Full paper

STI 2022 Conference Proceedings

Proceedings of the 26th International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Proceeding Editors

Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

Citation: Mutz, R. (2022). Why simply summing up any bibliometric indicators does not justify a good composite indicator for individual researcher assessment - A measurement perspective. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti22136). https://doi.org/10.5281/zenodo.6975538

Collection: https://zenodo.org/communities/sti2022grx/

Why simply summing up any bibliometric indicators does not justify a good composite indicator for individual researcher assessment - A measurement perspective

Rüdiger Mutz*

*ruediger.mutz@uzh.ch
Center of Higher Education and Science Studies, University of Zurich, Plattenstrasse 54, 8032 Zurich (Switzerland)

Introduction

Bibliometrics and scientometrics are dominated by the notion of single indicators that comprehensively capture a certain content, expressed in a number. For example, the h-index simultaneously measures the quantity and citation impact of a researcher's research output in one number (Fraumann & Mutz, 2021). Indicators composed of various individual indicators are discussed only sporadically in the bibliometric literature, and if so, then only very critically (Glänzel & Debackere, 2009). This poor reputation apparently stems from their use in most international university rankings, which are the subject of very severe criticism (Ioannidis et al., 2007; Moed, 2017).

The following problems of composite indicators can be identified (Barclay, Dixon-Woods, & Lyratzopoulos, 2019; Glänzel & Debackere, 2009):

1. *Correlated components*: Single indicators may not be independent of each other, i.e. a change in one indicator has an effect on another indicator. Correlated indicators are weighted more heavily in the composite score than non-correlated ones.
2. *Multidimensionality*: The possible multidimensionality of the research output is reduced to one dimension. Indicators are often chosen that do not fit together (comparing "apples with oranges").
3. *Arbitrary weightings*: The choice of weights for an individual component is often arbitrary. Rankings of individuals may depend on the choice of weights.
4. *Lack of transparency*: The construction of composite indicators requires a multitude of decisions (choice of indicators, choice of weights, treatment of missing values, ...). Often composite indicators are presented without information on their derivation.
5. *Simplistic policy conclusions*: "The simple big picture results which composite indicators show may invite politicians to draw simplistic policy conclusions." (Saisana & Tarantola, 2009, p. 5).
6. *Uncertainty*: "Composite indicators are not immune to chance variation: tiny differences in individual measures can translate into differences in the final rating, but will often be due to chance." (Barclay et al., 2019, p. 340).

Glänzel and Debackere (2009, p. 69), therefore, propose to completely abandon the concept of composite indicator in favour, for instance, of individual comparison among institutions.

Here, however, a different path is to be taken. It is argued that a central reason for the poor reputation of composite indicators is the lack of a concepts of aggregation that justify the construction of composite indicators from single indicators. Therefore, the first aim of this paper is to use measurement concepts from psychology, psychometrics for the construction of bibliometric composite indicators. Interestingly, Ioannidis, Boyack, and Baas (2020) are among the few prominent representatives in bibliometrics who favour composite indicators. The authors have compiled an extensive data set on excellent researchers from various scientific fields worldwide. Therefore, the second aim of this paper is to reanalyse their data to answer the question to what extent the developed composite indicator fulfils the psychometrical requirements of a measurement scale.

Measurement perspective

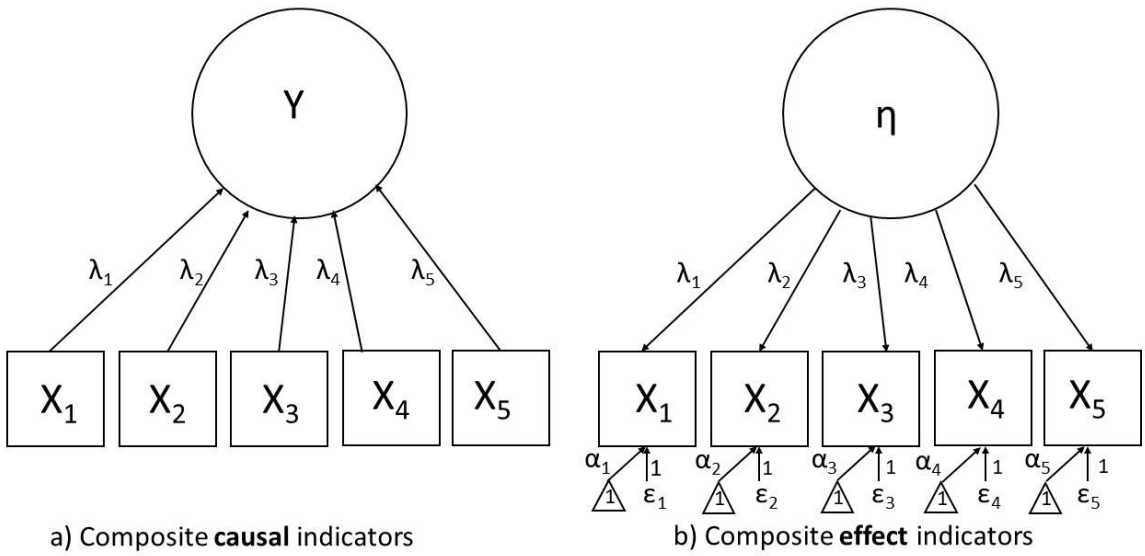
Basic concepts

The starting point should be a formal definition of composite indicators: «Different indicators X_k representing different aspects of quality, form the components of a composite indicator Y, the basis of the ranking; this composite indicator is usually a linear combination of the X_k`s, that is;

$$Y=\sum\lambda_k X_k \qquad (1)$$

where λ_k (k=1,2, …, p) are p pre-defined weightings and, without loss of generality, verify the equality $\sum\lambda_k=1$ (this last relation implies that Y is actually a weighted mean of the individual indicators X_k).” (Glänzel & Debackere, 2009, p. 66) (equation number added by the author). Usual composite indicators can also be represented graphically (Figure 1a), in which the indicators are the “causes” of the sum variable Y (see direction of arrows) (Bollen & Bauldry, 2011, p. 268).

Figure 1. Measurement models for different kinds of composite indicators



Saisana and Tarantola (2009) and Freudenberg (2003) provide methodological introductions to the construction of composite indicators. A common method to obtain weights for the individual indicators in particular is the Data Envelopment Analysis (DEA). Here, however,

we will present a measurement concept as the basis of composite indicators that is widely used in psychometrics, the so-called Classical Test Theory (CTT), which ultimately represents a measurement error theory (e.g., Mutz, Bornmann, & Daniel, 2016; Steyer, 1989).

CTT assumes that a theoretical construct is to be measured that is not directly observable but empirically manifests itself in indicators called items. Measurement is understood as the mapping of an empirical relative (objects) into a numerical relative, whereby relations in the empirical relative are mapped into the numerical relative (e.g., larger-smaller, addition). Here, the construct "scientific impact of researcher`s work in the scientific community" is to be measured, which is not directly observable (latent) but has an effect on the indicators. The starting point of CTT is the fact that a single indicatoris in principle subject to random errors, i.e. uncertainty, which is expressed in the following basic equation for a researcher *u* (Steyer, 1989, p. 28).

$$X_i=T_i+\varepsilon_i \qquad (2)$$

where X_i is the value of a single indicator *i* or item, T_i is the true score and ε_i is the random error. For instance, the total number of citations as first author can be a item for the scientific impact of a researcher's work. The true score T_i is ultimately the expected value of X for an individual researcher *u*: $E(X_i|u)=T_i$ with $E(\varepsilon_i|u)=0$. In order to determine the true score of a researcher, *independent replications* are needed, i.e. items that measure the construct "scientific impact" in the same way.

If, in addition, it is assumed that the true score and the error are uncorrelated, $Cov(T_i, \varepsilon_i)=0$, the errors are not correlated with each other, $Cov(\varepsilon_i, \varepsilon_j)=0$, and the error of one item is not correlated with the true score of another item, $Cov(T_i, \varepsilon_j)=0$, then a measurement scale is established. In this latent variable model, it is assumed that the latent variable is the cause of the observed correlations (Figure 1b), with the latent variable η representing the true score variable (Bollen & Bauldry, 2011, p. 268). The variance of an item is then composed as follows:

$$var(X_i)=var(T_i)+var(\varepsilon_i) \qquad (3)$$

Measurement models

Ultimately, a measurement model is not only about individual items and their respective true score, but about the correlation of the items in an overall score. The true score of an item *i* can be predicted from the true score of another item *j* that measures the same construct using the following regression model:

$$T_i=\alpha_{ij}+\lambda_{ij} T_j, \qquad (4)$$

where α_{ij} is the intercept and λ_{ij} is the slope of the regression or simplified by Eq. 5:

$$T_i=\alpha_i+\lambda_{1i} T_1, \qquad (5)$$

for items i=1 to k, where for one item e.g., i=1 $\alpha_1=1$ and $\lambda_1=1$. In Figure 1b the weight for the first item λ_1 is fixed. In this case, we have the most general and least restricted case of *congeneric measurements*, where the error variances of the items, $var(\varepsilon_i)$, can also be

different. A stronger restriction is associated with the assumption that the slope of each item is $\lambda_i=1$. In this case, *essential tau-equivalent measurements* are present, where $T_i=\alpha_i+T_1$. If, in addition, the intercepts, α_i, and the error variances, var(ε_i), are equal across all items, then tau-equivalence is present (exact parallel measurements). These assumptions are testable via a structural equation model by restricting the parameters. For the example above (Figure 1b), the model can be formulated as follows (e.g., Steyer, 1989, p. 30):

$$Y_1=0+1\ T+\varepsilon_1 \qquad (6)$$

$$Y_2=\alpha_2+\lambda_2\ T+\varepsilon_2$$

$$Y_3=\alpha_3+\lambda_3\ T+\varepsilon_3$$

$$Y_4=\alpha_4+\lambda_4\ T+\varepsilon_4$$

$$Y_5=\alpha_5+\lambda_5\ T+\varepsilon_5,$$

where T equals T_1. The model is statistically identified, i.e. all parameters can be estimated from the data.

Quality criteria

Associated with the measurement concept are certain measurement quality criteria that a scale must fulfil and which can be empirically testable to provide a (sum) scale:

1. *One-dimensionality*: The composite indicator as scale should be one-dimensional. This can be tested using factor analysis, a multivariate procedure for identifying basic dimensions in multivariate relationships between variables.
2. *Reliability*: The scale should accurately differentiate between high and low scorers. Reliability is generally defined by $\rho=\text{var}(T)/\text{var}(X)$, i.e. the proportion of the true score variance to the total variance (Eq. 3). A common reliability coefficient in the case of at least tau-equivalent measures is Cronbach's α (Sijtsma & Pfadt, 2021, p. 846):

$$Cronbach`s\quad \alpha=\frac{k}{k-1}\frac{\sum\sum_{i\neq j}\text{cov}(X_i,X_j)}{\text{var}(X)}, \quad (7)$$

where $\text{cov}(X_i, X_j)$ is the covariance of the items X_i and X_j and var(X) is the variance of the sum score of the k observed items or single indicators.

3. *Measurement invariance*: The measurement model should be invariant (e.g., same λ-coefficients, intercepts) across different groups, e.g., scientific fields. This assumption can be tested by multigroup structural equation modeling.
4. *Validity*: The scale should measure what is intended to measure. This can be tested, for example by correlation of the scale with external criteria. For example, a scale for measuring "scientific impact of a researcher`s work» should be highly correlated with researcher`s reputation subjectively assessed by peers of the scientific community.
5. *Fairness*: A test score should be independent of influences that have nothing to do with the trait or construct being measured. For example, the measurement of scientific impact should be independent of age and gender.

Data and Methods

Ioannidis, Klavans, and Boyack (2016, p. 1) had published in 2016 "a composite score summing standardised values of these six log-transformed indicators" for a total of 84,116

influential researchers. For all citation analyses, the authors used the Scopus bibliographic database. In 2020, they published an update of this database, which was further updated in 2021 (Baas, Boyack, & Ioannidis, 2021). This database, in particular the data for the entire career of a researcher, was used for the reanalysis. A total of 186,177 researchers from different fields were assessed. The composite indicator for "career impact" was obtained by adding the ratios of 6 log-transformed bibliometric indicators (Table 1), not excluding self-citations (Ioannidis et al., 2020, p. 2). To obtain indicators that vary exactly between 0 and 1, the following formula was applied, for example for the indicator NC:

$$t\_NC = \frac{\log(NC+1) - Min(\log(NC+1))}{Max(\log(NC+1)) - Min(\log(NC+1))} \qquad (8)$$

This transformation slightly differs from the original transformation used by Ioannidis et al. (2020), who have omitted the Minimum.

Table 1. Single indicators of the composite indicator (Ioannidis et al., 2020, p. 2)

Variable label	Explanation
NC9620	total cites 1996-2020
H20	h-index as of end-2020
HM20	hm-index as of end-2020
NCS	total cites to single authored papers
NCSF	total cites to single+first authored papers
NCSFL	total cites to single+first+last authored papers

The statistical analyses were carried out with the software SAS and the R-package “lavaan” (Rosseel, 2012).

Results

One-dimensionality

First, mean values, standard deviations of the items and correlations were calculated (Table 2). With the exception of NCS, the mean values are around 0.50, the variances around 0.11 except for NCS and NCSF. With the exception of NCS, there were moderately high positive correlations among the items.

Table 2. Mean values (M), standard deviation (STD) and correlations among log-transformed indicators (N= 186,177 researchers)

Item	M	STD	NC9620	H20	HM20	NCS	NCSF	NCSFL
NC9620	0.48	0.11	1.00	.92	.68	-.08	.48	.84
H20	0.53	0.11		1.00	.78	-.12	.38	.78
HM20	0.56	0.09			1.00	.19	.42	.77
NCS	0.40	0.16				1.00	.32	.10
NCSF	0.58	0.07					1.00	.62
NCSFL	0.43	0.11						1.00

An explorative factor analysis revealed 2 factors according to the Scree test, which explain 83.5% of the total variance (Table 3). One-dimensionality is given if the item NCS, as an indicator of the second factor, was eliminated. Therefore, a scale or composite indicator without NCS is assumed in the following.

Table 3. Varimax-rotated factor loading matrix

Item	Factor 1	Factor 2
NC9620	**.95**	-0.04
H20	**.95**	-0.10
HM20	**.83**	.23
NCS	-.09	**.93**
NCSF	**.53**	**.61**
NCSFL	**.90**	.23
Variance explained	3.72 (62.5%)	1.24 (20.7%)

Note. Loadings greater than .50 in bold face.

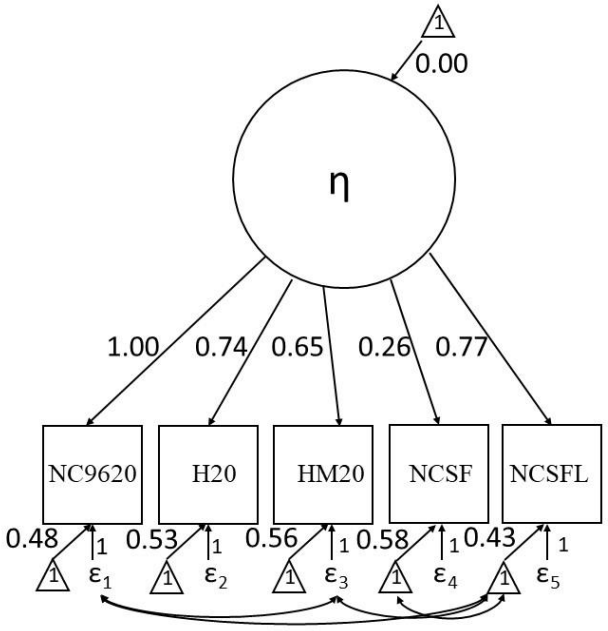
Reliability and measurement models

A comparison was made of the different measurement models, which can be represented as structural equation models (Table 4). The smaller the Bayes information criterion (BIC), the better the model fits the data. It turns out that the measurement model M_3 "congeneric measurements" fits best, whereby the assumption that the errors should be uncorrelated is violated (Figure 2).

Table 4. Model comparison regarding four different measurement models

	Measurement model	Bayes information criterion
M_0	Tau-equivalence (parallel items)	-159,431.3
M_1	Essential tau-equivalence	-234,446.2
M_2	Congeneric measurements	-248,106.6
M_3	Congeneric measurements + correlations among errors	-261,261.0

Figure 2. Estimated final congeneric measuring model (M_3) without the values of the error variances and covariances (Intercept of η is constrained to be Zero)



Cronbach's α as a measure of the lower limit of reliability, if congeneric measurements are present, is .91 for the values of the summation scale without the NCS item (with the NCS item α=.78). Thus, the scientific impact of a researcher is measured very accurately.

Measurement invariance

The measurement instrument must measure the same construct in different groups (here scientific fields). Invariance is strongest when the same structure (e.g. one-dimensionality), the same intercepts and the same weights (λ) are present in all fields. This can also be tested with structural equation models and parameter restrictions. The model comparison (Table 5) showed that "configural invariance" (M_0) is present, i.e. only the structure remains the same, the λ-coefficients and the intercepts α_i vary across the fields.

Table 5. Model comparison regarding measurement invariance

	Measurement invariance	Bayes information criterion
M_0	Configural invariance: The same factor structure is valid for all fields.	-248,106.6
M_1	Weak invariance: The factor loading λ_i are constrained to be equal across fields.	-234,446.2
M_2	Strong invariance: The intercepts α_i and factor loadings λ_i are constrained to be equal across groups.	-159,431.3

Validity and fairness

Since real external criteria are missing, the question of validity cannot be answered in this reanalysis. It is at least possible to examine the correlation between the true-score variable (η) and the composite indicator c of Ioannidis et al. and between the corresponding rankings of the researchers. The true-score variable correlates with the composite indicator to .69 (Pearson-Bravais), the corresponding rankings of the researchers to .60. Thus, the scale created here is not redundant to the original composite indicator of Ioannidis.

Unfortunately, the fairness of the measurements with regard to academic age (“year of last publication” minus “year of first publication”) is compromised. Thus, the correlation of academic age with the true score variable η amounts to .25. The higher the age, the higher the scientific impact tends to be. Researchers from different fields are also not comparable, as the lack of measurement invariance has already shown.

Discussion

Compared to single indicators, composite indicators have been little discussed in bibliometrics and scientometrics, and if so, then only very critically, which might be explained by their proximity to the world university rankings which are heavily under attack. A central problem of the use of composite indicators so far is the lack of concepts on how to create composite indicators from individual indicators. Unfortunately, simply summing up any bibliometric indicators does not justify a good composite indicator. A measurement perspective is needed. Using the example of Ioannidis' data, it was possible to show which measurement-theoretical requirements the data must fulfil in order to justify a measurement scale as a composite indicator (one-dimensionality, reliability, invariance, validity, fairness). Although the available data cannot fully meet the quality criteria, in principle the approach of Ioannidis et al. (2016) is to be welcomed. The common bibliometric indicators usually cannot even provide information on one of the mentioned quality criteria. Highly reliable and valid

composite indicators or scales are required, especially for use in empirical support of individual researcher assessment.

References

Baas, J., Boyack, K. W., & Ioannidis, J. P. A. (2021). *August 2021 data-update for "Updated science-wide author databases of standardized citation indicators, Mendeley Data, V3, .*

Barclay, M., Dixon-Woods, M., & Lyratzopoulos, G. (2019). The problem with composite indicators. *BMJ Quality and Safety, 28*(4), 338-344. http://dx.doi.org/10.1136/bmjqs-2018-007798

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal Indicators, composite Indicators, and covariates. *Psychological Methods, 16*(3), 265-284. http://dx.doi.org/10.1037/a0024448 M4 - Citavi

Fraumann, G., & Mutz, R. (2021). The h-index. In R. Ball (Ed.), *Handbook Bibliometrics* (pp. 169-180). Berlin: De Gruyter Saur.

Freudenberg, M. (2003). *Composite indicators of country performance: A critical assessment*. Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/405566708255

Glänzel, W., & Debackere, K. (2009). On the "multi-dimensionality" of ranking and the role of bibliometrics in university assessment. . In C. Dehon, D. Jacobs, & C. Vermandele (Eds.), *Ranking Universities*. Bruxelles: Université de Bruxelles.

Ioannidis, J. P. A., Boyack, K. W., & Baas, J. (2020). Updated science-wide author databases of standardized citation indicators. *PLoS biology, 18*(10), e3000918. https://doi.org/10.1371/journal.pbio.3000918

Ioannidis, J. P. A., Klavans, R., & Boyack, K. W. (2016). Multiple citation indicators and their composite across scientific disciplines. *PLoS biology, 14*(7). https://doi.org/10.1371/journal.pbio.1002501 M4 - Citavi

Ioannidis, J. P. A., Patsopoulos, N. A., Kavvoura, F. K., Tatsioni, A., Evangelou, E., Kouri, I., et al. (2007). International ranking systems for universities and institutions: A critical appraisal. *BMC Medicine, 5*. https://doi.org/10.1186/1741-7015-5-30

Moed, H. F. (2017). A critical comparative analysis of five world university rankings. *Scientometrics, 110*(2), 967-990. https://doi.org/10.1007/s11192-016-2212-y

Mutz, R., Bornmann, L., & Daniel, H. D. (2016). Funding decision-making systems: An empirical comparison of continuous and dichotomous approaches based on psychometric theory. *Research Evaluation, 25*(4), 416-426. https://doi.org/10.1093/reseval/rvw002

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2). http://www.jstatsoft.org/v48/i02/

Saisana, M., & Tarantola, S. (2009). *State-of-the-art Report on Current Methodologies and Practices for Composite Indicator Development*. Italy: Ispra.

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's Alpha: Discussing lower bounds and correlated errors. *Psychometrika, 86*(4), 843-860. https://doi.org/10.1007/s11336-021-09789-8

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika, III*, 25-60.