Research Synthesis 2019

incl. Pre-Conference Symposium

Big Data in Psychology

May 27-31, 2019

Abstract Collection

Pre-Conference Symposium
Big Data in Psychology

Table of contents
(by lead author)

	Page
Batzdorfer, V.	2
Christ, A., Penthin, M., Kröner, S.	4
Gordoni, G., Steinmetz, H., Schmidt, P.	7
Park, S., van Deun, K., Ceulemans, E.	11
Yuan, S., De Roover, K., Dufner, M., Denissen, J., van Deun, K.	14

Author:
Veronika Batzdorfer [1]

[1] Leibniz Institute for Psychology Information (ZPID)

Title:
The Pitfalls of Investigating Radicalization on Social Media

Session & Time:
Using Big Data: Applications. Tuesday, May 28th, 11:30 am - 12:00 pm

Abstract:

Background

Virtual communities as a facilitating arena for broadcasting radical beliefs, as well as connecting and recruiting sympathizers are broadly established in research, yet also seen as a panacea for predicting radicalization of specific ideological content. A plethora of research in the realms of terrorism research and radicalization has focused on social media, particularly Twitter to characterize from a network analysis perspective accounts and interactions or to detect individuals at risk or already radicalized. Though a desideratum as to the significance of virtual platforms such as 4chan or Reddit, as potential radicalization ecologies persists (see Schmid & Forest, 2018), a considerable amount of publications irrespectively focuses on collecting data from Twitter. Importantly, approaches used for data collection are prone to sampling bias (e.g. platform-specific biases along proxy population biases) jeopardizing the reliability of results, irrespective of the platform.

Objectives

The present work, on investigating social media data sampling practices and associated pitfalls, builds on a framework of data collection developed by Parekh et al. (2018). Their general model comprises four phases: initialization, expansion, filtering and validation. In an attempt to replicate their framework, their procedure leads the way to identifying prevalent data sampling methods (of user data, group data and interaction data) used in existing research on online radicalization, making use of social network analysis. In a similar vein, the objective is to investigate data collection limitations and implications in accordance to the outlined phases. As an extension, also studies dealing with other radical manifestations than ISIS sympathizers, such as extreme right-wing are considered.

Research Questions(s)/ hypothesis/es

Building on the results from Parekh et al. (2018), the question poses whether considered publications do sufficiently validate and filter their sampled data and consider possible sources of bias and whether the results compare to the work of the latter.

Method

Departing from the approach by Parekh et al. (2018), the impact of expansion and filtering on the quality of the data is tested by constructing an own dataset from

Twitter during one month, via the Twitter API. However, in contrast, not jihadist accounts are crawled, but white supremacist/ right-wing sympathizers' accounts and metadata. In the realm of the expansion phase (i.e. adding more accounts to the seed accounts) the friend and follower relationship of Twitter accounts are exploited and two data sets thereof created, which in turn are each subject to the filtering phase and non-filtering (comprising exclusion of accounts based on the number of followers and activity status). A random sample from each of the data sets is manually annotated as either neutral, radical right-wing, ambiguous, irrelevant or containing insufficient information, in order to establish the composition of the data set and validity.

Results/Findings

Preliminary findings are to be discussed.

Conclusions and Implications (expected)

By replicating the work of Parekh et al. (2018) and appraising and comparing existing data collection procedures of empirical studies in the realm of radicalization research, lenses are offered to improve the methodological founding of psychological research facing new possibilities with non-obtrusive insights into human behavior.

References

Parekh, D., Amarasingam, A., Dawson, L., & Ruths, D. (2018). Studying Jihadists on Social Media: A Critique of Data Collection Methodologies. *Perspectives on Terrorism*, *12*(3), 5-23.

Schmid, A. P., & Forest, J. J. (2018). Research Desiderata: 150 Un-and Under-Researched Topics and Themes in the Field of (Counter-) Terrorism Studies–a New List. *Perspectives on Terrorism*, *12*(4), 68-76.

Authors:
Alexander Christ [1], Marcus Penthin [1], Stephan Kröner [1]

[1] Friedrich-Alexander-University Erlangen-Nürnberg

Title:

Big Data and Digital Aesthetic, Arts and Cultural Education: Hot Spots of Current Quantitative Research

Session & Time:
Using Big Data: Applications. Tuesday, May 28th, 12:00 pm - 12:30 pm

Abstract:
While the digital transformation of society has substantial effects on cultural activities and corresponding education processes (Jörissen, Kröner, & Unterberg, 2019; Marres, 2017), it may also change our perspective on how to conduct research syntheses in this field. Considering scientific databases as bibliographic big data, approaches that have been previously applied to other big data sources may be transferred to research syntheses (Kröner, Penthin & Christ, 2019). This includes the application of text mining to discover main research topics ("hot spots") and explore relations between studies (O'Mara-Eves et al., 2015; Wu et al., 2014). The need to transfer big data analysis methods to research syntheses is particularly evident for the field of "digital aesthetic, arts and cultural education" (D-ACE, Jörissen, Kröner, & Unterberg, 2019), due to its high loadings on the three V's of big data (Diebold, 2012; Fan & Bifet, 2013): First, D-ACE's volume is high with a huge amount of relevant studies. Second, increasing velocity emerges from the recent exponential growth of publication output (Bornmann & Mutz, 2014). Third, large variety results from D-ACE's nature as an overarching concept connecting all spheres of cultural activities, ranging from classical arts to videogames. Variety is further increased as publications are tied to specific spheres of activity rather than being assigned to the overarching field of D-ACE (Keuchel, 2016; Fink et al., 2012). Accordingly, researchers in the field's subdisciplines are lacking a strong common D-ACE community self-concept, leading to insufficient mutual exchange. Thus, comprehensive syntheses of current research harnessing big data analysis methods promises particular added value for research in D-ACE.
For our approach, we rely on Scopus, which is ranking among the largest international research databases, as a source of bibliographic big data. We apply a still ongoing iterative approach of text mining, manual screening and predictive modelling (Zhao, 2017). To this end, we started with a search query consisting of terms that are indicative for the facets of (a) digitalization, (b) culture and spheres of cultural activities, and (c) education. The spheres of activities were further differentiated using the classification of ACE by Liebau et al. (2013). The final search query was expanded with synonyms and applied to Scopus to retrieve all publications from 2007 to 2017 in the subject areas "psychology", "arts & humanities" and "social sciences" that included at least one of our search terms of each facet

regardless whether in title, abstract or keywords. This resulted in N = 55,553 publications providing the basis for further text mining and predictive modelling.
The publications under scrutiny contained large numbers of words in the objects of title, abstract, keywords and journal (e.g. n = 11,347,105 words in abstracts, resulting from multiple occurrences of n = 122,542 unique words). Before applying text mining methods, titles and abstracts were parsed to exclude irrelevant strings such as non-English original titles or copyright statements. Afterwards, tidy texts were produced by deleting filler words, stemming all remaining words and deleting all words with up to four occurrences. This drastically reduced the amount of total and unique words (e.g. to n = 5,441,520 total and n = 12,396 unique terms in abstracts).
Next, we determined the most frequent terms and n-grams in each object and screened them regarding their value as indicators of relevance for the facets (a) digital, (b) culture and spheres of activities, (c) education and (d) quantitative methods. Moreover, (e) words indicative for a score for exclusion were collected. Afterwards, these collections of terms were used to determine relevance scores for every publication in the aforementioned facets using a bag of words approach (Zhang & Zhou, 2010).
To determine cut-offs at the relevance scores that may be used to select those among the retrieved publications that should be manually screened, we applied an iterative algorithm. Each iteration consisted of (1) selecting publications for screening according to preliminary cut-offs, (2) manual screening and (3) refining cut-offs for the next iteration via predictive modelling and refining relevance scores via topic modelling (Zhao, 2017).
In a first cycle of this work in progress, we sorted the retrieved publications according to the resulting scores and subjected the top 1400 highest scoring plus 1200 random publications to manual priority screening (O'Mara-Eves et al., 2015) so that they might be used as a training sample in predictive modelling. It turned out that n = 545 publications met the inclusion criteria.
Next we applied predictive modelling within the training sample to determine first empirically grounded estimates for cut-offs at the relevance scales. These were chosen such that publications subjected to manual screening could be expected to be judged relevant with a probability of .50. To this end, a logistic regression with positive and negative relevance scores as predictors of binary rater inclusion judgments was applied. In the first iteration, this resulted in McFadden R^2 = .47.
Applying the cut-offs from predictive modelling to the test sample implied manual screening of n = 2314 additional publications in the next cycle. Moreover, topic modelling was applied, showing that there was no need to change the terms on which the relevance scores were based.
In the next cycle, the publications (together with an additional subset of random publications to avoid a skewed distribution of hits) will be screened and added to the training sample. Predictive modelling will be reiterated and the cut-offs will be refined as long as screening will result in the identification of a considerable amount of additional relevant papers.

Finally, the resulting relevant publications will be categorized according to the spheres of activities to determine hot spots of research on D-ACE. Preliminary results from the first iteration cycle suggest video games as a major hot spot. Taken together, this study shows the benefits of conceiving the conducting research syntheses as an analysis of big data. Apart from economic and pragmatic advantages, this facilitates the detection of both underlying structures between relevant papers. However, while much tedious routine work in research syntheses might be automated, manual screening is still needed to filter between relevance and noise.

References

Bornmann, L., & Mutz, R. (2014). *Growth rates of modern science: A bibliometric analysis* (No. arXiv: 1402.4578).

Diebold, F. X. (2012). *On the Origin (s) and Development of the Term 'Big Data'*.

Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter, 14*(2), 1-5.

Fink, T., Hill, B., Reinwand, V.-I., & Wenzlik, A. (2012). Begrifflich, empirisch, künstlerisch: Forschung im Feld der Kulturellen Bildung. In T. Fink (Ed.). *Kulturelle Bildung: Vol. 29. Die Kunst, über kulturelle Bildung zu forschen* (pp. 9–21). München: kopaed.

Jörissen, B., Kröner, S., & Unterberg, L. (Eds.). (2019). *Forschung zur Digitalisierung in der Kulturellen Bildung*. München: kopaed.

Keuchel, S. (2016). Different Definitions and Focus on Arts Education. An Explorative International Empirical Study. In A. Berggraf Sæbø (Ed.), *International Yearbook for Research in Arts Education Vol. 4.: At the Crossroads of Arts and Cultural Education: Queries Meet Assumptions* (1st ed., pp. 31–40). Münster: Waxmann.

Kröner, S., Penthin, M., Christ, A. (2019). Forschungssynthesen zur Digitalisierung in der Kulturellen Bildung. In Jörissen, B., Kröner, S., & Unterberg, L. (Eds.). *Forschung zur Digitalisierung in der Kulturellen Bildung*. München: kopaed.

Liebau, E., Jörissen, B., Hartmann, S., Lohwasser, D., Werner, F., Klepacki, L., Schelter, F. (2013). Forschung zur Kulturellen Bildung in Deutschland: Bestand und Perspektiven: Projektbericht.

Marres, N. (2017). *Digital sociology: The reinvention of social research*. Cambridge: Policy Press.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, *4*(1), 5.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics, 1*(1-4), 43-52.

Zhao, Y. (2017). *Text Mining with R*.

Authors:
Galit Gordoni [1], Holger Steinmetz [2], Peter Schmidt [3]

[1] The Academic College of Tel Aviv-Yaffo; [2] Leibniz Institute for Psychology Information (ZPID); [3] University of Giessen

Title:
Usability of web scraping of open-source discussions for identifying key beliefs

Session & Time:
Using Big Data: Applications. Tuesday, May 28th, 11:00 am - 11:30 am

Abstract:
Background
The recent years has brought tremendous interest in the collection and use of Big Data. While in the first phase of interest, the discussion largely focused on practical and societal issues, researchers have begun to consider the use of Big Data for scientific uses. In Psychology, there is an increasing interest in the usability of user-generated data for addressing psychological research questions (Adjerid & Kelley, 2018; Harlow & Oswald, 2016). As a prominent data collection method, web scraping (i.e., an automated tool for finding and extracting data from online sources) has been used for research on eating disorders (Moessner, et al., 2018), mental toughness (Gucciardi, 2017) and personality (Farnadi et al., 2016).
One frequent characteristic of common Big Data analytics is its exploratory nature. In contrast, researchers increasingly demand to use it for theory-relevant research (e.g., Shmueli, 2010). Although web scraping is increasingly applied it is still not clear whether posts, can serve as a valuable data source in theory-driven empirical studies.
In this study we address the lack of knowledge on usability of user-generated data for assessing research questions concerning beliefs of people (Eagly & Chaiken, 1993). As a relevant, theoretical framework that focuses on the fundamental role of beliefs in interventions, we draw on the well replicated social psychological theory—the Theory of Planned Behavior (TPB; Ajzen, 1991). The theory integrates the cognitive foundation of motivational and decision processes (i.e., the beliefs) with attitudes, perceptions of social legitimization, efficacy, and feasibility of the behavior in question (Fishbein & Ajzen, 2010). Briefly, the theory claims that deliberate behavior is mainly determined by the intention to perform the behavior. The intention, in turn, is a function of the attitude towards the behavior (i.e., the perceived attractiveness of the behavior), the subjective norm (i.e., the perceived expectations of important others towards conducting the behavior), and the perceived behavioral control (i.e., the perceived feasibility and control with regard to the behavior). Furthermore, the theory claims that these motivationally relevant factors are based on beliefs about positive and negative consequences of the behavior, the opinions of specific others and barriers and facilitators. The TPB serves as a central theoretical framework for understanding and changing behaviors. Since changing beliefs is the

essence of intervention approaches, knowledge about potent beliefs of potential benefits, costs, social expectations, barriers, and facilitators of the behavior, is not only of theoretical value but provides the basis for practical endeavours to change behaviors (Steinmetz et al. 2016).
The initial stage in a TPB driven study includes identifying motivationally relevant key beliefs via a qualitative pilot study. While this procedure (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 2010) has been fruitful for identifying relevant beliefs for decades of TPB research, it has the limitation that the number of respondents is very small and that the approach runs the danger of reactive responses. Especially in cases with a non-familiar behavior, the comments may lack validity and will not concern those beliefs which occur in a natural decision process. In this study we focus on the potential of open-source discussions to serve as an additional data source that resembles the pitfalls of self-reported answers. Users comments are produced by individuals concerned with consequences of the behavior in question or expected difficulties of conducting the behavior, formulated in a natural setting, with no potential response bias due to factors, such as, interviewer effect, topic complexity and topic sensitivity.

Objectives
We aim to advance the knowledge on the usability of integrating web scraping of web discussions in the initial stage of theory-driven belief study, for identifying key beliefs underlying behaviors under interest.

Research questions
We use the behavior of Big Data adoption in organizations as an illustrative case for testing the following questions:
1. What are the key beliefs concerning Big Data adoption (behavioral beliefs, normative beliefs and control beliefs)?
2. Do key behavioral, normative and control beliefs concerning Big Data adoption identified in user-generated posts differ from those identified in self-report surveys?

Method
We conducted web scraping study of discussion boards on Big Data usage in Israel, generated between June and August 2018. Discussions appeared mainly after online articles (41%), in social networks (25%) and forums (19%). Unit of analysis was the complete discussion beginning with the opening post up to the closing one. 353 authentic discussions (i.e., containing at least 2 comments) were scraped. Content analysis was conducted, manually for a sample of 148 authentic discussions. We applied the methodology used for identifying key beliefs in TPB driven studies (de Leeuw et al., 2015) for counting the number of times a given category of comment content appeared across discussions. Second, following Landers et al. (2016), we compared the beliefs found via web scraping with representative surveys in French companies (Raguseo, 2018) and in German companies (Commerzbank AG, 2018). These external data sources serve as a base rate for testing the replicability of key beliefs found in the web scraping data. For

comparison we used for example the response distribution of the following multiple response question "*What are the benefits to companies from the systematic use of digital data?*" asked in the German companies survey (n=2004) conducted in 2017.

Results

Initial and descriptive results will be presented. Content analysis resulted in classification of the 148 discussions into semantic units representing the advantages and disadvantages of big data adoption, list of potential stakeholders, and factors that could impede or facilitate it. Initial results show similarity in the content of beliefs and frequency rank across the independent data sources. For example, the most frequently cited advantage, in both data sources, German survey and web scraping, was better decision making (cited by 58% of survey participants and in 41% of scraped discussions that cited advantages).

Conclusions and expected implications

Drawing upon web scraping of open-source discussions, we demonstrated initial results supporting the usefulness of using web scraping as an observational data collection method in first stages of identifying key beliefs underlying specific behaviors for a theory-driven belief-scale development.

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist, 73*(7), 899-917. http://dx.doi.org/10.1037/amp0000190

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.*Englewood Cliffs, NJ: Prentice-Hall.

Commerzbank Initiative Unternehmerperspektiven (2017). The Raw Material of the 21st century: Big Data, Smart Data – Lost Data? Retrieved from https://www.unternehmerperspektiven.de/portal/media/unternehmerperspektiven/up-startseite/2018_04_18_FL_UP_Studie_online_2018_EN.pdf.

De Leeuw, A., Valois, P., Ajzen, I., & Schmidt, P. (2015). Using the theory of planned behavior to identify key beliefs underlying pro-environmental behavior in high-school students: Implications for educational interventions. Journal *of Environmental Psychology, 42*, 128-138.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., ... & De Cock, M. (2016). Computational personality recognition in social media. *User Modeling and User-Adapted Interaction*, *26*(2-3), 109-142.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. Psychology Press.

Gucciardi, D. F. (2017). Mental toughness: progress and prospects. *Current Opinion in Psychology*, *16*, 17-23.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to special issue. *Psychological Methods*, *21*(4), 447–457. http://doi.org/10.1037/met0000120.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, *21*(4), 475-492.

Moessner, M., Feldhege, J., Wolf, M., & Bauer, S. (2018). Analyzing big data in social media: Text and network analyses of an eating disorder forum. *International Journal of Eating Disorders*, 51(7), 656-667.

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, *38*(1), 187-195.

Shmueli, G. (2010). To explain or to predict?. *Statistical Science*, *25*(3), 289-310.

Steinmetz, H., Knappstein, M., Ajzen, I., Schmidt, P., & Kabst, R. (2016). How effective are behavior change interventions based on the theory of planned behavior?. *Zeitschrift für Psychologie*, 224(3), 216–233.

Authors:
Soogeun Park [1], Katrijn Van Deun [1], Eva Ceulemans [2]
[1] Tilburg University; [2] KU Leuven

Title:
Ignoring the differences in model structures of PCA and sparse PCA can misguide practice

Session & Time:
Methodology. Tuesday, May 28th, 3:30 pm - 4:00 pm

Abstract:
Principal component analysis (PCA) represents data through principal components by linear combinations of the original variables that successively maximize variance (Jolliffe and Cadima, 2016). With an intrinsic link to factor analysis, the technique has a central position among psychological methods.
As PCA only returns non-zero loadings, components extracted by PCA are often complicated to interpret, especially when the number of variables is large. In order to improve the interpretability of the loadings, rotation techniques from factor analysis have been adopted to PCA. These techniques find simple structures, before near-zero loadings are neglected for interpreting the components.
However, although this practice to subset certain loadings according to the size has been conventional in psychology, Cadima and Jolliffe (1995) exhibited that it can be misleading. They demonstrated that the magnitude of the loadings is unrepresentative of the variable-component association: even a variable with a small-sized loading for a component can be a strong predictor for the component. Moreover, rotation techniques entail another weakness that the choices of a rotation criterion and a threshold value to truncate the loadings involve subjectivity. With increasing usage of big and high-dimensional data in psychology, the difficulty in interpreting PCA solutions is growing and these limitations obstruct the use of PCA.
To circumvent these limitations, sparse PCA methods, that constrain the loadings to contain zero-elements directly within the estimation process, has been proposed (Zou, Hastie and Tibshirani, 2006; Shen and Huang, 2008). With Trendafilov and Adachi (2015) presenting that sparse PCA can derive the desired solution of simple structure from 24 psychological tests reported in Harman (1976), the method is established as a suitable alternative to rotation techniques in psychological applications. Moreover, sparse PCA serves as an upgrade over rotation not only because they resolve the aforementioned weaknesses, but also since they are better-suited for modern data circumstances characterized by big and high-dimensional data. An influential study by Johnstone and Lu (2009) showed that sparse PCA produces consistent loadings under a high-dimensionality setup while the estimates from PCA were inconsistent. As so, sparse PCA methods are being widely used for applications involving high-dimensional datasets, such as gene expression data.

Our current study concerns the research practices carried out within the sparse PCA literature. As sparse PCA methods are adaptations of PCA, the literature appears to have been largely overlooking the differences in the underlying model structures imposed by PCA and sparse PCA. The structure imposed by PCA entails uncorrelated components. Also, it has a special property that the loadings, which represent variable-component association, are identical to weights, which are coefficients that transform the variables to the components. In contrast, sparse PCA introduces zero-elements to either of the weights or loadings, at the cost of this property: they are no longer equal within sparse PCA. The components derived by sparse PCA may also be correlated. However, most researchers appear to have maintained an inattentive conception that sparse PCA imposes the same underlying structures as PCA. This had led to misguided practices in research, such as generating data from simplistic structures based on traditional PCA for numerical simulations and naive usage of PCA-based initial values for the algorithms.

Our investigation presents the consequences of these erroneous practices. The benchmark data generation strategies in the literature for simulation studies entail mismatch between the structure that the data is generated from and the structure imposed by the method. On top of leading to less relevant results, such mismatch also provides optimistic insights about novel methods proposed. We provide a comparison of the results obtained from these benchmark strategies alongside with those resulted using alternative strategies that better reflect the underlying structure imposed by sparse PCA methods, and present that the former strategy leads to much more optimistic outcome.

Moreover, we also demonstrate that starting values employed by most sparse PCA methods need improvement. Most methods naively use values based on PCA for the algorithms. Such values would me suitable, if the PCA structure actually underlies the data in hand. However, in reality, the true underlying structure of the data is never fully known and therefore it is important for sparse PCA methods to adopt a multi-start approach that considers multiple starting values. We report that the multi-start approach results in smaller loss value than the conventional PCA-based starting values. The interpretation of the results from multi-start is also often different from the PCA-based values, which further emphasizes the importance of the starting values. An illustration through an empirical dataset is also given alongside.

With this paper, we hope to highlight erroneous practices and contribute in ameliorating them in forthcoming sparse PCA research and applications. Our aim is to shift the literature towards the necessary attention on the model structures of sparse PCA. As sparse PCA is a modern upgrade of PCA suitable for emerging big and high-dimensional data in psychology, improvements in these practices is of great importance.

References

Cadima, J. and Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principal components. *Journal of Applied Statistics,* 22(2):203–214.

Harman, H. H. (1976). *Modern factor analysis*. University of Chicago press.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*,104(486):682–693.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Phil. Trans.* R. Soc. A, 374(2065):20150202.

Trendafilov, N. T. and Adachi, K. (2015). Sparse versus simple structure loadings. *psychometrika*, 80(3):776–790.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis.*Journal of computational and graphical statistics,* 15(2):265–286.

Authors:
Shuai Yuan [1], Kim De Roover [1], Michael Dufner [2], Jaap Denissen [1], Katrijn Van Deun [1]

[1] Tilburg University; [2] University of Leipzig

Title:

Discovering Structure of Unknown Groups in Multi-block Data - A Novel Clustering Method Based on Sparse Component Loadings

Session & Time:
Methodology. Tuesday, May 28th, 4:00 pm – 4:30 pm

Abstract:

Background and objectives

The availability of large-scale multi-block data, containing information on the same set of respondents but coming from various sources (i.e., blocks), nowadays thrives in social and behavioral science research. For example, personality researchers gather self-report, peer-report, EMG, daily diary, and lab-task performance data from the same respondents to test the affective contingencies hypothesis (Dufner et al, 2015). Such multi-block data have the potential to suggest complex behavioral mechanisms where several influencing factors are at play: On the one hand, behavior may be the result of the concerted action of several factors (e.g., a combination of personal traits and situation factors), or they may be the result of a single influencing factor (e.g., self-report bias in questionnaire data). The former type of mechanism may be advised by *common variation* that links variables throughout multiple data blocks while the latter may be indicated by *distinct variation* that shows up only among the variables in one single block. The two sources of variation should be disentangled in multi-block data analysis.

A further complication in understanding social and behavioral mechanisms is their heterogeneity: behavior may be influenced in different ways for different individuals. For example, the fact that subgroups of depressive patients experience different problems links up with the observation that a one-size-fits-all treatment is ineffective (Fried, 2016). However, often these subgroups are not known to the researchers beforehand – especially when novel types of behavioral markers are studied. Hence, a *clustering* method for multi-block data is needed that has the potential to detect the subgroups together with their associated common and distinct variation. Such requirement is clearly beyond the capability of existing clustering methods for multi-block data (e.g. iCluster in Shen, Olshen, & Ladanyi, 2009), which determines subgroups based solely on *mean-level difference.* In addition, because such multiple data blocks are often untargeted and may include irrelevant variables, the ideal clustering method should preferably retain only the most relevant information.

Method

Model

We present a novel clustering method, CSSCA (Cluster-wise Spare Simultaneous Component Analysis). CSSCA is a Principal Component Analysis (PCA) based method that captures the variation by a few components; with variable loadings indicating how they components arise from data. CSSCA imposes structure on the loadings to reflect the common and distinctive components; it also penalizes the loadings to yield zero-loadings and – in this way – removes irrelevant information. In CSSCA, as in PCA, each component may suggest a construct (process or mechanism), that is associated to the variables having non-zero loadings. Simultaneously, CSSCA partitions the respondents into clusters in such a way that only respondents belonging to the same cluster have the same loading matrix.

Algorithm

We have developed an efficient procedure for the estimation of CSSCA. The implementation code will be publicly available as an R package on Github.

Model selection

To determine the optimal level of sparsity and optimal number of clusters, we propose a model selection procedure that is based on the convex hull (Wilderjans et al., 2013). We advise to set the number of (common and distinctive) components based on theoretical knowledge.

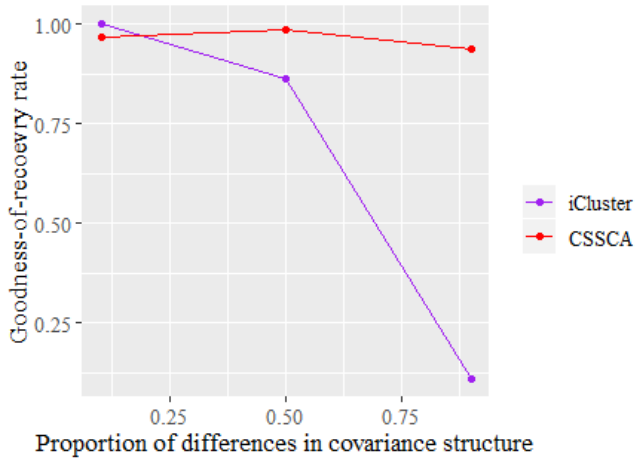
Results and Discussion

Simulation 1

The first simulation study assesses how well CSSCA recovers the true cluster partitions. The performance of CSSCA is assessed and compared to the performance of iCluster, a popular clustering method in multi-block data analysis which only detects mean-level differences.

As shown in Graph 1, CSSCA clearly outperforms iCluster when 10% or 50% of the cluster differences is stems from differences in covariance structure; iCluster performs slightly better when this is only 10%

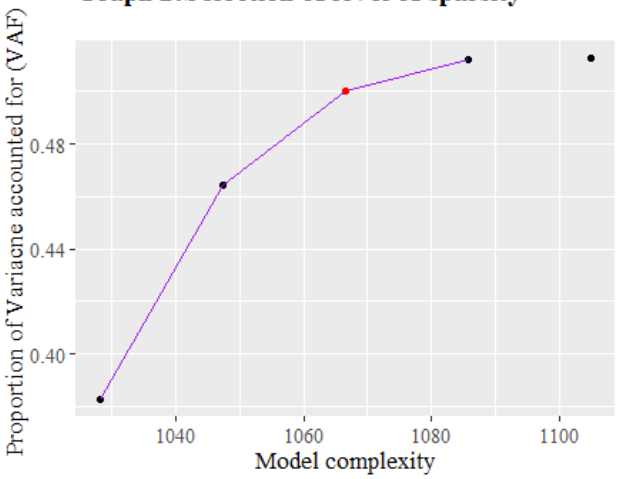
Graph 1:Performance of CSSCA and Cluster



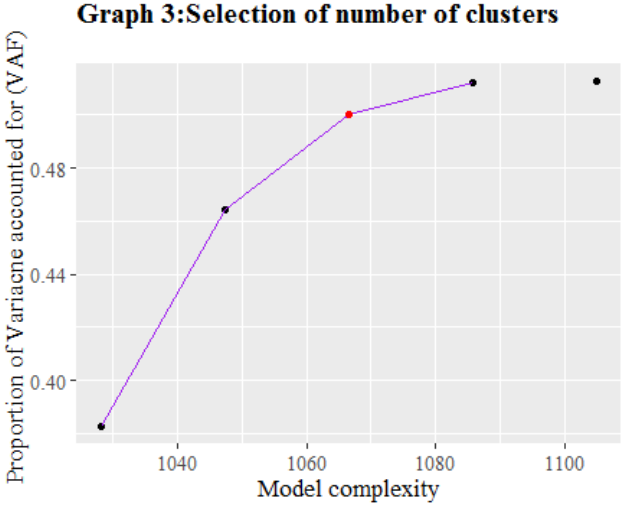
Simulation 2

The second simulation study evaluates the model selection procedure of CSSCA in six conditions. The selection of the level of sparsity is successful in 5 conditions (83.3%) while the selection of the number of clusters is successful in all 6 conditions (100%). The results for one of the conditions are shown in Graph 2 and Graph 3. The purple line represents the hull, and the red dot represents the selected value (both selections are successful)

Graph 2:Selection of level of sparsity



Graph 3:Selection of number of clusters



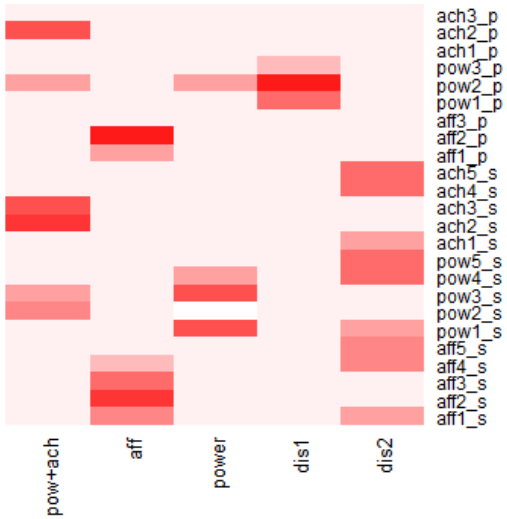
Analysis on empirical data

Finally, CSSCA is applied to an empirical data that includes self-report and peer-report measurements on three kinds of motivation dispositions: affiliation, achievement and power (Dufner et al., 2015). The data block of self-report measurements includes 15 variables (from 5 scales) while the data-block of other-report measurements include 9 variables (from 3 scales).

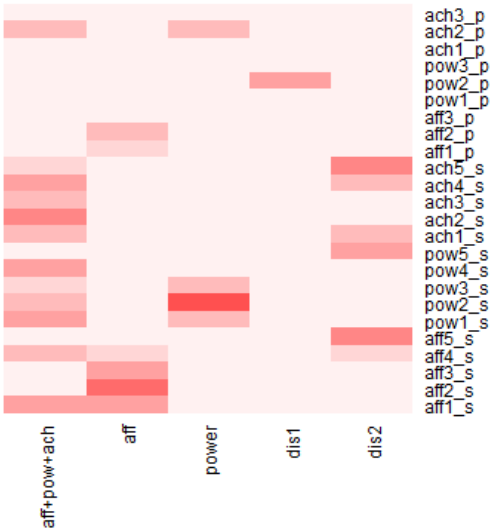
With the number of components set to five (of which three common components to represent the types of motivation and two distinctive components represent two types of reports), the CSSCA analysis and model selection procedure results in the optimal solution of 3 clusters and 70% sparsity. Graph 4 and Graph 5 are the one-cluster heat-map and three-cluster heat-maps (one for each cluster) that represent the (cluster-specific) loading matrices. In both graphs, rows represent variables while

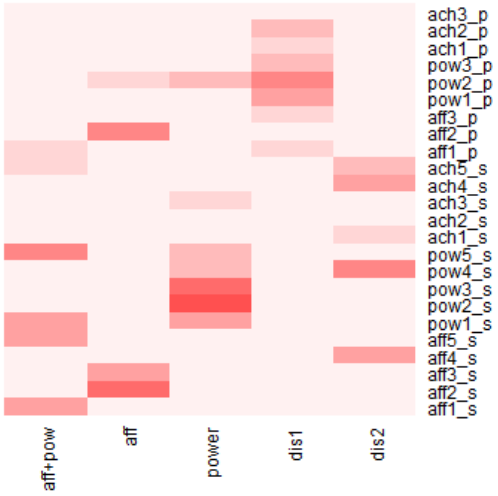
columns represent components, and (only) the non-zero loadings are filled with red (darker color represents larger absolute value of the loading)
Briefly, taking all respondents in one cluster (Graph 4), the affiliation motivation as well as the power motivation appears to be two apparent distinctive constructs, while the achievement motivation mixes up with the power motivation. However, in three-cluster solution (Graph 5), power and affiliation motivation still stands out as two distinctive constructs, while the constitution of the third component differs in three clusters.

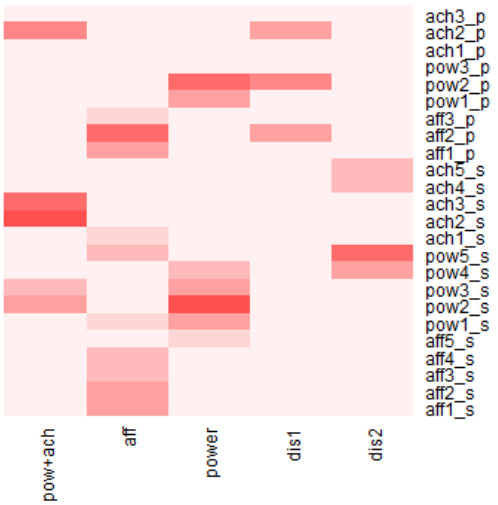
Graph 4: The heat-map for one-cluster solution (sparsity = 0.7)



Graph 5: The heat-map (of loadings) for three-cluster solution (sparsity = 0.7)







Conclusions and implications

In this paper we presented CSSCA, a novel and promising clustering method for the analysis of multi-block data. CSSCA is able to detect subgroups that differ in structural variation and also suggests cluster-specific mechanisms. Because CSSCA has variable selection properties, it can be used to address the challenges of high-dimensional problems. CSSCA and its associated model selection procedure have been validated in two simulations, demonstrating a clear advantage over existing methods. The empirical analysis of CSSCA demonstrated its utility in uncovering group structure of variables in psychological research.

References

Dufner, M., Arslan, R. C., Hagemeyer, B., Schönbrodt, F. D., & Denissen, J. J. (2015). Affective contingencies in the affiliative domain: Physiological assessment, associations with the affiliation motive, and prediction of behavior. *Journal of Personality and Social Psychology*, *109*(4), 662.

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR* D study. *Journal of affective disorders*, *172*, 96-102.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, *25*(22), 2906-2912.
Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). CHull: A generic convex-hull-based model selection method. *Behavior research methods*, *45*(1), 1-15.

Research Synthesis 2019

Table of contents
(by lead author)

	Page
Bartl, G. J., Blackshaw, E., Crossman, M., Allen, P., Sandrini, M.	22
Batzdorfer, V.	25
Burgard, T., Bosnjak, M., Kasten, N.	27
De Jonge, H., Jak, S., Kan, K. J.	31
Declercq, L., Jamshidi, L.	36
Fernández Castilla, B.	39
Haensch, A.-C.	43
Kasten, N., Burgard, T., Wedderhoff, O., Bosnjak, M., Gnambs, T.	46
Kedzior, K. K., Kaplan, I.	48
Kossmeier, M., Tran, U. S., Voracek, M.	51
López-Ibáñez, C., Blázquez-Rincón, D. M., Sánchez-Meca, J.	56
López-López, J. A.	59
Marker, C., Gnambs, T., Appel, M.	61
Matthias, K., Rissling, O., Nocon, M., Jacobs, A., Morche, J., Pieper, D., Wegewitz, U., Lorenz, R.	65
Metwaly, S. S., Fernández Castilla, B., Kyndt, E., van den Noortgate, W., Barbot, B.	68
Rubio-Aparicio, M., Sánchez-Meca, J., Núñez-Núñez, R. M., López-Pina, J. A., Martín-Martínez, F., López-López, J. A.	71
Sánchez-Meca, J., López-Pina, J. A., Rubio-Aparicio, M., Martín-Martínez, F., Núñez-Núñez, R. M., López-García, J. J., López-López, J. A.	74
Stasielowicz, L., Suck, R.	79

Tsuji, S., Cristia, A., Frank. M. C., Bergmann, C. 83

Tran, U. S., Burzler, M. A., Hegewisch, U. J. C., Voracek, M. 85

Voracek, M., Kossmeier, M., Slowik, A., Tran, U. S. 91

Authors:
Gergely Janos Bartl [1], Emily Blackshaw [1], Margot Crossman [1], Paul Allen [1], Marco Sandrini [1]

[1] University of Roehampton

Title:

Systematic review and network meta-analysis of tDCS effects on verbal memory: Modelling heterogeneity of stimulation locations

Session & Time:

Network Meta-Analysis. Thursday, May 30th, 8:00 pm - 8:30 pm

Abstract:

Background

There has been growing interest in the use of transcranial direct current stimulation (tDCS) in enhancing memory (Sandrini & Cohen, 2014), with a view to possible future applications in pathological aging. It has been argued that cathodal tDCS decreases motor cortical excitability while anodal tDCS increases motor cortical excitability (Dayan, Censor, Buch, Sandrini & Cohen, 2013). Thus, when tDCS is applied over associative areas of the cortex involved in memory formation, decreased or improved memory performance would be expected depending on stimulation polarity.

However, this behavioural dissociation is not often attempted or found at the single study level. Research synthesis efforts, whilst beneficial, also encounter the problem of how to model the wide variety of stimulation setups within the constraints of pairwise meta-analyses. A number of solutions have been attempted, for example running independent meta-analyses, grouping very different montages as similar, or including electrode location as a moderator variable. This may lead to the loss of a direct statistical comparison, and/or useful information when addressing the research question.

Objectives

Network Meta-Analysis (NMA) has been used increasingly over the last decade, predominantly in the field of clinical trials to make use of a combination of available direct and indirect comparisons in decision making (Riley et al., 2017). We applied NMA principles to a group of studies investigating the effect of different tDCS setups on verbal memory. Our aim was to synthesise evidence regarding differences in memory enhancement as a function of stimulation site. We also intended to test the feasibility of using this meta-analytic approach in the field and evaluate its ability to synthesise evidence on the effect of brain stimulation on other cognitive domains.

Research question(s) and/or hypothesis/es

Two questions were addressed. If tDCS is applied during a learning phase, (1) which anode placement location is the most effective in enhancing verbal memory, and (2) what degree of enhancement is likely to occur in these most effective setups?

Method/Approach
A systematic review of studies was carried out according to PRISMA guidelines, using databases PsychInfo, Web of Science and PubMed. Search terms included the following phrases and their truncated and alternative forms: 'tDCS', 'memory', and 'verbal'. Studies were screened based on the following criteria: (1) randomised control trials (2) applying tDCS during learning (3) with a subsequent retrieval task (4) involving verbal stimuli. Electrode location, stimulation intensity, population (age), and retrieval task type were key parameters extracted from original studies. The Cochrane Collaboration's guidelines were adopted for the evaluation of risk of bias. Hedges' g was calculated as a standardised mean difference measure (SMD) for all comparisons using the Metacont package in R. Corrections were made as necessary for repeated measures designs. In case of studies with multiple memory measures, a composite (pooled) effect size was calculated. A network meta-analysis was performed in R using the Netmeta package (Rücker & Schwarzer, 2015). Consistency of direct and indirect effects was evaluated using node-splitting (Dias et al., 2010). Ranking of treatment efficacy was performed, and the contrast of each electrode montage versus placebo stimulation was evaluated using 95% confidence intervals (CI).

Results/Findings
14 experiments from 10 different studies were included in the analysis with data from 372 participants. Most studies used young, healthy samples, with a few additional papers tested elderly groups. Electrode setups were placed into 18 different categories based on the location of anode. tDCS with the anode over the left prefrontal cortex (PFC) had a high certainty of being more successful in inducing memory enhancement than other setups (P=0.82). The cumulative effect of F3 anode placement (3 studies) was positive: g = 0.38 [0.06 – 0.7]. Two further contrasts suggested significant effects, although based on single studies only. Stimulation using opposite polarity (cathode) over F3 led to a marginally non-significant negative effect, g = – 0.5 [-0.94; 0], and anodal placement over F7 had a positive effect g = 1.04 [0.63; 1.45].

Conclusions and implications
Our analysis suggests there is a possibility of electrode-location specific modulation of verbal memory when tDCS is applied during the learning phase. Left PFC stimulation (anode over F3 or F7) appeared to have a positive effect compared to placebo/sham tDCS. The evidence so far comes from a relatively low number of comparisons, and predominantly from experiments testing memory in relatively short time periods (same day or next day) in healthy populations. Clinical trial endpoints in patients with memory impairments could provide a more robust test of the possibility that neuro-modulation can be an effective intervention to reduce memory decline. Effect size expectations of typical tDCS studies may be exaggerated. Sample size calculations based on the more successful electrode montages from our NMA still have a relatively high degree of uncertainty. Using SMD from these contrasts as a

proxy for population effect size suggests that larger sample sizes than those currently employed are needed in order to produce more reliable estimates. The possibility of a polarity-specific effect (over F3) is an interesting prospect, although the comparison is based on a narrow evidence base and would need further testing, particularly in head-to-head comparisons. It has been questioned recently whether tDCS is able to achieve significant effects on neural function at currently used (<2mA) stimulation intensities (Underwood, 2016; Voroslakos et al., 2018). Synthesis of behavioural data on location- and polarity-specific modulation could contribute to this debate, and NMA may be well placed to model the heterogeneity of stimulation protocols used in memory as well as other cognitive domains, and stimulation methods (e.g. Transcranial Magnetic Stimulation).

References:

Dayan, E., Censor, N., Buch, E. R., Sandrini, M., and Cohen, L. G. (2013). Non-invasive brain stimulation: From physiological mechanisms to network dynamics and back. *Nature Neuroscience,* 16, 838-844.

Dias, S., Welton, N. J., Caldwell, D. M., & Ades, A. E. (2010). Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine*, *29*(7-8), 932-944.

Riley, R. D., Jackson, D., Salanti, G., Burke, D. L., Price, M., Kirkham, J., & White, I. R. (2017). Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ*, *358*, j3932.

Rücker, G., & Schwarzer, G. (2015). Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Medical Research Methodology*, *15*(1), 58.

Sandrini, M, & Cohen, L. G. (2014). Effects of brain stimulation on declarative and procedural memories. *The Stimulated Brain: Cognitive Enhancement Using Non-Invasive Brain Stimulation (Edited by Roi Cohen-Kadosh, Elsevier)*.

Underwood, E. (2016). Cadaver study challenges brain stimulation methods. *Science* 352:397.

Vöröslakos, M., Takeuchi, Y., Brinyiczki, K., Zombori, T., Oliva, A., Fernández-Ruiz, A., ... & Berényi, A. (2018). Direct effects of transcranial electric stimulation on brain circuits in rats and humans. *Nature Communications*, *9*(1), 483.

Authors:
Veronika Batzdorfer [1]

[1] Leibniz Institute for Psychology Information (ZPID)

Session & Time:
Applications. Friday, May 31st, 1:30 pm - 2:00 pm

Title:
Systematic Literature Review of Conceptual Approaches and Operationalizations of Radicalization Facets

Abstract:
Background
The popularity of the term 'radical' and its derivatives bears no relation to its actual explanatory value (Mandel, 2009). Starting with 2004/05 the term has risen to importance in academia and policy-making, offering in particular a lens on 'homegrown' Islamist political violence as well as investigating root causes, whilst epitomizing the war against terrorism. Notwithstanding a wealth of theoretical approaches ranging from rational-choice, psychopathology, quest for significance, terror management theory, social movements or social network theory, to name a view, eventually, disagreement prevails as to how the construct and its determinants are defined in relation to, and distinguished from other related concepts, as well as operationalized.

Objectives
The present systematic literature review identifies definitions and conceptualizations of radicalization and its determinants paired with investigating respectively, valid measures, quantitative empirical research has brought up to capture the phenomenon in question. Determinants of radicalization (comprising psychological, social-psychological or environmental dimensions), as well as outcomes, along the axis of violent, non-violent radicalization, besides non-radicalization are considered eligible. Furthermore, this review characterizes how well the identified instruments capture radicalization and discerns future avenues.

Research question(s) and/or hypothesis/es
a) How is radicalization conceptually defined and operationalized in past studies?
b) How are determinants of radicalization defined and operationalized?
c) How well is radicalization explained based on determinants considered?
d) Which gaps in past research and avenues for the future can be identified?

Method/Approach
This study is part of a pre-registration. In the scope of this project empirical articles are rigorously screened (first based on title and abstract and then full-text screened) according to inclusion and exclusion criteria relating to the relevance, population, setting and availability of research. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting standards are adopted to record

the subsequent results of the literature searches and selection decisions in a flow diagram. For the sake of transparency, any changes to the search strategy will be detailed and justified. Retrieved search results will also be saved for subsequent re-analysis (if applicable). Instruments identified are appraised to establish their reliability and validity. Thereby, extracted data of concepts and operationalizations inform an evidence structure regarding self-reported, experimental and unobtrusive trace data and reveal gaps in evidence.

Results/Findings

As the data collection is still in progress, as of this moment no results are available. In order to prevent publication bias, the full paper will be published regardless of the results.

Conclusions and Implications (expected)

This systematic review of quantitative evidence and the identification and characterization of research gaps can guide evidence-informed choices for further empirical investigation in the field of radicalization.

References

Mandel, D. R. (2009). Radicalization: What does it mean? In T. Pick, & A. Speckhard (Eds.), *Indigenous terrorism: Understanding and addressing the root causes of radicalization among groups with an immigrant heritage in Europe*. Amsterdam: IOS Press

Authors:
Tanja Burgard [1], Michael Bosnjak [1], Nadine Kasten [2]

[1] Leibniz Institute for Psychology Information (ZPID); [2] University of Trier

Title:
Participation rates in psychological studies over time - A meta-analysis.

Session & Time:
Quality Appraisal. Thursday, May 30th, 10:30 am - 11:00 am

Abstract:
Background and objectives
Nonresponse is one of the most severe problems in survey research (Hox & De Leeuw, 1994). If nonresponse is completely at random, it only reduces the amount of data collected. But in the case of nonrandom nonresponse, it can cause biased results, as the final respondents are no longer representative for the population of interest (Groves and Peytcheva 2008).
The main question of the meta-analysis is, whether the initial participation rate in psychological studies has decreased over time. Moreover, possible moderators of this time effect will be addressed: The design of an invitation letter, the contact protocol, the topic, the data collection mode, the burden of participating in the study and the incentives given to participants.

Research questions and hypotheses
As the participation of psychological studies is presumed to be influenced by values, culture and communication habits, changes of these factors over time are expected to have contributed to a decrease of participation rates during the last three decades. Thus, the first hypothesis stated is:
H1: The initial participation rate in psychological studies has decreased over time.

In individualistic cultures, decisions are rather based on an individual cost-benefit-calculus. Thus, the burden of the participation, incentives and interest in the topic are more important to convince potential participants to comply (Esser 1986). Due to the higher importance of the cost-benefit-calculus through individualization, over time it can be expected that longer studies suffer more from the decrease in participation than shorter ones.
H2: A higher announced time duration of the study aggravates the decline in response rates.

An intensively researched topic in the area of survey participation is the effect of incentives. It is rather unambiguous that incentives have a positive effect on response rates (e.g. Cook 2000), thus they can also be expected to compensate for the trend of decreasing response rates, especially taking into consideration the assumed higher importance of individual costs and benefits in decision-making. Several studies have already concluded, that monetary incentives are more effective

than non-monetary incentives (Dykema et al. 2012, Lee and Cheng 2006). Moreover, it is plausible, that a higher incentive has a stronger effect in reducing response rates than a smaller one. Halpern et al. (2002), as well as Murdoch et al (2014) provide evidence from randomized controlled trials for this assumption. These findings from cross-sectional research indicate, that monetary incentives and higher incentives should lessen the decrease in response rates.
H3: The decrease in participation rates is less pronounced for monetary incentives relative to other kinds of incentives.
H4: The higher the incentive, the smaller the decrease in participation over time.

Depending on the content and style of an invitation letter, there is considerable variation of the effect on response rates (de Leeuw et al. 2007). A method to get more attention is the personalization of the invitation letter (Cook et al. 2000). Due to the higher amount of communication, this measure should have become more important to reduce nonresponse.
H5: The personalization of the invitation letter reduces the decrease of participation rates.

Another method to get more attention and to make the participation in a study more attractive, is the salience of the topic.
H6: The decrease in participation rates is less pronounced for more salient topics.

The mode of the study conduction also plays a role for the survey response. Hox & De Leeuw (1994) found the highest response rate for face-to-face interviews, followed by telephone surveys. Mail surveys suffered from the lowest response rates. Yet, mail surveys were found to be preferred over web surveys by most respondents, as the meta-analysis of Shih & Fan (2007) showed.
More than ten years later now and for the area of psychological studies, it would be interesting, to what extent the further diffusion of the internet has reduced the reservation towards online surveys. The overall increase of communication makes the easy access and fast processing of online surveys more attractive. This leads to the conclusion, that the preferences for study conduction modes may have changed.
H7: The decrease is less pronounced for online surveys than for other survey modes.

Method/Approach
Of interest are psychological studies reporting initial participation rates and at least one of the following study design characteristics already mentioned. Student samples will be excluded due to differing motivation structure and incentives. In the case of panel studies, only the first wave is taken due to panel mortality in later waves. Studies have to be published in the three decades between 1988 and 2018. Publication language has to be either English or German. Editorials or texts reviewing results of original articles will not be included.

Data is collected on two levels. At the level of the study report, general information on the publication is retrieved. Within the study reports, there may be different characteristics of study conduction, for example to compare a group not offered an incentive with a group offered one. For each kind of treatment, there is one single initial participation rate. Thus, all the information on the treatment and the sample is retrieved at the level of the effect sizes:

A multilevel meta-analysis will be conducted. The dependent outcome will be the participation rate. The relevant independent variable for all tests is the time of sampling. The moderating effects of the survey design will be tested using the characteristics of study conduction as moderator variables. As the effects of the study design characteristics on the time effect are of interest, random slopes models are used.

Conclusions and implications (expected)

There is plenty of evidence on declining response rates in the last decades. This trend can aggravate the possible bias due to nonresponse. It is of interest what factors may moderate this trend to be able to guide survey operations by empirical evidence to optimize survey response. Due to the change in the willingness to participate in scientific studies, the continuous updating of the cumulative evidence is of importance.

References

Cook; Heath; Thompson (2000): A meta-analysis of response rates in web- or internet-based surveys. Educational and psychological measurement, 60(6), 821-836.

De Leeuw; Callegaro; Hox; Korendijk; Lensvelt-Mulders (2007): The influence of advance letters on response in telephone surveys. A meta-analysis. Public Opinion Quarterly, 71(3), 413-443

Dykema, Jennifer; Stevenson, John; Kniss, Chad; Kvale, Katherine; González, Kim; Cautley, Eleanor (2012): Use of Monetary and Nonmonetary Incentives to Increase Response Rates Among African Americans in the Wisconsin Pregnancy Risk Assessment Monitoring System. Maternal and child health journal, Vol. 16(4), 785-791.

Esser (1986): Über die Teilnahme an Befragungen. ZUMA-Nachrichten 18: 38-46.

Groves, Robert; Peytcheva, Emilia (2008): The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. Public Opinion Quarterly, Volume 72, Issue 2, Pages 167–189.

Halpern, Scott; Ubel, Peter; Berlin, Jesse; Asch, David (2002): Randomized trial of 5 dollars versus 10 dollars monetary incentives, envelope size, and candy to increase physician response rates to mailed questionnaires. Medical care, Vol. 40(9), 834.

Hox; de Leeuw (1994): A comparison of nonresponse in mail, telephone and face to face surveys. Quality and Quantity, 28 (4), 319-344.

Lee, Soo-Kyung; Yu-Yao, Cheng (2006): Reaching Asian Americans: Sampling Strategies and Incentives. Journal of Immigrant and Minority Health, Vol. 8(3), 245-250.

Murdoch, Maureen; Simon, Alisha Baines; Polusny, Melissa Anderson; Bangerter, Ann Kay; Grill, Joseph Patrick; Noorbaloochi, Siamak; Partin, Melissa Ruth (2014): Impact of different privacy conditions and incentives on survey response rate, participant representativeness, and disclosure of sensitive information: a randomized controlled trial. BMC Medical Research Methodology, Vol. 14 (1).

Shi; Fan (2007): Response rates and mode preferences in web-mail mixed-mode surveys: a meta-analysis. International Journal of Internet Science, 2(1), 59-82.

Authors:
Hannelies de Jonge [1], Suzanne Jak [1], Kees-Jan Kan [1]

[1] University of Amsterdam

Title:
Dealing with Artificially Dichotomized Variables in Meta-Analytic Structural Equation Modeling

Session & Time:
Methods in Meta-Analysis. Wednesday, May 29th, 6:00 pm - 6:30 pm

Abstract:
Background
Meta-analysis (Glass, 1976) is a commonly used statistical technique to aggregate sample effect sizes of different independent primary studies in order to draw inferences concerning population effects. To extend the range of research questions that can be answered, new meta-analytic models have been developed, such as meta-analytic structural equation modeling (MASEM) (Becker, 1992, 1995; Cheung, 2014, 2015a; Cheung & Chang, 2005; Jak, 2015; Viswesvaran & Ones, 1995). In primary studies, an effect size may represent the strength and direction of the association between any two variables of interest. Such an effect size can be expressed in different ways, for example as Pearson product-moment correlation, Cohens' *d*, biserial correlation, and point-biserial correlation. How an effect size is expressed depends on the nature of the variables (e.g., continuous or dichotomous), but also on the way the variables are measured or analyzed.

If one of the two continuous variables is artificially dichotomized, one may express the effect size as a point-biserial correlation. However, this typically provides a negatively biased estimate of the true underlying Pearson product-moment correlation (e.g., Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002). The biserial correlation on the other hand should generally provide an unbiased estimate (Soper, 1914; Tate, 1955). Bias in the effect size of any primary study may affect meta-analytic results in the same direction (Jacobs & Viechtbauer, 2017). Therefore, we may expect that the use of the point-biserial correlation for the relationship between an artificially dichotomized and continuous variable also biases MASEM-parameters. In the current study we will evaluate how using point-biserial correlations versus biserial correlations from primary studies may affects path coefficients, their standard errors, and model fit in MASEM. Based on the results, we expect to be able to inform researchers about which of the two investigated effect sizes is the most appropriate to use in MASEM-applications and under which conditions.

Aim
Our aim is to investigate the effects of using (1) the point-biserial correlation and (2) the biserial correlation for the relationship between an artificially dichotomized

variable and a continuous variable on MASEM-parameters and model fit. Specifically, our interest lies in path coefficients, standard errors of these coefficients, and model fit.

Method

We simulated meta-analytic data according to a full mediation (hence overidentified) population model (see Figure 1), with a continuous predictor variable X, continuous mediator M, and a continuous variable Y as outcome. Depending on the condition, the predictor variable X is artificially dichotomized in all or a given percentage of the primary studies. We chose this population model because in educational research the median number of variables in a 'typical' meta-analysis is three (de Jonge & Jak, 2018) and because mediation is a popular research topic.

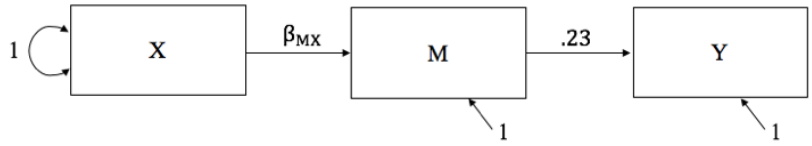


Figure 1. Population model with fixed parameter values.

Under this population model, random meta-analytic datasets were generated under different conditions. We systematically varied the following: (1) the size of the (standardized) path coefficient between X and M (.16, .23, .33), (2) the percentage of primary studies in which X was artificially dichotomized (25%, 75%, 100%), and (3) the cut-off point at which X was artificially dichotomized (at the median value, so a proportion of .05, or when groups become unbalance, at a proportion of .01). These choices were mainly based on typical situations in educational research. The size of the path coefficient, reflect the minimum, mean/median, and maximum pooled Pearson product-moment correlations in a 'typical' meta-analysis in educational research (de Jonge & Jak, 2018). The 75% primary studies that artificially dichotomize the variable X, is based on a comparable example of a meta-analysis in educational research (Jansen, Elffers, & Jak, 2019). We used between-study variances of .01. The number of primary studies in a meta-analysis was fixed at the median number of a 'typical' meta-analysis, which is 44 (de Jonge & Jak, 2018). Because we use a random-effects MASEM-method, the assumption is thus that the population comprises 44 subpopulations from which the 44 samples are drawn, and that the weighted mean of the subpopulation parameters equals the population parameter. Given a specific condition and the fixed number of 44 primary studies, we randomly sampled the within primary study sample sizes from a positively skewed distribution as used in Hafdahl (2007) with a mean of 421.75, yielding 'typical' sample sizes (de Jonge & Jak, 2018) for every iteration. We imposed 39% missing correlations (Sheng, Kong, Cortina, & Hou, 2016) by (pseudo) randomly deleting either variable M or Y from 26 of the 44 studies.

In each condition, we generated 2000 meta-analytic datasets drawn from the 44 subpopulations, which we analyzed using (1) the point-biserial and (2) the biserial

correlation as effect size between the artificially dichotomized predictor X and continuous mediator M. The full mediation model was fitted using random-effects two stage structural equation modeling (TSSEM) (Cheung, 2014) within the R-package 'metaSEM' (Cheung, 2015b).

As recommended (Becker, 2009; Hafdahl, 2007), we used the weighted mean correlation across the included primary studies to estimate the sampling variances and covariances of the correlation coefficients in the primary studies.

Next, over the converged simulated datasets, we (1) estimated the relative percentage bias in both path coefficients (less than 5% bias was considered negligible; Hoogland & Boomsma, 1998), (2) calculated the relative percentage bias of the standard errors of these path coefficients (less than 10% bias was considered acceptable; Hoogland & Boomsma, 1998), (3) calculated the rejection rates of the chi-square statistic of the model of Stage 2 (df = 1, α = .05) and tested whether the rejection rate significantly differed from the nominal α-level with the proportion test, and (4) compared the theoretical chi-square distribution (df = 1) with the empirical chi-square distribution (by means of QQplots and the Kolmogorov-Smirnov test).

Main Results

When the point-biserial correlation for the relation between an artificially dichotomized predictor and a continuous mediator was used, the path coefficient of this relationship in the population (β_{MX}) seems systematically underestimated. When the biserial correlation was used instead of the point-biserial correlation, this path coefficient could be considered unbiased in each condition. The estimated path coefficient between the two continuous variables (β_{YM}) could also be considered unbiased in all conditions, no matter if the biserial or point-biserial correlation was used. The relative percentage bias in the standard errors of all path coefficients could be considered as not substantial according to the criteria that were applied. However, we noticed that the relative percentage bias in the standard error of the path coefficient between the predictor and mediator (β_{MX}) seems systematically negatively biased when the biserial correlation was used. We also found that the relative percentage bias in the standard error of the path coefficient between the continuous variables Y and M (β_{YM}) seems systematically negative, regardless if the point-biserial or biserial correlation was used.

In most conditions, the rejection rate of the chi-square test of model fit at Stage 2 of the random-effects TSSEM was slightly above the nominal α-level, no matter if the point-biserial or biserial correlation was used. The results of the Kolmogorov-Smirnov test and QQplots show that when the biserial correlation was used, there was a statistically significant difference between the empirical chi-square distribution and the theoretical chi-square distribution in five of the 18 conditions. When the point-biserial correlation was used, there was a significant difference in the same five

conditions plus in three other conditions. There seems to be no clear pattern in which conditions the distributions differed significantly or not.

Expected Conclusions and Implications

We advise researchers who want to apply MASEM and want to investigate mediation to convert the effect size between any dichotomized predictor and continuous variable to a biserial correlation, not to a point-biserial correlation.

References

Becker, B. J. (1995). Corrections to “using results from replicated studies to estimate linear models”. *Journal of Educational and Behavioral Statistics, 20*, 100–102. doi:10.2307/1165390

Becker, B. J. (2009). Model-based meta-analysis. In H. Cooper, L. V. Hedges, & J.C. Valentine (Eds.) *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 377–395). New York: Russell Sage Foundation.

Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics, 17*, 341–362. doi:10.2307/1165128

Cheung, M. W.-L. (2014). Fixed-and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavior Research Methods*, *46*, 29–40. doi:10.3758/s13428-013-0361-y

Cheung, M. W.-L. (2015a). *Meta-analysis: A structural equation modeling approach*. Chichester, United Kingdom: John Wiley & Sons.

Cheung, M. W.-L. (2015b). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*, [1521]. https://doi.org/10.3389/fpsyg.2014.01521

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: a two-stage approach. *Psychological methods*, *10*, 40-64. doi:10.1037/1082-989X.10.1.40

Cohen, J. (1983). The cost of dichotomization. *Applied psychological measurement*, *7*, 249-253. https://doi.org/10.1177/014662168300700301

de Jonge, H., & Jak, S. (2018, June). *A Meta-Meta-Analysis: Identifying Typical Conditions of Meta-Analyses in Educational Research.* Paper presented at the conference Research Synthesis 2018 of Leibniz Institute for Psychology Information, Trier, Germany. http://dx.doi.org/10.23668/psycharchives.853

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *The Educational Researcher, 10*, 3–8. doi:10.3102/0013189X005010003

Hafdahl, A. R. (2007). Combining correlation matrices: Simulation analysis of improved fixed-effects methods. *Journal of Educational and Behavioral Statistics*, *32*, 180–205. doi:10.3102/1076998606298041

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–367. doi:10.1177/0049124198026003003

Jacobs, P., & Viechtbauer, W. (2017). Estimation of the biserial correlation and its sampling variance for use in meta-analysis. *Research synthesis methods*, *8*, 161-180. doi:10.1002/jrsm.1218

Jak, S. (2015). *Meta-analytic structural equation modelling*. Springer International Publishing. Jansen, D., Elffers, L., & Jak, S. (2019). *The functions of shadow education in school careers: a systematic review*. Manuscript submitted for publication.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*, *7*, 19–40. doi:10.1037//1082-989X.7.1.19

Tate, R. F. (1955). The theory of correlation between two continuous variables when one is dichotomized. *Biometrika*, *42*, 205–216. doi:10.2307/2333437

Sheng, Z., Kong, W., Cortina, J. M., & Hou, S. (2016). Analyzing matrices of meta-analytic correlations: current practices and recommendations. *Research synthesis methods*, *7*, 187-208. doi:10.1002/jrsm.1206

Soper, H. E. (1914). On the probable error of the bi-serial expression for the correlation coefficient. *Biometrika*, *10*, 384–390. doi:10.2307/2331789

Viswesvaran, C., & Ones, D. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology, 48*, 865–885. doi:10.1111/j.1744-6570.1995.tb01784.x

Authors:
Lies Declercq [1], Laleh Jamshidi [1]

[1] KU Leuven

Title:

Using effect sizes to reduce model complexity in a multilevel meta-analysis of single-case experimental design data

Session & Time:

Multilevel and IPD Meta-Analysis. Friday, May 31st, 11:00 am - 11:30 am

Abstract:

Background

In a single-case experimental design (SCED), a dependent variable is manipulated and repeatedly measured within a single subject or unit, to verify the effect of the manipulations ('treatments') on that variable (Onghena & Edgington, 2005). Typically, reports on SCED studies include scatterplots of the time series for one or more observed cases, making the raw SCED data readily available for meta-analysis. In raw SCED data obtained from multiple cases in one or more SCED studies, dependency is present in the data due to a nested hierarchical structure: measurements are nested within cases, which in turn are nested within studies. To account for this nesting, Van den Noortgate and (2003) proposed a hierarchical linear model with three levels to synthesize raw SCED data across cases. If the raw data are not available, Van den Noortgate and Onghena (2008) illustrate an alternative approach to statistically combine effect sizes from SCED studies. They propose an alternative standardized mean difference as an effect size to express the effect of the treatment for a particular case. These effect sizes are then combined in a three-level meta-analytical model.

Objectives, research questions and hypotheses

In a simulation study, we want to compare both multilevel approaches for synthesizing SCED data: the multilevel analysis of SCED raw data (RD approach) versus the multilevel analysis of SCED effect sizes (ES approach). For three models of increasing complexity, we simulate datasets and apply both approaches. For more complex models, the three-level models involve more regression coefficients and therefore more parameters to estimate. As such, the ES approach has an important potential benefit over the RD approach: the multilevel model estimated based on the effect sizes is reduced, so there are less parameters to estimate. This might result in faster estimation procedures and better convergence rates compared to the RD approach.

However, a drawback of the ES approach is the loss of information by reducing the rich raw data to effect sizes. It is not clear if the reduction in data combined with the smaller model in the ES approach will result in better or worse performance compared to the RD approach. Therefore we compare the performance of both

approaches in this simulation study by assessing the quality of the estimations, the convergence rate and the efficiency of both.

Method

A basic single-case design involves two phases, a baseline phase and a treatment phase. The most basic multilevel model for this type of data models a constant baseline level and an effect of the treatment on that level. Both coefficients are assumed to vary randomly around an overall mean at three levels due to 1) random sampling, 2) variation across participants and 3) variation across studies.
Alternatively to applying such a three-level model to the raw data (RD approach), a three-level model can also be applied to SCED effect sizes (ES approach). To calculate such effect sizes, we follow the approach proposed by Van den Noortgate and Onghena (2008) where we first obtain case-specific effect sizes, which are subsequently used in a three-level meta-analytic model to estimate the overall treatment effect.
In this simulation study, we generate raw SCED data from three models: the simple intercept-only model described above (model 1), a linear time trend model with a slope in both phases (model 2), and a quadratic time trend model (model 3). For each model we simulate 1000 datasets and apply the two approaches: we fit a three-level model directly onto the raw data (RD approach) and we use the raw data to first calculate effect sizes and then we fit a three-level model onto the effect sizes (ES approach). Note that for models 2 and 3, where the treatment has an effect on not only on the intercept (the constant) but also on the linear coefficient (models 2 and 3) and the quadratic coefficient (model 3). Therefore the ES approach requires a multivariate three-level model to simultaneously model two or three effect sizes.

Results

In terms of convergence, the ES approach performs well for all three models with convergence rates of 98% or higher. The RD approach performs slightly worse for model 2 but really breaks down for model 3, were only about half of the simulations converge. Convergence is especially bad for datasets with larger sample sizes.
In terms of absolute speed the comparison between both approaches depends of course on the software and the system used. The simulation was implemented in R with lme4 (Bates, Mächler, Bolker, & Walker, 2014) for the RD approach and metafor (Viechtbauer, 2010) for the ES approach. With identical settings for the optimizer and the maximum number of function evaluations for both approaches, the RD approach was faster for fitting complex models to small datasets. However, a single model fit took almost always less than a minute, so any difference between approaches might be negligible in practice.
In terms of quality of the estimations, the fixed effect estimations were unbiased for both approaches and they had identically small mean squared errors (MSE's). However, the ES approach resulted in CI's which were consistently too narrow and Type I error rates which were consistently too high. For the variance components, the ES approach estimations where less biased than those from the RD approach.

Conclusions

Both approaches provide reliable point estimates no matter the underlying model complexity. However, when using effect sizes in a three-level meta-analytic model, inference results might be unreliable. This is in line with previous research and several different adjustments and alternative testing procedures have been proposed and compared to accommodate this problem (Sánchez-Meca & Marín-Martínez, 2008). With more complex models the raw data approach tends to throw convergence warnings and errors. Based on our findings, we can confirm that the effect size approach is a reasonable alternative when SCED raw data are not available. Caution is however advised when performing unadjusted Wald-type *z*- or t-tests on the overall effect sizes when effect sizes were used instead of raw data, because these tests lead to unreliable confidence intervals and *p*-values.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4, *67*(1). https://doi.org/10.18637/jss.v067.i01

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: single-case design and analysis. *The Clinical Journal of Pain*, *21*(1), 56–68. https://doi.org/10.1097/00002508-200501000-00007

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence Intervals for the Overall Effect Size in Random-Effects Meta-Analysis. *Psychological Methods*, *13*(1), 31–48. https://doi.org/10.1037/1082-989X.13.1.31

Van den Noortgate, W., & Onghena, P. (2003). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, *18*(3), 325–346. https://doi.org/10.1521/scpq.18.3.325.22577

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 142–151. https://doi.org/10.1080/17489530802505362

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.1103/PhysRevB.91.121108

Authors:
Belén Fernández Castilla [1]

[1] KU Leuven

Title:
Multilevel Models in Meta-Analysis: A Systematic Review of Their Application and Suggestions

Session & Time:
Multilevel and IPD Meta-Analysis. Friday, May 31st, 10:30 am - 11:00 am

Abstract:
Introduction
Meta-analysis can be conceptualized as a multilevel analysis: effect sizes are nested within studies. Effect sizes vary due to sampling variance at Level 1, and possibly also due to systematic differences across studies at Level 2. Therefore, multilevel models and software can be used to perform meta-analysis. An advantage of using the multilevel framework for doing meta-analyses is the flexibility of multilevel models. For instance, additional levels can be added to deal with dependent effect sizes within and between studies. In primary studies, it is common to report multiple effect sizes extracted from the same sample. Also, studies might belong to different higher-level clusters, as countries or research groups. These two scenarios generate dependency among effect sizes, and for appropriately accounting for this dependency (and therefore avoid inflated Type I errors), additional levels can be added that explicitly account for the variation among effect sizes within and/or between studies. Besides hierarchical models, other non-purely hierarchical models have been also proposed for meta-analysis, such as Cross-Classified Random Effects models (CCREMs, Fernández-Castilla et al., 2018). Although multilevel models are very flexible, we suspect that applied researchers do not take advantage of all possibilities that these models offer. In fact, most published meta-analyses are restricted to three-level models despite some meta-analytic data require other model specifications, such as four- or five- level models or CCREMs. Therefore, the goal of this study is to describe how multilevel models are typically applied in meta-analysis and to illustrate how, in some meta-analyses, more sophisticated models could have been applied that accounts better for the (non) hierarchical data structure.

Method
Meta-analyses that applied multilevel models with more than one random component were searched in June, 2018. We looked at the meta-analyses that cited the studies of Cheung (2014), Hox and De Leeuw (2003), Konstantopoulos (2011), Raudenbush and Bryk (1985), and Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca (2013, 2014). We also searched in six electronic databases, using the strings “three-level meta-analysis” OR “multilevel meta-analysis” OR “multilevel meta-analytic review”. No date restriction was imposed. Meta-analysis were included if: a) effect sizes were combined using a multilevel model with more than one random

component; b) The meta-analysis was included in a journal article, conference presentation or a dissertation; c) The meta-analysis was written in English, Spanish or Dutch.

Results

The initial search resulted in 1,286 studies. After applying the inclusion criteria, we finally retrieved 178 meta-analyses. From these, 162 meta-analysis fitted a three-level model, 9 fitted a four-level model, 5 applied CCREMs, and 2 reported a five-level model. We could distinguish five situations in which other models different from the three-level model would have been more appropriate given the (non) hierarchical data structure:

1. A fourth level could have been added to model dependency within studies. For instance, Fischer and Boer (2011) specified a three-level model, were effect sizes (Level 1) were nested within studies (Level 2), nested within countries (Level 3). There were several effect sizes within studies, but this within-study variance was ignored. Therefore, it would have been appropriate to add an additional level to model between-outcomes (within-study) variance.

2. A fourth level could have been specified to deal with more sources of within-study dependencies. For instance, in O'Mara (2006), there were several interventions within studies, and that is why a three-level model was specified: Sampling variance (Level 1), between-interventions variance (Level 2), and between-studies variance (Level 3). However, there were 200 interventions and 460 effect sizes in total, meaning that each intervention led to multiple effect sizes, and that the dependency between these outcomes (within interventions) was not taken into account. A more appropriate model would have been a four-level model: Sampling variance (Level 1), between-outcomes variance (Level 2), between-comparisons variance (Level 3) and between-studies variance (Level 4).

3. A fourth level could have been added to take into account dependency across studies. In the study of Klomp and Valckx (2014), a three-level model was fitted because there were multiple outcomes within studies. In this case, some studies made use of the same big dataset, so a fourth level could have been added to model between-datasets variance.

4. A five-level model could have been applied to model additional within-study and between-study dependencies. In Rabl, Jayasinghe, Gerhart, and Kühlmann (2014), a three-level model was fitted, where effect sizes were nested within studies, nested within countries. There were several effect sizes within studies, so an additional level could have been added to model within-study variance. Furthermore, some studies used the same dataset, so another level could have been specified to estimate the between-datasets variance. The inclusion of these two additional levels would have led to a five-level model.

5. CCREM's could have been applied instead of three-level models. In the study of Fisher, Hanke and Sibley (2012), effect sizes were nested within studies, nested within countries. However, studies were not completely nested within countries, but rather studies and countries were two cross-classified factors: in one study, effect sizes could come from different countries, and effect sizes from the same country could belong to different studies. Therefore, a CCREM model would have accounted better for this cross-classified data structure.

Discussion

This systematic review shows how researchers using multilevel model typically apply three-level models to account for dependent effect sizes, although alternative model specifications, such as four- or five- level models or CCREMs, might be more correct given the nature of the data. We have given some examples of how alternative models could have been used for meta-analysis, and we encourage researchers to carefully consider the underlying data structure before selecting a specific multilevel model. Omitting levels in a multilevel analysis might increase the possibility of committing a Type I error. Therefore, the proper specification of the model is the only way to guarantee appropriate estimates of the combined effect size, standard errors, and variance components.

References

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*, 211-229.

Fernández-Castilla, B., Maes, M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2018). A demonstration and evaluation of the use of cross-classified random-effects models for meta-analysis. *Behavior Research Methods*, 1-19.

Fischer, R., & Boer, D. (2011). What is more important for national well-being: money or autonomy? A meta-analysis of well-being, burnout, and anxiety across 63 societies. *Journal of Personality and Social Psychology*, *101*, 164-184.

Fischer, R., Hanke, K., & Sibley, C. G. (2012). Cultural and institutional determinants of social dominance orientation: A cross-cultural meta-analysis of 27 societies. *Political Psychology*, *33*, 437-467.

Hox, J. J., & de Leeuw, E. D. (2003). Multilevel models for meta-analysis. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 90–111). Mahwah, NJ: Erlbaum.

Klomp, J., & Valckx, K. (2014). Natural disasters and economic growth: A meta-analysis. *Global Environmental Change*, *26*, 183-195.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2,* 61-76.

O'Mara, A. J., Marsh, H. W., & Craven, R. G. (July, 2006). A Comprehensive Multilevel Model Meta-Analysis of Self-Concept Interventions. In *Fourth International Biennial SELF Research Conference,* Ann Arbor.

Rabl, T., Jayasinghe, M., Gerhart, B., & Kühlmann, T. M. (2014). A meta-analysis of country differences in the high-performance work system–business performance relationship: The roles of national culture and managerial discretion. *Journal of Applied Psychology*, *99*, 1011-1041.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*, 75-98.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*, 576-594.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2014). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods, 47*, 1274-1294.

Authors:
Anna-Carolina Haensch [1]

[1] GESIS Leibniz Institute for the Social Sciences

Title:
IPD Meta-Analysis of Complex Survey Data: A Simulation Study

Session & Time:
Multilevel and IPD Meta-Analysis. Friday, May 31st, 11:30 am - 12:00 pm

Abstract:
When Glass coined the term meta-analysis (MA) in 1976, he exclusively referred to a type of meta-analysis that today is known as aggregate person data (APD) meta-analysis. In recent years, another type of meta-analysis has gained popularity that is referred to as individual person data (IPD) meta-analysis (Riley et al., 2010; Burke et al., 2017). IPD meta-analysis utilizes the raw, participant-level data by pooling multiple datasets, e.g., original data from different trials in medicine or surveys in the social sciences.

So far, IPD meta-analysis has been utilized in the medical sciences (Jeng et al.,1995; McCormack et al., 2004; Palmerini et al., 2015; Rogozinska et al., 2017) or psychology (Cuijpers et al., 2014; Gu et al., 2015; Karyotaki et al., 2015). In these disciplines, most original studies focus on some sort of treatment or intervention effect and apply experimental research designs to come to causal conclusions. In contrast, many epidemiological, sociological or economic studies are non-experimental, i.e., observational studies or based on survey data. When analyzing non-experimental data, researchers have to take into account confounding bias and cannot rely on simple bivariate effect sizes. Instead, the focus shifts to more sophisticated methods, e.g., regression models. The “effect sizes” of interest are now regression slopes of focal predictors on an outcome variable (Becker and Wu, 2007; Aloe and Thompson, 2013). However, it poses a challenge to estimate IPD meta-analyses of regression coefficients with survey-based data. In contrast to experimental data, survey-based data is subject to complex sampling like stratification of the population and cluster sampling.

To account for complex sampling schemes or endogenous sampling, survey-based data often comes with survey weights ranging from design-based weights to nonresponse weights, as well as post-stratification weights. These weights can be used to receive approximately unbiased populations estimates. Survey-weighted regressions are located between the two classical inferential frameworks, model- (Fisher, 1922) and design-based (Neyman, 1934) inference. Until now, the literature on IPD meta-analysis with complex survey data is sparse. So, even though IPD meta-analysis can be considered the “gold standard” in evidence-driven research, it is yet unclear how to deal with non-experimental, survey-based data that is subject to complex sampling. We systematically explore when and how to use survey

weighting in regression-based analyses in combination with different IPD meta-analytical approaches.

We will build up on the work done by DuMouchel and Duncan (1983) and Solon et al. (2013) for survey weighted regression analysis. We will show through Monte Carlo simulations that endogenous sampling and heterogeneity of effects models require survey weighting to receive approximately unbiased estimates in the meta-analytical case. Even though most researchers primarily aim for approximately unbiased estimates, it is not recommended to use weights "just in case." Weights can increase the variance of meta-analytical estimates quite dramatically.

Second, we focus on a list of methodological questions: Do survey weighted one-stage, and two-stage meta-analysis perform differently? How do we deal with weighted surveys which have different observation numbers – is it necessary to transform the weights? Is it possible to include random effects into survey weighted meta-analysis, especially if we have to assume study heterogeneity? Another challenging methodological question is the inclusion of random effects in a one-stage meta-analysis.

Our simulations show that two-stage IPD meta-analysis will be biased if the variation in the weights is high, whereas one-stage IPD meta-analysis remains unbiased. We show that researchers can improve the efficiency of their one-stage IPD analysis if they transform their weights with one of the transformations Korn and Graubard (1999) proposed. The scaling is beneficial in the case of surveys with different sample sizes. We also show that the inclusion of random effects in a one-stage meta-analysis is challenging but doable. Transformation of weights is needed in most cases.

References

Aloe, A. M. and Thompson, C. G. (2013). The Synthesis of Partial Effect Sizes. Journal of the Society for Social Work and Research, 4(4):390–405.

Burke, D. L., Ensor, J., and Riley, R. D. (2017). Meta-analysis Using Individual Participant Data: One-stage and Two-stage Approaches, and Why They May Differ. Statistics in Medicine, 36(5):855–875.

Becker, B. J. and Wu, M.-J. (2007). The Synthesis of Regression Slopes in Meta-Analysis. Statistical Science, 22(3):414–429.

Cuijpers, P., Weitz, E., Twisk, J., Kuehner, C., Cristea, I., David, D., DeRubeis, R. J., Dimidjian, S., Dunlop, B. W., Faramarzi, M., Hegerl, U., Jarrett, R. B., Kennedy, S. H., Kheirkhah, F., Mergl, R., Miranda, J., Mohr, D. C., Segal, Z. V., Siddique, J., Simons, A. D., Vittengl, J. R., and Hollon, S. D. (2014). Gender as Predictor and Moderator of Outcome in Cognitive Behaviour Therapy and Pharmacotherapy for Adult Depression: An "Individual Patient Data" Metaanalysis. Depression and Anxiety, 31(11):941–951.

DuMouchel, W. H. and Duncan, G. J. (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. Journal of the American Statistical Association, 78(383):535–543.

Fisher, R. (1922). On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 222(594-604):309–368.

Gu, J., Strauss, C., Bond, R., and Cavanagh, K. (2015). How do Mindfulness Based Cognitive Therapy and Mindfulness-based Stress Reduction Improve Mental Health and Wellbeing? A Systematic Review and Meta-analysis of Mediation Studies. Clinical Psychology Review, 37:1 – 12.

Jeng, G., Scott, J., and Burmeister, L. (1995). A Comparison of Meta-analytic Results Using Literature vs Individual Patient Data: Paternal Cell Immunization for Recurrent Miscarriage. JAMA, 274(10):830–836.

Karyotaki, E., Kleiboer, A., Smit, F., Turner, D. T., Pastor, A. M., Andersson, G., Berger, T., Botella, C., Breton, J. M., Carlbring, P., and et al. (2015). Predictors of treatment dropout in self-guided web-based interventions for depression: an "individual patient data" meta-analysis. Psychological Medicine, 45(13):2717–2726.

Korn, E. L. and Graubard, B. I. (1999). Analyses Using Multiple Surveys. In Korn, E. L. and Graubard, B. I., editors, Analysis of Health Surveys, chapter 8, pages 278–303. Wiley-Blackwell.

McCormack, K., Grant, A., and Scott, N. (2004). Value of Updating a Systematic Review in Surgery Using Individual Patient Data. BJS, 91(4):495–499.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society, 97(4):558–625.

Palmerini, T., Sangiorgi, D., Valgimigli, M., Biondi-Zoccai, G., Feres, F., Abizaid, A., Costa, R. A., Hong, M.-K., Kim, B.-K., Jang, Y., Kim, H.-S., Park, K. W., Mariani, A., Riva, D. D., Genereux, P., Leon, M. B., Bhatt, D. L., Bendetto, U., Rapezzi, C., and Stone, G. W. (2015). Short- Versus Long-term Dual Antiplatelet Therapy After Drug-eluting Stent Implantation: An Individual Patient Data Pairwise and Network Meta-analysis. Journal of the American College of Cardiology, 65(11):1092 – 1102.

Riley, R. D., Lambert, P. C., and Abo-Zaid, G. (2010). Meta-analysis of Individual Participant Data: Rationale, Conduct, and Reporting. BMJ, 340:c221.

Rogozinska, E., Marlin, N., Thangaratinam, S., Khan, K. S., and Zamora, J. (2017). Meta-analysis Using Individual Participant Data from Randomised Trials: Opportunities and Limitations Created by Access to Raw Data. BMJ Evidence-Based Medicine, 22(5):157–162.

Solon, G., Haider, S. J., and Wooldridge, J. (2013). What Are We Weighting For? Working Paper 18859, National Bureau of Economic Research.

Authors:
Nadine Kasten [1], Tanja Burgard [2], Oliver Wedderhoff [2], Michael Bosnjak [2], Timo Gnambs [3]

[1] University of Trier; [2] Leibniz Institute for Psychology Information (ZPID); [3] Leibniz Institute for Educational Trajectories

Title:
A meta-analytic investigation of the factor structure of the Positive and Negative Affect Schedule (PANAS)

Session & Time:
Health Psychology. Friday, May 31st, 3:30 pm - 4:00 pm

Abstract:
The 20-item Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) is a self-report measure to assess two global measures of psychological well-being, namely *positive affect* (PA) and *negative affect* (NA). Its brevity and repeated evidence of sufficient levels of reliability and validity has contributed to a frequent use in all areas of psychology. Moreover, the PANAS has been translated into various languages and is administered all over the world. Despite its widespread use, there is still an ongoing discussion with regard to the internal structure of the PANAS. Though originally developed to provide distinct and independent measures of PA and NA, empirical studies identified different factor structures including two- and three-factor models, second order models, and bifactor models. Additionally, there is few information on the robustness of the internal structure of the PANAS across, for example, different application contexts and questionnaire characteristics. In light of the ongoing discussion, the present study evaluates the nature and the generalizability of the PANAS factor structure by means of a meta-analytic structural equation modeling approach (MASEM; Cheung & Chan, 2005). In a first step, inter-item correlation matrices from 76 independent samples (total N = 54,976) were pooled. Then, popular factor models for the PANAS were compared using confirmatory factor analysis. Overall, the originally proposed orthogonal two-factor model exhibited a rather inferior fit (CFI = .884, TLI = .871, RMSEA = .052). In contrast, a bifactor model was the most appropriate representation of the PANAS (CFI = .930, TLI = .912, RMSEA = .043). This model included two specific factors for PA and NA as well as a general factor that represents a fundamental approach or withdrawal tendency (i.e., affective polarity). Moderator analysis revealed profound differences in the internal structure of the PANAS between the original English version and translated versions, leaving some doubts on the appropriateness of the application of the PANAS in cross-cultural research.

References

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods*, *10*(1), 40–64. https://doi.org/10.1037/1082-989X.10.1.40

Authors:
Karina Karolina Kedzior [1], Ilkay Kaplan [1]

[1] University of Bremen

Title:
Is AMSTAR2 an appropriate tool to assess the quality of systematic reviews in psychology?

Session & Time:
Quality Appraisal. Thursday, May 30th, 11:30 am - 12:00 pm

Abstract:
Background
Systematic reviews are frequently used in psychology to guide future research and to summarise the empirical evidence for decision making. However, the quality of such reviews is not always acceptable (Kedzior & Seehoff, 2018) leading to poor reproducibility of conclusions and outcomes of statistical meta-analysis (Lakens et al., 2016).
One method of assessing the quality of systematic reviews is 'A MeaSurement Tool to Assess Systematic Reviews' (AMSTAR) (Shea et al., 2007). AMSTAR is an 11-item scale designed to evaluate the quality of various aspects of systematic reviews, including the literature search, the data coding, the risk of bias assessment, and the data synthesis. Although frequently used, the psychometric properties of AMSTAR were criticised (Wegewitz et al., 2016) and a new version of the instrument (AMSTAR2) was developed (Shea et al., 2017). AMSTAR 2 consists of 16 items, including seven being critical for high quality.

Objective
The objective of the current study is to investigate if AMSTAR2 is a better tool to assess the quality of systematic reviews than AMSTAR. For this purpose we compare the scores on both scales that we have applied to the same systematic reviews in one specific field (the effects of Tai Chi on psychological well-being in Parkinson's Disease, PD).

Research question
The research question in the current study is: Is AMSTAR2 an appropriate tool to assess the quality of systematic reviews in psychology?

Method
The literature search, selection of systematic reviews, and quality assessment using AMSTAR and AMSTAR2 were done by each author independently and any inconsistencies were resolved by consensus during discussion.
Inclusion and exclusion criteria. We have searched for systematic reviews (with or without meta-analysis) regarding the effects of Tai Chi on symptoms of PD. The exclusion criteria for the current study were: 1) narrative (non-systematic) review, 2) primary study.

Search strategy. The search strategy is already described elsewhere (Kedzior & Kaplan, 2018). Briefly, the electronic literature search of PubMed and PsycInfo (on 14.02.2018) identified *k*=21 studies (Title/Abstract: 'Parkinson's Disease' AND Tai Chi AND review). Inclusion criteria were met by *k*=10 systematic reviews that were included in the current study.
Coding procedures. The data in the *k*=10 systematic reviews were coded using a self-developed form and the review quality was assessed using AMSTAR (in March 2018) and AMSTAR2 (in June 2018). AMSTAR outcomes vary between 11 (maximum quality) to 0 (minimum quality). AMSTAR2 outcomes vary between high quality (no critical weaknesses) to critically low quality (> one critical weakness).

Results
Overall quality assessment. The *k*=10 systematic reviews on Tai Chi in PD had a mean (±*SD*) AMSTAR score of 7±2 (range: 3-9, mode: 9, score<6 in 3/10 reviews). Therefore, most reviews (70%) had acceptable to high quality on AMSTAR. However, AMSTAR2 evaluation showed that the same reviews had 1-5/7 critical weaknesses. Therefore, all reviews had a low to critically low quality according to AMSTAR2.
Agreement between AMSTAR and AMSTAR2. The inspection of individual items revealed that there was a high agreement between both scales regarding the assessment of most items, including the review protocol, the literature search, the duplicate data extraction, the data coding and synthesis, the risk of bias assessment, the publication bias assessment, and the conflict of interest in the review. Our results also confirm that the quality of AMSTAR2 items has improved. For example, two double-barrelled items on AMSTAR (Item 2 regarding the duplicate study selection *and* data coding and Item 5 regarding the list of included *and* excluded studies) are listed as four separate items on AMSTAR2 (Items 5-6 and Items 7-8, respectively).
Disagreement between AMSTAR and AMSTAR2. The disagreement between the scales is due to the interpretation of the overall scores (too lenient in AMSTAR and too conservative in AMSTAR2) as well as the focus on critical items that may not have been routinely required/reported in the past reviews. Such items include the presence of the review protocol and the list of excluded studies with justification for exclusion. Since all *k*=10 systematic reviews had at least one critical weakness (either did not have *a priori* protocol and/or have not reported the list of excluded studies), they were classified as having low to critical low quality on AMSTAR2.

Conclusions and implications
AMSTAR2 may not be a valid tool for assessing the quality of the past systematic reviews because some critical items required for high quality have not been routinely included in journal requirements in the past. However, AMSTAR2 provides excellent guidelines for conducting of future systematic reviews and should be incorporated in journal guidelines for authors. Providing the AMSTAR2 evaluation of own systematic reviews (including the locations where specific items were addressed in own review)

could help the authors to conduct high quality reviews and the journal editors and readers to quickly assess the quality of such reviews.

References

Kedzior, K., & Kaplan, I. (2018). Scientific quality of systematic reviews on the effects of Tai Chi on well-being in Parkinson's disease (PD). *Systematic Reviews (submitted)*.

Kedzior, K. K., & Seehoff, H. (2018). *Common problems with meta-analysis in published reviews on major depressive disorders (MDD): a systematic review*. Paper presented at the Research Synthesis Conference 2018 (June 10-12, 2018, Trier, Germany).

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: six practical recommendations. [journal article]. *BMC Psychology, 4*(1), 24.

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., Porter, A. C., Tugwell, P., Moher, D., & Bouter, L. M. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7*(1), 1-7.

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ, 358*, j4008.

Wegewitz, U., Weikert, B., Fishta, A., Jacobs, A., & Pieper, D. (2016). Resuming the discussion of AMSTAR: What can (should) be made better? *BMC Medical Research Methodology, 16*(1), 111.

Authors:
Michael Kossmeier [1], Ulrich S. Tran [1], Martin Voracek [1]

[1] University of Vienna

Title:
Power-enhanced funnel plots for meta-analysis: The sunset funnel plot

Session & Time:
Methods in Meta-Analysis. Wednesday, May 29th, 5:30 pm – 6:00 pm

Abstract:
Background and Objectives
The funnel plot is the most widely used diagnostic plot for meta-analysis. Numerous variants exist to visualize small-study effects, heterogeneity, and the sensitivity of the meta-analytic summary estimates to new evidence (Langan, Higgins, Gregory, & Sutton, 2012). What is currently missing is a funnel plot variant which incorporates information on statistical study-level power to detect an effect of interest. To fill this gap, we here introduce the sunset funnel plot, which, in essence, is a power-enhanced funnel plot (Figure 1).
Visual funnel plot examinations for small-study effects include checks whether smaller studies in particular (i.e., those with larger standard errors and associated lower analytic power) tend to yield larger effect sizes. When such an association evidently is driven by conventional criteria of statistical significance, then publication bias is considered to be a likely explanation for the phenomenon, and preferred to other causes, such as true heterogeneity or chance alone (Peters, Sutton, Jones, Abrams, & Rushton, 2008).
Information on the power of studies can further support such evaluations of potential publication bias. The test for excess significance (Ioannidis & Trikalinos, 2007) is a widely used evidentiality test to check whether there is a higher number of statistically significant studies than expected, considering their power to detect an effect of interest. Such an excess of significant findings indicates bias in the set of studies under consideration. In the same spirit, if an implausible excess of significant, but at the same time underpowered, studies is visible and potentially drives small-study effects in the funnel plot, this can further weaken the credibility of these results and indicate bias.
In addition, significant effects observed in low-powered studies more likely are false positive findings (Forstmeier, Wagenmakers, & Parker, 2017). Power can therefore be seen as an indicator for the replicability of research findings. Indeed, for a set of studies, the deviation of (or, gap between) the proportion of actually observed significant studies and twice the median study power has been proposed as the R-index of replicability (Schimmack, 2016).
All in all, study-level power is one useful information to assess the credibility and evidentiality of a set of studies potentially included in a meta-analysis. Consequently, a power-enhanced funnel plot is one means to visualize and communicate this

information by incorporating information on study-level power in the well-known, classic funnel plot display.

Methods

The sunset (power-enhanced) funnel plot assumes normally distributed effect sizes and regards variances of these effect sizes as known. These assumptions are common in the context of meta-analysis and standard effect sizes for meta-analysis are suitable for the sunset funnel plot as well (e.g., Cohen *d*, Hedges *g*, log *OR*, Fisher's *z*-transformed *r*).

For a true population effect size δ, the power of a two-sided Wald test with significance level α testing the null hypothesis $\delta = 0$ is given by

$$\text{Power} = 1 - \Phi(z_{1-\alpha/2} - \delta/SE(d)) + \Phi(-z_{1-\alpha/2} - \delta/SE(d))$$

with Φ the cumulative distribution function of the standard normal distribution, $z_{1-\alpha/2}$ the 1-α/2 quantile of the standard normal distribution, and SE(*d*) the standard error of the study effect size *d*.

The sunset (power-enhanced) funnel plot visualizes these power estimates corresponding to specific standard errors on a second y-axis and with color-coded power regions (Figure 1). Color regions range from an alarming dark red for highly underpowered studies to a relaxing dark green for appropriately powered studies to detect the underlying true effect of interest. The color palette used in the graphic display is vividly remindful of a colorful sunset; hence, the denomination sunset funnel plot.

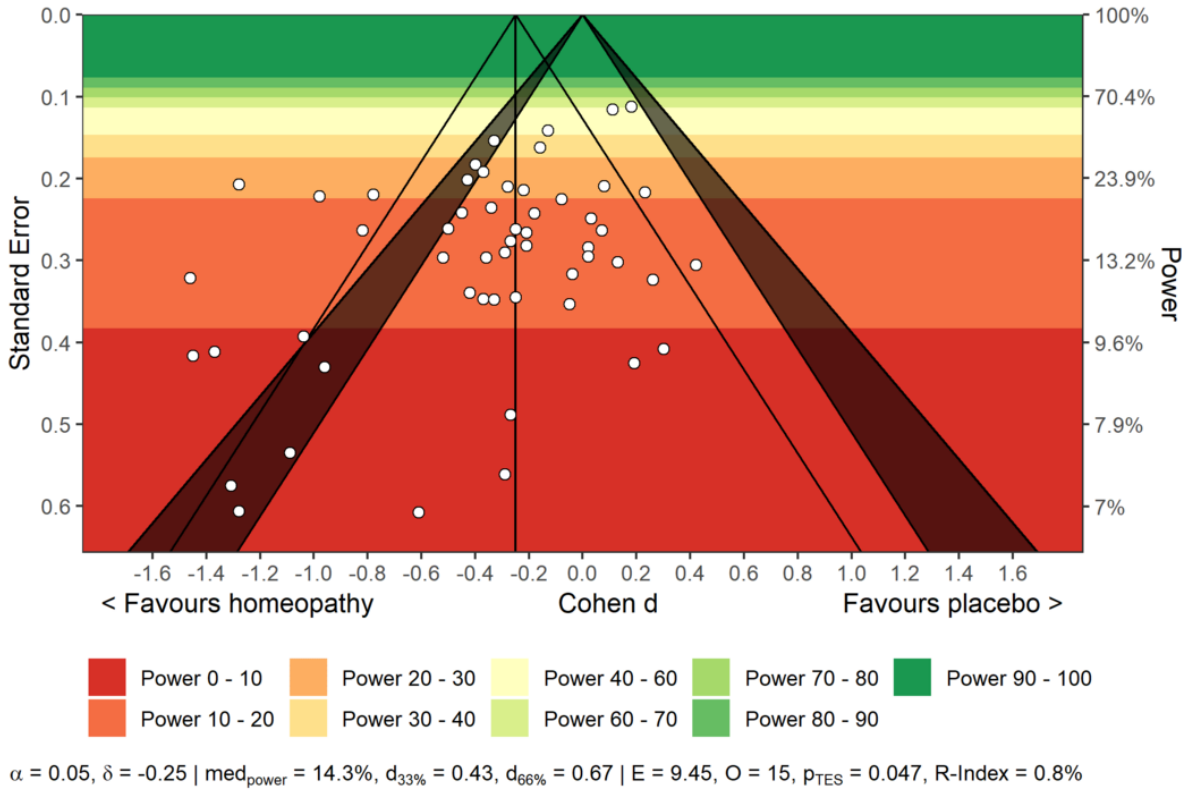


Figure 1: Sunset (power-enhanced) funnel plot, using data from a published meta-analysis (Mathie et al., 2017) comparing homeopathic treatment with placebo. 95% confidence contours are shown, with the black vertical reference line marking the observed summary effect (fixed-effect model) used for power analysis. Significance contours at the .05 and .01 levels are indicated through dark shaded areas. Power estimates are computed for a two-tailed test with significance level .05. *R* code to reproduce the figure:
https://osf.io/967bh/?view_only=e659e4eb1cfa46c2bfe4c8ceb622e922
The underlying true population effect size can either be determined theoretically (e.g., by assuming a smallest effect of interest), or empirically, using meta-analytic estimates of the summary effect. For the latter, the fixed-effect model estimator is one natural default choice, giving less weight (and therefore being less sensitive) to small, biased studies, as compared to random-effects meta-analytic modeling.
A number of related power-based statistics can be presented alongside the power-enhanced funnel plot and support its evaluation. These include (i) the median power of studies, (ii) the true underlying effect size necessary for achieving certain levels of median power (e.g., 33% or 66%), (iii) the results of the test for excess significance (Ioannidis & Trikalinos, 2007), and (iv) the R-index as measure for the expected replicability of findings (Schimmack, 2016).
To create sunset (power-enhanced) funnel plots and to compute statistics related to these, we provide the tailored function viz_sunset in the package metaviz (Kossmeier, Tran, & Voracek, 2018) within the statistical software *R* (R Core Team, 2018), and a corresponding online application available at https://metaviz.shinyapps.io/sunset/.

Results

For the following illustration example, we use data from a recent published meta-analysis on the effect of homeopathic treatment vs. placebo for numerous medical conditions (Mathie et al., 2017). In this systematic review and meta-analysis, bias assessment suggested high risk of bias for the majority of the 54 randomized controlled trials (RCTs) considered for meta-analysis; only three RCTs were judged as reliable evidence. For illustration purposes, we use the totality of these 54 effect sizes.
Visual examination of the corresponding funnel plot shows clear small-study effects, such that imprecise, smaller studies (those with larger standard errors) report larger effects in favor of homeopathy than more precise, larger studies (those with smaller standard errors). This association seems to be driven by studies reporting imprecise, but significant estimates, in particular. Incorporating power information in these considerations (with the fixed-effect estimate δ = -0.25 in favor of homeopathy) additionally reveals that a non-trivial, implausible high, and thus worrisome, number of the significant studies evidently are drastically underpowered (with power values lower than 10%) to detect this effect of interest, thus further suggesting bias (Figure 1). Accordingly, there is an excess of significant findings among the primary studies included in this meta-analysis (15 nominally significant studies observed, but, under

these circumstances, only 9.45 significant studies expected; p = .047). The median power of this set of primary studies merely amounts to 14.3% (*IQR*: 11.1-20.6%), and the true effects needed to reach typical (i.e., median) power levels of 33% or 66% would be substantial (absolute δ values of 0.43 or 0.67, respectively). The expected replicability of findings, as quantified with the R-Index, is extremely low (0.8%).

Conclusions and Implications

We introduce the sunset (power-enhanced) funnel plot as a new, useful display for the meta-analytic visualization toolbox. First and foremost, the sunset funnel plot allows to incorporate power considerations into classic funnel plot assessments for small-study effects. In the same spirit as testing for an excess of significant findings (Ioannidis & Trikalinos, 2007), the credibility of findings can further be critically examined by checking whether small-study effects are especially driven by an implausible large number of significant, but at the same time underpowered, studies. Second, the display allows to visually explore and communicate the distribution and typical values of study power for an effect of interest. This visualization is not only informative for meta-analyses, but also in the broader context of meta-scientific investigations into the power of studies of whole scientific fields (e.g., Szucs, & Ioannidis, 2017). Third, changes of power values for a set of studies can be visually examined by varying the true underlying effect. This directly corresponds to questions regarding the necessary true effect size, such that the power of individual or typical studies would reach desired levels. Software to create sunset (power-enhanced) funnel plots is provided.

References

Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive finding: A practical guide. *Biological Reviews*, *92*, 1941-1968.

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245-253.

Kossmeier, M., Tran, U. S., & Voracek, M. (2018). metaviz [R software package]. Retrieved from https://github.com/Mkossmeier/metaviz

Langan, D., Higgins, J. P., Gregory, W., & Sutton, A. J. (2012). Graphical augmentations to the funnel plot assess the impact of additional evidence on a meta-analysis. *Journal of Clinical Epidemiology*, *65*, 511-519.

Mathie, R. T., Ramparsad, N., Legg, L. A., Clausen, J., Moss, S., Davidson, J. R., ... McConnachie, A. (2017). Randomised, double-blind, placebo-controlled trials of non-individualised homeopathic treatment: Systematic review and meta-analysis. *Systematic Reviews*, *6*, 63.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Schimmack, U. (2016). The replicability-index: Quantifying statistical research integrity. Retrieved from https://replicationindex.wordpress.com/2016/01/31/a-revised-introduction-to-the-r-index/

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology, 15,* e2000797.

Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive finding: A practical guide. *Biological Reviews*, *92*, 1941-1968.

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245-253.

Kossmeier, M., Tran, U. S., & Voracek, M. (2018). metaviz [R software package]. Retrieved from https://github.com/Mkossmeier/metaviz

Langan, D., Higgins, J. P., Gregory, W., & Sutton, A. J. (2012). Graphical augmentations to the funnel plot assess the impact of additional evidence on a meta-analysis. *Journal of Clinical Epidemiology*, *65*, 511-519.

Mathie, R. T., Ramparsad, N., Legg, L. A., Clausen, J., Moss, S., Davidson, J. R., ... McConnachie, A. (2017). Randomised, double-blind, placebo-controlled trials of non-individualised homeopathic treatment: Systematic review and meta-analysis. *Systematic Reviews*, *6*, 63.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Schimmack, U. (2016). The replicability-index: Quantifying statistical research integrity. Retrieved from https://replicationindex.wordpress.com/2016/01/31/a-revised-introduction-to-the-r-index/

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology, 15,* e2000797.

Authors:
Carmen López-Ibáñez [1], Desirée María Blázquez-Rincón [1], Julio Sánchez-Meca [1]
[1] University of Murcia

Title:
Reliability Generalization Meta-Analysis: A Comparison of Statistical Analytic Strategies

Session & Time:
Reliability Generalization Meta-Analysis (REGEMA). Thursday, May 30th, 6:00 pm - 6:30 pm

Abstract:
Background
An important psychometric property of the test is reliability which is defined as the scores' replicability. A common issue is to interpret it assuming that reliability is inherent to test instead of to consider reliability as a property of the sample data (Sánchez-Meca, López-Pina, & López-López, 2009; Sánchez-Meca, López-Pina, & López-López, 2012). The Reliability Generalization Meta-Analytical (hereafter RG) approach has proven to solve that question (Vacha-Haase, 1998). RG aims to analyze the variability of reliability coefficients in the different applications of a test, with the objective of investigating the extent to which reliability of a test scores can be generalized to different applications (Sánchez-Meca et al., 2012).
Specifically, an RG research comprises both the reliability coefficients found in different studies about the same test, and study characteristics of the study as predictors of variability of reliability coefficients (dependent variable) (Sánchez-Meca et al., 2012). Thus, one of the main objectives of the RG studies is to obtain an average reliability coefficient. Feldt and Charter (2006) presented six different procedures to obtain it. All of them can be applied as unweighted or weighted by the sample size, so we have twelve different procedures for averaging reliability coefficients (Sánchez-Meca et al., 2012).
The first one is to average of the alpha coefficients directly untransforming them. The second, Feldt and Charter (2006) defined it as the value that doubles the average of typical measurement errors. Third method consists in transforming it to Fisher's Z to obtain the weighted average and then transforming it back to alpha coefficients (assuming that the alpha value is equivalent to that obtained by parallel forms).
The fourth proposed by Hakstian and Whalen (1976), consists of transforming it to the cubic root, normalizing the distribution. In the fifth procedure, the reliability index is used, making the square root of the reliability coefficient. By last, the sixth method uses Fisher's Z transformation of the reliability index, and then it is transformed back again, as in procedure 3.

To prove the variations between the different methods, Sánchez-Meca et al. (2012) carried out a simulation study where they tested each procedure in its weighted and unweighted form, finding differences among them: regarding both the mean square error and the bias of the estimator, the methods that yielded better results were the

procedures 2 and 4. In addition, they also observed better results when the coefficients were weighted by sample size of the empirical studies than when the coefficients were unweighted.

Objectives

This study aims to determine whether these differences are also found when applying these procedures to real RG meta-analyses. In addition, we also included a seventh transformation proposed by Bonett (2002), which consists of calculating the natural logarithm of the supplementary coefficient. We hope to find differences among the different methods to pool reliability coefficients and their corresponding 95% confidence intervals (Sánchez-Meca, López-López, & López-Pina, 2013).

Method

To carry out this study, all RG meta-analyses, published or not, that reported the database with the individual reliability coefficients, were selected for this study. The search is being accomplished through Google Scholar and Scopus search engines. In addition, since the reliability coefficient most commonly reported by empirical studies is usually Cronbach's alpha, we focused on meta-analyses that reported this type of reliability. To compare the different results of the procedures, we established two comparison measures: the differences between the average alpha values obtained with the different procedures and the width of the confidence interval around the average reliability coefficient. The confidence intervals were calculated according to different models assumed: the fixed-effect (FE) model (Hedges & Olkin, 1985; Konstantopoulos & Hedges, 2009), the random-effects (RE) model (Hedges & Vevea, 1998; Raudenbush, 2009), the varying-coefficient (VC) model advocated by Bonett (2008, 2009, 2010) and the improved method proposed by Knapp and Hartung (2003) under the random-effects model.

Conclusion

In order to be the most comprehensive as possible, the search for the RG meta-analyses to be included in this study will finish on December 31st 2018. Once finished the literature search, the results of applying the different methods for averaging reliability coefficients and for constructing confidence intervals will be compared. Finally, the results will be discussed and recommendations will be made for meta-analysts that can be interested in conducting RG meta-analyses.

References

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*(4), 335-340.

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, *15*(4), 368-385.

Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement*, *66*(2), 215-227.

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika*, *41*(2), 219-231.

López-Pina, J. A., Sánchez-Meca, J., & López-López, J. A. (2012). Métodos para promediar coeficientes alfa en los estudios de generalización de la fiabilidad. *Psicothema, 24,* 161-166.

Sánchez-Meca, J., López-López, J.A. y López-Pina, J.A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology, 66*, 402-425.

Sánchez-Meca, J., López-Pina, J. A. y López, J. A. (2009). Generalización de la fiabilidad: un enfoque metaanalítico aplicado a la fiabilidad. *Fisioterapia, 31*(6), 262-270.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*(1), 6-20.

Authors:
José A López-López [1]

[1] University of Bristol

Title:
Using network meta-analysis to identify effective components of complex mental health interventions

Session & Time:
Network Meta-Analysis. Thursday, May 30th, 7:30 pm - 8:00 pm

Abstract:
Network meta-analysis (NMA) allows pooling evidence on multiple interventions from a set of randomised controlled trials (RCTs), each of which compare two or more of the interventions of interest. This feature enables to address relevant questions for practitioners and policy makers across many health areas including mental health. Interventions designed to prevent or treat mental health problems tend to be complex, in the sense that they may include several active ingredients or "components". If each combination of components is considered a separate intervention, then NMA could be used to simultaneously compare the different interventions. However, NMA requires that the comparisons made by the RCTs form a connected network, in other words that there is a path of comparisons between any two included interventions. This is unlikely to be the case with complex interventions, due to the large number of possible component combinations, and even if such a network is connected, the resulting analysis may lead to imprecise estimates.
Recently, component-level NMA regression methods have been developed within a Bayesian framework to allow estimation of the additive contribution of components and/or combinations of components of complex interventions while fully respecting the randomised structure of the evidence. This approach allows meaningful conclusions on effectiveness of components of complex interventions, whilst overcoming issues with connected networks and low precision with standard NMA. In this presentation, we will illustrate the use of standard and component-level NMA with two examples in the context of mental health interventions. In the first example, we compared the effectiveness of different types of therapy, different components and combinations of components and aspects of delivery used in cognitive-behavioural therapy (CBT) interventions for adult depression. We included 91 RCTs and found strong evidence that CBT interventions yielded a larger short-term decrease in depression scores compared to treatment-as-usual, with a standardised difference in mean change of -1.11 (95% credible interval -1.62 to -0.60) for face-to-face CBT, -1.06 (-2.05 to -0.08) for hybrid CBT, and -0.59 (-1.20 to 0.02) for multimedia CBT, whereas wait list control showed a detrimental effect of 0.72 (0.09 to 1.35). We found no evidence of specific effects of any content components or combinations of components, and importantly, we found that multimedia and hybrid

CBT might be as effective as face-to-face CBT, although results need to be interpreted cautiously.
The second application that we will discuss is an ongoing systematic review where the overall aim is to identify the most effective intervention component(s), or combination of components, for universal, selective, and indicated prevention of anxiety and depression problems in children and young people. We will present results based on NMA models both at the therapy and at the component levels. Last, we will conclude the presentation with a summary of the advantages of component-level NMA methods to explore the impact of different components of complex interventions on mental health outcomes, alongside the challenges that researchers might find when implementing this approach.

Authors:
Caroline Marker [1], Timo Gnambs [2], Markus Appel [1]

[1] University of Würzburg; [2] Leibniz Institute for Educational Trajectories

Title:

Meta-analytic evidence on the relationship between sedentary video gaming and body mass

Session & Time:

Health Psychology. Friday, May 31st, 3:00 pm – 3:30 pm

Abstract:

Background

Video gaming has been widely discussed as one leisure activity that is positively associated with body mass and overweight (e.g., Borland, 2011; Inchley, Currie, Jewell, Breda, & Barnekow, 2017; Mazur et al., 2018). Empirical findings on the popular form of non-active video games (i.e., games that are played while sitting in front of a screen, sedentary video games), however, have been mixed. While some studies found positive associations between the intensity of playing sedentary games and indicators of overweight, such as the body mass index (BMI; e.g., Martinovic et al., 2015; Siervo, Cameron, Wells, & Lara, 2014), others found no relationships (Bickham, Blood, Walls, Shrier, & Rich, 2013; Scharrer & Zeller, 2014).

Objectives and research questions

The current meta-analysis had two goals. First, we wanted to provide an estimate of the average effect size of the relationship between body mass and video gaming that includes recent research from the last one and a half decades, and we acknowledged several context variables to gauge the stability of the average effect. Second, to provide additional evidence on processes, we tested the displacement effect of physical activity by video gaming time with the help of a meta-analytic structural equation model (MASEM; Cheung & Hong, 2017).

Method

Meta-Analytic Database

Relevant studies published until June 2018 were identified through databases (PsychINFO, MEDLINE, ProQuest), gray literature (e.g., unpublished reports, conference proceedings, or theses); Google Scholar, and from the references of all relevant articles. This resulted in 753 potentially relevant studies.

The studies were included in the meta-analysis if they met the following criteria: The study contained (a) a measure of body mass (i.e., body mass index, body fat percentage, waist circumference, or subscapular skinfold thickness), (b) a measure of video game use (e.g., frequency or duration of video game sessions), (c) data on their zero-order relationship (or respective statistics that could be used to approximate this relationship), and (d) the sample size. After applying all eligibility criteria, 20 publications met our criteria and were included in the meta-analysis.

Meta-Analytic Procedure
The meta-analysis was conducted following the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, Moher et al., 2015) and standard procedures and recommendations for the social and medical sciences (Lipsey & Wilson, 2001).

The focal effects concerned the zero-order relationship between video gaming and body mass. For studies that did not report respective correlation coefficients, we extracted any relevant statistic (e.g., odds ratio) that could be transformed into correlation coefficients. Inter-coder reliability between the two coders for the coded effect sizes showed an excellent Krippendorff's (1970) α = 1.00. The effect sizes were pooled using a random effects model with a restricted maximum likelihood estimator (Viechtbauer, 2005). To account for sampling error, the effect sizes were weighted by the inverse of their variances. Because some studies reported multiple effect sizes for two or more eligible associations (e.g., scores for two video gaming measures were each correlated with BMI), theses dependencies were accounted for by fitting a three-level meta-analysis to the data (Moeyaert et al., 2017; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). Analyses of the heterogeneity as well as analyses of possible publication bias were conducted. The meta-analytic models were estimated in R version 3.5.0 using the metafor package version 2.0-0 (Viechtbauer, 2010).

Sensitivity analyses and structural equation model
Sensitivity analyses were conducted for (1) publication year, (2) age groups, (3) gender ratio in the sample, (4) a sample-wise estimate of gender differences in body mass, (5) body mass measure, (6) continuous vs. dichotomous body mass measures, and (7) a study quality index. A possible mediating effect of physical activity was examined using MASEM following two steps (see Cheung & Hong, 2017).

Results
Across k = 24 samples and 32 effect sizes (total N = 38,097), the mean effect (corrected for sampling error) of the relationship between video gaming and body mass was , 95% CI [.03, .14]. Hence, higher video gaming was positively associated with higher body mass. This relationship was significant, but there remained significant total heterogeneity, Q (31) = 593.03, p < .001, I^2 = 95.13. In the sensitivity analyses, we found a significant moderation for the age groups; the omnibus test for age was $\chi 2$ (df = 2) = 6.56, p = .038. Compared to adults, children and adolescents showed a significantly lower relationship between video gaming and body mass.

The estimated mediation model is presented in Figure 1. The relationship between body mass and physical activity was significant with B = -.07, 95% CI [-.14, -.00]. Higher physical activity was associated with lower body mass. The average relationship between video gaming and physical activity was only marginally

significant with B = -.08, 95% CI [-0.16, 0.00]. The respective indirect effect was significant B = .01, 95% CI [.00, .02]; it explained 7 percent of the total effect of video gaming on body mass. However, this result should be interpreted with caution because of the small sample of primary studies.

Figure 1. Meta-analytic structural equation model. Standardized regression parameters (*p < 05) are presented.

Conclusions and implications

This meta-analysis investigated the relationship between non-active (sedentary) video gaming and body mass, contributing to the research base on the behavioral correlates of overweight and obesity. We identified a small significant correlation between video gaming and body mass overall. This relationship was qualified by participants' age. The focal association was identified for adult samples, but there was no significant association for samples of children or adolescents. Based on a smaller subset of primary studies we found a small indirect effect on body mass, indicating a displacement of physical activity by video gaming. In summary, sedentary video gaming is only weakly associated with overweight and obesity, physical activity might play a mediating role, and the effects vary with participants' age.

References

Bickham, D. S., Blood, E. A., Walls, C. E., Shrier, L. A., & Rich, M. (2013). Characteristics of screen media use associated with higher BMI in young adolescents. *Pediatrics*, *131*, 935-941. doi: 10.1542/peds.2012-1197

Borland, S. (2011). Playing computer games increases obesity risk in teens by making them hungry. *Daily Mail.* Retrieved from: http://www.dailymail.co.uk/health/article-1389096/Playing-games-encourages-obesity-teens-making-hungry.html

Cheung, M. W. L., & Hong, R. Y. (2017). Applications of meta-analytic structural equation modeling in health psychology: Examples, issues, and recommendations. *Health Psychology Review, 11*, 265-279. doi:10.1080/17437199.2017.134

Inchley, J., Currie, D., Jewell, J., Breda, J., & Barnekow, V. (2017). Adolescent obesity and related behaviours: trends and inequalities in the WHO European Region, 2002–2014. *Observations from the Health Behaviour in School-aged Children (HBSC) WHO collaborative cross-national study. Copenhagen, Denmark: World Health Organisation*.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*, 61-70. doi:10.1177/001316447003000105

Martinovic, M., Belojevic, G., Evans, G. W., Lausevic, D., ... & Boljevic, J. (2015). Prevalence of and contributing factors for overweight and obesity among

Montenegrin schoolchildren. *The European Journal of Public Health*, *25*, 833-839. doi:10.1093/eurpub/ckv071

Mazur, A., Caroli, M., Radziewicz-Winnicki, I., ... & Hadjipanayis, A. (2018). Reviewing and addressing the link between mass media and the increase in obesity among European children. *Acta Paediatrica*, *107*, 568-576. doi: 10.1111/apa.14136

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*, 1. doi:10.1186/2046-4053-4-1

Moeyaert, M., Ugille, M., Beretvas, S. N., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: a comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology, 20*, 559-572. doi:10.1080/13645579.2016.1252189

Scharrer, E., & Zeller, A. (2014). Active and sedentary video game time: Testing associations with adolescents' BMI. *Journal of Media Psychology, 26,* 39-49. doi:0.1027/1864-1105/a000109

Siervo, M., Cameron, H., Wells, J. C., & Lara, J. (2014). Frequent video-game playing in young males is associated with central adiposity and high-sugar, low-fibre dietary consumption. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, *19*, 515-520. doi:10.1007/s40519-014-0128-1

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*, 576-594. doi:10.3758/s13428-012-0261-6

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30*, 261-293. doi:10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36,* 1-48. doi:10.18637/jss.v036.i03/

Authors:
Katja Matthias [1], Olesja Rissling [1], Marc Nocon [1], Anja Jacobs [1], Johannes Morche [1], Dawid Pieper [2], Uta Wegewitz [3], Robert Lorenz [1]

[1] Federal Joint Committee (Germany), Medical Consultancy Department; [2] Witten/Herdecke University; [3] Federal Institute for Occupational Safety and Health (Germany)

Title:

Appraisal of the methodological quality of systematic reviews on pharmacological and psychological interventions for major depression in adults using the AMSTAR 2.

Session & Time:
Quality Appraisal. Thursday, May 30th, 11:00 am - 11:30 am

Abstract:
Trial registration number
International Prospective Register of Systematic Reviews (PROSPERO) registration number: CRD42018110214.

Background
Major depression is a common mental disorder with high prevalence and mortality. There is a high need for reliable and summarized information for clinicians as well as policy makers in the field. Whereas systematic reviews should provide a comprehensive and objective appraisal of evidence, poor reporting and flaws in methodological quality are often and impair the reliability of conclusions. Several standards have been developed to assess methodological quality of SR [2], widely used is the AMSTAR (A Measurement Tool to Assess SR, published in 2007) with 11 items. Recently, an updated version of AMSTAR - AMSTAR 2 [1] has been published, which allows a more detailed evaluation of SR in 16 items and the rating of the overall confidence in the results of the review.

Objectives
The present study will determine the methodological quality of SR in the treatment of adult major depression using the new AMSTAR 2 and identify potential predictive factors associated with the quality. To reflect the current quality we focus on SR published in the years of 2012 to 2017.

Methods
We conducted electronic searches in August 2017 in the bibliographic databases MEDLINE, EMBASE and the Cochrane Database of SR. We used a combination of Mesh terms and keywords to identify SR from 2012 to 2017 referring to the topic “Major Depression” and did not apply any restrictions on language or countries. Two authors independently screened the titles, abstracts and full texts of the retrieved literature to assess their eligibility according the a priori defined criteria and coded the bibliographic characteristics (e.g. corresponding author’s original region, number

of authors, Journal impact factor at year of publication) onto a data collection template in EXCEL.
All selected SR were appraised after a calibration phase with AMSTAR 2 by four independent evaluators. Each evaluator appraised 30 SR. A consensus for each of the 16 items was reached with majority rule. Furthermore, the rating of the overall confidence in the results of the review was performed with the critical domains as recommended by Shea et al. 2017 [1]. This was done by two evaluators independently. Any discrepancies were resolved through discussions.
To assess whether the intervention (pharmacological or psychological interventions), the type of review (Cochrane vs. non-Cochrane reviews), and/or Open Access status (yes vs. no) are associated with AMSTAR 2 scores, a sub-analysis of AMSTAR 2 scores will be performed. The associations between bibliographical characteristics and scoring on AMSTAR 2 items will be analysed using multivariate logistic regression or multi-nominal logistic regression depending on the scaling of the dependent variable.

Results

The electronic literature search detected 1,524 citations. 72 SR comprising 30 SRs with psychological and 42 SRs with pharmacological interventions met our eligibility criteria. 30 out of 42 pharmacological SRs were randomly drawn and served together with the identified 30 psychotherapeutic SRs as sample for this study.
Of the 60 SR evaluated in our sample, 42 SR included only randomized trials and 18 SR additionally non-randomized studies. Four out of the 60 SR were Cochrane Reviews. In four domains of AMSTAR 2 (item 1, 6, 14, 16) the majority (more than 50%) of the SR scores “yes”. The results according to all AMSTAR 2 domains are shown in figure 1.

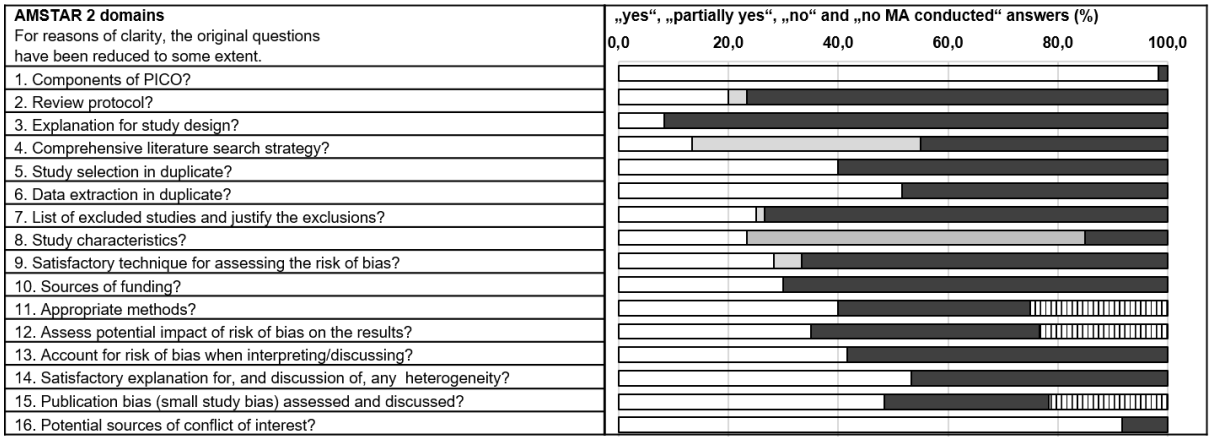


Figure 1: Methodological quality of 60 SR according to the 16 items of AMSTAR 2 yes: white colour, partially yes: light grey, no: dark grey, no meta-analysis (MA) conducted: striped

In rating overall confidence in the results of the SR only four reviews were considered as “high” (three of them Cochrane Reviews), two as “moderate”, one as “low” and 53 as “critically low”.

The analysis of subgroups and the evaluation of the predictors is currently in progress.

Conclusions and implications

According to AMSTAR 2 the overall methodological quality of our current and representative sample of SR on pharmacological and psychological interventions for major depression in adults is disappointing. In almost 90% of the sample of SR, overall confidence in the results of the SR was considered to be “critically low”, so the use of findings of these reviews should be limited.

Although there is a high need for reliable and summarized information for clinicians as well as policy makers in the field of mental disorders, this work demonstrates the need to critically assess SR before using their findings.

Possible suggestions for the improvement of the quality of SR are the following: Authors of future SR should establish review methods prior to the conduct of the review, give an explanation for study designs included in the review, use a satisfactory technique for assessing the risk of bias as well as publication bias, use appropriate meta-analytic methods, and consider the scientific quality when formulating conclusions.

References

[1] Shea et al. (2017). BMJ, 358, j4008.
[2] Zeng et al. (2015). J of Evidence-Based Medicine, 8: 2-10

Authors:
Sameh Said Metwaly [1], Belén Fernández-Castilla [1], Eva Kyndt [1,2], Wim Van den Noortgate [1], Baptiste Barbot [3,4]

[1] KU Leuven; [2] University of Antwerp; [3] Pace University; [4] Yale University

Title:

Developmental Trend of School-Age students' Divergent thinking: A Meta-analysis

Session & Time:

Applications. Friday, May 31st, 2:00 pm – 2:30 pm

Abstract:

Background

Over the past decades, there has been a great deal of research on the development of school-age students' divergent thinking. However, research findings regarding this issue have been inconsistent. Some studies have provided evidence for a continuous development of divergent thinking as grade level increases. Other studies have suggested a discontinuous developmental trajectory including one or more periods of significant drops. Torrance (1967) found in seven different cultures that a drop occurs in Grade 4, which has become widely known as the fourth grade slump in divergent thinking. The existence of the fourth grade slump has been reported in several subsequent studies. On the contrary, other studies have found no evidence of the fourth grade slump; some studies have found an increase or no decline in Grade 4, and other studies have found a slump but in other grades including Grades 1, 6, 7, and 9. In addition to the inconsistent results, most of the previous studies have been conducted on a small number of subjects and a limited grade range. Hence, the picture is less clear concerning whether divergent thinking slumps actually exist, how many there are, and when they occur.

Objectives

The purpose of this study was to meta-analyze previous research results regarding the development of school-age students' divergent thinking from Grades 1 to 12, with a particular focus on the fourth grade slump as it has sparked a major debate among researchers. We also examined whether the change in divergent thinking is affected by divergent thinking test, divergent thinking domain, intellectual ability, gender, and country of study.

Research questions

This study attempts to answer the following questions: (1) How does school-age students' divergent thinking change from Grades 1 to 12? (2) Does the fourth grade slump in divergent thinking exist? (3) Are there moderator variables that account for the observed variability across studies concerning the change in divergent thinking from Grade 3 to 4?

Method

We calculated for each study a standardized mean per grade, and combined these standardized means in a meta-analysis. A meta-analytic three-level model was employed in order to account for dependence within studies. To examine divergent thinking changes from Grades 1 to 12, we included 11 (number of grades - 1) dummy variables as predictors in the meta-analytic model. The first dummy variable takes the value 0 in the case of Grade 1 and 1 otherwise, and the second dummy variable takes the value 0 in the case of Grades 1 and 2 and 1 otherwise. Other dummy variables were coded using the same procedure, until the 11th dummy variable which is equal to 1 in the case of Grade 12 and 0 for the previous grades. In this way, the coefficient of the first dummy variable captures divergent thinking change from Grade 1 to 2, the second coefficient captures divergent thinking change from Grade 2 to 3, and so on. To avoid an excessively complicated model, the effects of the moderator variables were investigated only for divergent thinking change from Grade 3 to 4. To study the influence of each of the moderator variables, we included an additional term in the model, which represents the interaction between the dummy variable capturing divergent thinking change from Grade 3 to 4 and the suggested moderator variable.

Data sources

The present meta-analysis included divergent thinking literature published up to December 31st, 2017. The search process consisted of the following four steps: First, the following databases were searched: ERIC, Google Scholar, JSTOR, PsycARTICLES, Scopus, and Web of Science. Second, the reference lists of the papers identified in the first step were reviewed for other relevant references (i.e. “backward search”). Third, more recent references were retrieved by searching databases for papers that referred to the previously identified papers in steps 1 and 2 in their citations (i.e. “forward search”). Fourth, the relevant key journals were hand-searched. The papers identified using the search process were first screened for their relevance on the basis of their titles and abstracts. The remaining papers were included if they met the following two criteria: (1) reports on an original, empirical, and quantitative study, and (2) examines differences in divergent thinking between Grade 4 and other Grades (1-12). Moreover, we only included (1) journal articles, conference papers, or dissertations (2) that were written in English, and for which (3) the full text was available.

Results

A total of 2,139 standardized means from 41 studies were analyzed. Overall, the results showed an upward trend of divergent thinking across grades; however, there were some discontinuities. Also, there was no evidence of the fourth grade slump; instead a seventh grade slump was noted at both the overall and subscale (i.e., fluency, flexibility, and originality) levels (see Figure 1). Task domain significantly moderated the change of the overall divergent thinking in Grade 4. At the subscale level, intellectual ability moderated the change of fluency, as well as country of study moderated the change of originality in Grade 4.

Conclusions and implications

The results of this study inform the ongoing debate concerning the development of school-age students' divergent thinking. Furthermore, these results suggest a slump in divergent thinking in Grade 7. This might have valuable implications for parents, teachers, and other professionals working with students and could be used to guide interventions and training programs to promote divergent thinking development. Finally, our study revealed different developmental trends of divergent thinking in terms of task domain, intellectual ability and country of study. Hence, these factors need to be considered carefully when investigating divergent thinking development.

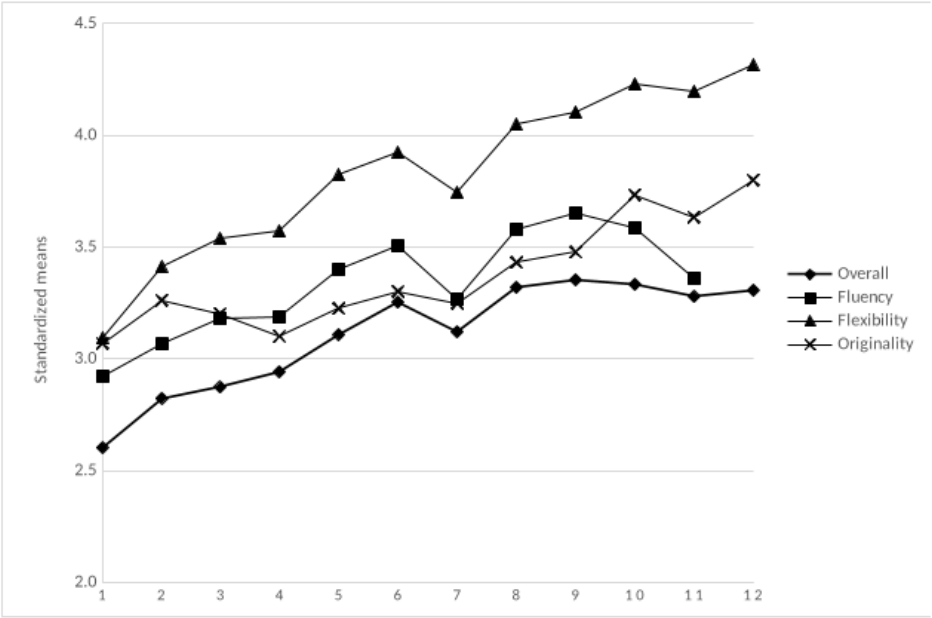


Figure 1. Developmental trends of divergent thinking by grade

Authors:
María Rubio-Aparicio [1], Julio Sánchez-Meca [2], Rosa Mª Núñez-Núñez [3], José Antonio López-Pina [2], Fulgencio Marín-Martínez [2], José Antonio López-López [4]

[1] University of Alicante; [2] University of Murcia; [3] University Miguel Hernández de Elche; [4] University of Bristol

Title:

Reliability Generalization Meta-Analysis of the Padua Inventory-Revised (PI-R)

Session & Time:

Reliability Generalization Meta-Analysis (REGEMA). Friday, May 30th, 6:30 pm - 7:00 pm

Abstract:

Background

Obsessive–compulsive disorder (OCD) is a mental disorder characterized by the presence of obsessions, compulsions, or both. The Padua Inventory (PI) of Sanavio is one of the measurement instruments most widely used to assess obsessive-compulsive symptoms (Sanavio, 1988). A number of shorter versions of the PI can also be found in the literature. This is the case of the Padua Inventory Revised (PI-R) developed by Van Oppen, Hoekstra, and Emmelkamp (1995), which consists of 41 items and five subscales adapted to Dutch language: Impulses (7 items), Washing (10 items), Checking (7 items), Rumination (11 items) and Precision (6 items). Higher scores indicate greater severity of obsessive–compulsive symptoms. Reliability of psychological tests depends on the composition and characteristics of the samples of participants and the application context. Since reliability varies in each test administration, meta-analysis is a suitable method to statistically integrate the reliability estimates obtained in different applications of a test. Vacha-Haase (1998) coined the term *reliability generalization* (RG) to refer to this type of meta-analysis.

Objectives

An RG meta-analysis of the empirical studies that applied the PI-R (Van Oppen et al. 1995) was carried out in order to: (a) estimate the average reliability (for the total scale and subscales); (b) examine the variability among the reliability estimates; and (c) search for characteristics of the studies (moderators) that can be statistically associated to the reliability coefficients.

Method

To be included in the meta-analysis, each study had to fulfil the following criteria: (a) to be an empirical study where the PI-R, or an adaptation maintaining the 41 items, was applied to a sample of at least 10 participants; (b) to report any reliability estimate based on the study-specific sample; (c) the paper had to be written in English or Spanish; (d) samples of participant from any target population were accepted (community, clinical of subclinical populations); and (e) the paper might be published or unpublished.

The search period of relevant studies covered from 1988 to September 2017 inclusive. The following databases were consulted: PROQUEST, PUBMED, and

Google Scholar. In the electronic searches, the keywords “Padua Inventory” were used to be found in the full-text of the documents.
Internal consistency was the type of reliability investigated in this RG meta-analysis, so that alpha coefficients reported in the primary studies were extracted. A random-effects model was assumed estimating the between-studies variance by restricted maximum likelihood (López-López, Botella, Sánchez-Meca, & Marín-Martínez, 2013; Sánchez-Meca, López-López, & López-Pina, 2013). The 95% confidence interval around each overall reliability estimate was computed with the improved method proposed by Hartung (1999). All statistical analyses were carried out with the *metafor* package in *R* (Viechtbauer, 2010).

Results
The search yielded a total of 1,335 references, out of which 1,234 were removed for different reasons. The remaining 101 references were empirical studies that had applied the PI-R and out of them, 24 were included in the meta-analysis.
The 24 estimates reported for the total scale yielded a mean coefficient alpha of .926 (95%CI: .913 and .937), ranging from .830 to .960. Subscales exhibited lower mean reliability coefficients than that of the total scale, with Washing yielding the largest estimates (mean = .889; 95%CI: .853 and .916), followed by Checking (mean = .879; 95%CI: .862 and .894), and Rumination (mean = .870; 95%CI: .845 and .890). Impulses (mean = .793; 95%CI: .762 and .820) and Precision (mean = .727; 95%CI: .678 and .768) were the subscales with the poorest average reliabilities.
Alpha coefficients presented a large heterogeneity, with *I2 Indices* over 80% in all cases. The large variability exhibited by the reliability coefficients obtained in different applications of the PI-R was investigated by analyzing the influence of potential moderator variables. Concretely, the standard deviation of test scores exhibited a statistically significant relationship with coefficient alpha and with a percentage of variance accounted for of 33%. In particular, this predictor exhibited a positive relationship with alpha coefficients, so that larger coefficients alpha were obtained as the standard deviation of the scores increased. Furthermore, statistically significant differences were found when comparing the mean alpha coefficients grouped by the test version (p = .034), with a 36% of variance of variance explained, the mean reliability being larger for Turkish studies.

Conclusions
Several guidelines have been proposed in the psychometric literature to assess the adequacy and relevance of reliability coefficients. In general, it is accepted that coefficients alpha must be over .70 for exploratory research, over .80 for general research purposes, and over .90 when the test is used for taking clinical decisions (Nunnally & Bernstein, 1994). Based on these guidelines, our findings demonstrated the good reliability of the PI-R total scores, both for screening and clinical purposes. The results also demonstrate how reliability depends on the application context and the composition and variability of the samples. In particular, as expected form psychometric theory, a strong positive relationship was found with the standard

deviation of test scores. Another characteristics of the studies that exhibited a statistical relationship with alpha coefficients was the test version.

References

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, *41*, 901-916.

López-López, J. A., Botella, J., Sánchez-Meca, & Marín-Martínez, F. (2013). Alternatives for mixed-effects meta-regression models in the reliability generalization approach: A simulation study. *Journal of Educational and Behavioral Statistics*, *38*, 443-469.

Nunnally J. C., & Bernstein I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Sanavio E. (1988). Obsessions and compulsions: The Padua Inventory. *Behaviour Research and Therapy*, *26*, 169–177.

Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization (RG) studies are conducted. *British Journal of Mathematical and Statistical Psychology*, *66*, 402-425.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20.

Van Oppen, P., Hoekstra, R.J., & Emmelkamp, P.M.G. (1995). The structure of obsessive-compulsive symptoms. *Behaviour Research and Therapy*, *33*, 15-23.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metaphor package. *Journal of Statistical Software*, *36*, 1–48.

Authors:
Julio Sánchez-Meca [1], José Antonio López-Pina [1], María Rubio-Aparicio [2], Fulgencio Marín-Martínez [1], Rosa Mª Núñez-Núñez [3], Juan José López-García [1], José Antonio López-López [4]

[1] University of Murcia; [2] University of Alicante; [3] University Miguel Hernández de Elche; [3] University of Bristol

Title:
REGEMA: Guidelines for Conducting and Reporting Reliability Generalization Meta-analyses

Session & Time:
Reliability Generalization Meta-Analysis (REGEMA). Friday, May 30th, 5:30 pm - 6:00 pm

Abstract:
Background
Reliability is one of the most important properties to assess psychometric quality of psychological measurement instruments. There is a mistaken idea, very extended among researchers, that reliability is an immutable property of a measurement instrument. However, reliability is not a property inherent to the test, but of the scores obtained when the test is applied to a given sample of participants in specific conditions (Gronlund & Linn, 1990). Inducing reliability from previous applications of a test is a phenomenon very extended among researchers that is appropriate only if the previous and the current study have samples of participants similar in composition and variability (Vacha-Haase et al., 2000). As it is very infrequent that studies use similar samples, then reliability induction becomes a malpractice that must be dismissed from research.

Fortunately, not all of the primary studies induce reliability from previous studies, but they report reliability coefficients with their own sample. If reliability varies from an application of a test to the next, then meta-analysis becomes a very useful methodology to statistically integrate the reliability estimates. With this purpose, Vacha-Haase (1998) coined the term ‘reliability generalization’ (RG) to refer to this kind of meta-analysis. An RG meta-analysis aims to investigate how measurement error of a test scores varies among different contexts, samples, and target populations. In particular, an RG meta-analysis enables: (a) to estimate the average reliability of a test scores, (b) to assess whether reliability coefficients are heterogeneous and, (c) in case of heterogeneity, to find characteristics of the studies that can explain, at least, part of the variability of the reliability coefficients (Henson & Thompson, 2002; Sánchez-Meca et al., 2013; Vacha-Haase et al., 2002).

From its inception in 1998, more than 120 RG meta-analyses have been published in psychology. This kind of meta-analysis presents distinctive characteristics that make it different in some aspects from typical meta-analyses to integrate effect sizes. In an RG meta-analysis the ‘effect size’ are the reliability coefficients reported in the primary studies. This circumstance makes that the typical guidelines proposed in the

meta-analytic arena for reporting meta-analyses does not adapt well to RG meta-analyses. Such guidelines as PRISMA (Moher et al., 2009), MARS (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008), AMSTAR-2 (Shea et al., 2017), MOOSE (Stroup et al., 2000), or the recent recommendations of the American Psychological Association (Appelbaum et al., 2018) include items that are not applicable to RG meta-analyses, and do not contain important items to be considered in RG meta-analyses.

Objectives

Up our knowledge, there have not been proposed specific guidelines for conducting and reporting RG meta-analyses that take into account their special features. Therefore, the purpose of this investigation was to elaborate a checklist specifically devised to help meta-analysts to conduct and report RG meta-analyses. The name for this checklist is REGEMA (**RE**liability **GE**neralization **M**eta-**A**nalysis).

Method

A first step consisted in a sound review of the items and criteria included in the most usually applied guidelines for systematic reviews and meta-analyses proposed in the meta-analytic literature: PRISMA, MARS, AMSTAR-2, and MOOSE. Based on this review, a second step consisted in elaborating a set of items or criteria that might be useful for REGEMA checklist. With this purpose, brainstorming meetings were held among the members of the Meta-analysis Unit team (University of Murcia) to obtain a first version of REGEMA checklist. Once elaborated a tentative REGEMA checklist, the third step consisted in sending the list to 30 researchers experts in meta-analysis. The criteria for selecting the researchers were: (a) to have large expertise in the methodology of meta-analysis, and/or (b) to have published several RG meta-analyses in psychology. Once received the comments, suggestions, and criticisms of the experts, the final step consisted in elaborating the definitive REGEMA checklist.

Results

The revision of PRISMA, MARS, AMSTAR-2, and MOOSE guidelines confirmed that none of them adapted well to be applied to RG meta-analyses. Once revised the items and criteria included in these guidelines, our research team carried out more than 20 brainstorming meetings to elaborate a first version of REGEMA checklist composed by 30 items. The tentative REGEMA checklist was electronically sent to 30 researchers with expertise in meta-analysis in order to obtain feedback on the adequacy of the checklist. Out of them, 12 experts answered and their interesting and useful comments and suggestions were added to the checklist. Finally, the REGEMA checklist was composed by 29 items structured shown in Table 1: one item for the Title, one for the Abstract, two for the Introduction, 14 for the Method, six for the Results, four for the Discussion, and one for Funding.

Table 1. REGEMA checklist.

Cluster
Item
Title/Abstract
1. Title
2. Abstract
Introduction
3. Background
4. Objectives

Method
5. Selection criteria
6. Search strategies
7. Data extraction
8. Reported reliability
9. Type of reliability induction
10. Data extraction of inducing studies
11. Reliability of data extraction
12. Transformation method
13. Statistical model
14. Weighting method
15. Heterogeneity assessment
16. Moderator analyses
17. Additional analyses
18. Software

Results
19. Results of the study selection process
20. Mean reliability and heterogeneity
21. Moderator analyses
22. Sensitivity analyses
23. Comparison of inducing and reporting studies
24. Data set

Discussion
25. Summary of results
26. Limitations
27. Implications for practice
28. Implications for future research
Funding
29. Funding

Conclusions

In order to bridging a gap in the meta-analytic literature, we have elaborated the REGEMA checklist, a list of guidelines for conducting and reporting RG meta-analyses that is adapted to the special characteristics of this kind of meta-analysis.

Based on the experience of Meta-analysis Unit's research team carrying out meta-analyses for more than 30 years, the REGEMA checklist have good construct validity. Future research must to assess its inter-coder reliability by applying it to RG meta-analyses already published. REGEMA checklist can be useful for meta-analysts interested in conducting RG meta-analysis, as well as for readers of these meta-analyses and even for editors of journals that may use it to assess the reporting quality of RG meta-analyses sent to publish.

References

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*, 839-851.

Appelbaum, M., Cooper, H., Kline, R.B., Mayo-Wilson, E., Nezu, A.M., & Rao, S.M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force report. *American Psychologist, 73*, 3-25.

Gronlund, N.E. y Linn, R.L. (1990). *Measurement and evaluation in teaching* (6ª ed.). Nueva York: Macmillan.

Henson, R.K. y Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-126.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology, 62*, 1006-1012.

Sánchez-Meca, J., López-López, J.A. y López-Pina, J.A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology, 66*, 402-425.

Shea, B.J., Reeves, B.C., Wells, G., Thuku, M., Hamel, C., Moran, J., …, & Henry, D.A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. British Medical Journal, 358:j4008. http://dx.doi.org/10.1136/bmj.j4008.

Stroup, D.F., Berlin, J.A., Morton, S.C., Olkin, I., Williamson, G.D., et al. (2000). *Journal of the American Medical Association, 283*, 2008-2012.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, 58, 6-20.

Vacha-Haase, T., Henson, R.K. y Caruso, J.C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62*, 562-569.

Vacha-Haase, T., Kogan, L.R. y Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals. *Educational and Psychological Measurement, 60*, 509-522.

Authors:
Lukasz Stasielowicz [1], Reinhard Suck [1]

[1] Osnabrück University

Title:

Distance correlation: Discovering meta-analytic relationships between variables when other correlation coefficients fail

Session & Time:
Methods in Meta-Analysis. Wednesday, May 29th, 4:30 pm - 5:00 pm

Abstract:

Background

Many meta-analysts use correlation coefficients in order to assess the strength of the relationship between selected variables across the studies. Usually the Pearson product-moment correlation is chosen. After all, it is implemented in most of the meta-analytic packages (Polanin, Hennessy, & Tanner-Smith, 2017). Furthermore, it is relatively easy to interpret as it ranges from -1 to 1 and researchers have proposed benchmarks to facilitate fast assessment of the practical relevance of the findings based on Pearson correlations (Bosco, Aguinis, Singh, Field, & Pierce, 2015; Gignac & Szodorai, 2016).

Notwithstanding the advantages it has to be noted that the Pearson correlation has several limitations, which need to be considered by people conducting meta-analyses, i.e. outliers can lead to biased estimates of the correlations. Furthermore, not every type of bivariate relationship can be discovered when utilizing Pearson correlations. Specifically, only linear relationships can be detected. This can be problematic, because it can lead to false conclusions when non-linear rather than linear relationships are present. To illustrate, it is well known that certain types of cognitive abilities – i.e. processing speed, memory (Li et al., 2004) – improve during the childhood and decline during the (late) adulthood. Due to the inverted-U relationship between age and cognitive abilities the value of the Pearson correlation will be close to zero, implying that there is no linear relationship. Unfortunately, people may be inclined to think that lack of linear relationship means that there is no relationship whatsoever, which in turn may lead to abandoning fruitful research questions. Although alternative well-established correlation coefficients are available (e.g. Kendall's tau, Spearman's rho) they are not adequate when assessing non-monotonic relationships. However, recently other measures of dependence emerged - i.e. distance correlation (Rizzo & Székely, 2016; Székely, Rizzo, & Bakirov, 2007) – which are not restricted to monotonic relationships. In contrast to the previously mentioned correlation coefficients the distance correlation ranges from 0 to 1. A value of zero implies lack of dependence.

Objectives

Although it has been suggested that distance correlations could be used in the meta-analytic context (Székely et al., 2007) to gauge the strength of the relationship

between variables such attempts were not undertaken. Thus, the main objective of the present study was to compare distance correlation to other correlation coefficients (Pearson correlation, Kendall's tau, Spearman's rho) by conducting separate meta-analyses for each effect size.

Research questions
We hypothesized that only by using the distance correlation one will be able to consistently detect meta-analytic dependence between the variables across several scenarios (e.g. linear relationship, non-linear monotonic relationship, non-linear non-monotonic relationship). In contrast, Kendall's tau and Spearman's rho will fail in the non-monotonic scenario and the Pearson correlation will fail even in the non-linear scenario.

Method
For each scenario (i.e. non-linear monotonic relationship) many samples of participants were simulated in order to mimic the meta-analytic procedure of reviewing different studies. Distance correlation, Pearson correlation, Kendall's tau and Spearman's rho were computed for each sample. Subsequently the mean effect size across the samples was calculated separately for each type of correlation coefficient. Finally, the respective mean effect sizes were compared.
The analyses were conducted using several R packages. The distance correlation was computed using the *energy* package. In order to compute the meta-analytic weights of each sample the variance of the distance correlation estimate was calculated by applying the jackknife technique within each sample (*bootstrap*package). The respective random-effect meta-analyses (REML estimator) were carried out using the *metafor* package.

Results
In general, the expected pattern of results could be confirmed. To illustrate, an inverted-U relationship y = -x*x, which could reflect the relationship between age and cognitive abilities, led to the following meta-analytic correlation estimates (k = 40, N = 2000): .01 (Pearson correlation), .03 (Kendall's tau), .02 (Spearman's rho), .33 (distance correlation). The reproducible R code will be made available upon publication.

Conclusions
Among the considered correlation coefficients only distance correlation could consistently yield evidence for the existing relationship between two variables (i.e. age and cognitive abilities). Thus, it could be fruitful to utilize distance correlations as the effect size in future meta-analyses. It would reduce the risk of wrongly concluding that there is no relationship when a non-linear non-monotonic relationship is present. Providing the evidence for usefulness of distance correlations in the meta-analytic context is the main contribution of the current study.
One important drawback that could stymie meta-analytic research based on distance correlations pertains to the fact that distance correlations cannot be derived from

other correlation coefficients. Thus, meta-analysts cannot compute it by utilizing summary statistics reported in relevant studies. Instead they need the access to raw data. However, considering the advances made by the open science movement (e.g. data repositories) it seems plausible to assume that in future meta-analyses the access to raw data stemming from new studies will be granted. Even nowadays small meta-analyses based on distance correlations could be feasible thanks to replication initiatives or multi-lab studies where several laboratories examine the same research question, conduct a mini meta-analysis and make their raw data available.

Nevertheless, further work on the use of dependence measures in meta-analyses is needed. In future studies one could try to examine the meta-analytic performance of distance correlations within the Bayesian framework (Bhattacharjee, 2014). Furthermore, one could simulate meta-analyses based on alternative measures of dependence within both the frequentist and Bayesian framework, e.g. Maximum Information Coefficient or Heller Heller Gorfine measure (de Siqueira Santos, Takahashi, Nakata, & Fujita, 2014).

References

Bhattacharjee, A. (2014). Distance correlation coefficient: An application with bayesian approach in clinical data analysis. *Journal of Modern Applied Statistical Methods*, *13*(1), 354–366. http://doi.org/10.22237/jmasm/1398918120

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. http://doi.org/10.1037/a0038047

de Siqueira Santos, S., Takahashi, D. Y., Nakata, A., & Fujita, A. (2014). A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in Bioinformatics*, *15*(6), 906–918. http://doi.org/10.1093/bib/bbt051

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. http://doi.org/10.1016/j.paid.2016.06.069

Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science*, *15*(3), 155–163. http://doi.org/10.1111/j.0956-7976.2004.01503003.x

Polanin, J. R., Hennessy, E. A., & Tanner-Smith, E. E. (2017). A review of meta-analysis packages in R. *Journal of Educational and Behavioral Statistics*, *42*(2), 206–242. http://doi.org/10.3102/1076998616674315

Rizzo, M. L., & Székely, G. J. (2016). Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, *8*(1), 27–38. http://doi.org/10.1002/wics.1375

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, *35*(6), 2769–2794. http://doi.org/10.1214/009053607000000505

Authors:
Sho Tsuji [1], Alejandrina Cristia [1], Michael C. Frank [2], Christina Bergmann [3]
[1] École Normale Superieure; [2] Stanford University; [3] Max Planck Institute for Psycholinguistics

Title:
Addressing publication bias in meta-analysis: Empirical findings from community-augmented meta-analyses of infant language development

Session & Time:
Methods in Meta-Analysis. Wednesday, May 29th, 4:00 pm - 4:30 pm

Abstract:
Meta-analyses have long been an indispensable research synthesis tool for characterizing bodies of literature and advancing theories. However, they have been facing the same challenges as primary literature in the context of the replication crisis: A meta-analysis is only as good as the data it contains,and which data end up in the final sample can be influenced at various stages of the process. Early on, the selection of topic and search strategies might be biased by the meta-analyst's subjective decision. Further,publication bias towards significant outcomes in primary studies might skew the search outcome, wheregrey, unpublished literature might not show up. Additional challenges might arise during data extraction from articles in the final search sample, for example since some articles might not contain sufficient detail for computing effect sizes and correctly characterizing moderator variables, or due to specific decisions of the meta-analyst during data extraction from multi-experiment papers.Community-augmented meta-analyses (CAMAs, Tsuji, Bergmann, & Cristia, 2014) have received increasing interest as a tool for countering the above-mentioned problems. CAMAs are open-access, online meta-analyses. In the original proposal, they allow the use and addition of data points by the research community, enabling to collectively shape the scope of a meta-analysis and encouraging the submission of unpublished or inaccessible data points. As such, CAMAs can counter biases introduced by data (in)availability and by the researcher. In addition, their dynamic nature serves to keep a meta-analysis, otherwise crystallized at the time of publication and quickly outdated, up to date.We have now been implementing CAMAs over the past four years in MetaLab(metalab.stanford.edu), a database gathering meta-analyses in Developmental Psychology and focused on infancy. Meta-analyses are updated through centralized, active curation.We here describe our successes and failures with gathering missing data, as well as quantify how the addition of these data points changes the outcomes of meta-analyses. First, we ask which strategies to counter publication bias are fruitful. To answer this question we evaluate efforts to gather data not readily accessible by database searches, which applies both to unpublished literature and to data not reported in published articles. Based on this investigation, we conclude that classical tools like database and citation searches can already contribute an important amount of grey literature. Furthermore, directly contacting authors is a fruitful way to get access to missing information. We then address

whether and how including or excluding grey literature from a selection of meta-analyses impacts results, both in terms of indices of publication bias and in terms of main meta-analytic outcomes. Here, we find no differences in funnel plot asymmetry, but (as could be expected) a decrease in meta-analytic effect sizes. Based on these experiences, we finish with lessons learned and recommendations that can be generalized for meta-analysts beyond the field of infant research in order to get the most out of the CAMA framework and to gather maximally unbiased dataset.

Authors:
Ulrich S. Tran [1], Matthias A. Burzler [1], Ulrich J. C. Hegewisch [1], Martin Voracek [1]
[1] University of Vienna

Title:
No specificity of psychometric mindfulness in accounting for the effects of mindfulness interventions on mental health: A systematic review and three-level meta-analysis of randomized controlled trials

Session & Time:
Health Psychology. Friday, May 31st, 4:00 pm – 4:30 pm

Abstract:
Background / Objectives / Research Questions
Mindfulness is a much-investigated topic in clinical psychology and intervention research. A large body of studies shows that mindfulness interventions, such as mindfulness-based stress reduction (MBSR; Kabat-Zinn, 1982) and mindfulness-based cognitive therapy (MBCT; Segal, Williams, & Teasdale, 2002), have positive effects on mental health in various psychiatric and non-psychiatric populations, and also with regards to chronic medical disease (e.g., Bohlmeijer, Prenger, Taal, & Cuijpers, 2010; Goldberg et al., 2018).
However, much less is known about the mechanisms of action of mindfulness interventions. There is a scarcity of empirical data and a lack of methodological rigour in the field of testing mediators and mechanisms of action (Alsubaie et al., 2017). As of yet, there is only one published formal meta-analytic account (Gu et al., 2015). Using two-stage meta-analytic structural equation modeling (TSSEM; Cheung & Chan, 2005), Gu et al. (2015) reported that increases in psychometric mindfulness (i.e., psychometrically assessed trait or dispositional mindfulness) and decreases in repetitive negative thinking mediate the effects of MBSR and MBCT on mental health in controlled treatment studies (both RCT and non-RCT) on the meta-analytic level. For psychometric mindfulness, the analysis of Gu et al. (2015) in essence showed that (1) treatment and control groups differed in their average change of psychometric mindfulness, and (2) that this difference accounted for (i.e., mediated) part of the difference in the change of mental health between treatment and control groups (results were similar for repetitive negative thinking).
While the study of Gu et al. (2015) was an important step, there still are a number of key methodological shortcomings. First, TSSEM requires the correlations between treatment and control groups with changes in the mediator, and changes in the outcome, to be known (i.e., these data need to be reported in primary studies). In practice, this is a very restrictive requirement – accordingly, the study sample of Gu et al. (2015) was limited to merely $k = 12$ (psychometric mindfulness) and $k = 6$ studies (repetitive negative thinking). Second, there are clear methodological advantages in using RCTs, rather than non-RCTs, in mediation analysis. However, the study sample meta-analyzed by Gu et al. (2015) confounded RCTs with non-RCTs. Third, there is a lack of empirical data on the specificity of investigated

mediators (Alsubaie et al., 2017); i.e., it is still not clear, whether mediators are specific for some type of mindfulness intervention (e.g., MBSR), patient populations (e.g., adults with recurrent depression), or outcomes (e.g., anxiety). Psychometric mindfulness is one of the most often investigated mediators and appears to be a universal (i.e., nonspecific, common) mediator (Alsubaie et al., 2017).
While this last result appears to be promising, the validity of psychometric mindfulness has been called into question (Van Dam et al., 2018). There is an absence of a normative definition of the mindfulness construct (Van Dam et al., 2018) and, what is more, evidence of trait overlap of psychometric mindfulness with neuroticism and negative effect (for a meta-analysis, see Giluk, 2009) and of unexpected increases of psychometric mindfulness in non-mindfulness treatments (e.g., Goldberg et al., 2015). Despite consistent evidence that increases of psychometric mindfulness are correlated with increases of mental health in intervention studies (e.g., Khoury et al., 2013), it remains an open question whether changes in psychometric mindfulness truly represent a specific mechanism of action of mindfulness interventions or rather reflect improvements of mental health in general.
To overcome the limitations of TSSEM and the previous meta-analysis (Gu et al., 2015), in the present study we applied for the first time a three-level meta-analytic approach (TLMA; Cheung, 2015a) to investigate the mediating role of psychometric mindfulness in accounting for the effects of mindfulness interventions on mental health. TLMA is akin to univariate random-effects meta-analysis, but allows the modeling of nonindependent effect sizes in hierarchical data structures (i.e., changes in psychometric mindfulness and mental health in treatment and control groups from the same studies). The degree of dependence (i.e., the covariance) between effect sizes needs not be known in TLMA, but rather is estimated from the data. Like TSSEM, TLMA follows a structural equation modeling approach. We utilized TLMA in a novel manner to investigate the mediational model in an indirect way, following the classic approach by Baron and Kenny (1986).
The application of TLMA enabled the utilization of a substantially larger study sample, 69 RCTs (total N = 4479), as compared to merely 12 studies (8 RCTs, total N = 1109) in Gu et al. (2015). Further, the TLMA allowed to investigate the moderating effects of study population (clinical vs. non-clinical), intervention type, treatment duration, and study quality as well. We examined a broad range of mindfulness interventions, including treatments like MBSR and MBCT, but also meditation trainings, as this allowed us to test the specificity of psychometric mindfulness in a more comprehensive way. We compared mindfulness interventions and control groups (active, treatment as usual, waiting list) concerning (1) the average change in psychometric mindfulness and mental health, and (2) the association between these two outcomes. We show that changes in psychometric mindfulness are a correlate of changes in mental health on the meta-analytic level, but – because of the observed generality of this association across treatment and control groups alike – may not represent a specific mechanism of action of

mindfulness interventions. We further show that TLMA is a useful tool to perform an indirect analysis of a mediational model in an RCT context.

Methods / Approach

We searched major databases (e.g., Web of Science, PsychINFO, PSYNDEX, PubMed), screened the reference lists of previous meta-analyses, and hand-searched the journal *Mindfulness* for relevant studies. Inclusion criteria were: (1) the study was an RCT; (2) the treatment group received a mindfulness intervention; (3) the control group was a waiting list, underwent treatment as usual (TAU) or received an alternative (non-mindfulness) treatment (active control); (4) psychometric mindfulness and mental health were assessed with standardized scales before and after the intervention; (5) the study reported sufficient data to compute pretest-posttest effect sizes. We categorized treatments as mindfulness-based therapy (MBT; including MBSR and MBCT), acceptance and commitment therapy (ACT; Hayes, Strosahl, & Wilson, 1999), and 'other interventions' (modifications and adaptations of MBSR and MCBT, and various forms of meditation trainings). Overall, 69 studies (33 clinical, 36 non-clinical) were included, with a total of 4479 participants. Twenty-one studies investigated MBT (13 MBSR, 8 MBCT), 11 ACT, and 37 other interventions. In 35 studies, mindfulness interventions were compared to waiting list control groups, 21 studies used active control groups, and 13 studies used TAU control groups. The package metaSEM (version 0.9.10, Cheung, 2015b) was used to conduct a three-level meta-analysis (TLMA; Cheung, 2015a). The individual effect sizes (changes of psychometric mindfulness and mental health) constituted level 1 of the data structure. These level-1 effects represented either treatment effects or control group effects within studies (level 2); studies themselves were on level 3. For both the changes in psychometric mindfulness and mental health, we tested moderating effects of: (1) clinical vs. non-clinical studies, (2) type of treatment and control group (MBT vs. ACT vs. other interventions vs. active controls vs. TAU vs. waiting list), (3) treatment duration, and (4) study quality. Finally, for the changes in mental health, we also tested moderating effects of (5) changes of psychometric mindfulness. The moderator analyses were used to perform an indirect assessment of the mediational model, following the classic approach by Baron and Kenny (1986).

Results / Findings

Increases in psychometric mindfulness were of similar size (namely, medium: g~ 0.50) for all mindfulness interventions (MBT; ACT; other interventions, including meditation trainings). Non-mindfulness active control treatments led to only slightly lower increases in psychometric mindfulness (g ~ 0.40); only among TAU and waiting list groups, no net increases were observed (g ~ 0). Effects on mental health were comparable (g ~ 0.60) for the various mindfulness interventions, lower for active control treatments (g ~ 0.40), small for TAU (g ~ 0.20), and negligible for waiting list groups (g ~ 0). Effect sizes of both outcomes were moderately heterogeneous on the within-study level (I^2 ~ 50%) and at most lowly heterogeneous

on the between-study level (I^2 ≤ 17%). Study population (clinical vs. non-clinical), treatment duration, and study quality did not impact substantially on the changes in psychometric mindfulness and mental health. However, changes in psychometric mindfulness were associated with changes in mental health, B = 0.51 (0.12), [0.28–0.74], p < .001; i.e., for a change of one unit in the effect estimate of psychometric mindfulness, the effect estimate of mental health changed by an increment of 0.51 units. Change of psychometric mindfulness also accounted for the mean differences between intervention and control groups, and in total for 70% of the within-study excess heterogeneity between the effect sizes. What is more, the linear association between the changes in psychometric mindfulness and mental health was not confined to the positive spectrum (i.e., concomitant increases). It also extended to the negative spectrum (i.e., concomitant decreases) for some of the inactive control groups.

Conclusions / Implications

Psychometric mindfulness may be no specific mechanism of action of mindfulness interventions. Instead, it either appears to be comparably trained by any other (non-mindfulness) intervention (because it is a universal mechanism of action) or is not a mechanism of action at all, but rather merely a correlate of improvements in mental health in general. It is emphasized that these two possible explanations are not necessarily exclusive and, in theory, could each account for a part of the presently observed patterns. Monitoring and acceptance theory (MAT; Lindsay & Creswell, 2017) cautiously conjectures that the mechanisms of action of mindfulness interventions might be equally relevant for non-mindfulness interventions. Research provides evidence for the tenets of MAT for mindfulness interventions (Lindsay, Young, Smyth, Brown, & Creswell, 2018), but evidence for non-mindfulness interventions is still needed.

On the other hand, the presently observed associations between increases, but also between decreases, of psychometric mindfulness and mental health, and the fact that change in psychometric mindfulness also accounted for the mean differences in the treatment and control groups, are suggestive of a mutual interaction between psychometric mindfulness and mental health, or even a reversed direction of causality for these entities. Mutual interactions, and reverse causality, of psychometric mindfulness and mental health are an understudied line of research. Yet, they could be related to the reported trait overlap of psychometric mindfulness with neuroticism and negative affect (Giluk, 2009) as well. Available evidence does not provide any indication of reverse causality (based on a single-subject analysis of n = 11 patients undergoing a mindfulness intervention; Snippe et al., 2015), but more data are certainly needed.

Based on the current findings, results of studies investigating psychometric mindfulness as a mediator of treatment efficacy of mindfulness interventions thus need to be interpreted with caution. A re-conceptualization and re-calibration of the construct of psychometric mindfulness may be needed in order for the field to move forward (see Van Dam et al., 2018). Alternative indicators of mindfulness (e.g.,

duration and frequency of training, behavioral measures) could be used alongside psychometric self-assessment methods in future studies as well. TLMA proved a useful tool to perform an indirect analysis of a mediational model in an RCT context. We recommend its application in other fields of clinical and non-clinical studies as well, where RCTs are commonly used.

References

Alsubaie, M., Abbott, R., Dunn, B., Dickens, C., Keil, T. F., Henley, W., & Kyuken, W. (2017). Mechanisms of action in mindfulness-based cognitive therapy (MBCT) and mindfulness-based stress reduction (MBSR) in people with physical and/or psychological conditions: A systematic review. *Clinical Psychology Review*, *55*, 74-91. doi:10.1016/j.cpr.2017.04.008

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *6*, 1173-1182.

Bohlmeijer, E., Prenger, R., Taal, E., & Cuijpers, P. (2010). The effects of mindfulness-based stress reduction therapy on mental health of adults with a chronic medical disease: A meta-analysis. *Journal of Psychosomatic Research*, *68*, 539-544. doi:10.1016/j.jpsychores.2009.10.005

Cheung, M. W. L. (2015a). *Meta-analysis: A structural equation modeling approach*. Chichester, UK: Wiley.

Cheung, M. W. L. (2015b). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*, 1521-1565. doi:10.3389/fpsyg.2014.01521

Cheung, M. W. L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two stage approach. *Psychological Methods*, *10*, 40-64. doi:10.1037/1082-989X.10.1.40.

Giluk, T. L. (2009). Mindfulness, Big Five personality, and affect: A meta-analysis. *Personality and Individual Differences*, *47*, 805-811.

Goldberg, S. B., Tucker, R. P., Greene, P. A., Davidson, R. J., Wampold, B. E., Kearney, D. J., & Simpson, T. L. (2018). Mindfulness-based interventions for psychiatric disorders: A systematic review and meta-analysis. *Clinical Psychology Review*, *59*, 52-60. doi:10.1016/j.cpr.2017.10.011

Goldberg, S. B., Wielgosz, J., Dahl, C., Shuyler, B., MacCoon, D. S., Rosenkranz, M., … Davidson, R. J. (2015). Does the Five Facet Mindfulness Questionnaire measure what we think it does? Construct validity evidence from an active controlled randomized clinical trial. *Psychological Assessment*, *28*, 1009-1014.

Gu, J., Strauss, C., Bond, R., & Cavanagh, K. (2015). How do mindfulness-based cognitive therapy and mindfulness-based stress reduction improve mental health and wellbeing? A systematic review and meta-analysis of mediation studies. *Clinical Psychology Review*, *37*, 1-12. doi:10.1016/j.cpr.2015.01.006

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (1999). *Acceptance and commitment therapy: An experiential approach to behavior change*. New York: Guilford.

Kabat-Zinn, J. (1982). An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: Theoretical considerations and preliminary results. *General Hospital Psychiatry*, *4*, 33-47. doi:10.1016/0163-8343(82)90026-3

Khoury, B., Lecomte, T., Fortin, G., Masse, M., Therien, P., Bouchard, V., … Hofmann, S. G. (2013). Mindfulness-based therapy: A comprehensive meta-analysis. *Clinical Psychology Review*, *33*, 763-771. doi:10.1016/j.cpr.2013.05.005

Lindsay, E. K., & Creswell, J. D. (2017). Mechanisms of mindfulness training: Monitor and acceptance theory (MAT). Clinical Psychology Review, 51, 48-59. doi: 10.1016/j.cpr.2016.10.011

Lindsay, E. K., Young, S., Smyth, J. M., Brown, K. W., & Creswell, J. D. (2018). Acceptance lowers stress reactivity: Dismantling mindfulness training in a randomized controlled trial. Psychoneuroendocrinology, 87, 63-73. doi:10.1016/j.psyneuen.2017.09.015

Segal, Z. V., Williams, J. M., & Teasdale, J. (2002). *Mindfulness-based cognitive therapy for depression: A new approach to preventing relapse.* London: Guilford.

Snippe, E., Bos, E. H., van der Ploeg, K. M., Sanderman, R., Fleer, J., & Schroevers, M. J. (2015). Time-series analysis of daily changes in mindfulness, repetitive thinking, and depressive symptoms during mindfulness-based treatment. Mindfulness, 6, 1053-1062. doi:10.1007/s12671-014-0354-7

Van Dam, N. T., van Vugt, M. K., Vago, D. R., Schmalzl, L., Saron, C. D., Olendzki, A., … Meyer, D. E. (2018). Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on Psychological Science*, *13*, 36-61. doi:10.1177/1745691617709589

Authors:
Martin Voracek [1], Michael Kossmeier [1], Agnieszka Slowik [1], Ulrich S. Tran [1]

[1] University of Vienna

Title:

When is a replication successful? A systematic evaluation of multiple replication outcome indicators across contemporary multi-sample replication studies

Session & Time:

Methods in Meta-Analysis. Wednesday, May 29th, 3:30 pm - 4:00 pm

Abstract:

Background and Objectives

One of the most noteworthy and central events of the current (2010s) debates on research reproducibility and trustworthiness, open science, and method reform in psychological science and other empirical science fields was the emergence of a strong emphasis put on replication and replicability of empirical research in general. For a long time, replication studies had low scholarly prestige and low publishing prevalence. Over the past few years, the stance of replication thinking and replication practice has fundamentally changed for the better, culminating in courageous statements such as “replication has more scientific value than original discovery” (Ioannidis, 2018). To this can be added that it might be even more appropriate to speak of initial findings and studies, instead of “original” ones. Journals now do publish replication research, or even have implemented replication sections; preregistered replication studies are now found in their thousands at online repositories like the Open Science Framework; important conceptual contributions, such as taxonomies of and recipes for replications, have been brought forward; and widely publicized, large-scale consortia for conducting replication research have been formed (Zwaan et al., 2018). The publication of the Reproducibility Project: Psychology (RPP; Open Science Collaboration, 2015) in particular is regarded as a watershed in the 2010s replicability debates.

On the other hand, it seems fair to say that one aspect definitely has not kept pace with these spectacular developments surrounding the concept of replication. It is not a trivial one, it evidently is a tricky question, and it boils down to the evaluation of replication studies: how do we tell successful from failed replications? This question naturally is intimately linked to the philosophy and methodology of research synthesis, because evaluating the outcome of replication studies, inevitably and invariably so, implies comparing, weighing, or summarizing the evidence from at least two studies (minimally, one initial finding replicated once). To be sure, there are a number of seemingly straightforward indicators for replication success which have been in use for quite a time. However, these are used interchangeably, do not yield identical conclusions, all have their pros and cons, and not a few have grave epistemic deficiencies. As for just one example, when the conventional criterion of statistical significance ($p < .05$) was applied to the 100 replication studies conducted

in the course of the RPP, the replication rate amounted to 36%. However, it was 68% in the same data, according to two-study meta-analytic estimates of the corresponding initial studies and their single replications combined. As is well known, this vast discrepancy (two-thirds of psychology studies were unreplicable vs. two-thirds replicated) gave rise to heated discussions in the profession, as well as in the news and online social media.

Different replication success indicators tell different stories about replication study outcomes, different replication project consortia have used overlapping (but neither identical, nor exhaustive) ad-hoc selections of replicability indicators, and there is a significant number of novel ideas for evaluating replication outcomes, which so far have not been used at all in the course of large-scale replication endeavours. Presently there is no agreement on what constitutes replication success, or which indicator (or indicators) for evaluating replication outcomes could or should be preferably used.

To address this scattered evidence and unorganized state of affairs, we systematically and exhaustively apply all known replication success indicators to the results of contemporary replication study projects. We focus on many-to-one and many-to-many replication studies (wherein one research finding is, or several research findings are, replicated several, if not multiple, times), as opposed to the one-to-one type of replication (wherein one, or each, research finding is only replicated once, such as in the RPP; Open Science Collaboration, 2015), because the former format epistemically and empirically is richer in information. Further, we focus on psychological research, rather than including additional evidence from other fields (e.g., economics: Camerer et al., 2016).

Methods and Approach

The sampling frame of this meta-evaluative investigation comprised all large-scale, multi-sample replication investigations in psychological science accessible to date, namely, nine Registered Replication Reports (RRRs; Alogna et al., 2014; Bouwmeester et al., 2017; Cheung et al., 2016; Eerland et al., 2016; Hagger et al., 2016; McCarthy et al., 2018; O'Donnell et al., 2018; Verschuere et al., 2018; Wagenmakers et al., 2016), the Social Science Replication Project (SSRP; Camerer et al., 2018), the Prepublication Independent Replication Project (PPIR, or "Pipeline Project"; Schweinsberg et al., 2016), the Human Penguin Project (IJzerman et al., 2018), and the series of Many Labs (ML) projects (ML 1: Klein et al., 2014; ML 2: Klein et al., 2018; ML 3: Ebersole et al., 2016).

In addition, we included individual reports known to us, wherein one effect has been attempted to replicate several times, namely, multiple replication studies of the Macbeth effect (Earp et al., 2014), of Bem's psi effects (Ritchie et al., 2012), of power posing (various reports: see Jonas et al., 2017), of loneliness/bathing habits associations (Donnellan et al., 2015), of the letters from the heart effect (Voracek et al., 2007), elderly priming (Doyen et al., 2012), intelligence priming (Shanks et al., 2013), of conception risk/prejudice associations (Hawkins et al., 2015), and seven further contemporary multi-sample replication studies, published in the same journal

issue (see Nosek & Lakens, 2014) as ML 1 (Blanken et al., 2014; Brandt et al., 2014; Calin-Jageman & Caldwell, 2014; IJzerman et al., 2014; Johnson et al., 2014; Lynott et al., 2014; Vermeulen et al., 2014).
We hope to identify further such replication studies of the latter type for additional inclusion. As well, we anticipate to be able to include the outcomes of the ML 4 (Klein et al., forthcoming) and ML 5 (Ebersole et al., forthcoming) projects and an additional RRR (Colling et al., forthcoming), currently underway, in due course. It is possible that results of the PPIR successor study PPIR-2 and the first multi-site replication piece of the Psychological Science Accelerator consortium (PSA; Moshontz et al., 2018) will become available in the first half-year of 2019.
We uniformly applied all proposed replication indicators known to us (old and novel ones, conventional vs. non-mainstream approaches, widely applied vs. untested ones; e.g., Mathur & VanderWeele, 2017, 2018) to the entirety of the above evidence from contemporary multi-sample replication studies. Altogether, we tested more than two dozen such indicators. The list included conventional ($p < .05$) and stricter ($p < .005$) significance criteria for replications, meta-analytic summaries of both initial and replication effects (or of the latter alone), confidence interval overlap between (and effect comparisons with) these, relative effect size between initial and replication effects, prediction intervals, Bayesian approaches (Bayes factors, mixture models, snapshot hybrid meta-analyses), equivalence tests, the safeguard sample ratio, Simonsohn's small-telescope test, the z-curve procedure, Gelman's type S and type M error concept, the Mathur-VanderWeele metrics, and (where applicable), p-curve, half p-curve, and p-uniform. We also propose and put to the test several new indicators, such as some derived from combinatorial (all-study-subsets) meta-analysis (Olkin et al., 2012): e.g., the majority vote from an all-subsets meta-analysis of replication study outcomes, and the percentile which an initial study's effect occupies within the distribution of these.

Results and Findings

Preliminary results for this multitude of replication indicators, as applied on a vast database of real-world replication study outcomes, show nuanced patterns of similarities, as well as differences, among replication indicators and replication projects alike.
Classic indicators, based on null-hypothesis significance testing, show higher agreement among themselves than with alternative indicators. Approaches based on prediction intervals draw a less optimistic picture than those based on confidence intervals, and Bayesian analyses and the small-telescope approach even pessimistic ones. Also, Bayesian approaches enable to chart the no-man's-land located between the classic statistical conclusion dichotomy of "significant" vs. "not significant".
Results for some novel indicators are noticeably different from the conclusions derived from conventional indicators and partly yield unexpected conclusions. These novel indicators, as a whole, agree less among themselves than conventional indicators do, mainly because they exploit quite different types of statistical information and adhere to different rationales for evidence-weighing. As well, there

are some groups of replication projects, for which replication indicators agree more (e.g., the RRR series), whereas less so for others (e.g., the ML series). We demonstrate and visualize the similarity vs. dissimilarity of conclusions based on the different replication indicators across different replication projects (and conversely, of the different replication projects and research effects, as evaluated by the indicators) via cluster-analytic and multidimensional scaling methods. We supplement this similarity-dissimilarity analysis of replication outcome indicators and replication projects alike with a thematic analysis of the replication outcome indicators and considerations of the statistical relations between these.

Conclusions and Implications

This is the first systematic and exhaustive trial of about 25 proposed indicators for replication success applied to about 30 retrievable contemporary multi-sample replication projects in psychology, with the latter comprised of scores of initial research effects and associated outcomes in replication samples. The findings show differential agreement with regards to these replication indicators and replication projects. We make no claim to have solved the problem of how to best evaluate empirical evidence of replications vis-à-vis initial study findings, which is a fundamental issue of current empirical science. Rather, the various (statistical, epistemic, practical) insights derived from this systematic investigation may serve as a fruitful intermediate result, which beneficially highlights areas of possible future consensus and holds promise to inspire further conceptual and statistical thinking along these lines. All in all, there appear to be many advantages associated with more fine-grained, continuous indicators for evaluating replication outcomes, as opposed to dichotomous ones (“success” vs. “failure”; Gelman, 2018).

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., ... Buswell, K. (2014). Registered Replication Report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9,* 556-578.

Blanken, I., van de Ven, N., Zeelenberg, M., & Meijers, M. H. (2014). Three attempts to replicate the moral licensing effect. *Social Psychology*, *45,* 232-238.

Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., ... Evans, A. M. (2017). Registered Replication Report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, *12,* 527-542.

Brandt, M. J., IJzerman, H., & Blanken, I. (2014). Does recalling moral behavior change the perception of brightness? A replication and meta-analysis of Banerjee, Chatterjee, and Sinha (2012). *Social Psychology*, *45,* 246-252.

Calin-Jageman, R. J., & Caldwell, T. L. (2014). Replication of the superstition and performance study by Damisch, Stoberock, and Mussweiler (2010). *Social Psychology, 45,* 239-245.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351,* 1433-1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... Altmejd, A. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, *2*, 637-644.

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., ... Carcedo, R. J. (2016). Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11,* 750-764.

Colling, J. L., Szücs, D., et al. (forthcoming). Registered Replication Report of Fischer, Castel, Dodd, & Pratt (2003). Available from https://osf.io/he5za/.

Donnellan, M. B., Lucas, R. E., & Cesario, J. (2015). On the association between loneliness and bathing habits: Nine replications of Bargh and Shalev (2012) Study 1. *Emotion*, *15,* 109-119.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLOS ONE*, *7*, e29081.

Earp, B. D., Everett, J. A., Madva, E. N., & Hamlin, J. K. (2014). Out, damned spot: Can the "Macbeth Effect" be replicated? *Basic and Applied Social Psychology*, *36,*91-98.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68-82.

Ebersole, C. R., Nosek, B. A., Kidwell, M. C., Buttrick, N., Baranski, E., …Hartshorne, J. (forthcoming). Many Labs 5: Testing pre-data collection peer review as an intervention to increase replicability. Available from https://osf.io/7a6rd/.

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., ... Crocker, C. (2016). Registered Replication Report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11,* 158-171.

Gelman, A. (2018). Don't characterize replications as successes or failures. *Behavioral and Brain Sciences*.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... Calvillo, D. P. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11,* 546-573.

Hawkins, C. B., Fitzgerald, C. E., & Nosek, B. A. (2015). In search of an association between conception risk and prejudice. *Psychological Science*, *26*, 249-252.

IJzerman, H., Blanken, I., Brandt, M. J., Oerlemans, J. M., Van den Hoogenhof, M. M., Franken, S. J., & Oerlemans, M. W. (2014). Sex differences in distress from infidelity in early adulthood and in later life: A replication and meta-analysis of Shackelford et al. (2004). *Social Psychology*, *45,* 202-208.

IJzerman, H., Lindenberg, S., Dalğar, İ., Weissgerber, S. S., Vergara, R. C., Cairo, A. H., ... Hall, C. J. (2018). The Human Penguin Project: Climate, social integration, and core body temperature. *Collabra: Psychology*, *4,* 1.

Ioannidis, J. P. (2018). Why replication has more scientific value than original discovery. *Behavioral and Brain Sciences*.

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? A direct replication of Schnall, Benton, and Harvey (2008). *Social Psychology, 45,* 209-215.

Jonas, K. J., Cesario, J., Alger, M., Bailey, A. H., Bombari, D., Carney, D., ... Jackson, B. (2017). Power poses: Where do we stand? *Comprehensive Results in Social Psychology*, *2,* 139-141.

Klein, R. A., Ebersole, C. L., Cook, C. L., Nosek, B. A., Redford, L., Levitan, C., … Harton, H. C. (forthcoming). Many Labs 4: Investigating effects of researcher expertise on replication outcomes. Available from https://osf.io/8ccnw/.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... Cemalcilar, Z. (2014). Investigating variation in replicability. *Social Psychology, 45,* 142-152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*.

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology, 45,* 216-222.

Mathur, M. B., & VanderWeele, T. J. (2017). New statistical metrics for multisite replication projects. Available from https://osf.io/apnjk/.

Mathur, M. B., & VanderWeele, T. J. (2018). New metrics for meta-analyses of heterogeneous effects. Available from https://osf.io/ksyq5/.

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., ... Barbosa, F. (2018). Registered Replication Report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science*, *1,* 321-336.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45,* 137-141.

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., ... Balatekin, N. (2018). Registered Replication Report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science*, *13,* 268-294.

Olkin, I., Dahabreh, I. J., & Trikalinos, T. A. (2012). GOSH: A graphical display of study heterogeneity. *Research Synthesis Methods, 3,* 214-223.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem‘s ‘retroactive facilitation of recall’ effect. *PlOS ONE*, *7,* e33423.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., ... Srinivasan, M. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55-67.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., ... Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLOS ONE, 8,* e56515.

Vermeulen, I., Batenburg, A., Beukeboom, C., & Smits, T. (2014). Breakthrough or one-hit wonder? Three attempts to replicate musical conditioning effects in advertising. *Social Psychology, 45,* 179-186.

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., ... Barbosa, F. (2018). Registered Replication Report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1,* 299-317.

Voracek, M., Tran, U. S., Schabauer, H., Koenne, G., & Glössl, B. (2007). On the elusive nature of the letters from the heart effect. *Perceptual and Motor Skills*, *104*, 803-814.

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... Bulnes, L. C. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11,* 917-928.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*.