Editorial

Why We Need Systematic Reviews and Meta-Analyses in the Testing and Assessment Literature

Dragos Iliescu[1], Andrei Rusu[1], Samuel Greiff[2], Marjolein Fokkema[3], and Ronny Scherer[4]

[1] Faculty of Psychology and Educational Sciences, University of Bucharest, Romania
[2] Department of Behavioural and Cognitive Sciences, University of Luxembourg, Luxembourg
[3] Department of Methods & Statistics, University Leiden, The Netherlands
[4] Centre for Educational Management, University of Oslo, Norway

The Utility of Systematic Reviews and Meta-Analyses

The cumulative nature of science is arguably one of the main pillars for the advancement of scientific knowledge which relies on the accumulation of empirical findings. This view can be tracked back at least to the 17th and 18th centuries, at first in the writings of classical empiricists, and later emphasized with the emergence of August Comte's positivism (Niiniluoto, 2019). Even though it had (and probably still has) its critics, the cumulative view passed the scrutiny of time and is both explicit and implicit in nowadays' scientific thinking in many different ways. Consider, for example, our current knowledge regarding life on Earth. It is hard to imagine we could have achieved such a complex grasp without the constant accumulation of findings in evolutionary biology in such areas as genetics or fossil record research (Zeigler, 2012). An example in the domain of psychological science is personality psychology, that is nowadays often explained through an integrative trait taxonomy – the five-factor model of personality (Condon et al., 2020; Goldberg, 1993; McCrae & John, 1992) – that has developed over time, through slow accumulation (Goldberg, 1993), and has been in the focus of a plethora of primary studies and research syntheses (e.g., Digman, 1997; van der Linden et al., 2010).

In fact, to combine the body of evidence on a specific topic in order to avoid the errors inherent in isolated findings is the standard approach in many fields, both for advancing our theoretical understanding and for informing practice and policy (see, for example, the evidence-based pyramid; e.g., Tomlin & Borgetto, 2011). Initially, this task was met almost exclusively through narrative reviews conducted by experts in a given field, in spite of the biasing nature of expert subjectivity. This is the motive for the emergence of methodologies suited for a more systematic (and more empirical) approach.

Biases (even for the best of experts) can slip through in many stages of the process, such as selection of relevant studies, focusing on what one specific expert considers important in a given study, or focusing rather on positive results in the publication process. No wonder that sometimes different reviewers reach contrasting conclusions on the same question. A memorable illustration is the debate started by Hans Eysenck's 1952 review that concluded that psychotherapy was not effective, and to which a compelling response was given only in 1978 by Gene Glass, who aggregated the findings on psychotherapy outcomes statistically and offered the opposite conclusion (Glass, 2015). Glass named this method *meta-analysis*, and it was one of the pivotal points in stepping towards a more systematic and quantitative process of literature review. In Glass's own words, a meta-analysis is "the statistical analysis of a large collection of analysis results from individual studies for purposes of integrating the findings" (Glass, 1976, p. 3). This definition can help us discern with more clarity between these two contingent terms: systematic review and meta-analysis. The first term labels the entire process of overviewing the knowledge related to a research question by exhaustively identifying, selecting, and synthesizing the body of evidence addressing that question, whereas the second, as quoted, refers exclusively to statistically combining the data resulted from the reviewed studies. A systematic review can be conducted without including a meta-analysis, but it is

harder to imagine a meta-analysis in the absence of a systematic review.

The drift towards systematic reviews and meta-analyses has noticeably accelerated in most fields of science in the 1990s (Borenstein et al., 2021). Nowadays systematic reviews are ubiquitous and are characterized by having clear sets of rules for searching the literature, inclusion of relevant studies, and coding of focal information. Even though subjectivity can still bias any of these stages, reporting standards emphasize transparency (e.g., PRISMA Statement: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Page et al., 2021). And when a systematic review provides a sufficient volume of statistical results, meta-analysis steps in, with a mature corpus of methods for data aggregation based on well-developed procedures (e.g., Borenstein et al., 2021; Hunter & Schmidt, 2004), and with continuing advancements in statistical sophistication (e.g., network meta-analysis; Bayesian meta-analysis, etc.).

Possibilities for Systematic Reviews and Meta-Analyses in the Assessment Literature

Despite its obvious strengths and comparative trendiness, the systematic approach in reviewing empirical findings is relatively rare in the psychological assessment literature – though some notable contributions are visible – e.g., studies in the tradition of reliability generalization (Sánchez-Meca et al., 2021; Scherer & Teo, 2020). There are reasons for this comparative scarcity of course – meta-analyses usually focus on the intensity of relationships between variables, or standardized differences between means, which by definition imply not one but several variables. For example, systematic reviews and meta-analyses in the tradition of validity generalization are arguably about the relationship of a predictor and criterion and less about the quality of measurement in either of these two.

Testing and assessment research usually focuses on the quality of measurement regarding one single variable, and when several variables appear, they are included rather for cross-validation purposes. In this respect, more obvious issues that may be investigated through meta-analytical procedures in the testing realm would be investigations into the relationships between items and (latent) variables, e.g., the investigation of the relationship between an item (e.g., "I have many friends") and a construct (e.g., extraversion), across different tests, samples, or cultures. Or, the already mentioned meta-analyses of test reliability (Scherer & Teo, 2020), and factor structures meta-analyses (e.g., Gnambs et al., 2018; Huang & Chen, 2015).

But Should They Be Used? If Yes, How?

We strongly believe that, at the very least, systematic reviews (and in many cases also full-fledged meta-analyses) can and should be used to summarize insights gathered over time with a specific test, a family of assessments, or a domain of assessment. For example, a systematic inquiry of the psychometric properties and other characteristics of a test, reported across different papers, samples, contexts, and cultures, may shed more light into that test than any specific study, no matter how well designed. A systematic overview of all the test adaptations developed for a specific test may provide the best possible indicator for the cross-cultural invariance of that specific test. A systematic review of all the assessments typical for a specific domain (say, burnout, or cognitive ability) may give incredible insight into the pros and cons of how that domain is currently approached in terms of assessment and, at the same time, highlight the gaps in the measurement of that construct, which, in turn, would mandate the development of new measures.

Additionally, when sufficient data are available, systematic reviews may be complemented with meta-analyses. These could offer, for example, comprehensive empirical overviews of the validity of different measures, as in matters of construct validity (e.g., meta-analytical nomological networks and factor structures) or of criterion validity (e.g., meta-analytical findings on assessment-criteria relationships), that could go as deep as item-level meta-analysis (Carpenter et al., 2016). The latest developments in multivariate meta-analysis (e.g., meta-analytic structural equation modeling; Cheung, 2015) enable researchers to expand the validity evidence from single effect sizes (e.g., correlations between a construct score and a criterion) to multiple, dependent effect sizes (e.g., indicator-indicator correlations) and thus test more complex theories and assumptions on the measurement of constructs. Also, by quantifying and disentangling the heterogeneity in the various operationalizations of a specific construct (e.g., as reflected through variations in the relationships with other variables), a meta-analysis could provide insights for scrutinizing and even revisiting (failed) conceptual replications – that is, when the replication and the original study investigate the same construct but by means of different measures (Stanley et al., 2018).

Systematic reviews and meta-analyses can also be used to examine the performance of diagnostic tests. This is an important practical implication for clinical tests that can only be summarized through research that aggregates data on such attributes like the specificity and sensitivity of one or several measurement instruments that focus on

a specific diagnostic call. Hence, systematic reviews and meta-analyses could offer a well-needed overview, critical analysis, and necessary practical guidance on "*When?*", "*Where?*", on "*Whom?*", and "*How?*" specific measures can be used.

Notable Examples in the Literature

Some illustrative examples of good practices in using systematic reviews and meta-analyses can be spotted in the field of psychological testing and assessment.

For example, Elosua and Iliescu (2012) systematically reviewed the psychometric practice for the ten most used psychological tests in Europe from a decade ago. Their work looked at the practice of psychological assessment and reviewed the manuals of those commercially available instruments. They identified and elaborated on the gap between psychometric theory and practice pushing towards a stronger reconciliation of science with practice. Also with a cultural focus, Zeinoun et al. (2021) conducted a recent systematic review of the psychological tests that were published in the mental health domain in Arabic. Their effort identified 138 instruments for the assessment of psychological disorders and discussed the degree to which each of them aligns with the latest psychometric standards. Through this, they offered a comprehensive perspective both for researchers and mental health practitioners who study and work with Arabic-speaking populations.

Besides focusing on specific cultures, other scholars inquired the operationalization of specific constructs (especially if these are of more recent development). For example, King et al. (2020) systematically reviewed the available instruments for gaming disorder. Besides identifying the best existing measures, the authors also spotted large inconsistencies between the various operationalizations and signaled the need for a common understanding and for aligned measurement instruments.

Systematic reviews and meta-analyses can also be developed on narrower topics in testing – for example through a comprehensive focus on only one measure. Bardhoshi et al. (2016) offer such a case through their meta-analysis of the psychometric evidence regarding the Beck Anxiety Inventory (Beck & Steer, 1993). The study provided a comprehensive quantitative overview on different forms of reliability and validity, including diagnostic accuracy, being highly informative for the evidence-based approach. In this manner, we can address even narrower questions, such as a specific psychometric property of a measure. For example, Andrei et al. (2014) meta-analyzed the incremental validity of the Trait Emotional Intelligence Questionnaire, or Greenwald et al. (2009) meta-analyzed the predictive validity of the Implicit Association Test. Gnambs et al. (2018) examined the factor structure of the Rosenberg Self-Esteem Scale meta-analytically and tested several competing factor models.

Of note, the very same researcher that came up with the name meta-analysis also proposed its retirement 25 years later, in favor of analyzing participant-level data from multiple studies (Glass, 2000). Glass (2000) argued, that instead of publishing studies and aggregating statistics at the study level, involving often untestable assumptions and precluding assessment of individual-level effects, the original datasets should be aggregated. This is often referred to as individual participant data meta-analysis (IPDMA) and could even be considered the state-of-the-art in evidence accumulation. Although IPDMAs are already common in the medical field, in psychological assessment they are still quite rare. Examples of IPDMAs in the field of assessment include factor analyses of tests for cognitive functioning (van Rentergem et al, 2020) and criterion validity of a self-report measure for depression (Fischer et al., 2021). IPDMAs involve pooling datasets from multiple studies. Mixed-effects models allow for analyzing such data, and even allow for combining results of studies from which individual-level, as well as from which only study-level data are available (Cooper & Patall, 2009; Sutton, Kendrick & Coupland, 2008). Along these lines, we would highly value submissions involving IPDMA. In order to allow for IPDMA, we also strongly encourage the publishing of data and code used in single- or multi-study reports.

One should bear in mind that we do not claim these to be the only or the best expressions of the current scholarly literature, just some examples of the type of future contributions that we could imagine for the *European Journal of Psychological Assessment* as well.

Conclusion and a Call to Action

We encourage submissions of contributions of systematic reviews and meta-analyses in the vein described in this editorial. We especially believe that three types of papers will contribute massively to closing extant gaps in the literature:

First, we recommend work on systematic reviews and meta-analyses that investigate the psychometric qualities of a specific measure, and we would recommend in this context that studies focus on acknowledged "gold standard" measures that are intensively used in practice and research.

Second, we believe in the utility of systematic reviews that focus on a specific construct and the various approaches and measures used for its assessment. In this regard we recommend that researchers focus on highly used constructs that are often characterized by lack of unity in conceptualization and measurement; examples such as

"the assessment of burnout", or "testing for student wellbeing" or "the assessment of culture-bound clinical syndromes" jump to mind.

Third, we suggest that systematic reviews that focus on measures that have been developed or adapted, and are available, in a specific cultural space will be useful. Especially understudied and non-WEIRD cultures such as sub-Saharian Africa, the Arab cultural space, or South-East Asia are of interest in this regard and could contribute to a lucid view of the current state of psychological assessment and the extant gaps to be addressed by future efforts in those cultural spaces.

The editorial office will in the immediate future streamline the submission and review of these papers and will provide the possibility of a dedicated type of submission ("Systematic Review"), with up to 10,000 words, which we strongly recommend contributors to consider.

We should probably note here that we especially encourage pre-registration of these reviews: through pre-registration, a systematic review becomes even stronger by promoting the spirit of open science this journal adheres to (Greiff et al., 2020). The "Registered Report" route is also open for these initiatives, and we strongly encourage authors to use this route (Greiff & Allen, 2018). Finally, a special issue featuring such systematic reviews is in preparation, and a call for proposals will go out presently.

References

Andrei, F., Siegling, A. B., Aloe, A. M., Baldaro, B., & Petrides, K. V. (2014). The incremental validity of the Trait Emotional Intelligence Questionnaire (TEIQue): A systematic review and meta-analysis. *Journal of Personality Assessment, 98*(3), 261–276. https://doi.org/10.1080/00223891.2015.1084630

Bardhoshi, G., Duncan, K., & Erford, B. T. (2016). Psychometric meta-analysis of the English version of the Beck Anxiety Inventory. *Journal of Counseling & Development, 94*(3), 356–373. https://doi.org/10.1002/jcad.12090

Beck, A. T., & Steer, R. A. (1993). *Beck Anxiety Inventory Manual*. Psychological Corporation.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). Wiley.

Carpenter, N. C., Son, J., Harris, T. B., Alexander, A. L., & Horner, M. T. (2016). Don't forget the items: Item-level meta-analytic and substantive validity techniques for reexamining scale validation. *Organizational Research Methods, 19*(4), 616–650. https://doi.org/10.1177/1094428116639132

Cheung, M. W.-L. (2015). *Meta-analysis: A structural equation modeling approach*. Wiley. https://doi.org/10.1002/9781118957813

Condon, D. M., Wood, D., Mõttus, R., Booth, T., Costantini, G., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Ziegler, M., & Zimmermann, J. (2020). Bottom up construction of a personality taxonomy. *European Journal of Psychological Assessment, 36*(6), 923–934. https://doi.org/10.1027/1015-5759/a000626

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14*(2), 165–176. https://doi.org/10.1037/a0015565

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 73*(6), 1246–1256. https://doi.org/10.1037/0022-3514.73.6.1246

Elosua, P., & Iliescu, D. (2012). Tests in Europe: Where we are and where we should go. *International Journal of Testing, 12*(2), 157–175. https://doi.org/10.1080/15305058.2012.657316

Fischer, F., Levis, B., Falk, C., Sun, Y., Ioannidis, J. P., Cuijpers, P., Shrier, I., Benedetti, A., & Thombs, B. D., Depression Screeing Data (DEPRESSED) PHQ Collaboration. (2021). Comparison of different scoring methods based on latent variable models of the PHQ-9: An individual participant data meta-analysis. *Psychological Medicine*. Advance online publication. https://doi.org/10.1017/s0033291721000131

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3–8. https://doi.org/10.3102/0013189X005010003

Glass, G. V. (2000, March). *The future of meta-analysis*. Paper presented at the University of California, Berkeley–Stanford University Colloquium on Meta-Analysis, Department of Psychology, University of California, Berkeley. http://www.gvglass.info/papers/meta25.html

Glass, G. V. (2015). Meta-analysis at middle age: A personal history. *Research Synthesis Methods, 6*(3), 221–231. https://doi.org/10.1002/jrsm.1133

Gnambs, T., Scharl, A., & Schroeders, U. (2018). The structure of the Rosenberg Self-Esteem Scale: A cross-cultural meta-analysis. *Zeitschrift für Psychologie, 226*(1), 14–29. https://doi.org/10.1027/2151-2604/a000317

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*(1), 26–34. https://doi.org/10.1037/0003-066X.48.1.26

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41. https://doi.org/10.1037/a0015575

Greiff, S., & Allen, M. S. (2018). *EJPA* Introduces Registered reports as new submission format. *European Journal of Psychological Assessment, 34*(4), 217–219. https://doi.org/10.1027/1015-5759/a000492

Greiff, S., van der Westhuizen, L., Mund, M., Rauthmann, J. F., & Wetzel, E. (2020). Introducing new open science practices at *EJPA*. *European Journal of Psychological Assessment, 36*(5), 717–720. https://doi.org/10.1027/1015-5759/a000628

Huang, C., & Chen, J.-H. (2015). Meta-analysis of the factor structures of the Beck Depression Inventory-II. *Assessment, 22*(4), 459–472. https://doi.org/10.1177/1073191114548873

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Sage Publications.

King, D. L., Chamberlain, S. R., Carragher, N., Billieux, J., Stein, D., Mueller, K., Potenza, M. N., Rumpf, H. J., Saunders, J., Starcevic, V., Demetrovics, Z., Brand, M., Lee, H. K., Spada, M., Lindenberg, K., Wu, A. M. S., Lemenager, T., Pallesen, S., Achab, S., … P H., Delfabbro, P. H. (2020). Screening and assessment tools for gaming disorder: A comprehensive systematic review. *Clinical Psychology Review, 77*, Article 101831. https://doi.org/10.1016/j.cpr.2020.101831

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115. https://doi.org/10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Niiniluoto, I. (2019). Scientific progress. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.). The Metaphysics Research Lab Center for the Study of Language and Information. https://plato.stanford.edu/entries/scientific-progress/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Blázquez-Rincón, D. M., López-Ibáñez, C., & López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods, 12*(4), 516–536. https://doi.org/10.1002/jrsm.1487

Scherer, R., & Teo, T. (2020). A tutorial on the meta-analytic structural equation modeling of reliability coefficients. *Psychological Methods, 25*(6), 747–775. https://doi.org/10.1037/met0000261

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Sutton, A. J., Kendrick, D., & Coupland, C. A. (2008). Meta-analysis of individual-and aggregate-level data. *Statistics in Medicine, 27*(5), 651–669.

Tomlin, G., & Borgetto, B. (2011). Research pyramid: A new evidence-based practice model for occupational therapy. *American Journal of Occupational Therapy, 65*, 189–196. https://doi.org/10.5014/ajot.2011.000828

van der Linden, R., Nijenhuis, J., & Bakker, A. B. (2010). The General factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*(3), 315–327. https://doi.org/10.1016/j.jrp.2010.03.003

van Rentergem, J. A. A., Nathalie, R., Schmand, B. A., Murre, J. M., Staaks, J. P., & Huizenga, H. M. (2020). The factor structure of cognitive functioning in cognitively healthy participants: A meta-analysis and meta-analysis of individual participant data. *Neuropsychology Review, 30*(1), 51–96.

Yarkoni, T. (2020). Implicit realism impedes progress in psychology: Comment on Fried (2020). *Psychological Inquiry, 31*(4), 326–333.

Zeigler, D. (2012). Evolution and the cumulative nature of science. *Evolution: Education and Outreach, 5*(4), 585–588. https://doi.org/10.1007/s12052-012-0454-6

Zeinoun, P., Iliescu, D., & El Hakim, R. (2021). Psychological tests in Arabic: A review of methodological practices and recommendations for future use. *Neuropsychology Review*, 1–19. https://doi.org/10.1007/s11065-021-09476-6

Published online March 7, 2022

Dragos Iliescu
Faculty of Psychology and Educational Sciences
University of Bucharest
Sos. Panduri 90
050657 Bucharest
Romania
dragos.iliescu@fpse.unibuc.ro

