric counts, but that combination is not very satisfactory. It includes only three variables, it is not very internally consistent, and it accounts for only moderate amounts of its variables' variance.

Correlation between the composite altmetric scores and citations

Correlation between the AAS and components presented in Tables 3 and 6 with mean citation counts are shown in Table 7. Correlations are calculated separately for all six HCGs and HAGs and for the combined data.

Judging from the correlation values and their 95% confidence intervals (CIs) it is obvious that there is little consistency in correlations between the groups. Values range from trivial to large (Cohen 1992), and from negative to positive. There are several

Table 6 Principal component analysis with added Mendeley and CiteULike (the final iteration)

Variables	Loadings						
	HCG_{2015_ph1}	HCG_{2015_ph2}	HAG_{2015_ph1}	HAG_{2015_ph2}	HCG_{2016}	HAG_{2016}	Comb.
Blogs	.564	.476	.624	.778	.716	.776	.612
Mendeley	.697	.766	.837	.696	.645	.842	.702
CiteULike	.777	.805	.800	.719	.777	.825	.833
Number of components suggested by MAP	1	1	1	1	1	1	1
Number of components suggested by PA	1	1	1	1	1	1	1
Explained variance	46.90%	48.69%	57.68%	53.54%	51.08%	66.43%	52.02%
Internal consistency (ω)	.47	.51	.65	.57	.53	.75	.62
Correlation with the AAS (ρ)	.55*** [.38, .69]	.45*** [.24, .62]	.49*** [.32, .64]	.48*** [.31, .62]	.66*** [.52, .78]	.35*** [.14, .53]	.46*** [.37, .55]

HCG_{2015_ph1} = highly cited group from 2015, based on data from 2016. HCG_{2015_ph2} = highly cited group from 2015, based on data from 2017. HAG_{2015_ph1} = high AAS group from 2015, based on data from 2016. HAG_{2015_ph2} = high AAS group from 2015, based on data from 2017. HCG_{2016} = highly cited group from 2016, based on data from 2017. HAG_{2016} = High AAS group from 2016, based on data from 2017. Comb. = Combined data ($N = 387$ articles) for the non-repeated groups: $HCG_{2015\_ph1} + HAG_{2015\_ph1} + HCG_{2016} + HAG_{2016}$. Reliability of internal consistency is measured by McDonald's ω coefficient (McDonald 1999; Zinbarg et al. 2005). ρ = Spearman's rank correlation. Correlations of the component scores with the 'raw' AAS values was calculated with the AAS' of the same year (values given in [ ] are 95% confidence intervals of ρ, based on 5000 bootstrap samples). ***$p < .001$

Table 7 Correlations of the AAS and altmetric components with mean citations

	HCG_{2015_ph1}	HCG_{2015_ph2}	HAG_{2015_ph1}	HAG_{2015_ph2}	HCG_{2016}	HAG_{2016}	Comb.
AAS	.05 [− .16, .26]	.10 [− .12, .30]	.25 [.03, .45]	.35 [.13, .52]	.16 [− .04, .35]	.30 [.10, .48]	− .53 [− .60, − .45]
Table 3 component	.09 [− .12, .29]	.09 [− .13, .30]	− .18 [− .36, .02]	.41 [.22, .58]	.10 [− .10, .28]	.58 [.42, .71]	− .47 [− .54, − .39]
Table 6 component	.31 [.13, .49]	.31 [.10, .49]	− .07 [− .26, .13]	.47 [.29, .62]	.03 [− .17, .23]	.62 [.46, .75]	− .06 [− .15, .04]

HCG_{2015_ph1} = highly cited group from 2015, based on data from 2016. HCG_{2015_ph2} = highly cited group from 2015, based on data from 2017. HAG_{2015_ph1} = high AAS group from 2015, based on data from 2016. HAG_{2015_ph2} = high AAS group from 2015, based on data from 2017. HCG_{2016} = highly cited group from 2016, based on data from 2017. HAG_{2016} = high AAS group from 2016, based on data from 2017. Comb. = Combined data ($N = 387$ articles) for the non-repeated groups: $HCG_{2015\_ph1} + HAG_{2015\_ph1} + HCG_{2016} + HAG_{2016}$. Correlations are Spearman's rank correlation (ρ) and values given in [] are 95% confidence intervals of ρ, based on 5000 bootstrap samples. Correlations are always calculated with the corresponding mean citation counts instances of CIs very noticeably not overlapping (which suggests significant differences in correlation values) or overlapping very slightly. This occurs both between the groups, but on the same altmetric measure (e.g., HAG_{2015_ph2} or HAG_{2016} vs. combined data for all three altmetric composites/components), and within the groups, but on different altmetric variables (e.g., correlation of the AAS with mean citations vs. correlation of the Table 3 component with mean citations, within the HAG_{2015_ph1}).

The most important trend to point out is that correlations of altmetric measures with mean citation counts within individual HCGs and HAGs are mostly positive, while, in contrast, correlations for the combined data group are moderately to strongly negative (Cohen 1992) for two out of three variables. Taking a closer look at the data patterns again suggested a curvilinear nature of the associations. However, due to high concern for the correlations being artifactual and misleading considering the non-congruence (Reise et al. 1993), we did not explore this issue any further at this point.[2]

Discussion

This research was conducted in order to examine if it is justified to group different altmetric measures into one composite score, as it is done in increasingly popular Altmetric.com's (Adie and Roe 2013) 'Altmetric Attention Score'—AAS. Some critics raised their concerns over such practice (and the AAS specifically) on a conceptual level, suggesting that this approach carries the risk of becoming an altmetric version of the impact factor (with accompanying potential for misuse) and that it likely simplifies the multidimensional nature of the altmetric data with no good justification (Gumpenberger et al. 2016). To the best of our knowledge, our research is the first formal empirical examination of this issue.

We conducted a series of analyses on the five most frequent AAS composites, i.e., on the five key AAS altmetric indicators: News, Blogs, Facebook, Twitter, and Google+ counts. These analyses were done in order to answer two specific questions. First, is a combination of these individual altmetrics functionally unidimensional—which is a necessary condition if they are to be aggregated into a single composite score? Second, are the altmetrics interrelated in a similar fashion across conditions, i.e., is their structure congruent between different article groups—which is a necessary property if comparisons are to be made based on their composite scores (Drasgow 1984; Meredith 1993; Reise et al. 1993)?

For the groups of highly cited articles (HCGs), five examined altmetric variables grouped together strongly, reliably, and a unidimensional structure seems to be justified. In other words, these five altmetric indicators can be aggregated into a single principal component (composite score) without any substantial loss of variance or misrepresentation of a nature of the data on these groups. Furthermore, the structures of all the HCGs' component loadings are functionally equivalent according to Tucker's congruence coefficients—ϕ (Tucker 1951; Lorenzo-Seva and Ten Berge 2006), including an equivalence over time. This means that HCGs' scores can be safely compared among themselves. However, this was not the case for the groups of articles with high AAS (HAGs), where the individual altmetric variables seem to be very loosely, and even inversely interrelated, with a lot of variance/information being lost when a single component was extracted.

[2] In fact, even several examinations of differences between groups in the AAS and citations done at the beginning of the Results is, strictly speaking, not justified due to later established incongruence.

Furthermore, the structures of the HAGs' principal components were mostly incongruent, i.e., they differed substantially when compared among themselves (in every single comparison), when compared to the HCGs' components (with only two comparison pairs being exceptions), and when compared to the combined groups' data (with one exception). Not even the structures of the same HAG in two different time points were equivalent. Thus, component scores from half of the examined conditions are empirically not suitable for group comparison purpose.

For one of the HAG's, correlation with its corresponding 'raw' AAS value, which should supposedly represent the same general underlying structure, was only .10. This illustrates how much of a discrepancy can exist between the 'wrongly' presupposed structure, i.e., the 'raw' AAS, and the actual empirically observed structure, i.e., the structure obtained by the principal component analysis. In this particular case, it was clear that observed altmetric variables should not be forcefully combined together, as the AAS does, because the data was clearly suggesting that there are two, inversely, rather than positively related subcomponents.

One of the major sources of the incongruence between the components was a curvilinear relationship between Twitter and News counts, which became obvious when the data was combined. Curvilinearity likely explains why both bivariate correlations and loadings of these variables were 'flip-flopping' between positive and negative values, depending on the article group. This implies that an underlying structure of at least some of the altmetric data is likely much more complex than a simple unidimensional configuration. Thus, it is probably not wise to aggregate altmetric data into a single composite score, such as the AAS, even if the unidimensonality and congruence assumptions appear to hold for a particular article group—as the same might not be true across the board.

Most of our findings point to a conclusion that the AAS scores should not be calculated and used for any kind of comparison purpose, at least until a reliable unidimensional, congruent, and linear structure of a particular combination of individual altmetrics is found. Such structure should also be confirmed on sufficiently large datasets and under multiple conditions, including low, medium, and high values of altmetric counts. We made a small ad hoc attempt to identify a congruent and reliable altmetric composite structure on our data. However, this was not particularly successful, even when additional altmetrics, namely Mendeley and CiteULike, which are controversially not included in the AAS (Gumpenberger et al. 2016), were added to the pool of the examined variables. This resulted only in Blogs, Mendeley and CiteULike variables congruently grouping together for all the examined conditions/groups, but with only moderate amounts of explained variance and fairly low internal consistencies.

We also briefly explored the associations of aggregated/composite altmetric data (i.e., the 'raw' AAS and two versions of the altmetric principal components) with citation counts. While several correlations do seem to converge to previously reported value of .30, obtained between the AAS and citations on a large sample of Economics and Business Studies articles (Nuredini and Peters 2016), coefficients are generally 'hectic', varying from positive to negative, in a trivial to large intensity magnitude (Cohen 1992). While inconsistent correlations between aggregated altmetric values and citations are not in and of itself a sign of incongruence (as group memberships can moderate correlations between otherwise congruent measures), they do add up to a list of potential causes of concern regarding the integrity and usability of the composite altmeric measures. This is especially true since there were signs of altmetric—citations curvilinearity, which we did not explore in detail, precisely due to high levels of previously determined incongruency, which does render correlation patterns potentially artifactual and misleading (Reise et al. 1993). Thus, this issue should be reexamined if and when a congruent altmetric composite measure is obtained.

Several criticisms can be raised regarding our conclusion that a unidimensional composite altmetric measure, such as the AAS, should not be computed and used for comparison purposes. For example, we did not examine all the parts of the AAS specifically, due to some individual altmetric variables having too low counts to be taken into consideration. However, this does not invalidate the conclusion, as the five altmetrics that we did consider are arguably the most important ones (and certainly the most popular, i.e., the most frequently occurring ones). The fact that they do not seem to interrelate in a unidimensional fashion, with signs of non-linearity, is enough to point to a problem.

There is also a potential concern related to our data gathering methodology. Articles that we had selected in phase one of data collection were published during 2015, but the data were collected in mid-2016 and again in mid-2017, while the data for articles published in 2016 were collected in mid-2017. This means that there was a delay of about 0.5-1.5 years from articles' publication to data collection (and twice that for the phase two data for 2015 articles). This delay, albeit arguably still short, was necessary in order to give the articles enough time to receive a certain amount of citations. However, this also means that altmetric data for the HAGs from 2015 and 2016 had the same delay, i.e., altmetric counts that we have gathered are not the original counts based on which these articles were initially ranked in their respective '[The] Altmetric Top 100' lists, from which we selected them. This, however, is not a severe limitation from the perspective of this study, given that the only important selection requirement was that such articles, on average, have substantially higher altmetric counts in comparison to the HCG's articles. Our research focus was not on the individual altmetric mean values and their changes, but rather on measurement invariance, i.e., on the consistency in the structure of relationships between different altmetric variables when they are aggregated together. The mean values of articles have obviously changed from the time they were initially awarded their 'Top 100' status, but this is not important, as we wanted to study their congruence and not their means specifically. As Drasgow (1984, p. 134) points out: "[…] in the definition of measurement equivalence there is no requirement of equal means, variances, skewnesses, or kurtoses for either the observed test scores or the latent trait. Measurement equivalence requires an identical relation between test scores and the latent trait." On a related note, we did not specifically match the articles and groups based on the subject fields/disciplines, article types, or anything along these lines. However, we argue that this is also largely irrelevant in the context of our research problem. If the AAS (or any other aggregate/composite altmetric measure, for that matter) is to be used as a basis for comparison of scientific articles in general, it should have the same underlying structure regardless of the subject fields, article types, etc. (and regardless of potential differences in altmetric means or medians between groups of articles from different fields or of different types). Note that under these same 'unmatched' conditions, the structure of citation data is perfectly congruent, while the structure of altmetric data is not.

Another point of concern is the fact that the most problems were observed in the HAGs, whose average altmetric values are by definition higher than those of HCGs, in which no particular problems were observed, as both unidimensionality and congruence do seem to hold. In other words, HAGs represent groups of mostly high-end values. While this is true to an extent, the same can be stated for the HCGs with regards to citation counts from different sources (WoS, Scopus, and Google Scholar), which we also examined mainly for this precise reason. Note that differences in mean AAS—which favor HAGs, are virtually of the same magnitude as are the differences in mean citations—which favor HCGs.

However, there are no issues for citation variables, in either HCGs or HAGs, as unidimensionality holds in every group, internal consistencies are high, and there is a high congruence for every single group comparison, both within and between HCGs and HAGs. This also includes congruence over time. Furthermore, when altmetic data of four groups were combined, a reasonably representative range of low, middle, and high values was obtained, which, in turn, enabled us to identify the additional curvilinearity issue described earlier—none of which was observed for citation counts. Thus, under very similar conditions, citation data aggregates very well, while altmetric data does not.

It is also important to clarify why we opted to focus on the 'extreme' groups. The reason is multifaceted. Given a concern that the AAS might become misused 'just like the impact factor' (Gumpenberger et al. 2016), it is likely going to be used to compare the top articles, researchers, institutions, etc., i.e., to differentiate 'good from the best'. This is, essentially, what 'The Altmetric Top 100' does right now, on an article level. Comparisons of articles based on altmetrics and 'traditional citations' are also intuitively attractive, as 'new school versus old school' comparisons arguably often are. In such cases, groups with high altmetric counts are likely to be put against groups with high citation counts. Therefore, we argue that it is very important for measurement invariance to be present in such 'extreme' cases. Also, given that 'The Altmetric Top 100' lists are publicly available, they make a good basis for the follow-up comparisons by other researchers.

Concluding recommendations

Naturally, our findings should necessarily be replicated and extended, ideally using much larger data sets, more groups and time points, which should all become feasible reasonably soon, as time passes and further data accumulates. However, based on our current findings, we would caution against the usage of the AAS for any group comparisons, including both research and practical purposes (such as policy and decision making). We also extend our warning to any other aggregate/composite altmeric measures for which the issue of measurement invariance is not successfully solved. Measurement invariance appears to be too big of a problem to allow for the composites of currently most popular altmetrics to be safely used for any consequential purpose. The underlying pattern of associations between these individual altmetrics is likely too complex, too inconsistent, and too unreliable across different conditions to justify them being aggregated into a coherent single score. If this issue is ignored, and comparisons based on composite altmetric measures are conducted nevertheless—such comparisons would likely have no actual meaning. Until such time when the measurement invariance issue is hopefully solved (i.e., some reliable and congruent linear unidimensional altmetric composite structure is found) it is safer to rely upon individual altmetrics instead of their ad hoc composites. Furthermore, we would appeal to Altmetric.com to consider and address our and other constructive criticisms (e.g., Gumpenberger et al. 2016) and try and re(de)fine the idea behind the AAS, as it arguably still has a potential worth exploring.

References

Adie, E., & Roe, W. (2013). Altmetric: Enriching scholarly content with article-level discussion and metrics. *Learned Publishing, 26*(1), 11–17.

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). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics, 103*(3), 1123–1144.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Development Core Team, R. (2005). *R: A language and environment for statistical computing [Computer Software].* Austria: R Foundation for Statistical Computing.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95*(1), 134–135.

Dunlap, W. P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin, 116*(3), 509–511.

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, 109*(2), 1117–1166.

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164417719308.

Galligan, F., & Dyas-Correia, S. (2013). Altmetrics: Rethinking the way we measure. *Serials Review, 39*(1), 56–61.

Glänzel, W., & Gorraiz, J. (2015). Usage metrics versus altmetrics: Confusing terminology? *Scientometrics, 102*(3), 2161–2164.

Gorraiz, J., Gumpenberger, C., & Schlögl, C. (2014). Usage versus citation behaviours in four subject areas. *Scientometrics, 101*(2), 1077–1095.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430–450.

Gumpenberger, C., Glänzel, W., & Gorraiz, J. (2016). The ecstasy and the agony of the altmetric score. *Scientometrics, 108*(2), 977–982.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics, 101*(2), 1145–1163.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185.

Kline, P. (2010). *Handbook of psychological testing* (2nd ed.). London: Routledge.

Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multidiscipline exploratory analysis. *Journal of the American Society for Information Science and Technology, 58*(7), 1055–1065.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38*(1), 88–91.

Lorenzo-Seva, U., & Ferrando, P. J. (2013). FACTOR 9.2: A comprehensive program for fitting exploratory and semiconfirmatory factor analysis and IRT models. *Applied Psychological Measurement, 37*(6), 497–498.

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*(2), 57–64.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*(2), 361–365.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics, 106*(1), 213–228.

Nuredini, K., & Peters, I. (2016). Enriching the knowledge of altmetrics studies by exploring social media metrics for Economic and Business Studies journals. *EconStor Conference Papers, ZBW—German National Library of Economics.* Retrieved from http://EconPapers.repec.org/RePEc:zbw:esconf:146879.

O'Reilly, T. (2005). What is web 2.0? Design patterns and business models for the next generation of software. Retrieved from http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1.

Ortega, J. L. (2015). Relationship between altmetric and bibliometric indicators across academic social sites: The case of CSIC's members. *Journal of Informetrics, 9*(1), 39–49.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. Retrieved from http://altmetrics.org/manifesto/.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566.

Ronald, R., & Fred, Y. Y. (2013). A multi-metric approach for research evaluation. *Chinese Science Bulletin, 58*(26), 3288–3290.

Subotić, S. (2013). Pregled metoda za utvrđivanje broja faktora i komponenti (u EFA i PCA) [Review of methods for determining the number of factors and components to retain (in EFA and PCA)]. *Primenjena Psihologija, 6*(3), 203–229.

Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. *Scientometrics, 98*(2), 1131–1143.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLoS ONE, 8*(5), e64841.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70.

Vaughan, L., & Shaw, D. (2005). Web citation data for impact assessment: A comparison of four science disciplines. *Journal of the Association for Information Science and Technology, 56*(10), 1075–1087.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321–327.

Wang, X., Fang, Z., & Sun, X. (2016). Usage patterns of scholarly articles on Web of Science: A study on Web of Science usage count. *Scientometrics, 109*(2), 917–926.

Wouters, P., & Costas, R. (2012). *Users, narcissism and control—Tracking the impact of scholarly publications in the 21st century*. Utrecht: SURF foundation.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment Research and Evaluation, 12*(3), 1–26.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432–442.