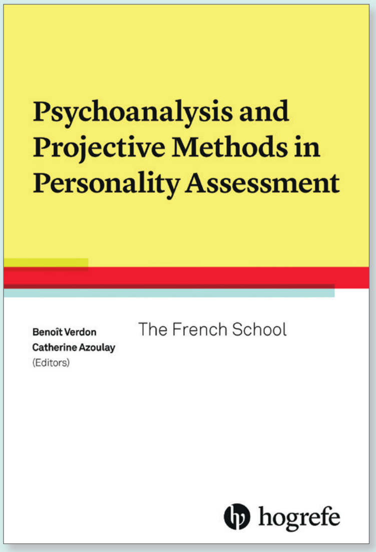
Holger Steinmetz
Nadine Wedderhoff
Michael Bošnjak
(Editors)

Hotspots in Psychology 2022

Zeitschrift für Psychologie
Founded in 1890
Volume 230 / Number 1 / 2022

hogrefe



Psychoanalysis and Projective Methods in Personality Assessment

Benoît Verdon
Catherine Azoulay
(Editors)

The French School

hogrefe

Holger Steinmetz
Nadine Wedderhoff
Michael Bošnjak
(Editors)

Hotspots in Psychology 2022

Zeitschrift für Psychologie
Volume 230 / Number 1 / 2022

hogrefe

Hogrefe Publishing
Incorporated and registered in the Commonwealth of Massachusetts, USA, and in Göttingen, Lower Saxony, Germany

Cover image ©Adobe Stock/iamchamp

Printed and bound in Germany

ISBN 978-0-88937-622-9

The *Zeitschrift für Psychologie*, founded by Hermann Ebbinghaus and Arthur König in 1890, is the oldest psychology journal in Europe and the second oldest in the world. Since 2007, it appears in English and is devoted to publishing topical issues that provide convenient state-of-the-art compilations of research in psychology, each covering an area of current interest.

The *Zeitschrift für Psychologie* is available as a journal in print and online by annual subscription and the different topical compendia are also available as individual titles by ISBN.

Editor-in-Chief
Michael Bošnjak, Robert Koch Institute, Department of Epidemiology and Health Monitoring, General-Pape-Str. 62-66, 12101 Berlin, Germany, Tel +49 30 18754 3103, michael@bosnjak.eu

Publisher
Hogrefe Publishing, Merkelstr. 3, 37085 Göttingen, Germany, Tel. +49 551 999 50 0, Fax +49 551 999 50 425, publishing@hogrefe.com
North America: Hogrefe Publishing Corporation, 44 Merrimac Street, Suite 207, Newburyport, MA 01950, USA, Tel. +1 978 255 3700, customerservice@hogrefe.com

Production
Irina Rau, Hogrefe Publishing, Merkelstr. 3, 37085 Göttingen, Germany,
Tel. +49 551 999 50 430, Fax +49 551 999 50 425, production@hogrefe.com

Subscriptions
Hogrefe Publishing, Herbert-Quandt-Str. 4, 37081 Göttingen, Germany, Tel. +49 551 999 50 900,
Fax +49 551 999 50 998

Advertising/Inserts
Hogrefe Publishing, Merkelstr. 3, 37085 Göttingen, Germany, Tel. +49 551 999 50 423,
Fax +49 551 999 50 425, marketing@hogrefe.com

ISSN
ISSN-L 2151-2604, ISSN-Print 2190-8370, ISSN-Online 2151-2604

Publication
Published in 4 topical issues per annual volume.

Subscription Prices
Calendar year subscriptions only. Rates for 2021: Institutions - from US $353.00 / € 272.00 (print only; pricing for online access can be found in the journals catalog at hgf.io/journalscatalog); Individuals US $195.00 / €139.00 (all plus US $16.00 / €12.00 shipping & handling; € 6.80 in Germany).
Single issue US $49.00 / € 34.95 (plus shipping & handling).

Payment
Payment may be made by check, international money order, or credit card, to Hogrefe Publishing, Merkelstr. 3, 37085 Göttingen, Germany. US and Canadian subscriptions can also be ordered from Hogrefe Publishing, 361 Newbury Street, 5th Floor, Boston, MA 02115, USA

Electronic Full Text
The full text of *Zeitschrift für Psychologie* is available online at https://econtent.hogrefe.com and in PsycARTICLES™.

Abstracting Services
Abstracted/indexed in Current Contents/Social and Behavioral Sciences (CC/S&BS), Social Sciences Citation Index (SSCI), Research Alert, PsycINFO, PASCAL, PsycLit, IBZ, IBR, ERIH, and PSYNDEX.
Impact Factor (2020): 2.633
2020 Impact Factor 2.633, 5-year Impact Factor 3.309, Journal Citation Reports (Clarivate 2021)

Contents

Editorial

Hotspots in Psychology – 2022 Edition

Holger Steinmetz[1], Nadine Wedderhoff[1], and Michael Bošnjak[1,2]

[1]University of Trier, Germany

[2]Department of Epidemiology and Health Monitoring, Robert Koch Institute, Berlin, Germany

Abstract: This editorial introduces the five articles included in the sixth "Hotspots in Psychology" of the *Zeitschrift für Psychologie*. With the new edition, the format enlarges it focus beyond meta-analyses and systematic reviews to new developments such as big data in psychology. The included five articles span a diverse array of topics, that is, the application of individual participants meta-analyses as a way to replicate studies, the role of the degree of anthropomorphism ("human-likeness") in human-robot interactions, the challenge of multiple dependent effect sizes when conducting a meta-analytical structural equation model, the value of using log data of online platform as a way to predict learning outcomes, and the utility of a block-wise fit evaluation in structural equation models with many longitudinally measured variables. To promote the open science philosophy, the papers present supplemental material that can be accessed via the repository PsychArchives (www.psycharchives.org).

The "Hotspots in Psychology" topical issue has a tradition in presenting articles focusing on reviews and meta-analyses in research-active (i.e., hotspot) fields (Bošnjak & Erdfelder, 2018; Bošnjak & Gnambs, 2019; Bošnjak & Wedderhoff, 2020; Bošnjak, Wedderhoff, & Steinmetz 2021; Erdfelder & Bošnjak, 2016). While the focus on meta-analyses is still the backbone of this year's edition, we decided to widen the focus to also include articles discussing the fruitfulness and challenges of big data, for instance, of intensive longitudinal data (e.g., time series, experience sampling), data provided by means of nonobtrusive methods (e.g., sensors, log data), and approaches related to the treatment and handling of such data or their statistical analysis.

The first article by Maria Klose, Diana Steger, Julian Fick, and Cordula Artelt (2022) focuses on the utility and challenges of analyzing log data on learning activities in the field of learning analytics. The authors present a meta-analysis using 41 studies (N = 28,986), revealing a positive relationship between total login time and login frequency on a learning platform and grades. Beyond the specific implications of the usefulness of log data for analyzing learning behavior and outcomes, the paper fits in with the increasing research on the value of big, technically measured data for psychology and related fields.

The article by Isidora Stolwijk, Suzanne Jak, Veroni Eichelsheim, and Machteld Hoeve (2022) addresses a challenge most researchers testing meta-analytical structural equation models (MASEM) are confronted with: how to proceed when primary studies present several estimates of a target correlation. While this problem can easily be handled through a three-level approach when the target is a simple bivariate correlation, issues become tricky when the goal is a MASEM based on a matrix of correlations. Stolwijk and colleagues compare the differences among several approaches (i.e., ignoring, aggregation, elimination, and applying a multilevel approach). Based on the severe differences among the analytical approaches, the authors recommend relying on the multilevel approach as the approach that fully considers the nested structure of the data.

The paper by Martina Mara, Markus Appel, and Timo Gnambs (2022) reports a meta-analysis in the field of human-robot interaction, especially on the Uncanny Valley Hypothesis, which claims that there is a curvilinear relation-ship between the evaluation of a robot as likable and the extent of anthropomorphism (i.e., the creation of the robot as human-like). According to the hypothesis, individuals will increasingly react with positive feelings to robots with their increasing human-likeness but – beyond a certain point – will tend to negatively react to even increasing likeness. At the extreme value of human-likeness, however, reactions will turn positive again. Using data from 49 (N = 3,556) studies and focusing of studies that used the most widely used instruments (i.e., the Godspeed scales), the polynomial meta-regression analysis shows a nonlinear but monotonic relationship between anthropomorphism and likeability. While these results could be interpreted as counterevidence to the uncanny valley hypothesis, the authors conclude that the limited amount of data addressing the extreme part of the anthropomorphism dimension is the most direct conclusion.

The article by Julia Norget and Axel Mayer (2022) addresses the question of how to evaluate the fit of structural equation model applied to longitudinal data, for instance, from experience sampling designs. Based on their

https://doi.org/10.1027/2151-2604/a000491

argument that common fit indices perform poorly in models estimated with many variables measured in experience sampling, the authors propose and analyze the value of a block-wise fit assessment for such models. The authors present two simulation models that support the value of their approach but also shows some limitation such that some misspecified models cannot be detected and that the assessment of fit naturally is focused and, thus, limited on misspecifications in the respective block.

In the fifth and final article, Robbie van Aert (2022) discusses the value and benefits of conducting an individual participants meta-analysis when analyzing multiple replications of studies conducted in different labs that have become prominent in the *many labs* project (Ebersole et al., 2016). He points out to the weaknesses of meta-analytical approaches to aggregate the results of these studies, most notably the lower statistical power when analyzing moderator effects (in the form of differences across studies) and – more importantly – the danger of aggregation biases, that is, falsely concluding differences between studies to differences between individuals. As a remedy, he provides an introduction and tutorial with R to individual participant meta-analysis, in which the primary data of all studies are modeled in a multilevel model in which differences across labs can be represented in the form of random effects.

Overall, we believe that the broadened methodological scope has contributed to presenting an illustration of a broad array of topics highlighting developments and emerging themes in psychology.

References

Bošnjak, M., & Erdfelder, E. (2018). Hotspots in Psychology – 2018 Edition. *Zeitschrift für Psychologie, 226*(1), 1–2. https://doi.org/10.1027/2151-2604/a000323

Bošnjak, M., & Gnambs, T. (2019). Hotspots in Psychology – 2019 Edition. *Zeitschrift für Psychologie, 227*(1), 1–3. https://doi.org/10.1027/2151-2604/a000350

Bošnjak, M., & Wedderhoff, N. (2020). Hotspots in Psychology – 2020 Edition. *Zeitschrift für Psychologie, 228*(1), 1–2. https://doi.org/10.1027/2151-2604/a000398

Bošnjak, M., Wedderhoff, N., & Steinmetz, H. (2021). Hotspots in Psychology – 2021 Edition. *Zeitschrift für Psychologie, 229*(1), 1–2. https://doi.org/10.1027/2151-2604/a000438

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., ... Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, Article 68015082. https://doi.org/10.1016/j.jesp.2015.10.012

Erdfelder, E., & Bošnjak, M. (2016). Hotspots in Psychology: A new format for special issues of the *Zeitschrift für Psychologie*. *Zeitschrift für Psychologie, 224*(3), 141–144. https://doi.org/10.1027/2151-2604/a000249

Klose, M., Steger, D., Fick, J., & Artelt, C. (2022). Decrypting log data: A meta-analysis on general online activity and learning outcome within digital learning environments. *Zeitschrift für Psychologie, 230*(1), 3–15. https://doi.org/10.1027/2151-2604/a000484

Mara, M., Appel, M., & Gnambs, T. (2022). Human-like robots and the uncanny valley: A meta-analysis of user responses based on the Godspeed scales. *Zeitschrift für Psychologie, 230*(1), 33–46. https://doi.org/10.1027/2151-2604/a000486

Norget, J., & Mayer, A. (2022). Block-wise model fit for structural equation models with experience sampling data. *Zeitschrift für Psychologie, 230*(1), 47–59. https://doi.org/10.1027/2151-2604/a000482

Stolwijk, I., Jak, S., Eichelsheim, V., & Hoeve, M. (2022). Dealing with dependent effect sizes in MASEM: A comparison of different approaches using empirical data. *Zeitschrift für Psychologie, 230*(1), 16–32. https://doi.org/10.1027/2151-2604/a000485

Van Aert, R. C. M. (2022). Analyzing data of a multilab replication project with individual participant data meta-analysis: A tutorial. *Zeitschrift für Psychologie, 230*(1), 60–72. https://doi.org/10.1027/2151-2604/a000483

Published online February 2, 2022

Holger Steinmetz
Faculty of Management
University of Trier
Universitätsring 15
54296 Trier
Germany
steinmetzh@uni-trier.de

Review Article

Decrypting Log Data

A Meta-Analysis on General Online Activity and Learning Outcome Within Digital Learning Environments

Maria Klose[1], Diana Steger[2], Julian Fick[3], and Cordula Artelt[1,4]

[1]Leibniz Institute for Educational Trajectories (LIfBi), University of Bamberg, Germany
[2]Department of Psychological Assessment, University of Kassel, Germany
[3]Institute for Communication Science and Institute of Educational Psychology, Technische Universität Braunschweig, Germany
[4]Department of Longitudinal Educational Research, University of Bamberg, Germany

Abstract: Analyzing log data from digital learning environments provides information about online learning. However, it remains unclear how this information can be transferred to psychologically meaningful variables or how it is linked to learning outcomes. The present study summarizes findings on correlations between general online activity and learning outcomes in university settings. The course format, instructions to engage in online discussions, requirements, operationalization of general online activity, and publication year are considered moderators. A multi-source search provided 41 studies (N = 28,986) reporting 69 independent samples and 104 effect sizes. The three-level random-effects meta-analysis identified a pooled effect of r = .25 p = .003, 95% CI [.09, .41], indicating that students who are more active online have better grades. Despite high heterogeneity, $Q(103)$ = 3,960.04, p < .001, moderator analyses showed no statistically significant effect. We discuss further potential influencing factors in online courses and highlight the potential of learning analytics.

Keywords: online learning, log data, learning analytics, academic achievement, meta-analysis

Until recently, face-to-face teaching and on-site exams were considered the gold standard in formal higher education. Universities face increased demand as the number of students constantly grows (Araka et al., 2020), and because of the COVID-19 pandemic, the need for online learning in higher education increased drastically (Ali, 2020). However, only little is known about how students use online classes and how their learning behavior is linked to learning outcomes. Online learning environments, so-called Learning Management Systems (LMS), make it easier to obtain information about students' learning behavior by analyzing automatically tracked log data from students' interactions with the LMS (Gašević et al., 2016). Although interest in learning analytics constantly grows (Dawson et al., 2014) and uses of trace data are increasing (Winne, 2020), the potentials of using process generated data for purposes of summative and formative assessments is yet an emerging research field in which common concerns about the potential limitations are addressed. Concerns relate to the weak relation between overt learning behavior operationalized via log data with complex learning behavior. Henrie and colleagues (2018) describe them as strongly simplified proxies for complex learning behavior, which ultimately questions their usefulness for the evaluation of online classes. Consequently, there is an ongoing debate whether log data are a valid predictor for learning outcomes (e.g., Agudo-Peregrina et al., 2014; Campbell, 2007), or whether they fail to predict learning outcomes – either because it is difficult linking log data to learning behavior, or because courses with substantial online elements are too heterogeneous to draw a general conclusion (e.g., Gašević et al., 2016; Macfadyen & Dawson, 2010). Although added value of using fine-grained and event-related process data for certain unobtrusive assessment purposes has been demonstrated (e.g., measuring intra-individual change; Barthakur et al., 2021), the use of log data for platform-, domain- and demand independent assessment of successful digital learning activities is still a matter of debate. Since systematic reviews about the value of broad log indicators are missing, the present meta-analysis summarizes findings on the relationship between these online activities derived from log data and learning outcomes within LMS. We focus on two broad log data indicators of general online activity (i.e., total login time and login frequency), which can be classified as access-related log events (Kroehne & Goldhammer, 2018) and are commonly used as measures linked to students' achievement (see You, 2016 for a detailed discussion of this issue). Accordingly, in the case of learning outcomes, we focus on indicators of learning success (i.e., course grade or course score).

https://doi.org/10.1027/2151-2604/a000484

General Online Activity and Learning Outcomes

One major advantage of online learning for educational research is the availability of a vast amount of information that can be derived automatically and unobtrusively. For example, courses offered via LMS allow collecting an enormous amount of log data about interactions with the platform (Campbell, 2007). However, the primary use of explorative approaches for analyzing log data results in a lack of theoretical grounding (Winne, 2020). Besides the number of studies comprising data-driven approaches for the decision which log data to examine, several researchers have considered pedagogical theories (see Tempelaar et al., 2015, for a review). For example, general online activities, such as *total login time* or *login frequency*, are considered indicators for learning engagement (Beer et al., 2010). Furthermore, according to Carroll's (1963) *model of school learning*, time spent on learning is one of the crucial factors for students' performance. Thus, time spent learning online might also be a crucial factor for web-based achievement (Jo et al., 2015). Besides, login frequency is associated with Engeström's (1987) *activity theory*, which states that mere activity produces meaningful learning, leading to higher learning outcomes (Kupczynski et al., 2011). Finally, total login time and login frequency are considered proxies for time management strategies, as they are indicators for sufficient time investment and, thus, key factors for performance (Jo et al., 2015). However, there is a debate on the type of log data suitable to measure learning behavior (Agudo-Peregrina et al., 2014). Critics suggest focusing on quality, not the quantity of online learning behavior (You, 2016): As active participation is crucial for success, indicators that do not distinguish between active and passive engagement are problematic (Ransdell & Gaillard-Kenney, 2009). Overall, these contradictory assumptions on the usefulness of broad log data indicators go along with inconsistent findings on the association between those indicators of general online activity and learning outcomes: some studies reported no association (Broadbent, 2016), negative correlations (e.g., Ransdell & Gaillard-Kenney, 2009; Strang, 2016), or positive correlations (e.g., Liu & Feng, 2011; McCuaig & Baldwin, 2012; Saqr et al., 2017). Other studies that examined various online courses simultaneously obtained mixed results across different courses (e.g., Conijn et al., 2017; Gašević et al., 2016), indicating that online courses might be too heterogeneous to draw a general conclusion about the link between general online activity and learning outcomes.

The Present Study

Our aim was to systematically review findings on the relationship between two log data indicators of general online activity and learning outcomes within LMS. To guarantee a minimum level of comparability between classes, we focused on online university courses because they are more structured than informal online courses (Song & Bonk, 2016) and because they are usually graded, offering a measure of learning outcomes. Regarding general online activity, we focused on total login time and login frequency to assess the applicability of these broad measures derived from log data as a proxy of online learning behavior and examine how they are linked to educational outcomes (i.e., course grade or course score). Moreover, we examined the impact of several moderators to explain the inconsistent findings reported in previous literature, since in the course of recent technological developments (Palvia et al., 2018), teaching tools became more sophisticated and online courses became more diverse.

First, we use the term "online course" to describe all courses that include substantial online elements, that is, courses that are taught exclusively online (*fully online* courses), and courses that combine online and face-to-face delivered content (*blended* courses; Allen & Seaman, 2014). Compared to fully online courses, blended courses offer more structure through regular face-to-face sessions (Means et al., 2013), as well as the opportunity to easily get in touch with peers regularly (Broadbent, 2016). However, the varying parts of face-to-face teaching in blended courses cannot be tracked via log data (Mwalumbwe & Mtebe, 2017). Therefore, we expect a stronger relationship between general online activity and learning outcomes within fully online than within blended learning courses, as for the latter, substantial parts of learning might remain unreflected by log data (Hypothesis 1).

Second, online courses vary with respect to the emphasis put on the use of online discussion boards. An explicit instruction to use discussion boards implemented within the LMS might foster deeper learning while being online and lead to better achievement (Song et al., 2019). Although interactions in discussion forums are considered an essential part of learning (Uijl et al., 2017), the mere existence of a discussion board is not enough for promoting active participation within the LMS (Lee & Martin, 2017). Since active content engagement is crucial for students' achievement (Ransdell & Gaillard-Kenney, 2009), we expect higher correlations between general online activity and learning outcomes for courses with instruction for discussion board usage (Hypothesis 2).

Third, online courses differ in their grading systems. For the present study, it is important to what extent online activities are explicitly incentivized. While some courses do not incentivize online participation at all, other courses either offer bonus points for regular online participation or even require online activities within the LMS (e.g., online group discussions, quizzes, or online assignments) as part of the final grade. These incentives encourage active online engagement (Tempelaar et al., 2019) and offer guidance for students to effectively use the tools and materials provided within the LMS. Therefore, we expect a higher correlation between general online activity and learning outcomes for courses that incentivize certain online activities through their grading systems (Hypothesis 3).

Fourth, we compare the operationalization of general online activity as total login time or as login frequency. Ever since the growing popularity of investigating log data, there has been the challenge of capturing log data that might be translated into psychologically meaningful variables (Seifert et al., 2018). As both operationalizations are theoretically reasoned with either Carroll's (1963) *model of school learning* or Engeström's (1987) *activity theory*, we want to explore if general online activity operationalized as *total login time* versus *login frequency* differs in their relationship with learning outcomes (Hypothesis 4).

Lastly, we considered publication year as a potential moderator. In the face of rapid technological advancements (Palvia et al., 2018), we expect changes in how LMS provides education. Through technological change, LMS tools become more advanced, and multiple types of learning tools can be implemented to foster students' active engagement within the LMS (Kebritchi et al., 2017). Therefore, students ought to be enabled to benefit from a more interactive online learning experience, and online activity within recent studies might result in higher achievements than within older studies (Hypothesis 5).

Method

In accordance with common open science practices, we provide all additional materials (i.e., coding manual, syntax, data, PRISMA20-checklist, and supplemental figures and tables) online within the Open Science Framework (Center for Open Science, 2021).

Literature Search and Study Selection

We identified 33,724 potentially relevant studies from electronic searches in major scientific databases (PsycINFO, PsycArticles, PSYNDEX, and ERIC) and Google Scholar using the Boolean search term (*online learning OR online course* OR web-based learning OR e-learning OR elearning OR learning management system* OR LMS OR learning analytics*) *AND* (*achievement OR performance OR outcome*) in February 2021. We retrieved five further studies by calls for unpublished work (via mailing list of the German Psychological Society, ResearchGate, and Twitter). Additionally, we performed a "rolling snowball" search and identified 17 further studies by screening the reference lists of all eligible studies and by conducting forward citation tracking using Google Scholar. Finally, we contacted 19 corresponding authors of studies not reporting bivariate correlations and received them for three studies. We included all published or unpublished types of studies. See Figure S1 for the detailed literature search process, including specifications of all sources that were searched. Subsequently, these studies were included in the analysis depending on the following inclusion criteria: (a) The study investigated a fully online or blended course in an institutional setting; (b) General online activity was measured using log data and operationalized as total login time or login frequency (i.e., number of single logins or number of days with at least one login); (c) Learning outcome was measured as course grade or course score; (d) The study consisted of a sample comprising university students; (e) The study was published between 1969 (year of the first connection of the Internet) and 2021, and (f) was written in English or German; (g) The study reported at least one correlation between general online activity and learning outcome or appropriate statistics that could be transformed into correlations. Studies were excluded if: (h) General online activity was measured as a self-report as we focused on the usefulness of log data indicators of general online activity, or (i) measured as the duration or frequency of single activities in the LMS (e.g., time spent on quizzes, number of forum postings) because these types of log data fall within different categories (such as response-related or process-related log data; see Kroehne & Goldhammer, 2018), and (j) the study had a commercial e-learning course as setting as we focused on the specific context of higher education. After applying these criteria, 41 primary studies remained (see Table S1). Study selection followed a two-stage process. First, two researchers reviewed titles and abstracts of the first 50 records and discussed disagreements about eligibility until consensus was reached. Then, one researcher screened all titles and abstracts of all studies retrieved. In cases in which eligibility was unclear, the study was considered for the second stage in the form of a full-text review. A sample of full-text studies (~15%; 36/241) was independently screened by two researchers. The remaining full-text studies were screened by one researcher. Finally, the second researcher independently reviewed all included studies and those with uncertain eligibility. Again, disagreements about eligibility were resolved through discussion.

Coding Process

We developed a standardized coding manual and data extraction sheet for the data collection process (see Center for Open Science, 2021, for detailed information). The coding manual comprised eligibility criteria, guidelines for selecting effect sizes and coding, and definitions of all outcomes and other variables for which data were sought (i.e., name and description of the respective variables, guidelines regarding the format of coding, and coding examples). For each study, we extracted all relevant effect sizes for the association between general online activity and learning outcomes. Moreover, we collected data on study and sample and online course characteristics covering especially moderator variables (i.e., course format, emphasis of discussion, course activities as part of grading, operationalization of general online activity, and publication year) and general study information.

All studies were coded twice using the coding manual and data extraction sheets by two independent raters. To evaluate the coding process, Cohen's (1960) κ for categorical variables and intraclass coefficients (ICC; Shrout & Fleiss, 1979) for continuous variables were calculated for the focal variables. Interrater reliability for the effect size was ICC = .89, 95% CI [.84, .92] and for the sample size and publication year ICC = 1, 95% CI [1, 1]. The Cohen's κ for the remaining categorical variables was .91, overall indicating strong to excellent intercoder agreement (LeBreton & Senter, 2008). All discrepancies were solved upon discussion by comparing extracted data.

Statistical Analyses

Effect Size

We used Pearson product-moment correlation as an effect size measure. Because transforming standardized weights from multiple linear regression analyses into correlation coefficients is problematic (Aloe, 2015), authors from studies reporting only regression weights were contacted to obtain correlations. If no correlation was available, the study was excluded from the analyses (k = 6). To standardize the direction of effects, we conversed effect sizes in cases where learning outcomes were conceptualized as smaller numbers indicating better achievement.

Meta-Analytic Model

We pooled effect sizes using a random-effects model with a restricted maximum likelihood estimator (Viechtbauer, 2010). A three-level meta-analysis was conducted to account for dependent effect sizes (Cheung, 2014) because some studies reported more than one effect size (e.g., provided correlations for both operationalizations of general online activity for a given sample). Dependencies between effect sizes derived from the same sample are acknowledged by decomposing the total random variance into two variance components: one reflecting the heterogeneity of effects within samples, and the other indicating heterogeneity of effect sizes between samples (see Gnambs & Appel, 2018 for a detailed description). We calculated I^2 statistics to quantify heterogeneity in observed effect sizes (Higgins et al., 2003). Considering I^2 is not an absolute measure of heterogeneity (Borenstein et al., 2017), we additionally report the Q-statistics. Since using sample size weights performs best for estimating the random-effects variance component in meta-analytic models with correlations as effect size measures (Brannick et al., 2011), we used this weighting procedure to account for sampling error. We reported our findings focusing on the size of the effect and its confidence and prediction interval. To visualize our meta-analysis, we used a forest plot (Viechtbauer, 2010). Lastly, we conducted subgroup and meta-regression analyses to examine moderating effects on the pooled effect size (Harrer et al., 2019), given the diversity of online courses being investigated. Therefore – apart from publication year –, we categorized the included studies along with dichotomous moderators: fully online vs. blended course format, instructed vs. not instructed discussion board usage, graded activities vs. no requirements, and total login time vs. login frequency as general online activity.

Sensitivity Analyses

First, we used the studentized deleted residuals (Viechtbauer & Cheung, 2010) to identify extreme correlations. Additionally, we conducted sensitivity analyses that removed the identified outliers from the analyses to examine the impact of these outliers. Moreover, the robustness of the presented meta-analysis was investigated by removing two particular studies that differed in their conceptualization (i.e., Lauría et al., 2012: correlation comprised data from an entire university; and Mödritscher et al., 2013: log data from only two weeks before the examination) from the meta-analytic database and comparing the pooled effect to the pooled effect from the full database.

Publication Bias

The presence of potential publication bias was investigated in two ways: First, we performed a meta-regression with the publication type as a moderator. Effect sizes from peer-reviewed sources were compared to effect sizes from other sources (i.e., theses or conference papers). A statistically significant difference between effect sizes extracted from both sources could result from a distortion in the peer-reviewed research literature due to systematic suppression of (e.g., nonsignificant) effects. Second, we conducted a modified regression test for asymmetry by including a precision measure (i.e., $1/n$) as a moderator in the meta-analytic model to account for dependent effect sizes.

Statistical Software

All analyses were conducted using *R* version 4.0.4 (R Core Team, 2021). Meta-analytic models were estimated with the *metafor* package version 2.4-0 (Viechtbauer, 2010).

Results

Descriptive Statistics

The meta-analysis is based on 41 studies published between 1997 and 2021, predominantly as peer-reviewed articles (73%). The remaining studies appeared as theses (5%) or conference papers (22%). The database covered 69 independent samples that provided 104 effect sizes, with each sample comprising between 1 and 3 effect sizes. Overall, the meta-analysis included scores from 28,986 students (range of samples' *n*s: 11–11,195, *Mdn* = 122). The mean age was 22.21 years, and 53.71% of the students were female, however, only 11 studies reported information on age, and 20 studies information on gender. The duration of the courses varied between 6 and 19 weeks (*Mdn* = 12 weeks), mainly covering one academic semester. Moreover, courses varied with respect to their format (24% fully online, 72% blended, 1% not reported separately, or 3% missing), emphasis of discussion (18% instructed use of discussion boards, 69% discussion boards available within the LMS without further instructions, 10% not mentioned, or 3% missing), and requirements (45% online activities within the LMS as part of grading, 54% none, or 1% missing). In 44% of the cases, general online activity was operationalized as total login time and 56% as login frequency (number of single logins or number of days with at least one login).

Overall Pooled Correlation

In total, the three-level random-effects meta-analysis identified a pooled correlation of $r = .25$ $p = .003$, 95% CI [.09, .41], indicating that students who are more active online also have a better learning outcome (Figure 1). The result of the pooled correlation was robust and replicated in the separate moderator analyses (Table 1).

Overall, these findings indicate a small but statistically significant positive association between general online activity and learning outcomes. Yet, the high random variance resulted in an exceedingly large prediction interval around the pooled effect, 80% PI [−.10, .59]. Hence, we conducted sensitivity analyses to examine the impact of certain studies on the prediction interval. Further, the studies showed higher between-cluster heterogeneity (I^2 = .92; see also Figure 1 for an illustration of the variability between samples) compared to within-cluster-heterogeneity (I^2 = .05), indicating pronounced unaccounted differences between samples that might be explained by moderator analyses, but negligible variability within samples (Higgins et al., 2003).

Moderator Analyses

We conducted meta-regression analyses to examine the effects of course format, emphasis of discussion, requirements, operationalization of general online activity, and publication year on the pooled effect (Table 2). On effect size level, correlations between moderators ranged from −.54 to .46, indicating negligible multicollinearity. None of the moderators was statistically significant. This result remained the same even when each moderator was examined separately (Table 1). Overall, moderator analyses showed no effect, indicating that our data do not provide evidence in favor of a moderating effect of course format, emphasis of discussion, requirements, operationalization of general online activity, or publication year on the relationship between general online activity and learning outcome.

Sensitivity Analyses

We performed sensitivity analyses to examine the robustness of our findings. In a first step, the robustness of the presented results was investigated by removing nine extreme correlations (i.e., outliers with $z > 1.96$; Viechtbauer & Cheung, 2010; Figure S2) from the database to compare this pooled effect to the original pooled effect. After eliminating these effects from the database, the pooled effect was $r = .24$, $p < .001$. The 80% PI decreased from [−.10, .59] to [.04, .43], indicating a reduced random variance. However, the outliers did not distort the pooled effect. Similar patterns appeared for all subgroup analyses (see Table S2). Overall, the outlier analyses provided evidence for the robustness of the correlation between general online activity and learning outcomes. In a second step, sensitivity analyses with respect to two studies that differ in their conceptualizations from the other included studies resulted in negligible differences: $r = .29$, $p < .001$, 80% PI [−.05, .62] (Lauría et al., 2012), and $r = .24$, $p = .011$, 80% PI [−.11, .60] (Mödritscher et al., 2013), also indicating the robustness of the present meta-analysis (see Table S3).

Publication Bias

First, the meta-regression analysis that we conducted to examine publication bias indicated no statistically significant difference between effect sizes extracted from peer-reviewed versus other sources ($\gamma = -0.06$, $SE = 0.13$, $p = .624$). Second, the modified regression test for asymmetry ($\gamma = 7.93$, $SE = 8.57$, $p = .355$) revealed no statistically

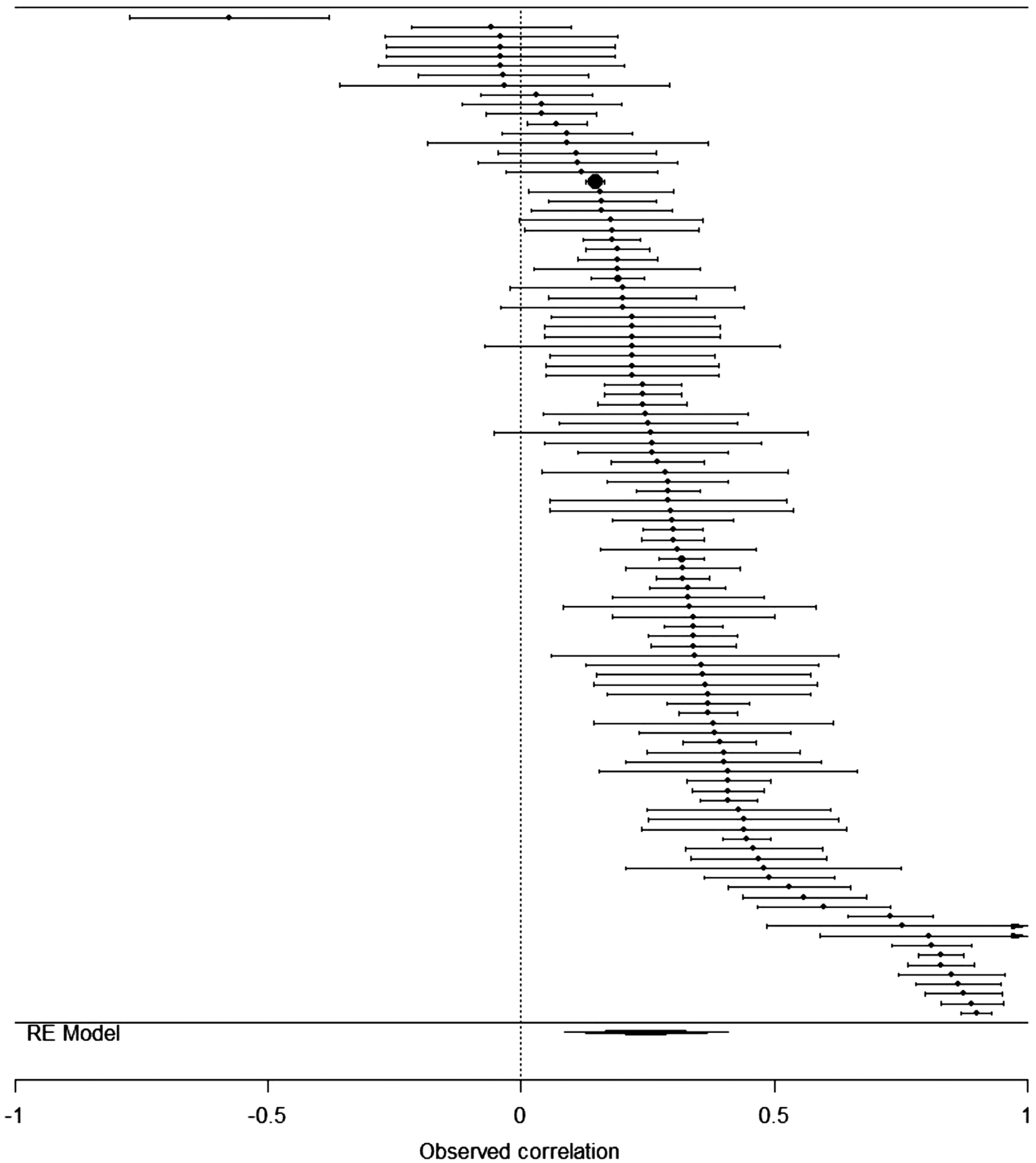


Figure 1. Forest plot. *Note.* Effect sizes are ordered by increasing magnitude. Larger symbols illustrate larger sample sizes.

significant effect for measurement precision. Overall, we did not find evidence of publication bias.

Discussion

Online courses are more important than ever (Ali, 2020), and they provide the possibility to conveniently and unobtrusively record log data (Dawson et al., 2014). Yet, it is unclear how log data contribute to explaining the linkage of learning behavior to academic achievement. Although several studies have examined the association between general online activities and learning outcomes, their findings are ambiguous (e.g., Broadbent, 2016; Campbell, 2007; Gašević et al., 2016). Hence, we provided a systematic review of existing findings and investigated several potential moderators to explain ambiguity in previous literature: In a first step, we identified a small – yet statistically significant – pooled correlation of r = .25 between general

Table 1. Meta-analysis on general online activity and learning outcome and separate moderator analyses

	k_1	k_2	N	r	SE_r	95% CI	Q_M	Q	$\sigma^2_{(2)}$	$\sigma^2_{(3)}$	$I^2_{(2)}$	$I^2_{(3)}$	80% PI
Overall	104	69	28,986	.25*	.08	[.09, .41]		3,960.04*	.06	< .01	.92	.05	[−.10, .59]
Course format						[−.27, .18]	0.16						
Fully online	25	20	4,816	.32*	.08	[.17, .48]		125.03*	.04	< .01	.91	< .01	[.05, .60]
Blended	75	46	12,543	.28*	.07	[.15, .41]		2,919.10*	.07	< .01	.93	.04	[−.08, .63]
Discussion board usage						[−.20, .27]	0.07						
Instructed	19	14	1,077	.28*	.08	[.11, .44]		112.98*	.06	< .01	.86	< .01	[−.05, 61]
Not instructed[a]	82	52	16,475	.29*	.06	[.17, .40]		3,052.80*	.06	< .01	.91	.06	[−.05, .63]
Requirements						[−.22, .20]	0.01						
Graded activities	47	35	7,673	.29*	.10	[.10, .48]		2,313.94*	.10	< .01	.95	.03	[−.13, .72]
None	56	33	10,118	.28*	.05	[.19, .37]		420.52*	.03	< .01	.80	.10	[.06, .51]
General online activity						[−.29, .18]	0.21						
Total login time	46	46	11,825	.29*	.07	[.16, .40]		2,320.63*	.04	.04	.49	.49	[−.09, .66]
Login frequency	58	53	24,842	.23*	.07	[.09, .37]		585.41*	.02	< .01	.81	.11	[.01, .45]
Publication year[b]						[−.03, .05]	0.23						
Older than 2017	43	31	19,860	.22*	.09	[.04, .40]		501.64*	.03	< .01	.88	.07	[−.05, .49]
Newer than 2017	61	38	9,126	.29*	.08	[.14, .44]		2,495.45*	.08	< .01	.93	.05	[−.09, .67]

Note. k_1 = Number of effect sizes; k_2 = Number of samples; r = Pooled correlation; SE_r = Standard error of r; 95% CI = 95% confidence interval of r; Q = test of heterogeneity ($df = k_1 - 1$); Q_M = test statistic for the omnibus test of coefficients (df = 1); $\sigma^2_{(2)}$ = Random effect of r between samples; $\sigma^2_{(3)}$ = random effect of r within samples, $I^2_{(2)}$ = Proportion of between-cluster heterogeneity; $I^2_{(3)}$ = proportion of within-cluster heterogeneity; 80% PI = 80% prediction interval of r. [a]Includes the categories of discussion board usage *available* and *not mentioned*. [b]For illustrative purposes, subgroup analyses are reported for older versus newer studies based on a median split. *p < .05.

Table 2. Moderator analysis including all five moderator variables simultaneously

	Moderator analysis			Correlations			
	γ	SE_γ	z	(1)	(2)	(3)	(4)
Intercept	−5.55	25.28	−0.22				
(1) Course format (1 = blended; 0 = fully online)	−0.07	0.13	−0.50				
(2) Discussion board usage (1 = instructed; 0 = available, not mentioned)	−0.04	0.13	−0.28	−.54			
(3) Requirements (1 = none; 0 = graded activities)	−0.03	0.10	−0.25	.12	−.27		
(4) General online activity (1 = login frequency; 0 = total login time)	0.01	0.05	0.12	−.16	.12	.12	
(5) Publication year (metric)	< 0.01	0.01	0.23	.46	−.45	.12	−.28
Q_M		0.31					
$\sigma^2_{(2)}/\sigma^2_{(3)}$		0.06/< 0.01					
k_1/k_2		100/66					

Note. Phi coefficients for dichotomous moderator variables and point-biserial coefficients for dichotomous and metric moderator variables on effect size level are displayed. The correlations are based on 100–104 effect sizes. γ = Fixed effects regression weight; SE_γ = Standard error of γ; Q_M = test statistic for the omnibus test of coefficients (df = 5); $\sigma^2_{(2)}$ = Random effect of r between samples; $\sigma^2_{(3)}$ = Random effect of r within samples; k_1 = Number of effect sizes; k_2 = Number of samples.

online activity and learning outcome. This finding indicates that students who are online for a longer time (or more often) within the LMS also tend to have better course grades. This effect might seem small at first, but it remained robust across sensitivity analyses even though we used very broad indicators of general online activity. Additionally, academic success in itself is extremely complex and, therefore, difficult to predict (see Alyahyan & Düştegör, 2020, for a review). Comparing our results to a meta-analysis examining multiple psychological correlates of university students' academic performance (Richardson et al., 2012), log data indicators perform better in predicting academic success as compared to demographics (i.e., gender, age, and socioeconomic status) and personality traits (except for conscientiousness), but perform slightly worse than prior academic performance and academic self-efficacy. The strongest correlate of all 50 measures was performance self-efficacy (r = .59). Against this background, the present analysis demonstrates the potential of log data, given that even two broad log data indicators of online learning behavior are associated with the learning outcome. However, the meta-analytic model revealed high heterogeneity between studies that could not be explained by moderator analyses. Therefore, we discuss reasons why our moderator variables might have failed to explain high heterogeneity and other possible sources of variance.

Limitations of the Included Moderators

First, our potential moderator variables were restricted to broad course characteristics, which can be illustrated by the variable *course format.* The dichotomous classification of *blended* versus *fully online format* might be too coarse as there exist flowing transitions depending on the portion of content delivered online (Allen & Seaman, 2014) and the share of online elements in a certain course might be better depicted as a continuous variable rather than a dichotomous one. However, most studies on online courses only provide very superficial characteristics. Given the fact that faculties struggle with the transition to online teaching (Kebritchi et al., 2017), more evidence is needed so practitioners who design online courses are able to make informed choices to improve quality of online learning.

Second, there is a lack of information on contextual variables (e.g., instructional design, interactive tools, or synchronicity) reported in primary studies. Varying contexts and tools affect the learning process by providing different learning opportunities which are decisive for improved learning (Lust et al., 2012). The consideration of contextual factors might help explain ambiguous findings in the current literature (Gašević et al., 2016) as they enable more comprehensive moderator analyses. Our moderator analyses were limited to basic information about how the LMS was implemented. In the following, we discuss which aspects might help explain differences between settings.

Potential Other Moderators

An overall structure offered by online courses might help reduce individual differences in online learning behavior, as it provides guidance for students to engage in the most beneficial activities at a certain point of course processing (Winne, 2004). On the individual level, instructors can help to reduce existing heterogeneity within the association between general online activity and learning outcome, as not all learners seem to benefit equally from learning opportunities (Lust et al., 2012). Learners need to be instructed how to use LMS (Kebritchi et al., 2017) to exploit the full potential of online courses. Other forms of structure are, for example, shares of synchronous methods and applications in online courses (e.g., teaching sessions, collaborative learning, or support and monitoring by a tutor; see Kinshuk & Chen, 2006), which provide a structured schedule for learning behavior, or any form of online assessment across the course duration to monitor students' learning progress or to provide personalized feedback (Knight, 2020), or to encourage students' engagement (Tempelaar et al., 2019). In the present literature, systematic information on the extent of the structuredness of online courses is unfortunately limited. Future research might document the effects of these course characteristics and their impact on learning behavior.

Another aspect of online course design comprises the incentives that are used to ensure students' participation. Apart from the extent to which participation is included in course grading, only little is known about how instructors use incentives for constant participation throughout the course. One example for these incentives is gamification – the implementation of game-design principles and elements in non-game environments (Deterding et al., 2011) – which can promote motivation (see Mora et al., 2017 for a review) for example by providing visualized immediate feedback to the learner on goal completion or students' learning progress compared to other students. Gamification for educational purposes can be associated with increased activity (e.g., Hamari, 2017; Huang & Hew, 2015) or general engagement in online programs (Looyestyn et al., 2017). But it remains unclear if a game-based induced increase in online activity automatically leads to improvement in learning. If practitioners systematically provide online courses with and without different types of gamification, future research could examine differences in the online learning activity and its impact on learning outcomes.

Finally, our meta-analysis was based on log data indicators that specify the extent of total login time or login frequency within LMS over an entire academic semester. Differences in the distribution of online activity across the course duration could not be considered. As distributed learning is a more efficient learning strategy than cramming before examinations for an equal amount of time (Dunn et al., 2013), future research should address the mere amount of activity and the distribution of total login time or logins in a more fine-grained way. Additionally, students' diversity and consistency of online activities might provide substantial insights into how students' activity affects learning (Lust et al., 2012).

Log Data in Educational Research

The present meta-analysis can be seen as a starting point of how log data can be used to contribute to our understanding of complex variables like academic achievement by linking course outcomes to broad log data indicators of online learning behavior. However, given the increase of online education in higher education and the recent technological development (Kebritchi et al., 2017) there undoubtedly exist more fine-grained data in research than overall participation measures comprising platform usage. One big advantage of the use of log data is that large amounts of data are easily and immediately accessible so that learning analytics can draw on detailed and extensive log data about learners' studying activities within modern software tools (Winne, 2010). Although log data are a more accurate reflection of the quantity of media use than self-reports (Parry et al., 2021), researchers' degrees of freedom in data tracking, collection and analysis persist and thereby limiting the objectiveness of log data indicators (Avella et al., 2016). Moreover, the biggest issue is the connection with existing educational theories and the resulting necessity to consider the reliability of log data as well as its role in claims about validity based on this kind of data (Winne, 2020).

But how can learning analytics meet the vision to help improving learning and teaching by using generated data as people engage in learning? On the student-centered level, learning analytics facilitate predictive modeling of course completion (Clow, 2013). Information on student's previous educational experience and demographics, as well as data on online activity and formative and summative assessment, are combined and then used to develop interventions designed to improve retention and performance. This enables early intervention systems and personalized learning, as students receive real-time feedback on their learning progress (Arnold & Pistilli, 2012). Otherwise, instructors can take advantage of learning analytics by using them to identify areas in need of improvements regarding the curriculum as well as their own performance (Avella et al., 2016). Finally, the implementation of new tools or mechanisms can be checked (Song, 2018). Multiple types of learning tools enhance the learning experience (Hathaway, 2014), but it is not always *the more, the better*. Instructors have to consider which tool, design element, or multimedia will add to the learning process and which ones are distracting (Kebritchi et al., 2017).

Limitations of the Present Study and Implications for Future Research

Research on learning analytics is a promising approach for an advanced understanding of the learning process (Gašević et al., 2015). This meta-analysis provides an initial insight into the value of broad log data indicators of learning behavior. However, the present analyses come with limitations. Specifically, recent developments in meta-analytic methods suggest that it might be more adequate to model the hierarchical structure of the data by including the covariances of effect sizes derived from the same sample in the model, rather than using the default model, which assumes no covariances. On a more general stance, data dependency is an important issue in meta-analytic research that is often neglected (e.g., Rodgers & Pustejovsky, 2021). While these advanced models might be even better suited to model the data structures, many issues arise not from inappropriate modeling choices but rather from shortcomings in the primary studies. How should future research look like in order to contribute to a more advanced understanding of online learning?

While the number of studies using log data increases steadily, only a few of these studies transparently describe their methodologies for data collection and cleaning, utilized measures or analyses (Bergdahl et al., 2020). In general, learning analytics has to face challenges of heterogeneous data sources and the lack of unified vocabulary (Papamitsiou & Economides, 2014). Future meta-analyses could make use of quality assessments of primary studies, something that is already common for assessing the methodological quality of intervention effectiveness research (e.g., Valentine & Cooper, 2008). Scheffel and colleagues (2014) have proposed a framework of quality indicators for learning analytics earlier on. Future meta-analyses would benefit from a standardized procedure that allows taking the methodological quality of learning analytics studies into account for weighting procedures as well as the decision whether to include or exclude a primary study. However, learning analytics can benefit from a unified framework for the use of terms and definitions, operationalizations, and methodological procedures.

Another promising development to overcome central issues (i.e., data access and transparency) in meta-analyses is the open data movement (Gurevitch et al., 2018). As soon as researchers follow standards regarding an open scientific process, design standards would reduce unclearly reported methodologies, and data sharing standards would enable to directly generate effect sizes from open data (Nosek et al., 2015).

Moreover, open science practices facilitate multi-level analyses based on raw data in the form of meta-analysis of individual participant data (IPD; e.g., Riley et al., 2010). IPD meta-analyses are considered the gold standard as they prevent aggregation biases and enable to look at the impact of heterogeneity that originates from differences within studies (Kaufmann et al., 2016). Due to the COVID-19 pandemic, educational institutions were forced to shift teaching to online learning (De' et al., 2020), potentially with an increase in the usage of interactive tools (such as video conferencing tools) that could facilitate online discussions or group work, which may also lead to an increase in the quality of online courses. Hopefully, these changes will be documented in forthcoming research on online courses. Future research might comprise large collaborations and centrally coordinated data collections within online courses to benefit from the incoming data due to the digital surge so that they might gain deeper insights to improve the quality of online learning.

Apart from that, this review focused on formal higher education. However, informal learning gains more attention in the field of educational research (Zheng et al., 2019), and therefore is a promising extension. Especially the change toward online learning promotes informal learning (i.e., learner-directed and independent learning outside of formal educational contexts; Song & Bonk, 2016). However, due to the absence of external assessment within informal learning (Callanan et al., 2011), studies with an informal context could not be included in the present meta-analysis. Even though formal and informal learning lead to gains in knowledge and skills (Cerasoli et al., 2018), it is difficult to combine them in meta-analyses as their outcomes are operationalized differently since for informal learning, educational success is traditionally defined as course completion (Henderikx et al., 2017). Accordingly, it seems worthwhile to examine whether the positive association between general online activities and learning outcomes can be transferred to an informal context.

Conclusion

In summary, we identified an association between broad log data indicators of general online activity and learning outcomes. Although several sensitivity analyses indicated the robustness of the present meta-analysis, the high heterogeneity between studies could not be explained by our moderator variables, which were limited to basic information on course implementation. We recommend for future research to form bigger collaborations and centrally collect data to conduct IPD meta-analyses to gain deeper insight into online learning. Learning analytics have the potential to provide more fine-grained data, but it is necessary to connect generated data to existing educational theories.

References

Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á., & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior, 31*, 542–550. https://doi.org/10.1016/j.chb.2013.05.031

Ali, W. (2020). Online and remote learning in higher education institutes: A necessity in light of COVID-19 pandemic. *Higher Education Studies, 10*(3), 16–25. https://doi.org/10.5539/hes.v10n3p16

Allen, I. E., & Seaman, J. (2014). *Grade change: Tracking online education in the United States* (Annual Report No. 11). Babson Survey Research Group and Quahog Research Group, LLC. https://eric.ed.gov/?id=ED602449

Aloe, A. M. (2015). Inaccuracy of regression results in replacing bivariate correlations. *Research Synthesis Methods, 6*(1), 21–27. https://doi.org/10.1002/jrsm.1126

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*, Article 3. https://doi.org/10.1186/s41239-020-0177-7

Araka, E., Maina, E., Gitonga, R., & Oboko, R. (2020). Research trends in measurement and intervention tools for self-regulated learning for e-learning environments – Systematic review (2008–2018). *Research and Practice in Technology Enhanced Learning, 15*, Article 6. https://doi.org/10.1186/s41039-020-00129-5

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge – LAK '12*, 267–270. https://doi.org/10.1145/2330601.2330666

Avella, J. T., Kebritchi, M., Nunn, S. G., & Kanai, T. (2016). Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. *Online Learning, 20*(2), 13–29. https://eric.ed.gov/?id=EJ1105911

Barthakur, A., Kovanovic, V., Joksimovic, S., Siemens, G., Richey, M., & Dawson, S. (2021). Assessing program-level learning strategies in MOOCs. *Computers in Human Behavior, 117*, Article 106674. https://doi.org/10.1016/j.chb.2020.106674

Beer, C., Clark, K., & Jones, D. (2010). Indicators of engagement. In C. H. Steel, M. J. Keppell, P. Gerbic, & S. Housego (Eds.), *Curriculum, technology & transformation for an unknown future. Proceedings ascilite Sydney 2010* (pp. 75–86). https://ascilite.org/conferences/sydney10/procs/Beer-full.pdf

Bergdahl, N., Nouri, J., Karunaratne, T., Afzaal, M., & Saqr, M. (2020). Learning analytics for blended learning: A systematic review of theory, methodology, and ethical considerations. *International Journal of Learning Analytics and Artificial Intelligence for Education, 2*(2), 46–79. https://doi.org/10.3991/ijai.v2i2.17887

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Brannick, M. T., Yang, L.-Q., & Cafri, G. (2011). Comparison of weights for meta-analysis of *r* and *d* under realistic conditions. *Organizational Research Methods, 14*(4), 587–607. https://doi.org/10.1177/1094428110368725

Broadbent, J. (2016). Academic success is about self-efficacy rather than frequency of use of the learning management system. *Australasian Journal of Educational Technology, 32*(4), 38–49. https://doi.org/10.14742/ajet.2634

Callanan, M., Cervantes, C., & Loomis, M. (2011). Informal learning. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(6), 646–655. https://doi.org/10.1002/wcs.143

Campbell, J. P. (2007). *Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study* (Publication No. 3287222) [Doctoral dissertation]. Purdue University. ProQuest Dissertations & Theses.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723–733. https://psycnet.apa.org/record/1963-08222-001

Center for Open Science. (2021). *Open Science Framework*. https://osf.io/wy2px/

Cerasoli, C. P., Alliger, G. M., Donsbach, J. S., Mathieu, J. E., Tannenbaum, S. I., & Orvis, K. A. (2018). Antecedents and outcomes of informal learning behaviors: A meta-analysis. *Journal of Business and Psychology, 33*, 203–230. https://doi.org/10.1007/s10869-017-9492-y

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*(2), 211–229. https://doi.org/10.1037/a0032968

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education, 18*(6), 683–695. https://doi.org/10.1080/13562517.2013.827653

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies, 10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Dawson, S., Gašević, D., Siemens, G., & Joksimovic, S. (2014). Current state and future trends: A citation network analysis of the learning analytics field. *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge – LAK '14*, 231–240. https://doi.org/10.1145/2567574.2567585

De', R., Pandey, N., & Pal, A. (2020). Impact of digital surge during Covid-19 pandemic: A viewpoint on research and practice. *International Journal of Information Management, 55*, Article 102171. https://doi.org/10.1016/j.ijinfomgt.2020.102171

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, 9–15. https://doi.org/10.1145/2181037.2181040

Dunn, D. S., Saville, B. K., Baker, S. C., & Marek, P. (2013). Evidence-based teaching: Tools and techniques that promote learning in the psychology classroom. *Australian Journal of Psychology, 65*(1), 5–13. https://doi.org/10.1111/ajpy.12004

Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research* [Doctoral dissertation (monograph)]. Orienta-Konsultit.

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education, 28*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Gnambs, T., & Appel, M. (2018). Narcissism and social networking behavior: A meta-analysis. *Journal of Personality, 86*(2), 200–212. https://doi.org/10.1111/jopy.12305

Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature, 555*(7695), 175–182. https://doi.org/10.1038/nature25753

Hamari, J. (2017). Do badges increase user activity? A field experiment on the effects of gamification. *Computers in Human Behavior, 71*, 469–478. https://doi.org/10.1016/j.chb.2015.03.036

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing meta-analysis in R: A hands-on guide*. https://doi.org/10.5281/zenodo.2551803

Hathaway, K. L. (2014). An application of the seven principles of good practice to online courses. *Research in Higher Education Journal, 22*, 1–12. https://eric.ed.gov/?id=EJ1064101

Henderikx, M. A., Kreijns, K., & Kalz, M. (2017). Refining success and dropout in massive open online courses based on the intention – behavior gap. *Distance Education, 38*(3), 353–368. https://doi.org/10.1080/01587919.2017.1369006

Henrie, C. R., Bodily, R., Larsen, R., & Graham, C. R. (2018). Exploring the potential of LMS log data as a proxy measure of student engagement. *Journal of Computing in Higher Education, 30*, 344–362. https://doi.org/10.1007/s12528-017-9161-1

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Jorunal, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Huang, B., & Hew, K. F. (2015). Do points, badges and leaderboard increase learning and activity: A quasi-experiment on the effects of gamification. *Proceedings of the 23rd International Conference on Computers in Education*, 275–280. https://www.researchgate.net/profile/Khe-Hew/publication/286001811_Do_points_badges_and_leaderboard_increase_learning_and_activity_A_quasi-experiment_on_the_effects_of_gamification/links/5665404708ae15e746333d22/Do-points-badges-and-leaderboard-increase-learning-and-activity-A-quasi-experiment-on-the-effects-of-gamification.pdf

Jo, I.-H., Kim, D., & Yoon, M. (2015). Constructing proxy variables to measure adult learners' time management strategies in LMS. *Journal of Educational Technology & Society, 18*(3), 214–225. https://www.jstor.org/stable/jeductechsoci.18.3.214

Kaufmann, E., Reips, U.-D., & Maag Merki, K. (2016). Avoiding methodological biases in meta-analysis: Use of online versus offline individual participant data (IPD) in educational psychology. *Zeitschrift für Psychologie, 224*(3), 157–167. https://doi.org/10.1027/2151-2604/a000251

Kebritchi, M., Lipschuetz, A., & Santiague, L. (2017). Issues and challenges for teaching successful online courses in higher education: A literature review. *Journal of Educational Technology Systems, 46*(1), 4–29. https://doi.org/10.1177/0047239516661713

Kinshuk & Chen, N.-S. (2006). Synchronous methods and applications in e-learning. *Campus-Wide Information Systems, 23*(3). https://doi.org/10.1108/cwis.2006.16523caa.001

Knight, S. (2020). Augmenting assessment with learning analytics. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), *Re-imagining university assessment in a digital world* (Vol. 7, pp.129–145). Springer International Publishing. https://doi.org/10.1007/978-3-030-41956-1_10

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika, 45*(2), 527–563. https://doi.org/10.1007/s41237-018-0063-y

Kupczynski, L., Gibson, A. M., Ice, P., Richardson, J., & Challoo, L. (2011). The impact of frequency on achievement in online courses: A study from a South Texas university. *Journal of Interactive Online Learning, 10*(3), 141–149. http://www.ncolr.org/jiol/issues/pdf/10.3.3.pdf

Lauría, E. J. M., Baron, J. D., Devireddy, M., Sundararaju, V., & Jayaprakash, S. M. (2012). Mining academic data to improve college student retention: An open source perspective. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 139–142. https://doi.org/10.1145/2330601.2330637

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lee, J., & Martin, L. (2017). Investigating students' perceptions of motivating factors of online class discussions. *The International Review of Research in Open and Distributed Learning, 18*(5), 148–172. https://doi.org/10.19173/irrodl.v18i5.2883

Liu, Y., & Feng, H. (2011). An empirical study on the relationship between metacognitive strategies and online-learning behavior & test achievements. *Journal of Language Teaching and Research, 2*(1), 183–187. https://doi.org/10.4304/jltr.2.1.183-187

Looyestyn, J., Kernot, J., Boshoff, K., Ryan, J., Edney, S., & Maher, C. (2017). Does gamification increase engagement with online programs? A systematic review. *PLoS One, 12*(3), Article e0173403. https://doi.org/10.1371/journal.pone.0173403

Lust, G., Juarez Collazo, N. A., Elen, J., & Clarebout, G. (2012). Content Management Systems: Enriched learning opportunities for all? *Computers in Human Behavior, 28*(3), 795–808. https://doi.org/10.1016/j.chb.2011.12.009

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

McCuaig, J., & Baldwin, J. (2012). Identifying successful learners from interaction behaviour. *Proceedings of the 5th International Conference on Educational Data Mining*,160–163. https://eric.ed.gov/?id=ED537220

Means, B., Toyama, Y., Murphy, R., & Baki, M. (2013). The effectiveness of online and blended learning: A meta-analysis of the empirical literature. *Teachers College Record, 115*, 1–47. https://agronomy.unl.edu/online/documents/Effectiveness_of_online_learning.pdf

Mödritscher, F., Andergassen, M., & Neumann, G. (2013). Dependencies between e-learning usage patterns and learning results. *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies – i-Know '13, 24*, 1–8. https://doi.org/10.1145/2494188.2494206

Mora, A., Riera, D., González, C., & Arnedo-Moreno, J. (2017). Gamification: A systematic review of design frameworks. *Journal of Computing in Higher Education, 29*, 516–548. https://doi.org/10.1007/s12528-017-9150-4

Mwalumbwe, I., & Mtebe, J. S. (2017). Using learning analytics to predict students' performance in moodle learning management system: A case of Mbeya University of Science and Technology. *The Electronic Journal of Information Systems in Developing Countries, 79*(1), 1–13. https://doi.org/10.1002/j.1681-4835.2017.tb00577.x

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Palvia, S., Aeron, P., Gupta, P., Mahapatra, D., Parida, R., Rosner, R., & Sindhi, S. (2018). Online education: Worldwide status, challenges, trends, and implications. *Journal of Global Information Technology Management, 21*(4), 233–241. https://doi.org/10.1080/1097198X.2018.1542262

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society, 17*(4), 49–64. https://www.jstor.org/stable/10.2307/jeductechsoci.17.4.49

Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour, 5*, 1535–1547. https://doi.org/10.1038/s41562-021-01117-5

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ransdell, S., & Gaillard-Kenney, S. (2009). Blended learning environments, active participation, and student success. *The Internet Journal of Allied Health Sciences and Practice, 7*(1), 1–4. https://nsuworks.nova.edu/ijahsp/vol7/iss1/9/

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353–387. https://doi.org/10.1037/a0026838

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *British Medical Journal, 340*, Article c221. https://doi.org/10.1136/bmj.c221

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods, 26*(2), 141–160. https://doi.org/10.1037/met0000300

Saqr, M., Fors, U., & Tedre, M. (2017). How learning analytics can early predict under-achieving students in a blended medical education course. *Medical Teacher, 39*(7), 757–767. https://doi.org/10.1080/0142159X.2017.1309376

Scheffel, M., Drachsler, H., Stoyanov, S., & Specht, M. (2014). Quality indicators for learning analytics. *Educational Technology & Society, 17*(4), 117–132. https://www.jstor.org/stable/jeductechsoci.17.4.117

Seifert, A., Hofer, M., & Allemand, M. (2018). Mobile data collection: Smart, but not (yet) smart enough. *Frontiers in Neuroscience, 12*, Article 971. https://doi.org/10.3389/fnins.2018.00971

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Song, D. (2018). Learning analytics as an educational research approach. *International Journal of Multiple Research Approaches, 10*(1), 102–111. https://doi.org/10.29034/ijmra.v10n1a6

Song, D., & Bonk, C. J. (2016). Motivational factors in self-directed informal learning from online learning resources. *Cogent Education, 3*(1), Article 1205838. https://doi.org/10.1080/2331186X.2016.1205838

Song, D., Rice, M., & Oh, E. Y. (2019). Participation in online courses and interaction with a virtual agent. *The International Review of Research in Open and Distributed Learning, 20*(1), 43–62. https://doi.org/10.19173/irrodl.v20i1.3998

Strang, K. D. (2016). Do the critical success factors from learning analytics predict student outcomes? *Journal of Educational Technology Systems, 44*(3), 273–299. https://doi.org/10.1177/0047239515615850

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior, 47*, 157–167. https://doi.org/10.1016/j.chb.2014.05.038

Tempelaar, D. T., Rienties, B., & Nguyen, Q. (2019). Learning engagement, learning outcomes and learning gains: Lessons from LA. *Proceedings of the 16th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2019)*, 257–264. https://doi.org/10.33965/celda2019_201911L032

Uijl, S., Filius, R., & Ten Cate, O. (2017). Student interaction in small private online courses. *Medical Science Educator, 27*, 237–242. https://doi.org/10.1007/s40670-017-0380-x

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods, 13*(2), 130–149. https://doi.org/10.1037/1082-989X.13.2.130

Viechtbauer, W. (2010). Conducting meta-analyses in *R* with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Winne, P. H. (2004). Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environments. *International Journal of Educational Research, 41*(6), 466–488. https://doi.org/10.1016/j.ijer.2005.08.012

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist, 45*(4), 267–276. https://doi.org/10.1080/00461520.2010.517150

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior, 112*, 106457. https://doi.org/10.1016/j.chb.2020.106457

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet and Higher Education, 29*, 23–30. https://doi.org/10.1016/j.iheduc.2015.11.003

Zheng, L., Zhang, X., & Gyasi, J. F. (2019). A literature review of features and trends of technology-supported collaborative learning in informal learning settings from 2007 to 2018. *Journal of Computers in Education, 6*(4), 529–561. https://doi.org/10.1007/s40692-019-00148-2

History
Received April 30, 2021
Revision received October 2, 2021
Accepted November 2, 2021
Published online February 2, 2022

Conflict of Interest

We have no known conflict of interest to disclose.

Open Data

The review was not preregistered. Instead of a review protocol, we submitted a structured abstract for the *Zeitschrift für Psychologie* and prepared a project on the Open Science Framework (OSF). Therefore, additional study material, including all articles involved in the meta-analysis, the structured abstract, as well as descriptions and explanations of any amendments, and data are available online at the Center for Open Science, 2021 (https://osf.io/wy2px/).

Funding
Open access publication enabled by the University of Bamberg.

ORCID
Maria Klose
https://orcid.org/0000-0002-1252-6937

Maria Klose
Leibniz Institute for Educational Trajectories (LIfBi)
University of Bamberg
Wilhelmsplatz 3
96047 Bamberg
Germany
maria.klose@lifbi.de

Review Article

Dealing With Dependent Effect Sizes in MASEM

A Comparison of Different Approaches Using Empirical Data

Isidora Stolwijk, Suzanne Jak, Veroni Eichelsheim, and Machteld Hoeve

Faculty of Social and Behavioural Sciences, University of Amsterdam, The Netherlands

Abstract: The objective of the present study was to examine whether different methods for dealing with dependency in meta-analytic structural equation modeling (MASEM) lead to different results. Four different methods for dealing with dependent effect sizes in MASEM were applied to empirical data, including: (1) ignoring dependency; (2) aggregation; (3) elimination; and (4) a multilevel approach. Random-effects two-stage structural equation modeling was conducted for each method separately, and potential moderators were examined using subgroup analysis. Results demonstrated that the different methods of dealing with dependency in MASEM lead to different results. Thus, the decision on which approach should be used in MASEM-analysis should be carefully considered. Given that the multilevel approach is the only approach that includes all available information while explicitly modeling dependency, it is currently the theoretically preferred approach for dealing with dependency in MASEM. Future research should evaluate the multilevel approach with simulated data.

Keywords: meta-analytic structural equation modeling (MASEM), structural equation modeling, dependent effect sizes, meta-analysis, subgroup analysis

Meta-analytic structural equation modeling (MASEM) is an increasingly popular technique for summarizing findings from multivariate correlational research (Becker, 1992; Cheung & Chan, 2005; Viswesvaran & Ones, 1995). The goal of MASEM is to fit and interpret structural equation models in order to explain the (synthesized) correlations between variables. For most MASEM methods, the first step involves the estimation of a synthesized correlation matrix based on the studies' observed correlation matrices.

An important assumption related to synthesizing effect sizes is that each effect size is independent of the other (e.g., Cheung, 2019). In MASEM, this implicates that each study may only provide one correlation coefficient for each cell (each relationship between variables) in the correlation matrix. This assumption often does not hold as dependence among effect sizes can occur for a variety of reasons (e.g., Ahn et al., 2012). For instance, multiple informants (e.g., mother- and father-report on parenting practices) or multiple measurement occasions (e.g., pre-and post-test measures) will lead to multiple correlation coefficients for the same relationship in a study. Failure to properly deal with dependency can lead to over- or underestimation of the available information, which has important implications for the statistical inferences (Cheung, 2019; Moeyaert et al., 2017; Wilson et al., 2016).

Dependency of effect sizes is a common issue in meta-analytic research (Cheung, 2019; Moeyaert et al., 2017). There have been several (methodological) reviews on the occurrence of dependent effect sizes. A recent review of 28 meta-analyses from educational research found that 57% of the studies reported dependent effect sizes (Rios et al., 2020). This is similar to the findings of Ahn and colleagues (2012), who found that of the 56 meta-analyses on educational research they reviewed, 62% reported multiple (dependent) effect sizes, and a review of 44 meta-analyses on randomized controlled trials reported that 70% of the studies included dependent effect sizes (Page et al., 2015).

Over time, several (ad hoc) solutions have arisen to overcome the issue of dependency in MASEM, which have not always been justified or well-examined for their statistical properties (Wilson et al., 2016). The objective of the present study was to examine whether applying different methods for dealing with dependent effect sizes to empirical data leads to different results when conducting MASEM analysis. Four methods were compared, including: (1) ignoring dependency; (2) aggregation; (3) elimination; and (4) a recently developed multilevel approach by Wilson and colleagues (2016), further referred to as the WPL-approach.

The next section describes the concept of MASEM in more detail. The section thereafter further elaborates on the issue of dependency and provides descriptions of the four methods for dealing with dependency, including a discussion of their (dis)advantages. The final section describes the application of the different methods for dealing with

https://doi.org/10.1027/2151-2604/a000485

dependency to empirical data, including a comparison of the results.

Meta-Analytic Structural Equation Modeling

MASEM combines meta-analysis (MA) and structural equation modeling (SEM) and thereby overcomes some of the disadvantages of the separate techniques. SEM allows for testing more complex research questions, and MA provides sufficiently large samples to test these complex theories in SEM with sufficient statistical accuracy. There are many different ways to combine MA and SEM, but it mostly consists of two stages: (1) effect sizes from primary studies are synthesized to obtain a pooled correlation matrix; and (2) a structural equation model is fitted to the pooled correlation matrix from Stage 1 (e.g., Cheung & Chan, 2005; Jak, 2015; Viswesvaran & Ones, 1995). MASEM can be conducted using two-stage structural equation modeling (TSSEM). TSSEM was first developed for fixed-effects models (Cheung & Chan, 2005) and later extended to fit random-effects models by including study-specific random-effects (Cheung, 2014a), which is very similar to the GLS-approach by Becker (1992, 1995). Nowadays, random-effects models are preferred because fixed-effects models assume homogeneity of effect sizes which is often unrealistic (e.g., Cheung, 2014a; Yuan, 2016).

In Stage 1 of TSSEM, the correlation coefficients are weighed by their sampling variance (v_i) and study-level variance (τ^2). The random-effects model for the correlation vectors r_i = vechs(R_i) in the ith correlation matrix R_i is

$$r_i = \rho_{\text{Random}} + u_i + \varepsilon_i, \tag{1}$$

with ρ_{Random} as a vector of the means of the correlation coefficients over studies, u_i describing the study-specific random effects in study i, and ε_i the sampling deviation study i from its study-specific population coefficients, with Cov(u_i) = T^2 representing the estimated between-study variance and Cov(ε_i) = V_i representing the sampling covariance matrix in the ith study (Cheung, 2014a). The model is fitted using maximum likelihood (ML) estimation.

In Stage 2, the structural equation model is fitted to the pooled correlation matrix R (consisting of the estimates of ρ_{Random}) of Stage 1 using weighted least squares (WLS) estimation. The weight matrix used in WLS-estimation is the inverse asymptotic covariance matrix of the Stage 1 estimates (Cheung & Chan, 2005). These weights ensure that correlation coefficients based on more information (on more studies and/or studies with larger sample sizes) get more weight in the estimation of the Stage 2 parameters. Since the between-studies variance is filtered out at Stage 1, it does not play a direct role at Stage 2 (Cheung, 2014a).

Different Methods for Dealing With Dependent Effect Sizes in MASEM

There are different ways of dealing with the dependency of effect sizes in MASEM. When not properly dealt with, dependent effect sizes may lead to under-or overestimation of standard errors (*SE*s) of the average effect sizes, which could result in inflation of Type I errors or reduced statistical power (Cheung, 2019; López-López et al., 2017). In the following section, the (potential) advantages and disadvantages of four different approaches for dealing with dependent effect sizes are described, including: (1) ignoring dependency; (2) aggregation; (3) elimination; and (4) the WPL-approach.

Ignoring Dependency of Effect Sizes

Ignoring dependency is a known-to-be incorrect strategy that is likely to bias results, to the extent that it threatens the validity of the inferences (Moeyaert et al., 2017; López-López et al., 2018). For one, studies with just one effect size will have a smaller influence on the resulting average effect size than studies with multiple effect sizes, which may result in biased estimates (Cheung, 2014b; Van den Noortgate et al., 2013). Second, simulation studies showed that the estimated *SE*s of the average effect sizes are underestimated, resulting in an increased likelihood of significant results (i.e., inflation of Type I errors; López-López et al., 2017; Moeyaert et al., 2017). One might incorrectly assume that the estimates are very precise and statistical inferences are more likely to be wrong (Cheung, 2019). The approach of ignoring dependency is a non-acceptable practice in meta-analysis and is merely presented in the current study to emphasize its inappropriateness and underline its (negative) implications.

Aggregation of Effect Sizes

Aggregation is a commonly used approach that involves averaging dependent effect sizes within a study before pooling effect sizes across studies (Cheung & Chan, 2004; Cheung, 2014b; Marín-Martínez & Sánchez-Meca, 1999). There are different ways to aggregate effect sizes. One option is simple aggregation, which involves calculating the arithmetic mean. Simple aggregation may be appropriate when sample sizes are (close to being) equal and when it is likely that population effect sizes are the same (Marín-Martínez & Sánchez-Meca, 1999; Moeyaert et al., 2017). However, in practice, this is often unrealistic.

Another option involves weighted aggregation. Here, effect sizes are averaged using some weighting scheme (e.g., by the inverse of the sampling variance; Marín-Martínez & Sánchez-Meca, 1999; Moeyaert et al., 2017). Weighted aggregation essentially involves cell-by-cell sub-meta-analyses. For each study that contributes multiple (dependent) effect sizes per cell, a pooled correlation matrix is estimated with a single pooled estimate within each cell (Wilson et al., 2016). An advantage of weighted aggregation – over simple aggregation – is that more weight is assigned to more precise estimates and less weight to less precise estimates.

An advantage of aggregation – both simple and weighted – is that it is a relatively intuitive and simple procedure. Disadvantages of aggregation are that it ignores within-study variability (López-López et al., 2018), and the loss of information limits the possibility to examine characteristics that can be used to evaluate effect size variability (Wilson et al., 2016). Also, a recent simulation study showed that the aggregation approach is too conservative, especially when the level of dependency is relatively low (Moeyaert et al., 2017). Their results showed that *SE*s are overestimated, which could lead to an inflation of Type II errors.

Thus, even though the aggregation approach is appealing and intuitive, given its disadvantages, it is not considered a state-of-the-art approach for dealing with dependent effect sizes (López-López et al., 2018; Moeyaert et al., 2017; Wilson et al., 2016).

Elimination of Effect Sizes

With elimination, one effect size per study is randomly picked or chosen based on some a priori decision rule, resulting in independent effect sizes (Cheung, 2014b; Cheung, 2019; Wilson et al., 2016). Randomly picking one effect size could be appropriate when effect sizes are assumed to be truly equivalent. However, this is a very strong assumption that rarely holds in practice. To test the assumption, one could conduct sensitivity analyses to compare results from the initial randomly picked effect sizes to another set of effect sizes (López-López et al., 2018).

Elimination based on an *a priori* decision rule may be appropriate when there are substantive (or validity) considerations for preferring one effect size over the other. For example, if a study includes multiple measurements of child delinquency, if reliability is higher for self-reported delinquency than for parent-reported child delinquency, the effect size pertaining to the self-report measure may be preferred. The disadvantages of elimination are similar to those of aggregation in that it affects statistical power and excludes the possibility to examine study characteristics that can be used to evaluate effect size variability.

Additionally, if the effect size is chosen based on some a priori decision rule, the fixed-effect estimates will likely show some bias towards the characteristics of the decision rule (Cheung, 2019). For example, choosing only the first measurement from longitudinal studies may bias the results to samples of younger ages. This may – depending on the specific association of interest – lead to systematically larger or smaller effects for the specific associations. Still, both with randomly picking or choosing an effect size, the resulting effect sizes will be less efficient because the information is lost. Thus, elimination may be appropriate when relevant to the research question, but it is an inappropriate method for solving dependency issues (Cheung, 2014b).

The WPL-Approach

Wilson and colleagues (2016) developed an approach to deal with dependency in MASEM, which combines three-level meta-analysis and TSSEM. A three-level random-effects meta-analysis is used to account for dependency in which participants (Level 1) are nested within effect sizes (Level 2) and effect sizes within clusters (Level 3; Van den Noortgate et al., 2013). Information from all available (dependent) effect sizes per study is incorporated in the pooled correlation matrix, and dependency is explicitly modeled.

The most important advantage of the WPL-approach is that all available information is incorporated, thus it does not reduce statistical power. Additionally, both within- and between cluster variance are taken into account, allowing for examination of heterogeneity at different levels (Cheung, 2014b; Cheung, 2019). One potential disadvantage is that the approach is somewhat more complex and not yet widely used, thus may pose more of a challenge for researchers. However, examples of studies that incorporated the WPL-approach are available (e.g., Graf-Drasch et al., 2019; Loignon & Woehr, 2018).

Empirical Application

The empirical application examined whether using the four different methods for dealing with dependency in MASEM would lead to different results. In case no (or minor) differences are found, one could conclude that the differences are mainly theoretical with no important practical implications. Then, deciding on how to deal with dependency may be based on personal preferences. However, if (large) differences are found that affect statistical inferences, the decision on which method to use for dealing with dependency in MASEM is an important one and should be carefully considered.

The next section describes the empirical application in further detail. To start, some background information is provided on the empirical data consisting of a meta-analysis on the intergenerational continuity of criminal behavior.

Background

The intergenerational continuity of criminal behavior has been well established. For instance, a meta-analysis found that children of criminal parents are at two times higher risk for criminal behavior themselves than children of non-criminal parents (Besemer et al., 2017). Explanatory mechanisms are not yet well studied, but from the literature, potential explanations can be derived. A potential mechanism through which criminal parents affect their children may be that criminal parents use less efficient (or even problematic) parenting practices. Evidence for this comes from a longitudinal study that found that mothers with a history of antisocial behavior show increased odds for problematic parenting behaviors, when compared to mothers without a history of antisocial behavior (Johnson et al., 2004). Finally, these problematic parenting practices are associated with child delinquency, with moderate associations between both behavioral control and parental support and child delinquency (Hoeve et al., 2009).

The empirical application examined the underlying mechanisms through which parental crime is associated with child delinquency. It was hypothesized that the effect of parental crime on child delinquency was fully mediated by parental support and behavioral control. The hypothesized full mediation model was compared to a partial mediation model in which a direct effect of parental crime on child delinquency was added. Given that the hypothesized model involved a path model, MASEM was necessary for the analyses.

Procedure

Sample of Studies and Selection Criteria

The selection of studies was derived from a meta-analysis on the relation between parenting practices and child delinquency (Hoeve et al., 2009), and an additional selection of studies (Silva Pinho, 2018; Van den Berg, 2018), which are part of a larger project 'The potential mediating role of parenting on the intergenerational continuity of criminal behaviour'. The coding of studies and the manual search are still in progress; therefore, a subset of studies was included in this study.

Studies were selected using the following criteria: studies had to (1) focus on child delinquency, parental crime, and parenting behavior; (2) involve Western samples; and (3) report on bivariate associations. Child delinquency and parental crime were operationalized as all behavior prohibited by law. Broadly, parenting behaviors were defined such that all behaviors had to be directed at the child. Parental support includes all behaviou of the parent towards the child that makes the child feel comfortable and accepted. Behavioral control includes supervision, regulation, and active monitoring (excluding child disclosure and parental knowledge). Note that studies including negative support (e.g., rejection), and negative behavioral control (e.g., low supervision) were also included.

The articles were screened and coded for effect sizes on (1) parental crime and parenting behaviors, (2) parental crime and child delinquency, and (3) parenting behaviors. A more elaborate description of the search strategy, selection criteria, and the coding procedure can be found in the original meta-analysis of Hoeve and colleagues (2009) and the PRISMA flow diagram included in the Electronic Supplementary Material (ESM 1, Figure E3).

Classification and Computation of Effect Sizes

The Pearson product-moment correlation coefficient (*r*), further referred to as the correlation (coefficient), was used as the input effect size for the analyses because this is the only effect size suitable for conducting MASEM. Primary studies often report on a variety of effect sizes, be it due to different reporting standards across disciplines or differing nature of the variable included in the study (e.g., continuous versus categorical). The raw (non-correlation) effect sizes were converted to correlation coefficients using methods and formulae provided by Lipsey and Wilson (2001) and Borenstein and colleagues (2009). A total of 18 effect sizes were converted.

The directions of effect sizes were coded such that a positive effect indicated higher levels (e.g., more occurrences, increased severity) of child delinquency or parental crime. In case primary studies reported effect sizes that were not in line with the hypothesis of the current study, the effect sizes were reversed. For example, when support and behavioral control were negatively formulated, the effect sizes were reversed to indicate a negative association between parenting behavior and child delinquency.

Statistical Analyses

Evaluation of Publication Bias

Publication bias was evaluated using three-level funnel plots (Fernández-Castilla et al., 2020). The three-level funnel plot provides two graphs from which to evaluate publication bias: (1) a graph in which all effect sizes are plotted; and (2) a graph which plots the study-specific effects (i.e., amount of effect sizes reported per study, including their variability) against their meta-analytic standard errors.

Dealing With Dependent Effect Sizes

The procedures of the four approaches for dealing with dependency are described in the following section.

Ignoring Dependency

With ignoring dependency, no additional adjustments of the data or calculations were required. All effect sizes were included and treated as independent.

Aggregation

With simple aggregation, the arithmetic mean was calculated (i.e., the average of all effect sizes within a study). With weighted aggregation, the dependent effect sizes within a study were weighed using the inverse of the sampling variance (Cheung, 2014b). The sampling variance (v_i) was estimated using

$$v_i = \frac{\left(1 - r_i^2\right)^2}{n_i}, \tag{2}$$

with r_i representing the observed correlation coefficient of study i, and n_i representing the sample size of study i (Olkin & Siotani, 1976). Using the effect sizes and sampling variances, submeta-analyses were performed, resulting in one (weighted) effect size per study.

Elimination

With the elimination approach, one effect size per study was chosen based on a set of a priori decision rules. In case of multiple measurement occasions, only the effect size from the first measurement of child delinquency was included. In the case of both a boy and a girl sample, the girl sample was chosen because boys were overrepresented in the current sample of studies. In case of multiple samples or multiple informants, the sample or informant with the highest reliability was chosen. If no distinction could be made based on the described criteria, the first effect size that was reported was chosen.

WPL-Approach

With the WPL-approach, the synthesized correlation matrix was estimated using a three-level hierarchical model, thereby accounting for the statistical dependencies (Van den Noortgate et al., 2013; Wilson et al., 2016). Each unique effect size is coded with a unique effect size ID, and the effect sizes are nested within studies. Wilson and colleagues (2016) provide a nice illustration of how a dataset with such structure may be organized.

A random-effects no-intercept model was estimated using maximum likelihood (ML) estimation to synthesize correlations in each of the cells. Input required for the random-effects no-intercept model was the unique effect sizes and the variances of the effect sizes, which were calculated using simple sample size weighing (Schmidt & Hunter, 2014). Using a no-intercept model allows interpreting the regression coefficients as synthesized correlation coefficients, which are necessary for Stage 2 of the analysis. Also, the asymptotic covariance matrix of the pooled correlation matrix is available, which provides information on the precision of the pooled correlations (Wilson et al., 2016).

Random-Effect TSSEM Analysis

The hypothesized model was tested using random-effects TSSEM (Cheung, 2014a) and was overidentified with 1 *df*. For the WPL-approach, Stage 1 involved estimating a random-effects no-intercept model using ML estimation in which the effect sizes were nested within studies (Wilson et al., 2016). For the remaining approaches, a pooled correlation matrix was estimated in Stage 1 using ML estimation (Cheung, 2014a). The hypothesized model includes four variables, resulting in a pooled correlation matrix with six cells. Each cell contains a pooled estimate representing one of the associations of interest. In case the model did not reach convergence, the between-studies variance (τ^2) was fixed at zero for the associations that seemed to lack heterogeneity.

The degree of heterogeneity was qualified using I^2, which typically estimates how much of the total variance of effect sizes is due to between-study heterogeneity. Due to its three-level nature, the WPL-approach has the additional benefit of evaluating heterogeneity on both the within- and between-study level. The following rules of thumbs are used, with an I^2 of .25, .50, and .75 indicating low, medium, and high levels of heterogeneity, respectively (Higgins et al., 2003).

At Stage 2, the hypothesized model was fitted on the pooled correlation matrix obtained at Stage 1 using weighted least squares (WLS) estimation (Cheung, 2014a). Model fit was evaluated using the chi-squared difference ($\Delta\chi^2$) test, using an α = .05 criterion for indicating a significant discrepancy between the (saturated) partial mediation model and the (more parsimonious) full mediation model. Note that with the evaluation of model fit in SEM, it is common to report alternative fit indices (e.g., RMSEA, CFI) because the $\Delta\chi^2$-tests are known to be very sensitive to small discrepancies when working with large sample sizes (e.g., Barret, 2007). Therefore, the RMSEAs (including their 95% CIs) are reported, using the following guidelines for adequate- to a good fit, respectively: RMSEA $\leq$ .08 and $\leq$ .05 (Hu & Bentler, 1998).

Finally, the parameter estimates of the retained model were interpreted. Criteria used to evaluate the size of the effects were based on the guidelines provided by Funder and Ozer (2019, p. 166), with an *r* of .05 indicating a very small effect, *r* of .10 a small effect, *r* of .20 a medium effect, *r* of .30 a large effect, and *r* of .40 a very large effect. These guidelines were originally developed for interpreting the size of correlations coefficients but are deemed appropriate for the interpretation of standardized parameter estimates.

Moderator Analysis

The (hypothesized) moderator involved the type of sample on which the effect size was based, being either a sample from the general community or a high-risk sample (e.g., a sample coming from high-crime neighborhoods, an offender sample). The moderator analyses were conducted using subgroup analysis (Jak & Cheung, 2018), which tests whether the parameter estimates are equal across groups.

The retained model was fitted to the pooled correlation matrices of each group separately. To test for subgroup differences, a model in which the parameter estimates were constrained to equality across groups was compared to a model without equality constraints. In the case of a significant $\Delta\chi^2$-test, the constrained model fits significantly worse than the model without the equality constraints which indicates that there are significant subgroup differences.

Software

Analyses were performed using R (version 3.5.1.; R Core Team, 2020) with the metafor package (version 2.4.0.; Viechtbauer, 2010) for Stage 1 of the WPL-approach, and the metaSEM package (version 1.2.4.; Cheung, 2015) for the MASEM and the subgroup analyses.

Results

Study Descriptives

The current sample of studies consisted of 140 manuscripts, with 114 unique samples and a total sample size of N = 163,709. Of the studies, 72.1% (k = 101) reported multiple (dependent) effect sizes. The studies contained a total of 764 effect sizes (see Table 1 for the number of effect sizes and the total sample sizes per association). There was an almost equal number of longitudinal (k = 68) and cross-sectional (k = 72) studies. Most studies were conducted with

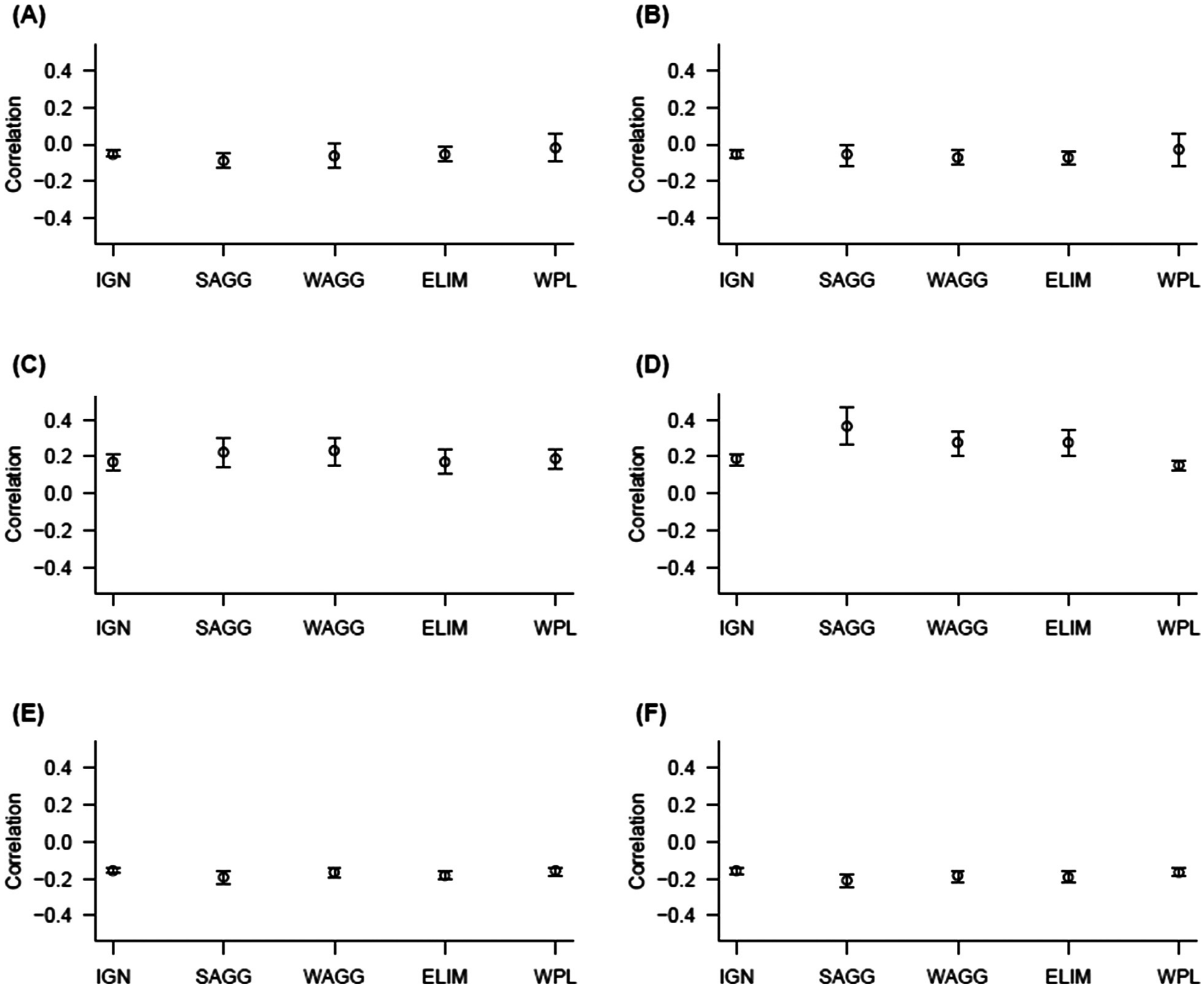


Figure 1. Pooled correlations including their 95% CIs for each association per approach of dealing with dependency. IGN = ignoring dependency; SAGG = simple aggregation; WAGG = weighted aggregation; ELIM = elimination; WPL = WPL-approach. (A) Parental Crime – Support; (B) Parental Crime – Behavioral Control; (C) Parental Crime – Child Delinquency; (D) Support – Behavioral Control; (E) Support – Child Delinquency; (F) Behavioral Control – Child Delinquency.

samples from North America (75%), with fewer studies conducted with European (22.9%) and Australian/New Zealand (2.1%) samples. With regard to sample type, 70.7% were general community samples, 22.9% were high-risk or delinquent samples, and 6.4% were other types of samples (e.g., combined samples of delinquents and non-delinquent). The studies included in the meta-analysis are listed in Table E1 (ESM 1), including some of their characteristics.

Table 1. Number of effect sizes and total sample sizes per association

	1	2	3	4
1. Parental crime		6,773	6,695	30,137
2. Support	20		53,081	108,720
3. Behavioral control	11	171		87,275
4. Child delinquency	40	286	244	

Note. Number of effect sizes are shown below the diagonal, and sample sizes above the diagonal.

Comparison of Results From the Different Approaches for Dealing With Dependency

TSSEM Analysis

Stage 1 analyses were conducted to allow for the evaluation of the heterogeneity of effect sizes and to obtain the pooled correlation matrices needed for Stage 2 (see Figure 1). With the simple- and weighted aggregation approaches, running the Stage 1 model led to some convergence issues, which were likely due to the lack of heterogeneity in the associations between parental crime and support and parental crime and behavioral control. Thus, with the aggregation approaches, it seemed that the loss of information contributed to a lack of heterogeneity, leading to convergence issues, which was not the case with the other approaches.

Evaluation of I^2 indicated large levels of heterogeneity (I^2 = .94 to I^2 = .97) for all approaches, with only small differences of .01 to .03. A benefit of the WPL-approach is the possibility to divide the overall heterogeneity into within- and between-cluster (i.e., studies) heterogeneity. Under the WPL-approach, 15% of the total variance was estimated to be due to between-study heterogeneity, and 81% due to within-study heterogeneity (with the remaining 4% due to random sampling variance). Note that under the other approaches, one may incorrectly infer that variability of effect sizes is mainly due to differences between studies, whereas the WPL-approach shows that most variability of effect sizes is due to differences within studies.

Next, the pooled correlation matrices for all approaches were compared, which are presented in Table E2 (ESM 1). Some differences were found in the size of the estimated pooled correlations. For example, with the simple aggregation approach, there is a large to the very large association between support and behavioral control (r = .37), which is small to moderate with the WPL-approach (r = .15). Also, there were differences regarding the significance of the associations. For example, the association between parental crime and behavioral control was non-significant with the WPL-approach but significant for the other approaches. Figure 1 presents the pooled correlation estimates, including their 95% confidence intervals (CIs) per approach. The width of the CIs of the ignoring dependency approach seems to be consistently smaller than the width of the CIs of the other approaches. In line with expectations, the ignoring dependency approach seems to overestimate the precision of the estimates, whereas the aggregation- and elimination approaches seem to underestimate their precision. Note that with the WPL-approach, the CIs of the associations coming from a larger number of effect sizes are also quite narrow and wider in association coming from less effect sizes. This is to be expected since a larger number of effect sizes should contribute to the precision of the estimates.

In Stage 2, both the hypothesized full mediation model and the partial mediation model were fitted to the pooled correlation matrices obtained at Stage 1. Inferences regarding model comparison were similar for all approaches. Model comparison showed significant differences between the full mediation model and the partial mediation model, indicating that the (more parsimonious) full mediation model fit significantly worse than the (saturated) partial mediation model, with $\Delta\chi^2$ = 47.33, Δdf = 1, p < .001, for the ignoring dependency approach, $\Delta\chi^2$ = 25.17, Δdf = 1, p < .001, for the simple aggregation approach, $\Delta\chi^2$ = 28.23, Δdf = 1, p < .001, for the weighted aggregation approach, $\Delta\chi^2$ = 19.56, Δdf = 1, p < .001, for the elimination approach, and, lastly, $\Delta\chi^2$ = 156.01, Δdf = 1, p < .001, for the WPL-approach. Each approach of dealing with dependency showed good fit of the full mediation model with RMSEAs ranging from .01 to .03, with RMSEA = .01, 95% CI [.01, .01] for the ignoring dependency approach, RMSEA = .01, 95% CI [.01, .02], for the simple aggregation approach, RMSEA = .01, 95% CI [.01, .01] for the weighted aggregation approach, RMSEA = .01, 95% CI [.01, .02], for the elimination approach, and RMSEA = .03, 95% CI [.03, .03] for the WPL-approach. Note that even though conclusions regarding model comparison are the same across approaches, the values of $\Delta\chi^2$-tests show seemingly large differences across approaches. Given the statistical power of the $\Delta\chi^2$-test, it may be that studies with smaller sample sizes would lead to different conclusions across the different approaches.

Next, the parameter estimates of the partial mediation model were compared, which are presented in Table E3 (ESM 1). Overall, the parameter estimates were quite similar in size across the approaches. Small differences in the point estimates were found, ranging from 0.003 to 0.071. For example, the effect of parental crime on child delinquency was small to moderate with the ignoring dependency approach (β = 0.16) and moderate with the weighted aggregation approach (β = 0.21). Also, differences were found regarding the statistical significance of the effects. For example, the effect of parental crime on support was non-significant with the weighted aggregation- and the WPL-approach but significant with the other approaches. This may be explained by this effect coming from the least amount of information (i.e., coming from the smallest number of effect sizes) and because of the relatively large amount of within-study heterogeneity, which is only accounted for by the WPL-approach. Thereby, the precision of the estimates may be smaller than portrayed by the other approaches.

Figure 2 presents plots of the parameter estimates, including their 95% CIs. It seems that with ignoring dependency, the CIs of the parameter estimates are consistently smaller, which is in line with expectations. The CIs of the simple- and weighted aggregation-, and elimination approaches seem consistently larger than, except for the CIs of the effects of the parenting behaviors on child delinquency (which come from the largest amount of effect sizes and largest sample sizes). Similar to the comparison of the pooled correlations, these results are somewhat in line with expectations. Again, it seems that the differences between the approaches are larger for effects coming from a smaller amount of information than for effects coming from a larger amount of information. This suggests that using the aggregation and/or elimination approach does not affect results as much if there is a sufficiently large dataset because then there will still be enough power.

The differences found in the parameter estimates across methods are also reflected in the residual variances. The residual (co)variances of the partial mediation model are presented in Table E4 (ESM 1). The variance in child delinquency explained by the partial mediation model was 6.7%, 10.1%, 9.3%, 7.7%, and 8% across the ignoring dependency-, simple aggregation-, weighted aggregation-, elimination-, and the WPL-approach, respectively. Figure 3 presents the final model estimated under the WPL-approach.

Comparison of Results From the Moderator Analyses

Moderator analyses were conducted using subgroup analysis with sample type (i.e., general community vs. high-risk) as the moderator. With the simple aggregation approach, it was impossible to conduct moderator analyses due to the lack of information on the association between parental crime and behavioral control for the general community subgroup. There were convergence issues when using the ignoring dependency approach for the high-risk subgroup. Additionally, with the weighted aggregation approach, there were convergence issues for both subgroups. In both cases, the between-studies variances (τ^2) for the associations between parental crime and support and parental crime and behavioral control were fixed to zero.

Subgroup analyses showed similar results across approaches, except for the ignoring dependency approach, $\Delta\chi^2 = 15.37$, $\Delta df = 5$, $p = .009$. With the weighted aggregation-, $\Delta\chi^2 = 4.12$, $\Delta df = 5$, $p = .532$, elimination-, $\Delta\chi^2 = 9.20$, $\Delta df = 5$, $p = .101$, and the WPL-approach $\Delta\chi^2 = 2.10$, $\Delta df = 5$, $p = .836$, results showed no significant differences between the regression coefficients from the general community versus the high-risk subgroup.

Evaluation of Publication Bias With the WPL-Approach

Evaluation of publication bias was conducted using three-level funnel plots, which are presented in Appendix C (ESM 1). Figure E1 (ESM 1) shows the graph in which all effect sizes are plotted. Visual inspection of the effect size plot shows one effect size in the lower-right part of the graph, whereas there is no result with similar precision at the lower-left part of the graph, which may be a sign of publication bias. Figure E2 (ESM 1) shows the plot in which the study-specific effects are plotted against their meta-analytic standard errors. The study-funnel plot shows some signs of asymmetry, especially at the bottom of the graph. Concluding from both graphs, there may be some signs of publication bias, which should be taken into account when interpreting the meta-analytic results.

Discussion

The aim of this study was to examine whether applying different methods for dealing with dependency to empirical data leads to different results when conducting MASEM analysis. The empirical application demonstrated that the different approaches for dealing with dependency in MASEM are not only theoretically different but also lead to different results with important practical implications. An overview of the (dis)advantages of the four approaches is presented in Table 2.

The most important differences lie in the *SE*s of the parameter estimates. The *SE*s of the parameter estimates with the ignoring dependency approach seemed consistently smaller, and the *SE*s of the aggregation- and elimination approaches seemed consistently larger. The *SE*s of the WPL-approach did not seem consistently higher or lower

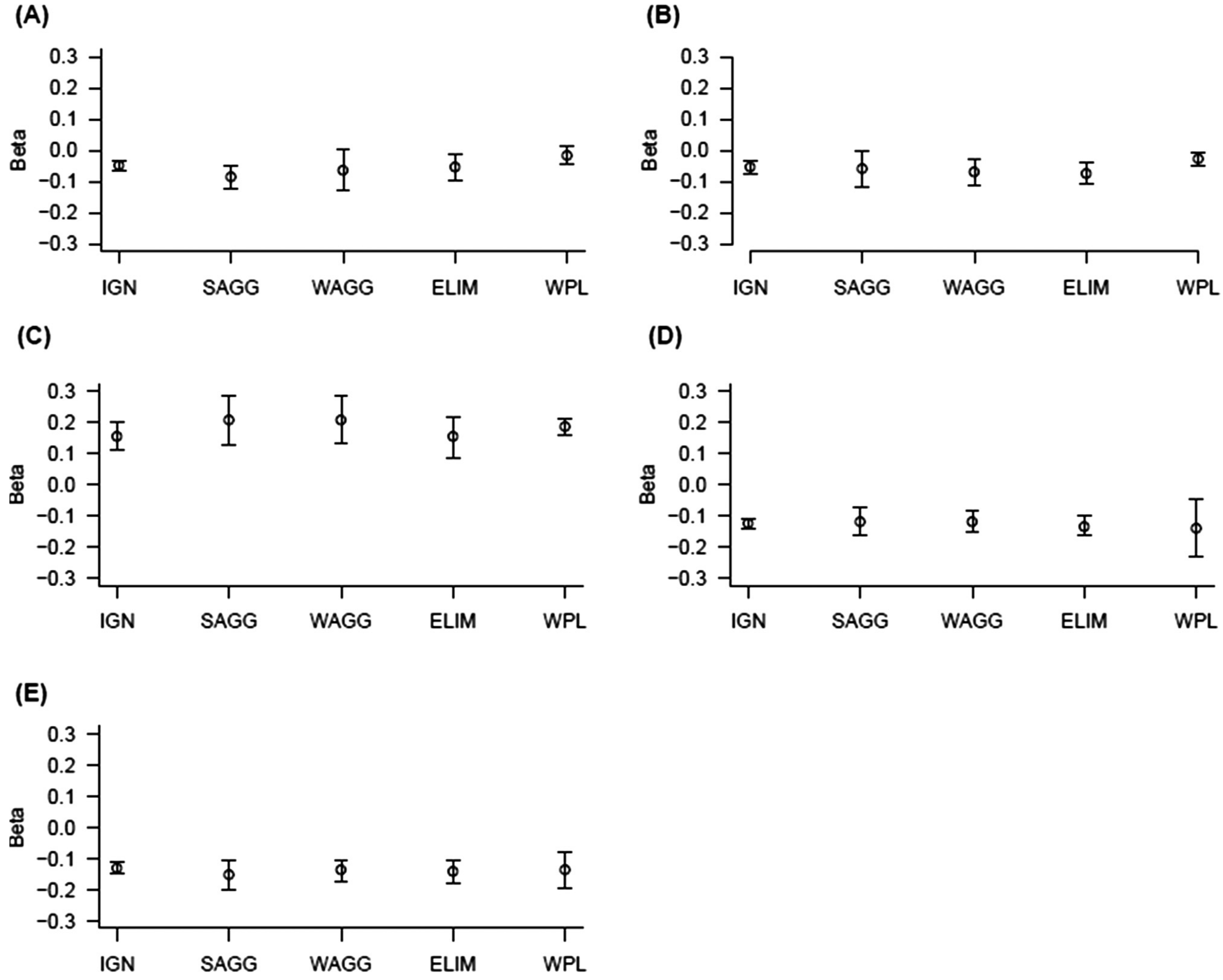


Figure 2. Parameter estimates of the partial mediation model including their 95% CIs for each effect per approach of dealing with dependency. IGN = ignoring dependency; SAGG = simple aggregation; WAGG = weighted aggregation; ELIM = elimination; WPL = WPL-approach. (A) Parental Crime – Support; (B) Parental Crime – Behavioral Control; (C) Parental Crime – Child Delinquency; (D) Support – Child Delinquency; (E) Behavioral Control – Child Delinquency.

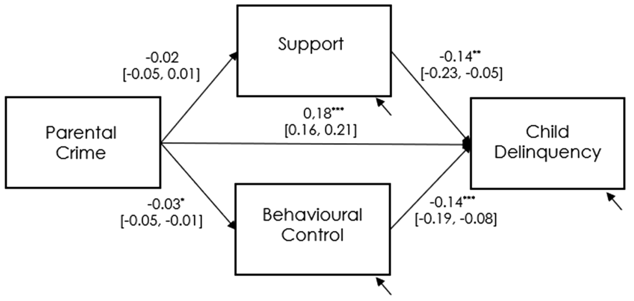


Figure 3. Partial mediation model with parameter estimates including their 95% CIs. Standardized parameter estimates are presented, with their corresponding 95% CIs between the brackets. **p* < .05; ***p* < .01; ****p* < .001.

across the different associations but, as one would expect, seemed to depend on both the amount of information available and the level of within- and between-study variability. Under- or overestimation of the *SE*s has important implications for statistical inferences. For instance, in the present study, with the WPL-approach, the effect of parental crime on parent support is not significant and therefore may be removed from the model, whereas this effect was significant with the other approaches.

Results from the subgroup analysis were also affected by the use of the different approaches for dealing with dependency. For one, with the simple aggregation approach, it was not possible to conduct subgroup analyses due to the lack of information available on the variable of interest. Second, with the ignoring dependency approach, significant differences were found between the subgroups. Given that this was the only approach showing significant differences, this may have been the result of overestimated precision of the estimates. Thus, using different methods for dealing with dependency in MASEM also has important practical

Table 2. Overview of the (dis)advantages of different approaches for dealing with dependent effect sizes in MASEM

	Short description		Advantages	Disadvantages
Ignoring Dependency	Each effect size is treated as independent		–	Standard errors are underestimated (affecting Type I errors; López-López et al., 2017; Moeyaert et al., 2017); Studies with less effect sizes contribute less to the resulting pooled estimate, than those with multiple effect sizes (Cheung, 2014b; Van den Noortgate et al., 2013).
Aggregation	Simple:	Calculate the arithmetic mean.	Relatively simple and intuitive approach.	Standard errors are overestimated (affecting Type II errors; Moeyaert et al., 2017);
	Weighted:	Average the effect size using some weighting scheme.	Advantage of weighted-over simple aggregation is that more weight assigned to more precise estimates than less precise estimates.	Loss of information limits the ability to examine effect size variability; Too conservative when level of dependency is relatively low (Moeyaert et al., 2017); Ignores within-study variability (López-López et al., 2018).
Elimination	One effect size per study is randomly picked or chosen based on some a priori decision rule.		Elimination based on an a priori decision rule may be appropriate when there are substantive (or validity) considerations.	Similar to those of aggregation; Elimination based on an a priori decision rule is likely to result in some bias towards the characteristics of the decision rule (Cheung, 2019).
WPL-approach	Thee-level random-effects meta-analysis allows for effect sizes to be nested within studies.		All available information is incorporated; Dependency is explicitly modelled; Examination of effect size variability is possible at both the within- and between-study level (Cheung, 2019).	Approach is somewhat more complex; More research needed to identify the strengths and weaknesses of the approach; Moderator analysis only available for grouping variables.

implications with regard to the evaluation of (potential) moderators.

These findings are in line with previous research. By ignoring dependency, the available information was overestimated, thereby increasing the likelihood of Type I errors (Cheung, 2019; López-López et al., 2017; Moeyaert et al., 2017). Hereby, one may incorrectly infer that the estimates are very precise and important subgroup differences. Therefore, ignoring dependency is deemed non-acceptable in meta-analytic research.

With aggregation- and elimination of effect sizes, a lot of information was lost by reducing the available information to one effect size per study. Even though the parameter estimates seemed to show no specific bias, the standard errors were consistently larger in comparison to the other approaches. Overestimation of the standard errors is problematic because it affects statistical power, thereby increasing the likelihood of Type II errors (Cheung, 2014b; Moeyaert et al., 2017). Given that most parameter estimates were significant in the current study, the loss of information did not seem to affect statistical inferences. However, this study had a relatively large dataset to work with. It may be the case that in meta-analyses with a smaller number of studies, the lack of statistical power does affect results and fails to identify a potential effect. Thus, using aggregation- and elimination of effect sizes may not be problematic if there is sufficient number of studies and the level of dependency is relatively low (Moeyaert et al., 2017). Still, aggregation- and elimination of effect sizes, even though simple and intuitive, is deemed suboptimal for dealing with dependency in MASEM because these are less efficient approaches.

The WPL-approach showed no consistently higher or lower *SE*s across associations. One explanation for this may be the amount of within-study variability of effect sizes, which is not accounted for by the other approaches. The ability to account for within-study variability is another important benefit of the WPL-approach, as it gives a more accurate representation of the data. Accounting for both within- and between-study variability of effect sizes can lead to different inferences than when one can only examine between-study variability. In this study, with the ignoring dependency-, aggregation-, and elimination approaches, one would infer that there is large significant variability in effect sizes due to between-studies differences. However, the WPL-approach paints a very different picture and shows that most of the heterogeneity is due to within-study differences, with a moderate amount due to between-study differences.

The WPL-approach is the only approach where all available information is included while also explicitly modeling dependency by nesting the effect sizes within studies (Van den Noortgate et al., 2013; Wilson et al., 2016). Because all the available information is used, statistical power is not affected. By nesting effect sizes within studies, the dependency is properly accounted for, and therefore the precision of the estimates is not overestimated. Limitations of the WPL-approach are that is has not been evaluated in a simulation study and that it is not yet frequently used in practice. However, the paper by Wilson and colleagues (2016) describes the procedure extensively and provides the syntax in the supplementary materials. Also, some examples are available (e.g., Graf-Drasch et al., 2019; Loignon & Woehr, 2018). Based on the findings of the current study, the WPL-approach is the theoretically preferred approach for dealing with dependency in MASEM analysis.

Strengths, Limitations, and Future Directions

A strength of the current study is that – to the author's knowledge – this study is the first to compare frequently used (ad hoc) methods for dealing with dependency in MASEM to the relatively new WPL-approach using empirical data. Additionally, this study aspires to facilitate the reproducibility of the analyses. Given that the WPL-approach may be viewed as somewhat more complex, the authors have provided the data (incl. the code book; Stolwijk et al., 2021a) and the R-script (Stolwijk et al., 2021b) for the WPL-approach in PsychArchives. Combined with the extensive description of the procedure by Wilson and colleagues (2016), this should aid interested researchers in conducting MASEM-analysis using the WPL-approach to handle dependency. Lastly, this study gives a comprehensive overview of commonly used approaches for dealing with dependency and shows its pitfalls. Providing an overview of the (dis)advantages hopefully aids researchers to decide on an appropriate method.

The current study is limited in that it offers a comparison based solely on empirical data, and inferences can stretch not much further than to the specifics of the current dataset. However, from this practical application, there is a basis from which to conduct a simulation study in order to examine the robustness of the WPL-approach under ideal and non-ideal conditions (e.g., Hallgren, 2013). For instance, it would be interesting to examine the effects of differences in the amount of overall heterogeneity that can be attributed to within-versus between-studies differences. Additionally, the level of dependency may be altered to evaluate the impact on the performance of the WPL-approach, relative to other approaches. Also, the minimum number of studies necessary to conduct the WPL-approach should be examined.

Conclusion

In summary, dependency is a non-avoidable issue in meta-analytic research. This study demonstrated that using different approaches for dealing with dependency in MASEM leads to different results, which can have important practical implications. Thus, the decision on which approach should be used in MASEM-analysis should be one that is carefully considered. Given that the WPL-approach is the only approach that includes all available information while explicitly modeling dependency, it is currently the theoretically preferred approach for dealing with dependency in MASEM. Future research should evaluate the multilevel approach with simulated data.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/2151-2604/a000485

ESM 1. Table E1: Characteristics of the studies included in the meta-analysis. Table E2: Pooled correlation matrices for all variables. Table E3: Parameter estimates of the Partial Meditation Model. Table E4: Residual (co)variances of the Partial Meditation Model. Figure E1: Funnel plot of all effect sizes. Figure E2: Study-funnel plot. Figure E3: PRISMA flow diagram.

References

*References marked with an asterisk are included in the meta-analysis.

*Aaron, L., & Dallaire, D. H. (2010). Parental incarceration and multiple risk experiences: Effects on family dynamics and children's delinquency. *Journal of Youth and Adolescence, 39*, 1471–1484. https://doi.org/10.1007/s10964-009-9458-0

*Adams, J. B. (2001). Self-control as a mediator for the effects of parental bonding, prosocial behavioural training and psychological autonomy on adolescent delinquency. *Dissertation Abstracts International, 62*, 4(A).

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research, 82*(4), 436–476. https://doi.org/10.3102/0034654312458162

*Allen, J. P., Porter, M. R., & McFarland, F. C. (2005). The two faces of adolescents' success with peers: Adolescent popularity, social adaptation, and deviant behaviour. *Journal of Child Development, 76*(3), 747–760. https://doi.org/10.1111/j.1467-8624.2005.00875.x

*Baldry, A. C., & Farrington, D. P. (2000). Bullies and delinquents: Personal characteristics and parental styles. *Journal of Community and Applied Social Psychology, 10*(1), 17–31. https://doi.org/b39pnk

*Banyard, V. L., Cross, C., & Modecki, K. L. (2006). Interpersonal violence in adolescence: Ecological correlates of self-reported perpetration. *Journal of Interpersonal Violence, 21*(10), 1314–1332. https://doi.org/10.1177/0886260506291657

*Barberet, R., Bowling, B., Junger-Tas, J., Rechea-Alberola, C., Van Kesteren, J., & Zurawan, A. (2004). *Self-reported juvenile delinquency in England and Wales, the Netherlands and Spain*. European Institute for Crime Prevention and Control.

*Barnow, S., Lucht, M., & Freyberger, H. (2005). Correlates of aggressive and delinquent conduct problems in adolescence. *Aggressive Behaviour, 31*, 24–39. https://doi.org/10.1002/ab.20033

Barret, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*, 815–824. https://doi.org/10.1016/j.paid.2006.09.018

*Bean, R. A., Barber, B. K., & Crane, D. R. (2006). Parental support, behavioural control, and psychological control among African American youth: The relationships to academic grades, delinquency, and depression. *Journal of Family Issues, 27*(10), 1335–1355. https://doi.org/10.1177/0192513X06289649

Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics, 17*, 341–362. https://doi.org/10.3102/10769986017004341

Becker, B. J. (1995). Corrections to "Using results from replicated studies to estimate linear models". *Journal of Educational Statistics, 20*, 100–102. https://doi.org/10.3102/10769986020001100

Besemer, S., Ahmad, S. I., Hinshaw, S. P., & Farrington, D. P. (2017). A systematic review and meta-analysis of the intergenerational transmission of criminal behaviour. *Aggression and Violent Behaviour, 37*, 161–178. https://doi.org/10.1016/j.avb.2017.10.004

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386.

*Bowman, M. A., Prelow, H. M., & Weaver, S. R. (2007). Parenting behaviours, association with deviant peers, and delinquency in African American adolescents: A mediated-moderation model. *Journal of Youth and Adolescence, 36*(4), 517–527. https://doi.org/10.1007/s10964-006-9117-7

*Brauer, J. R. (2011). *Autonomy-supportive parenting and adolescent delinquency* [Unpublished doctoral dissertation]. North Carolina State University.

*Brendgen, M., Vitaro, F., Tremblay, R. E., & Lavoie, F. (2001). Reactive and proactive aggression: Predictions to physical violence in different contexts and moderating effects of parental monitoring and caregiving behaviour. *Journal of Abnormal Child Psychology, 29*(4), 293–304. https://doi.org/10.1023/A:1010305828208

*Burton, V. S., Cullen, F. T., Evans, T. D., Dunaway, G. R., Kethineni, S. R., & Payne, G. L. (1995). The impact of parental controls on delinquency. *Journal of Criminal Justice, 23*, 111–126. https://doi.org/10.1016/0047-2352(95)00009-F

*Byrnes, H. F., Miller, B. A., Chen, M. J., & Grube, J. W. (2011). The roles of mothers' neighborhood perceptions and specific monitoring strategies in youths' problem behaviour. *Journal of Youth and Adolescence, 40*, 347–360. https://doi.org/10.1007/s10964-010-9538-1

*Caldwell, R. M., Beutler, L. E., Ross, S. A., & Silver, N. C. (2006). Brief report: An examination of the relationships between parental monitoring, self-esteem and delinquency among Mexican American male adolescents. *Journal of Adolescence, 29*(3), 459–464. https://doi.org/10.1016/j.adolescence.2005.07.005

*Campbell, A. (1987). Self-reported delinquency and home life: Evidence from a sample of British girls. *Journal of Youth and Adolescence, 16*(2), 167–177. https://doi.org/10.1007/BF02138918

*Capaldi, D. M., Pears, K. C., Patterson, G. R., & Owen, L. D. (2003). Continuity of parenting practices across generations in an at-risk sample: A prospective comparison of direct and mediated associations. *Journal of Abnormal Child Psychology, 31*(2), 127–142. https://doi.org/10.1023/A:1022518123387

*Cernkovich, S. A., & Giordano, P. C. (1987). Family relationships and delinquency. *Criminology, 25*(2), 295–321.

*Chapple, C. L. (2003). Examining intergenerational violence: Violent role modeling or weak parental controls? *Violence and Victims, 18*(2), 143–162. https://doi.org/10.1891/vivi.2003.18.2.143

Cheung, M. W.-L. (2014a). Fixed- and random-effect meta-analytic structural equation modelling: Examples and analyses in R. *Behaviour Research methods, 46*, 29–40. https://doi.org/10.3758/s13428-013-0361-y

Cheung, M. W.-L. (2014b). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modelling approach. *Psychological Methods, 19*(2), 211–229. https://doi.org/10.1037/a0032968

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

Cheung, M. W.-L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review, 29*, 387–396. https://doi.org/10.1007/s11065-019-09415-6

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

Cheung, S. F., & Chan, D. K.-S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology, 89*(5), 780–791. https://doi.org/10.1037/0021-9010.89.5.780

*Chhangur, R. R., Overbeek, G., Verhagen, M., Weeland, J., Matthys, W., & Engels, R. C. (2015). DRD4 and DRD2 genes, parenting, and adolescent delinquency: Longitudinal evidence for a gene by environment interaction. *Journal of Abnormal Psychology, 124*, 791–802. https://doi.org/10.1037/abn0000091

*Chung, H. L., & Steinberg, L. (2006). Relations between neighborhood factors, parenting behaviours, peer deviance, and delinquency among serious juvenile offenders. *Developmental Psychology, 42*(2), 319–331. https://doi.org/10.1037/0012-1649.42.2.319

*Conrad, J. B. (2015). *The relationship between parental warmth, the effects of childhood witnessed violence and pre-adolescent delinquency* [Unpublished doctoral dissertation]. Cleveland State University.

*Cook, A. (2009). *Parental competencies of juvenile probationers and adherence to court sanctions and recidivism rates*. Virginia Commonwealth University.

*Coughlin, C., & Vuchinich, S. (1996). Family experience in preadolescence and the development of male delinquency. *Journal of Marriage and the Family, 58*(2), 491–501. https://doi.org/10.2307/353512

*Crane, C. R. (2010). *Peer and neighborhood risk contexts, and adolescents' delinquent behaviours: The protective potential of family and neighborhood connectedness* [Unpublished doctoral dissertation]. Oklahoma State University.

*Criss, M. M. (2002). *How parents find out about their teenagers' activities: Validating an observational measure of monitoring as dyadic process* [Unpublished doctoral dissertation]. Auburn University.

*Crosswhite-Gamble, J. (2006). *Mediating mechanisms: Understanding the link between parenting and adolescent deviance* [Unpublished doctoral dissertation]. Auburn University.

*Dawes, K. J. (1976, April). *Parent-child relationships, parental role models and reference others–their joint impact on juvenile delinquency*. Paper presented at the Annual Meeting of the Midwest Sociological Society, St. Louis, MO.

*De Kemp, R. A. T., Scholte, R. H. J., Overbeek, G., & Engels, R. C. M. E. (2004). Opvoeding, delinquente vrienden en delinquent gedrag van jongeren [Parenting, delinquent friends and delinquent behaviour of adolescents]. *Pedagogiek, 24*(3), 262–278.

*De Vries, S. L., Hoeve, M., Stams, G. J. J., & Asscher, J. J. (2016). Adolescent-parent attachment and externalizing behaviour: The mediating role of individual and social factors. *Journal of Abnormal Child Psychology, 44*, 283–294. https://doi.org/10.1007/s10802-015-9999-5

*Deutsch, A. R., Crockett, L. J., Wolff, J. M., & Russell, S. T. (2012). Parent and peer pathways to adolescent delinquency: Variations by ethnicity and neighborhood context. *Journal of Youth and Adolescence, 41*, 1078–1094. https://doi.org/10.1007/s10964-012-9754-y

*Dishion, T. J., Owen, L. D., & Bullock, B. M. (2004). Like father, like son: Toward a developmental model for the transmission of male deviance across generations. *European Journal of Developmental Psychology, 1*, 105–126. https://doi.org/10.1080/17405620444000094

*Dishion, T. J., Patterson, G. R., Stoolmiller, M., & Skinner, M. L. (1991). Family, school, and behavioural antecedents to early adolescent involvement with antisocial peers. *Developmental Psychology, 27*(1), 172–180. https://doi.org/10.1037/0012-1649.27.1.172

*Dodge, K. A., Greenberg, M. T., & Malone, P. S. (2008). Testing an idealized dynamic cascade model of the development of serious violence in adolescence. *Child Development, 79*, 1907–1927. https://doi.org/10.1111/j.1467-8624.2008.01233.x

*Eaton, N. R., Krueger, R. F., Johnson, W., McGue, M., & Iacono, W. G. (2009). Parental monitoring, personality, and delinquency: Further support for a reconceptualization of monitoring. *Journal of Research in Personality, 43*, 49–59. https://doi.org/10.1016/j.jrp.2008.10.006

*Edens, J. F., Skopp, N. A., & Cahill, M. A. (2008). Psychopathic features moderate the relationship between harsh and inconsistent parental discipline and adolescent antisocial behaviour. *Journal of Clinical Child & Adolescent Psychology, 37*, 472–476. https://doi.org/10.1080/15374410801955938

*Estevez, E., Musitu, G., & Herrero, J. (2005). The influence of violent behaviour and victimization at school on psychological distress: The role of parents and teachers. *Adolescence, 40*(157), 183–196.

*Evans, S. Z., Simons, L. G., & Simons, R. L. (2012). The effect of corporal punishment and verbal abuse on delinquency: Mediating mechanisms. *Journal of Youth and Adolescence, 41*, 1095–1110. https://doi.org/10.1007/s10964-012-9755-x

*Farrington, D. P., Jolliffe, D., Loeber, R., Stouthamer-Loeber, M., & Kalb, L. M. (2001). The concentration of offenders in families, and family criminality in the prediction of boys' delinquency. *Journal of Adolescence, 24*, 579–596. https://doi.org/10.1006/jado.2001.0424

*Farrington, D. P., Loeber, R., Yin, Y., & Anderson, S. J. (2002). Are within-individual causes of delinquency the same as between-individual causes? *Criminal Behaviour and Mental Health, 12*(1), 53–68. https://doi.org/10.1002/cbm.486

*Farrington, D. P., Ttofi, M. M., Crago, R. V., & Coid, J. W. (2015). Intergenerational similarities in risk factors for offending. *Journal of Developmental and Life-Course Criminology, 1*, 48–62. https://doi.org/10.1007/s40865-015-0005-2

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). Visual representation of meta-analyses of multiple outcomes: Extensions to forest plots, funnel plots, and caterpillar plots. *Methodology, 16*(4), 299–315. https://doi.org/10.5964/meth.4013

*Finkenauer, C., Engels, R. C. M. E., & Baumeister, R. F. (2005). Parenting behaviour and adolescent behavioural and emotional problems: The role of self-control. *International Journal of Behavioural Development, 29*(1), 58–69. https://doi.org/10.1080/01650250444000333

*Flannery, D. J., Williams, L. L., & Vazsonyi, A. T. (1999). Who are they with and what are they doing? Delinquent behaviour, substance use, and early adolescents' after-school time. *American Journal of Orthopsychiatry, 69*(2), 247–253. https://doi.org/10.1037/h0080426

*Fletcher, A. C., Steinberg, L., & Williams-Wheeler, M. (2004). Parental influences on adolescent problem behaviour: Revisiting Stattin and Kerr. *Child Development, 75*(3), 781–796. https://doi.org/10.1111/j.1467-8624.2004.00706.x

Funder, D. C., & Ozer, D. K. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202

*Gainey, R. R., Catalano, R. F., Haggerty, K. P., & Hoppe, M. J. (1997). Deviance among the children of heroin addicts in treatment: Impact of parents and peers. *Deviant Behaviour, 18*(2), 143–159. https://doi.org/10.1080/01639625.1997.9968050

*Gault-Sherman, M. (2012). It's a two-way street: The bidirectional relationship between parenting and delinquency. *Journal of Youth and Adolescence, 41*, 121–145. https://doi.org/10.1007/s10964-011-9656-4

*Giever, D. M. (1996). An empirical assessment of the core elements of Gottfredson and Hirschi's general theory of crime. *Dissertation Abstracts International, 56*, 4155-A.

*Gold, J., Sullivan, M. W., & Lewis, M. (2011). The relation between abuse and violent delinquency: The conversion of shame to blame in juvenile offenders. *Child Abuse & Neglect, 35*, 459–467. https://doi.org/10.1016/j.chiabu.2011.02.007

Graf-Drasch, V., Gimpel, H., & Barlow, J. B. (2019, June). *Clarifying the structure of collective intelligence in teams: A meta-analysis*. Paper presented at the Proceedings of the Collective Intelligence Conference, Pittsburgh, PA.

*Gray-Ray, P., & Ray, M. C. (1990). Juvenile delinquency in the Black community. *Youth and Society, 22*(1), 67–84. https://doi.org/10.1177/0044118X90022001005

*Griffin, K. W., Botvin, G. J., Scheier, L. M., Diaz, T., & Miller, N. L. (2000). Parenting practices as predictors of substance use, delinquency, and aggression among urban minority youth: Moderating effects of family structure and gender. *Psychology of Addictive Behaviours, 14*(2), 174–184. https://doi.org/10.1037/0893-164X.14.2.174

*Guimond, F. A., Laursen, B., Vitaro, F., Brendgen, M., Dionne, G., & Boivin, M. (2016). Associations between mother–child relationship quality and adolescent adjustment: Using a genetically controlled design to determine the direction and magnitude of effects. *International Journal of Behavioural Development, 40*, 196–204. https://doi.org/10.1177/0165025415620059

*Haapasalo, J. (2000). Young offenders' experiences of child protection services. *Journal of Youth and Adolescence, 29*, 355–371. https://doi.org/10.1023/A:1005151809736

*Hair, E. C., Moore, K. A., Garrett, S. B., Ling, T., & Cleveland, K. (2008). The continued importance of quality parent–adolescent relationships during late adolescence. *Journal of Research on Adolescence, 18*, 187–200. https://doi.org/10.1111/j.1532-7795.2008.00556.x

*Halgunseth, L. C., Perkins, D. F., Lippold, M. A., & Nix, R. L. (2013). Delinquent-oriented attitudes mediate the relation between parental inconsistent discipline and early adolescent behaviour. *Journal of Family Psychology, 27*, 293–302. https://doi.org/10.1037/a0031962

Hallgren, K. A. (2013). Conducting simulation studies in the R programming environment. *Tutorials in Quantitative Methods for Psychology, 9*(2), 43–60. https://doi.org/10.20982/tqmp.09.2.p043

*Harris, C., Vazsonyi, A. T., & Bolland, J. M. (2017). Bidirectional relationships between parenting processes and deviance in a sample of inner-city African American youth. *Journal of Research on Adolescence, 27*, 201–213. https://doi.org/10.1111/jora.12267

*Hay, C. (2001). Parenting, self-control, and delinquency: A test of self-control theory. *Criminology, 39*, 707–736. https://doi.org/10.1111/j.1745-9125.2001.tb00938.x

*Hay, C. (2003). Family strain, gender, and delinquency. *Sociological Perspectives, 46*(1), 107–135. https://doi.org/10.1525/sop.2003.46.1.107

*Haynie, D. L. (2003). Contexts of risk? Explaining the link between girls' pubertal development and their delinquency involvement. *Social Forces, 82*(1), 355–397. https://doi.org/10.1353/sof.2003.0093

*Heaven, P. (1994). Family of origin, personality, and self-reported delinquency. *Journal of Adolescence, 17*, 445–459. https://doi.org/10.1006/jado.1994.1038

*Heaven, P. C. L., Newbury, K., & Mak, A. (2004). The impact of adolescent and parental characteristics on adolescent levels of delinquency and depression. *Personality and Individual Differences, 36*(1), 173–185. https://doi.org/10.1016/S0191-8869(03)00077-1

*Henneberger, A. K., Durkee, M. I., Truong, N., Atkins, A., & Tolan, P. H. (2013). The longitudinal relationship between peer violence and popularity and delinquency in adolescent boys: Examining effects by family functioning. *Journal of Youth and Adolescence, 42*, 1651–1660. https://doi.org/10.1007/s10964-012-9859-3

*Henneberger, A. K., Tolan, P. H., Hipwell, A. E., & Keenan, K. (2014). Delinquency in adolescent girls: Using a confluence approach to understand the influences of parents and peers. *Criminal Justice and Behaviour, 41*, 1327–1337. https://doi.org/10.1177/0093854814538624

*Herman, M. R., Dornbusch, S. M., Herron, M. C., & Herting, J. R. (1997). The influence of family regulation, connection, and psychological autonomy on six measures of adolescent functioning. *Journal of Adolescent Research, 12*(1), 34–67.

*Herrenkohl, T. I., Maguin, E., Hill, K. G., Hawkins, J., Abbott, R. D., & Catalano, R. F. (2000). Developmental risk factors for youth violence. *Journal of Adolescent Health, 26*(3), 176–186. https://doi.org/10.1177/0743554897121004

*Herrington, L. L. (2015). *Unique and combined contributions of callous-unemotional traits and parental incarceration on juvenile delinquency in an at-risk sample* [Unpublished doctoral dissertation]. University of Southern Mississippi.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557

*Hill, G. D., & Atkinson, M. P. (1988). Gender, familial control, and delinquency. *Criminology: An Interdisciplinary Journal, 26*(1), 127–147. https://doi.org/10.1111/j.1745-9125.1988.tb00835.x

Hoeve, M., Dubas, J. S., Eichelsheim, V. I., Van der Laan, P. H., Smeenk, W., & Gerris, J. R. M. (2009). The relationship between parenting and delinquency: A meta-analysis. *Journal of Abnormal Child Psychology, 37*, 749–775. https://doi.org/10.1007/s10802-009-9310-8

*Hoeve, M., Dubas, J. S., Gerris, J. R., van der Laan, P. H., & Smeenk, W. (2011). Maternal and paternal parenting styles: Unique and combined links to adolescent and early adult delinquency. *Journal of Adolescence, 34*, 813–827. https://doi.org/10.1016/j.adolescence.2011.02.004

*Hoeve, M., Smeenk, W. H., Loeber, R., Stouthamer-Loeber, M., Van der Laan, P. H., Gerris, J. R. M., & Semon-Dubas, J. (2007). Long term effects of parenting and family characteristics on delinquency of male young adults. *European Journal of Criminology, 4*(2), 116–194. https://doi.org/10.1177/1477370807074854

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modelling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

*Intravia, J., Jones, S., & Piquero, A. (2012). The roles of social bonds, personality, and perceived costs: An empirical investigation into Hirschi's "New" Control Theory. *International Journal of Offender Therapy and Comparative Criminology, 56*, 1182–1200. https://doi.org/10.1177/0306624X11422998

Jak, S. (2015). *Meta-analytic structural equation modeling. SpringerBriefs in research synthesis and meta-analysis*. Springer International Publishing.

Jak, S., & Cheung, M. W.-L. (2018). Testing moderator hypotheses in meta-analytic structural equation modeling using subgroup-analysis. *Behavioural Research Methods, 50*, 1359–1373. https://doi.org/10.3758/s13428-018-1046-3

*Janssen, H., Eichelsheim, V., Dekovic, M., & Bruinsma, G. (2016). How is parenting related to adolescent delinquency? A between- and within-person analysis of the mediating role of self-control, delinquent attitudes, peer delinquency, and time spent in criminogenic settings. *European Journal of Criminology, 13*, 169–194. https://doi.org/10.1177/1477370815608881

Johnson, J. G., Smailes, E., Cohen, P., Kasen, S., & Brook, J. S. (2004). Anti-social parental behaviour, problematic parenting and aggressive offspring behaviour during adulthood: A 25-year longitudinal investigation. *British Journal of Criminology, 44*, 915–930. https://doi.org/10.1093/bjc/azh041

*Johnson, R. E. (1987). Mother's versus father's role in causing delinquency. *Adolescence, 22*(86), 305–315.

*Jones, D. J., Forehand, R., & Beach, S. R. H. (2000). Maternal and paternal parenting during adolescence: Forecasting early adult psychosocial adjustment. *Adolescence, 35*(139), 513–530.

*Jug, V. (2015). *Personality traits, social intelligence, social support and juvenile delinquency*. [Conference session]. International Conference on Advances in Social Science, Economics and Management Study – SEM, Birmingham, United Kingdom.

*Juras, J. L. (2004). *Integrating models of risk and protection for the prevention of adolescent delinquency* [Unpublished doctoral dissertation]. Michigan State University.

Keijsers, L., Branje, S., Hawk, S. T., Schwartz, S. J., Frijns, T., Koot, H. M., Van Lier, P., & Meeus, W. (2012). Forbidden friends as forbidden fruit: Parental supervision of friendships, contact with deviant peers, and adolescent delinquency. *Child Development, 83*, 651–666. https://doi.org/10.1111/j.1467-8624.2011.01701.x

*Keijsers, L., Branje, S. J., VanderValk, I. E., & Meeus, W. (2010). Reciprocal effects between parental solicitation, parental control, adolescent disclosure, and adolescent delinquency. *Journal of Research on Adolescence, 20*, 88–113. https://doi.org/10.1111/j.1532-7795.2009.00631.x

*Kerr, M., Stattin, H., & Trost, K. (1999). To know you is to trust you: Parents' trust is rooted in child disclosure of information. *Journal of Adolescence, 22*(6), 737–752. https://doi.org/10.1006/jado.1999.0266

*Kjellstrand, J. M., & Eddy, J. M. (2011). Parental incarceration during childhood, family context, and youth problem behaviour across adolescence. *Journal of Offender Rehabilitation, 50*, 18–36. https://doi.org/10.1080/10509674.2011.536720

*Krohn, M., & Massey, J. L. (1980). Social control and delinquent behaviour: An examination of the elements of the social bonds. *Sociological Quarterly, 21*, 529–543. https://doi.org/10.1111/j.1533-8525.1980.tb00634.x

*Krohn, M., Stern, S. B., Thornberry, T., & Jang, S. J. (1992). The measurement of family process variables: An examination of adolescent and parent perception of family life and delinquent behaviour. *Journal of Quantitative Criminology, 8*, 287–315. https://doi.org/10.1007/BF01064550

*Kupanoff, K. M. (2002). Adolescent behavioural autonomy as a moderator between parenting practices and adolescent delinquency. *Dissertation Abstracts International, 63*, 5(B).

*Kwon, J. A., & Wickrama, K. A. S. (2014). Linking family economic pressure and supportive parenting to adolescent health behaviours: Two developmental pathways leading to health promoting and health risk behaviours. *Journal of Youth and Adolescence, 43*, 1176–1190. https://doi.org/10.1007/s10964-013-0060-0

*Lahlah, E., Lens, K. M., Bogaerts, S., & van der Knaap, L. M. (2013). When love hurts: Assessing the intersectionality of ethnicity, socio-economic status, parental connectedness, child abuse, and gender attitudes in juvenile violent delinquency. *Child Abuse and Neglect, 37*, 1034–1049. https://doi.org/10.1016/j.chiabu.2013.07.001

*Lahlah, E., Van der Knaap, L. M., Bogaerts, S., & Lens, K. M. (2014). Ethnic differences in the effect of perceived parenting on juvenile violent delinquency of Dutch and Moroccan-Dutch boys. *Journal of Child and Family Studies, 23*, 333–346. https://doi.org/10.1007/s10826-013-9725-2

*Laible, D., Carlo, G., Davis, A., & Karahuta, E. (2016). Maternal sensitivity and effortful control in early childhood as predictors of adolescents' adjustment: The mediating roles of peer group affiliation and social behaviours. *Developmental Psychology, 52*, 922–932. https://doi.org/10.1037/dev0000118

*Larzelere, R. E., & Patterson, G. (1990). Parental management: Mediator of the effect of socioeconomic status on early delinquency. *Criminology, 28*, 301–323.

Lipsey, M., & Wilson, D. (2001). *Practical Meta-analysis*. Sage.

Loignon, A. C., & Woehr, D. J. (2018). Social class in the organizational sciences: A conceptual integration and meta-analytic review. *Journal of Management, 44*(1), 61–88. https://doi.org/10.1177/0149206317728106

López-López, J. A., Van den Noortgate, W., Tanner-Smith, E. E., Wilson, S. J., & Lipsey, M. W. (2017). Assessing meta-regression methods for examining moderator relationships with dependent effect sizes: A Monte Carlo simulation. *Research Synthesis Methods, 8*, 435–450. https://doi.org/10.1002/jrsm.1245

López-López, J. A., Page, M. J., Lipsey, M. W., & Higgins, J. P. T. (2018). Dealing with dependent effect size multiplicity in systematic reviews and meta-analyses. *Research Synthesis Methods, 9*, 336–351. https://doi.org/10.1002/jrsm.1310

*Loukas, A., Suizzo, M. A., & Prelow, H. M. (2007). Examining resource and protective factors in the adjustment of Latino youth in low income families: What role does maternal acculturation play? *Journal of Youth and Adolescence, 36*(4), 489–501. https://doi.org/10.1007/s10964-006-9124-8

*Luo, Q. (2000). *Parenting and friend affiliation in adolescent development: A cross-cultural comparison* [Unpublished doctoral dissertation]. Wayne State University.

*Mak, A. S. (1994). Parental neglect and overprotection as risk factors in delinquency. *Australian Journal of Psychology, 46*(2), 107–111. https://doi.org/10.1080/00049539408259481

*Manders, W., Scholte, R., Janssens, J., & De Bruyn, E. (2006). Adolescent personality, problem behaviour and the quality of the parent-adolescent relationship. *European Journal of Personality, 20*, 237–254. https://doi.org/10.1002/per.574

*Mann, F., Kretsch, N., Tackett, J., Harden, K., & TuckerDrob, E. (2015). Person × environment interactions on adolescent delinquency: Sensation seeking, peer deviance and parental monitoring. *Personality and Individual Differences, 76*, 129–134. https://doi.org/10.1016/j.paid.2014.11.055

Marín-Martínez, F., & Sánchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology, 2*(1), 32–38. https://doi.org/10.1017/S1138741600005436

*Mason, C. A. (1996). Neither too sweet nor too sour: Problem peers, maternal control, and problem behaviour in African American adolescents. *Child Development, 67*(5), 2115–2130. https://doi.org/10.2307/1131613

*Mathis, C. W. (2013). *Children's delinquency after paternal incarceration* [Unpublished doctoral dissertation]. Texas A&M University.

*McCord, J. (1991). Family relationships, juvenile delinquency and adult criminality. *Criminology, 29*, 397–417.

Moeyaert, M., Ugille, M., Natasha Beretvas, S., 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*(6), 559–572. https://doi.org/10.1080/13645579.2016.1252189

*Mulvey, E. P., Steinberg, L., Piquero, A. R., Besana, M., Fagan, J., Schubert, C., & Cauffman, E. (2010). Trajectories of desistance and continuity in antisocial behaviour following court adjudication among serious adolescent offenders. *Developmental Psychopathology, 22*, 453–475. https://doi.org/10.1017/S0954579410000179

*Murray, J., Janson, C. G., & Farrington, D. P. (2007). Crime in adult offspring of prisoners: A cross-national comparison of two longitudinal samples. *Criminal Justice and Behaviour, 34*, 133–149. https://doi.org/10.1177/0093854806289549

*Nijhof, K. S., De Kemp, R. A. T., & Engels, R. C. M. E. (2009). Frequency and seriousness of parental offending and their impact on juvenile offending. *Journal of Adolescence, 32*, 893–908. https://doi.org/10.1016/j.socscimed.2008.10.005

*O'Connor, B. P., & Dvorak, T. (2001). Conditional associations between parental behaviour and adolescent problems: A search for personality-environment interactions. *Journal of Research in Personality, 35*(1), 1–26. https://doi.org/10.1006/jrpe.2000.2295

Olkin, I., & Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. In S. Ikeda (Ed.), *Essays in probability and statistics* (pp. 235–251). Shinko Tsusho Co., Ltd.

Page, M. J., McKenzie, J. E., Chau, N., Green, S. E., & Forbes, A. (2015). Methods to select results to include in meta-analyses deserve more consideration in systematic reviews. *Journal of Clinical Epidemiology, 68*, 1282–1291. https://doi.org/10.1016/j.jclinepi.2015.02.009

*Park, S., Morash, M., & Stevens, T. (2010). Gender differences in predictors of assaultive behaviour in late adolescence. *Youth Violence and Juvenile Justice, 8*, 314–331. https://doi.org/10.1177/1541204009361173

*Paschall, M. J., Ringwalt, C. L., & Flewelling, R. L. (2003). Effects of parenting, father absence, and affiliation with delinquent peers on delinquent behaviour among African-American male adolescents. *Adolescence, 38*(149), 15–34.

*Patouris, E., Scaife, V., & Nobes, G. (2016). A behavioural approach to adolescent cannabis use: Accounting for nondeliberative, developmental, and temperamental factors. *Journal* *of Substance Use, 21*, 506–514. https://doi.org/10.3109/14659891.2015.1076076

*Patterson, G. R., & Dishion, T. J. (1985). Contributions of families and peers to delinquency. *Criminology, 23*(1), 63–79. https://doi.org/10.1111/j.1745-9125.1985.tb00326.x

*Peterson, D. (2002). *"Don't forget the women": A multi-level analysis of individual and contextual effects on girls' and boys' delinquency* [Unpublished doctoral dissertation]. University of Nebraska.

*Pettit, G. S., Laird, R. D., Dodge, K. A., Bates, J. E., & Criss, M. M. (2001). Antecedents and behaviour-problem outcomes of parental monitoring and psychological control in early adolescence. *Child Development, 72*(2), 583–598. https://doi.org/10.1111/1467-8624.00298

*Prinzie, P., Onghena, P., Hellinckx, W., Grietens, H., Chesquière, P., & Colpin, H. (2004). Parent and child personality characteristics as predictors of negative discipline and externalizing problem behaviour in children. *European Journal of Personality, 18*, 73–102. https://doi.org/10.1002/per.501

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.R-project.org/

*Rankin, J. H., & Kern, R. (1994). Parental attachments and delinquency. *Criminology, 32*(4), 495–515. https://doi.org/10.1111/j.1745-9125.1994.tb01163.x

*Ray, J. V., Frick, P. J., Thornton, L. C., Wall Myers, T. D., Steinberg, L., & Cauffman, E. (2017). Callous–unemotional traits predict self-reported offending in adolescent boys: The mediating role of delinquent peers and the moderating role of parenting practices. *Developmental Psychology, 53*, 319–328. https://doi.org/10.1037/dev0000210

Rios, J. A., Ihlenfeldt, S. D., Dosedel, M., & Riegelman, A. (2020). A topical and methodological systematic review of meta-analyses published in the educational measurement literature. *Educational Measurement: Issues and Practice, 39*(1), 71–81. https://doi.org/10.1111/emip.12282

*Roberts, J. (2002). *Family rituals and deviant behaviour* [Unpublished doctoral dissertation]. University of North Texas.

*Samaniego, R. Y., & Gonzales, N. A. (1999). Multiple mediators of the effects of acculturation status on delinquency for Mexican American adolescents. *American Journal of Community Psychology, 27*(2), 189–210. https://doi.org/10.1023/A:1022883601126

*Sampson, R. J., & Laub, J. H. (1994). Urban poverty and the family context of delinquency: A new look at structure and process in a classic study. *Child Development, 65*(2), 523–540. https://doi.org/10.2307/1131400

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Sage.

*Scholte, E. M. (1999). Factors predicting continued violence into young adulthood. *Journal of Adolescence, 22*(1), 3–20. https://doi.org/10.1006/jado.1998.0197

Silva Pinho, A. F. (2018). *The effects of child personality traits on parenting and child delinquency: A meta-analytic approach* [Unpublished master's thesis]. University of Amsterdam.

*Simon, R., Simons, L., Chen, Y., Brody, G., & Lin, K. (2007). Identifying the psychological factors that mediate the association between parenting practices and delinquency. *Criminology: An Interdisciplinary Journal, 45*, 481–517. https://doi.org/10.1111/j.1745-9125.2007.00086.x

*Simons, R. L., Robertson, J. F., & Downs, W. R. (1989). The nature of the association between parental rejection and delinquent behaviour. *Journal of Youth and Adolescence, 18*(3), 297–310. https://doi.org/10.1007/BF02139043

*Skinner, K. B. (2000). Associations between parenting, acculturation, and adolescent functioning among Chinese families in North America. *International, 61*(2), 784(A).

*Spano, R., Vazsonyi, A. T., & Bolland, J. (2009). Does parenting mediate the effects of exposure to violence on violent behaviour? An ecological–transactional model of community violence. *Journal of Adolescence, 32*, 1321–1341. https://doi.org/10.1016/j.adolescence.2008.12.003

*Stewart, E. A., Simons, R. L., Conger, R. D., & Scaramella, L. V. (2002). Beyond the interactional relationship between delinquency and parenting practices: The contribution of legal sanctions. *Journal of Research in Crime and Delinquency, 39*(1), 36–59. https://doi.org/10.1177/002242780203900102

Stolwijk, I. J., Jak, S., Eichelsheim, V. I., & Hoeve, M. (2021). *Dealing with dependent effect sizes in MASEM: A comparison of different approaches using empirical data*. [Data set and code book]. PsychArchives. http://dx.doi.org/10.23668/psycharchives.5130

Stolwijk, I. J., Jak, S., Eichelsheim, V. I., & Hoeve, M. (2021). *Dealing with dependent effect sizes in MASEM: A comparison of different approaches using empirical data*. [Code]. PsychArchives. http://dx.doi.org/10.23668/psycharchives.5129

*Svensson, R., Weerman, F. M., Pauwels, L. J. R., Bruinsma, G. J. N., & Bernasco, W. (2013). Moral emotions and offending: Do feelings of anticipated shame and guilt mediate the effect of socialization on offending? *European Journal of Criminology, 10*, 22–39. https://doi.org/10.1177/1477370812454393

*Tilton-Weaver, L. (2014). Adolescents' information management: Comparing ideas about why adolescents disclose to or keep secrets from their parents. *Journal of Youth and Adolescence, 43*, 803–813. https://doi.org/10.1007/s10964-013-0008-4

*Torrente, G., & Vazsonyi, A. T. (2012). Personality, parenting and deviance among Spanish adolescents. *Anales de Psicologia, 28*, 654–664. https://doi.org/10.6018/analesps.28.3.155951

*Unnever, J. D., Cullen, F. T., & Agnew, R. (2006). Why is "bad" parenting criminogenic?: Implications from rival theories. *Youth Violence and Juvenile Justice, 4*(1), 3–33. https://doi.org/10.1177/1541204005282310

Van den Berg, T. (2018). *A comparison of meta-analytic structural equation modeling and univariate meta-analysis: An application in forensic child and youth care sciences* [Unpublished master's thesis]. University of Amsterdam.

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

*Van der Graaff, J., Branje, S., De Wied, M., & Meeus, W. (2012). The moderating role of empathy in the association between parental support and adolescent aggressive and delinquent behaviour. *Aggressive Behaviour, 38*, 368–377. https://doi.org/10.1002/ab.21435

*Van Voorhis, P., Cullen, F. T., Mathers, R. A., & Garner, C. C. (1988). The impact of family structure and quality on delinquency: A comparative assessment of structural and functional factors. *Criminology, 26*, 235–261. https://doi.org/10.1111/j.1745-9125.1988.tb00840.x

*Van Vugt, E., Loeber, R., & Pardini, D. (2016). Why is young maternal age at first childbirth a risk factor for persistent delinquency in their male offspring? Examining the role of family and parenting factors. *Criminal Behaviour and Mental Health, 26*, 322–335. https://doi.org/10.1002/cbm.1959

*Vazsonyi, A. T., & Flannery, D. J. (1997). Early adolescent delinquent behaviours: Associations with family and school domains. *Journal of Early Adolescence, 17*(3), 271–293. https://doi.org/10.1177/0272431697017003002

*Vazsonyi, A. T., Ksinan Jiskrova, G., Ksinan, A. J., & Blatny, M. (2016). An empirical test of self-control theory in Roma adolescents. *Journal of Criminal Justice, 44*, 66–76. https://doi.org/10.1016/j.jcrimjus.2015.12.004

*Vazsonyi, A. T., Trejos-Castillo, E., & Young, M. A. (2008). Rural and non-rural African American youth: Does context matter in the etiology of problem behaviours? *Journal of Youth and Adolescence, 37*, 798–811. https://doi.org/10.1007/s10964-007-9239-6

*Vega, W. A., Gil, A. G., Warheit, G. J., Zimmerman, R. S., & Apospori, E. (1993). Acculturation and delinquent behaviour among Cuban American adolescents: Toward an empirical model. *American Journal of Community Psychology, 21*(1), 113–125. https://doi.org/10.1007/BF00938210

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

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

*Vitaro, F., Brendgen, M., & Tremblay, R. E. (2000). Influence of deviant friends on delinquency: Searching for moderator variables. *Journal of Abnormal Child Psychology, 28*(4), 313–325. https://doi.org/10.1023/A:1005188108461

*Walker-Barnes, C. J., & Mason, C. A. (2001). Ethnic differences in the effect of parenting on gang involvement and gang delinquency: A longitudinal, hierarchical linear modeling perspective. *Child Development, 72*(6), 1814–1831. https://doi.org/10.1111/1467-8624.00380

*Walters, G. (2014). Pathways to early delinquency: Exploring the individual and collective contributions of difficult temperament, low maternal involvement, and externalizing behaviour. *Journal of Criminal Justice, 42*, 321–326. https://doi.org/10.1016/j.jcrimjus.2014.04.003

*Weintraub, K. J., & Gold, M. (1991). Monitoring and delinquency. *Criminal Behaviour and Mental Health, 1*(3), 268–281.

*Wells, L. E., & Rankin, J. H. (1988). Direct parental controls and delinquency. *Criminology, 26*, 263–285. https://doi.org/10.1111/j.1745-9125.1988.tb00841.x

*Werner, N. E., & Silbereisen, R. K. (2003). Family relationship quality and contact with deviant peers as predictors of adolescent problem behaviours: The moderating role of gender. *Journal of Adolescent Research, 8*(5), 454–480. https://doi.org/10.1177/0743558403255063

*Williams, A. J. (2004). Risk factors for selected health-related behaviours among American Indian Adolescents. *Dissertation Abstracts International, 65*, 3(B).

*Williams, L. R., & Steinberg, L. (2011). Reciprocal relations between parenting and adjustment in a sample of juvenile offenders. *Child Development, 82*, 633–645. https://doi.org/10.1111/j.1467-8624.2010.01523.x

Wilson, S. J., Polanin, J. R., & Lipsey, M. W. (2016). Fitting meta-analytic structural equation models with complex datasets. *Research Synthesis Methods, 7*(2), 121–139. https://doi.org/10.1002/jrsm.1199

*Wissink, I. B., Dekovic, M., & Meijer, A. M. (2006). Parenting behaviour, quality of the parent-adolescent relationship, and adolescent functioning in four ethnic groups. *Journal of Early Adolescence, 26*(2), 133–159. https://doi.org/10.1177/0272431605285718

*Wolfe, T., & Shoemaker, D. (1999). Actor, situation, and context: A framework for delinquency theory integration. *American Journal of Criminal Justice, 24*, 117–138. https://doi.org/10.1007/BF02887621

*Wolff, J. M., & Crockett, L. J. (2011). The role of deliberative decision making, parenting, and friends in adolescent risk behaviours. *Journal of Youth and Adolescence, 40*, 1607–1622. https://doi.org/10.1007/s10964-011-9644-8

*Worthen, M. G. F. (2011). Gender differences in parent–child bonding: Implications for understanding the gender gap in delinquency. *Journal of Criminal Justice, 34*, 3–23. https://doi.org/10.1080/0735648X.2011.554744

*Wright, J. P., & Cullen, F. T. (2001). Parental efficacy and delinquent behaviour: Do control and support matter? *Criminology, 39*(3), 677–705. https://doi.org/10.1111/j.1745-9125.2001.tb00937.x

Yuan, K.-H. (2016). Meta analytical structural equation modeling: Comments on issues with current methods and viable alternatives. *Research Synthesis Methods, 7*, 215–231. https://doi.org/10.1002/jrsm.1213

History
Received April 30, 2021
Revision received September 24, 2021
Accepted November 2, 2021
Published online February 2, 2022

Acknowledgments
This research was presented at the Research Synthesis & Big Data (virtual) Conference in May 2021. We thank Tessa van den Berg and Ana Silva Pinho for their efforts in the data collection.

Funding
Suzanne Jak was supported by the Dutch Research Council with the grant VI.Vidi.201.009.

ORCID
Isidora Stolwijk
https://orcid.org/0000-0003-4468-0354

Isidora Stolwijk
Methods and Statistics
Child Development and Education
University of Amsterdam
Nieuwe Achtergracht 127
1018 WS
Amsterdam
The Netherlands
i.j.stolwijk@uva.nl

Review Article

Human-Like Robots and the Uncanny Valley

A Meta-Analysis of User Responses Based on the Godspeed Scales

Martina Mara[1], Markus Appel[2], and Timo Gnambs[3]

[1]LIT Robopsychology Lab, Johannes Kepler University Linz, Austria
[2]Psychology of Communication and New Media, University of Würzburg, Germany
[3]Leibniz Institute for Educational Trajectories (LIfBi), University of Bamberg, Germany

Abstract: In the field of human-robot interaction, the well-known uncanny valley hypothesis proposes a curvilinear relationship between a robot's degree of human likeness and the observers' responses to the robot. While low to medium human likeness should be associated with increased positive responses, a shift to negative responses is expected for highly anthropomorphic robots. As empirical findings on the uncanny valley hypothesis are inconclusive, we conducted a random-effects meta-analysis of 49 studies (total N = 3,556) that reported 131 evaluations of robots based on the Godspeed scales for anthropomorphism (i.e., human likeness) and likeability. Our results confirm more positive responses for more human-like robots at low to medium anthropomorphism, with moving robots rated as more human-like but not necessarily more likable than static ones. However, because highly anthropomorphic robots were sparsely utilized in previous studies, no conclusions regarding proposed adverse effects at higher levels of human likeness can be made at this stage.

Keywords: uncanny valley, humanoid robot, anthropomorphism, likeability, meta-analysis

When people think of robots, they usually have an image of a human-like machine in their minds: an apparatus with arms, legs, and a head, covered in metal or possibly silicone skin (see Cave et al., 2020; Mara et al., 2020). Even though such robots hardly, if at all, exist in our everyday lives, media reports about engineering advancements and science fiction stories about the – sometimes more, sometimes less peaceful – relationship between humans and their robotic counterparts have long made us wonder what it would be like if humanoid machines were really among us. Given the diffuse mental pictures many people have about robots, representative survey data show that many people are skeptical regarding their use in everyday life (e.g., Gnambs, 2019; Gnambs & Appel, 2019). One of the most popular conceptual frameworks to speculate about human responses to human-like robots is the uncanny valley hypothesis (Mori, 1970). Its central proposition is that increasing anthropomorphism (i.e., human likeness) in artificial characters does not necessarily go hand in hand with increasing likeability but will result in negative responses when the degree of human resemblance is very high, yet not perfect. Over the past decade, the number of empirical investigations of human-robot relationships and determinants of robot acceptance has steadily increased, many of which have dealt with potentially aversive reactions to human-like machines. However, due to inconsistent empirical evidence, the existence of the uncanny valley effect and the conditions under which it is more or less pronounced are a matter of debate (see Kätsyri et al., 2015; Wang et al., 2015; Zhang et al., 2020). Given the great popularity of the uncanny valley hypothesis, it is surprising that its basic propositions still lack systematic empirical corroboration. We address this gap by conducting the first meta-analytic test of the curvilinear relationship between the human likeness and the likeability of robots as proposed by Mori (1970).

Human-Like Robots

From mythological figures such as the Golem to modern-day science fiction, stories about artificial replications of the human species were told throughout history. Starting in the 18th century, there have also been attempts to physically create human-like machines. Around the first industrial revolution, watchmakers and mechanical engineers constructed life-sized automatons in the shape of adult humans that appeared as if they could write, draw, or play chess (see Voskuhl, 2013). When the term "robot" was first ever used in the context of the 1920 theater play "Rossum's Universal Robots" (Čapek, 1920/2001), it was also

Zeitschrift für Psychologie (2022), *230*(1), 33–46
https://doi.org/10.1027/2151-2604/a000486

human-like automata that were shown on stage. Today, the imitation of the human body and mind constitutes an objective that is being pursued in subdisciplines of robotics and artificial intelligence. While the number of functional human-like robots is still quite small to date, some robotics labs specialize in developing human-like autonomous machines that can serve entertainment purposes (Johnson et al., 2016), answer questions to customers (Pandey & Gelin, 2018), facilitate telepresence (Ogawa et al., 2011), assist in healthcare (Yoshikawa et al., 2011), act as sex toys (Döring et al., 2020), or are used for research into human behavior and bodily functions (Hoffmann & Pfeifer, 2018). Depending on how easily they can be distinguished from real people, human-like robots are typically referred to either as *humanoids* or *androids*. Humanoid robots are easily recognized as robots by their overall mechanical look, even though they usually possess a head, torso, arms, and sometimes legs. In contrast, android robots are intended to mimic human appearance as realistically as possible, emphasized for example, by silicone skin, clothing, wigs, or highly realistic details such as eyelashes (see Ishiguro, 2016).

The Uncanny Valley Hypothesis

Many years before robotics could even draw near the development of real android robots, Japanese roboticist Masahiro Mori introduced the hypothetical model of the uncanny valley (Mori, 1970). Initially intended more as a philosophical contribution than a blueprint for empirical research, after many years of little attention, the uncanny valley turned into a much-discussed and much-studied concept in the past two decades. The popular uncanny valley graph (Figure 1), which was originally based only on Mori's personal experience and conjecture, proposes a nonlinear relationship between the human likeness of an artificial figure, for example, of a robot, and the valence it elicits in observers. Mori suggested that within a spectrum of a generally low to medium degree of visual anthropomorphism, increasing levels of human likeness are associated with increasing acceptance and likeability. Observers should therefore sympathize more strongly with a slightly humanoid robot than, for example, with a swivel-arm robot from the industry. However, after a first positive peak of the curve along the human likeness continuum, this effect should reverse as soon as a rather high level of nearly realistic human likeness is obtained. At this point, acceptance is expected to drop, and the android should evoke a negative and irritating feeling of uncanniness (eeriness, creepiness). As an inherent property of animated entities, motion is moreover assumed to moderate the uncanny valley effect, with moving robots eliciting more pronounced reactions than static objects (or static pictures of moving objects). Therefore, a moving, highly human-like android robot should be perceived as less likable than the corresponding still artifact. Ultimately, on the right side of the uncanny valley, the likeability curve is expected to go up again when a robot's design is so perfectly realistic that it becomes indistinguishable from a real person. At the upper end of the human likeness continuum, at which the real human constitutes the endpoint, the valence of associated affect and cognition should then reach a second positive peak (Mori, 1970; Mori et al., 2012).

Different perceptual, cognitive, or evolutionary explanations have been proposed to underly the uncanny valley phenomenon, including assumptions related to categorical uncertainty, difficulties in the configural processing of human-like artifacts, threat avoidance, or the role of android robots as salient reminders of human mortality (see Diel & MacDorman, 2021; Wang et al., 2015, for an overview of suggested mechanisms).

Research on the Uncanny Valley

Compared to other scientific fields, research on the uncanny valley is characterized by a great diversity of involved disciplines, ranging from robotics, computer science, and virtual reality to animation, design, philosophy, communication science, and psychology. It, therefore, comes as no great surprise that the available studies exhibit considerable methodological heterogeneity. While, for example, a number of researchers investigated the uncanny valley by presenting study participants with physical humanoid or android robots (e.g., Bartneck, Kanda, et al., 2009; Mara & Appel, 2015) or with media representations of actually existent robots (e.g., Kim et al., 2020), other scholars focused on computer-generated stimuli such as virtual faces and avatars (e.g., Kätsyri et al., 2019; Stein & Ohler, 2017) or self-created image morphs (e.g., Lischetzke et al., 2017). Independent of the visual appearance of robots, a more recent branch of uncanny valley research also deals with aversive reactions to purely behavioral human likeness, partly relying on textual descriptions of robots as stimuli (e.g., Appel et al., 2020). Different approaches also prevail in the operationalization of central variables and associated measurements. Single-item self-reports appear to be a common means in research on user responses to human-like robots. Regarding validated multi-item scales for investigations of the uncanny valley, it is, in particular, the Godspeed questionnaire by Bartneck, Kulić, and colleagues (2009) that can be regarded as a dominant instrument for the assessment of robot anthropomorphism (representing the x-axis in Figure 1) and robot likeability (representing the y-axis in Figure 1) (see Weiss & Bartneck, 2015). Another multi-item measure, the uncanny valley indices by Ho and MacDorman (2010, 2017), has been utilized in few studies.

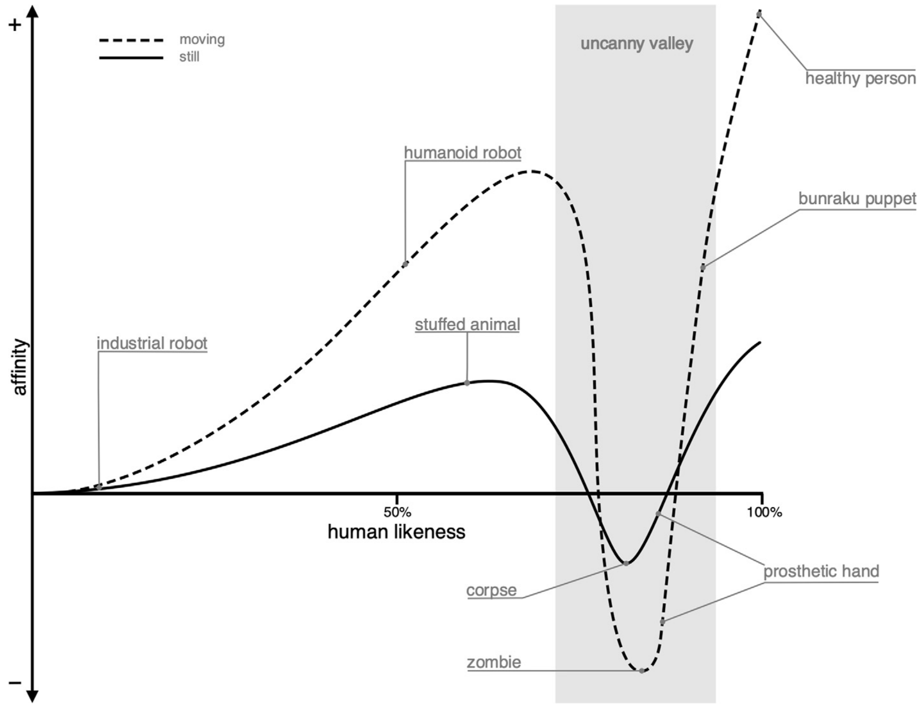


Figure 1. Uncanny Valley Hypothesis (after Mori, 1970).

Empirical support for the idea of the uncanny valley itself has been inconsistent. While results from some studies provide evidence for Mori's propositions (e.g., Mathur & Reichling, 2016) or found partial support (e.g., Bartneck et al., 2007), others failed to reveal a drop in acceptance for highly anthropomorphic machines (e.g., Bartneck, Kanda, et al., 2009) or even revealed an additional uncanny valley along the human likeness continuum (Kim et al., 2020). A literature review (Kätsyri et al., 2015) concluded that a bulk of studies supported a linear increase in affinity for more human-like robots, while evidence for nonlinear uncanny valley effects was scarce. Similarly, the assumption that robot motion should result in stronger uncanny valley effects (see Figure 1) was rarely corroborated (Piwek et al., 2014; Thompson et al., 2011). So far, a quantitative summary of uncanny valley effects is sorely missing.

The Present Study

One factor that contributes to the heterogeneity of study results on the uncanny valley might be the use of unstandardized measurements of the core constructs that exhibit unknown reliability and validity (see Wang et al., 2015). Therefore, the present meta-analysis focuses on the multi-item Godspeed questionnaire (Bartneck, Kulić, et al., 2009) that constitutes a widely used instrument for the assessment of both anthropomorphism and likeability in human-robot interaction research. It can be used to map values on both the *x*-axis and the *y*-axis of the uncanny valley graph. In the interest of ecological validity, we furthermore decided to only include studies in which participants were presented with actual robotic systems or media representations of such. To examine the central propositions of the uncanny valley effect as suggested by Mori (1970) in Figure 1, we hypothesized that (a), overall, with increasing human likeness attributed to a robot, it will be rated more positively (i.e., higher likeability).[1] Moreover, (b) the association between human likeness and likeability should show a nonlinear relationship, leading to (c) an inverted U-shaped function and thus a sharp decline of likeability ratings for highly but not perfectly anthropomorphic robots. Furthermore, (d) a second turning point at the end of the inverted U-shape at the bottom of the valley was expected to lead to more positive ratings for the most human-like robotic agents that are (nearly) indistinguishable from humans. Finally, we assumed (e) robot motion to have a moderating role because Mori (1970) speculated that motion, as an inherent property of animated objects, should amplify the uncanny valley effect.

[1] Nonlinear prediction models such as the Uncanny Valley hypothesis might exhibit an average linear trend, which is then specified in detail by nonlinear associations between the focal variables.

Method

Literature Search and Study Selection

In January 2021, we performed a literature search for studies in which at least one robot was evaluated with the help of the Godspeed questionnaire by identifying articles in Google Scholar, citing Bartneck, Kulić, and colleagues (2009). Initial search results provided 1,330 potentially relevant publications. After screening the titles, abstracts, and method sections of these articles, 95 records were subjected to detailed evaluations. To be included in the meta-analysis, a study had to meet the following criteria. First, it had to have administered the anthropomorphism and likeability scales of the Godspeed questionnaire without substantial changes to the item content. However, we considered short forms of the scale if they included at least two items, and we allowed for deviations in the number of response options (from the original 5-point ratings). Second, the respondents interacted with or viewed a real robot, a close reproduction of a real robot, or viewed a photograph or video of a robot. Virtual agents, avatars, morphed images, fictional representations (e.g., drawings, caricatures), or mere verbal descriptions of robots were not considered. No restrictions were applied on the size or the form of the robot to cover technical systems with a broad range of human likeness. Third, the study must have reported means, standard deviations, and sample sizes for both scales or provided information to derive these statistics (e.g., plots). Fourth, the study must have included healthy samples without psychological disorders. Finally, we acknowledged all studies published until December 2020. No restrictions were set on the publication type. After applying these criteria, 49 publications reporting on 93 independent samples were available (see the flow diagram in the Electronic Supplementary Material, ESM 1).

Data Extraction

From each article, we coded the mean, standard deviation, reliability (coefficient alpha), number of administered items, and number of response options for the anthropomorphism and likeability scales. For 19 studies that did not report numeric results, means and standard deviations were approximated from plots (e.g., histograms with standard errors) using the *R* package *metaDigitise* version 1.0.1 (Pick et al., 2019). In case a study reported on multiple robots, we coded each robot separately. In contrast, if different ratings were presented for the total sample and different subgroups (e.g., different experimental conditions), we only coded the results for the total sample (i.e., with the largest sample size). However, if the information was available for different values of the examined moderators (see below), then results for the different subgroups (i.e., whether the robot moved or talked) were coded separately. Additionally, we recorded the name of the evaluated robot, how it was presented (real, photo, video, virtual reality), whether it moved, and whether it communicated (e.g., talked or made sounds). Descriptive information on the sample included the sample size, the mean age of the respondents, the share of females, the country of origin of the participants, and the language of administration. Finally, we noted the publication year and the publication type (journal, proceedings, book chapter, thesis) of each study. All studies were coded by the last author and, independently, by three research assistants. Additionally, the risk of bias for each study was evaluated by two research assistants using eight items of the *Risk of Bias Utilized for Surveys Tool*, a checklist to code quality criteria such as the acceptability of exclusion rates or the sufficiency of sample sizes for primary studies used in meta-analyses (Nudelman & Otto, 2020).

For most coded variables, the interrater reliability (Krippendorff's alpha) indicated good agreement exceeding $\alpha_K \geq .85$ (*Mdn* = .90). However, the codings of the sample sizes ($\alpha_K = .63$) and whether the robot moved ($\alpha_K = .31$) or communicated ($\alpha_K = .66$) were less consistent. The interrater reliability of the risk of bias assessments was good with $\alpha_K = .91$. Discrepancies were solved by the first author. The characteristics of the samples, including the coded statistics, are summarized in ESM 1.

Analysis Plan

Because the uncanny valley hypothesis refers to a nonlinear association between anthropomorphism and likeability, the means of the likeability scale were the focal statistics that were pooled across studies. A random-effects meta-analysis was conducted using the *metafor* software version 2.4-0 (Viechtbauer, 2010) with a restricted maximum likelihood estimator. To account for sampling error, the means were weighted by the inverse of their sampling variances. Because some studies reported more than one evaluation (e.g., obtained for different robots), we estimated a three-level meta-analytic model that acknowledged dependencies between samples using a random-effects structure (see Cheung, 2019; Van den Noortgate et al., 2013). The uncanny valley effect was examined using polynomial meta-regression analyses that predicted likeability ratings from anthropomorphism scores. To model the hypothesized inflection points (see Figure 1) the regression also included higher-order polynomials of the anthropomorphism scores. In sensitivity analyses, we included several additional covariates (e.g., share of female respondents, risk of bias) and repeated the polynomial regression to determine the robustness of the observed effects. Moreover, we also repeated these analyses, excluding outliers (Viechtbauer & Cheung, 2010) and using robust meta-regression analyses (Hedges et al., 2010) to highlight the generalizability of results against different methodological choices (see Voracek et al., 2019). The homogeneity of the pooled scores was tested using the χ^2-distributed *Q*-statistic and quantified using I^2 that indicates the percentage of the total variance in observed scores due to random variance. Moderators were evaluated using the χ^2-distributed omnibus test statistic Q_m. The precision of the predicted nonlinear association between anthropomorphism and likeability was determined using a 95% confidence interval. All analyses were conducted in *R* version 4.03 (R Core Team, 2020).

Open Practices

The checklist for the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (Page et al., 2021) is provided in ESM 1. To foster transparency and reproducibility, we also provide the coding manual, extracted data, computer code, and analysis results at https://osf.io/t9rdk. The meta-analysis was not preregistered.

Results

Description of Meta-Analytic Database

The meta-analytic database included 49 studies that reported on 93 independent samples and included 131 evaluations of robots. Each sample contributed between 1 and 9 (*Mdn* = 1) evaluations of a robot using the Godspeed scales, predominantly in their original form, including five items and 5-point response scales. Both scales exhibited good reliabilities with median coefficient alphas of .86 for anthropomorphism and .89 for likeability. Results of respective reliability generalizations are summarized in ESM 1. Key characteristics of the included samples are also given in Table 1. The sample sizes ranged from 6 to 121 and included a median of 21 respondents. Most samples were from Germany (44%) and the United Kingdom (11%). The median proportion of female participants was 50%. Although the mean age of the samples spanned a broad range from 9 to 68 years, most samples were rather young (*Mdn* = 25 years) and dominated by students or university personnel (79%). Few studies included more diverse groups such as individuals with lower education (Trovato et al., 2015b), children (Meghdari et al., 2018; Shariati et al., 2018), or senior citizens (Rosenthal-von der Pütten et al., 2017). About 55% of studies were published in conference proceedings, while journal articles (33%) were less prevalent. The risk of bias assessments had a median of 3 (on a scale from 0 to 8) and, thus, indicated that many studies exhibited several designs or reporting weaknesses that might have limited the validity of the reported study results to some degree.

Evaluations of Robots

The studied robots came in different forms and sizes, representing a broad range of different models. Most available ratings pertained to the NAO robot by SoftBank Robotics (33%), the iCub robot by the Italian Institute of Technology (8%), and the Pepper robot by SoftBank Robotics (7%). In addition, various custom-built robots were examined, such as the bartender robot JAMES (Foster et al., 2012; Giuliani et al., 2013), the neuro-inspired companion robot NICO (Kerzel et al., 2020), the blessing robot BlessU2 (Löffler et al., 2019), a Sunflower housing robot (Syrdal et al., 2013), or the industrial robot ARMAR-6 (Busch et al., 2019). The distributions of the average anthropomorphism and likeability scores for these robots in Figure 2 highlight two intriguing results. First, the observed anthropomorphism scores ranged between 1.20 and 4.14, and most ratings fell in the lower middle range of possible scores (*Mdn* = 2.61). Thus, human likeness scores in the upper range were scarce. Second, the observed likeability scores ranged between 2.63 and 4.98 (*Mdn* = 3.92). This implies that most robots were rated moderately to very favorably, whereas only a few likeability ratings were in the low range.

However, there were notable differences in these evaluations between different robot models. Therefore, we pooled the anthropomorphism and likeability scores for selected robot models and summarized the meta-analytic estimates in Figure 3. Detailed meta-analytic results, based on calculations in which we used the robot model as a predictor in a meta-regression, are reported in ESM 1. For example, the bartender robot JAMES was rated significantly ($p < .05$) less human-like as compared to the average rating across all robots. In contrast, the iCub robot and Pepper received significantly higher anthropomorphism scores (see Table E2 in ESM 1). A rather similar picture emerged for the pooled likeability ratings. While the bartender robot JAMES was evaluated significantly less likable as compared to the average evaluation, the NAO robot was evaluated significantly more likable. Interestingly, the robot model explained about 20% in anthropomorphism scores, while it only accounted for about 4% in likeability ratings.

Tests of the Uncanny Valley Hypothesis

The association between the two Godspeed scales was examined using meta-regression analyses that predicted the likeability scores from the anthropomorphism ratings.

Table 1. Descriptive statistics for samples included in the meta-analytic database

Variable	*Mdn*/%	Min	Max	Valid	Missing
Sample size	21	6	121	93	0%
Number of evaluations per sample	1	1	9	93	0%
Country of origin				80	14%
Germany	44%				
Italy	5%				
Japan	5%				
The Netherlands	6%				
United Kingdom	11%				
Other	29%				
Publication year	2018	2011	2020	93	0%
Percentage females	50	0	81	89	4%
Mean age	25	9	68	80	14%
Sample type				65	30%
Students/university personnel	79%				
General public	9%				
Children	3%				
Other	9%				
Publication type				93	0%
Journal article	33%				
Proceedings	55%				
Book chapter	3%				
Thesis	6%				
Other	2%				
Response scales				52	44%
5-point	89%				
6-point	4%				
7-point	8%				
Number of items for anthropomorphism				40	57%
2 items	3%				
3 items	5%				
5 items	93%				
Number of items for likeability				39	58%
4 items	3%				
5 items	90%				
6 items[a]	8%				

Note. Valid = Number of samples that reported the respective information. Missing = Percentage of samples failing to report the respective information. [a]We suspect the studies by the research group claiming to have administered a sixth item (Foster et al., 2012; Giuliani et al., 2013) to be a reporting error because Bartneck, Kulić, and colleagues (2009) did not present a sixth item.

The nonlinear relationship suggested by the uncanny valley hypothesis (see Figure 1) could be modeled using higher-order polynomials of degree 3. To empirically determine the optimal number of higher-order terms, different meta-regression models were estimated and compared using the Bayesian information criterion (Schwarz, 1978). This suggested the inclusion of a linear term, a quadratic term, and a cubic term (see ESM 1). The respective meta-regression revealed a significant ($p < .05$) effect for anthropomorphism ($Q_m = 89.46$, $df = 3$, $p < .001$) that explained about 5% in the variance of likeability ratings between samples (see Table 2). These results were rather robust ($Q_m = 98.43$, $df = 3$, $p < .001$) and replicated after controlling for sample characteristics (i.e., mean age, share of women, publication year, country), robot characteristics (i.e., movement, communication), and methodological characteristics (i.e., presentation mode, risk of study bias). To study the effect in more detail, the likeability ratings predicted from this meta-regression model (including a 95% confidence interval) were plotted in Figure 4. Consistent with the assumption (a), these results confirmed more positive evaluations for more human-like robots overall. In accordance

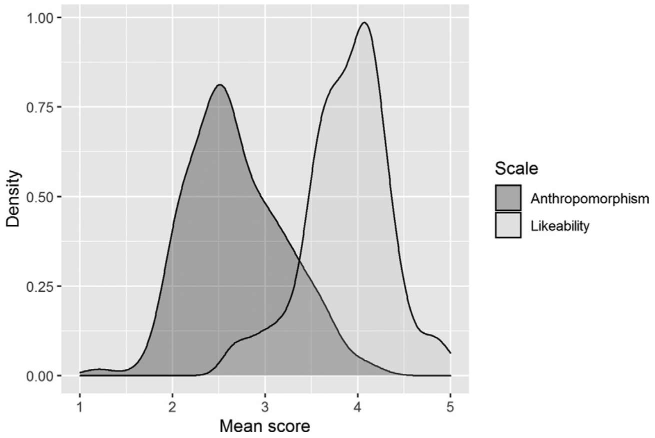


Figure 2. Average score distributions of the Godspeed anthropomorphism and likeability scales.

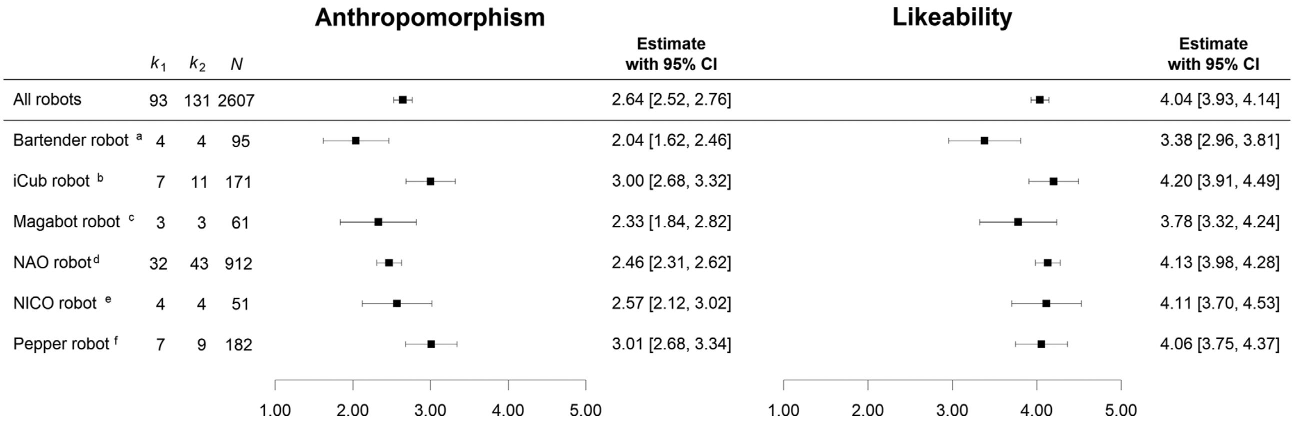
with the assumption (b), we also found evidence for a nonlinear effect. Although the effect approximated a sigmoid shape with a plateau in the region of the greatest anthropomorphism scores contained in the sample, we were unable to corroborate the hypothesized decline of likeability for highly realistic android robots as stated in assumption (c). Consequently, we were also unable to identify the rise of likeability at even higher scores of human likeness as expected in assumption (d). Again, these results were rather stable and replicated after controlling for various covariates (see Figure 4 and Table 2). The pooled association between anthropomorphism and likeability was also rather invariant toward various methodological choices and replicated after excluding outliers, children, or older samples and adopting robust meta-analytic models (see ESM 1).

Movement and Other Moderating Effects

In line with Mori's hypothesis (Mori et al., 2012), static robots were evaluated significantly less human-like as compared to moving robots (B = −0.35, 95% CI [−0.56, −0.14]). In contrast, the movement had no impact on likeability ratings (see Table E2 in ESM 1). Unexpectedly, communication had an opposite effect: For anthropomorphism, it was immaterial whether a robot was mute or communicated with the participants (B = 0.23, 95% CI [−0.07, 0.54]), whereas communicative robots were evaluated significantly ($p < .05$) more likable as compared to mute robots (B = −0.27, 95% CI [−0.47, −0.06]). To examine whether these effects also extended to the nonlinear association between anthropomorphism and likeability, we extended the previous meta-regression analyses and included respective interactions for the linear, quadratic, and cubic terms. However, inconsistent with the assumption (e), these interactions were not significant (see Table 2), thus, indicating that movement and communication did not moderate the predicted effects given in Figure 4. However, our database included only 19 results with static robots, while most of the robots exhibited some form of movement.

Discussion

Masahiro Mori's (1970) hypothetical graph on the uncanny valley has developed into a dominant influence on recent research into user perceptions of human-like robots. Complementing and extending insights gained from narrative reviews on the uncanny valley hypothesis (Kätsyri et al., 2015; Wang et al., 2015; Zhang et al., 2020), we presented



	k_1	k_2	N	Anthropomorphism Estimate with 95% CI	Likeability Estimate with 95% CI
All robots	93	131	2607	2.64 [2.52, 2.76]	4.04 [3.93, 4.14]
Bartender robot [a]	4	4	95	2.04 [1.62, 2.46]	3.38 [2.96, 3.81]
iCub robot [b]	7	11	171	3.00 [2.68, 3.32]	4.20 [3.91, 4.49]
Magabot robot [c]	3	3	61	2.33 [1.84, 2.82]	3.78 [3.32, 4.24]
NAO robot[d]	32	43	912	2.46 [2.31, 2.62]	4.13 [3.98, 4.28]
NICO robot [e]	4	4	51	2.57 [2.12, 3.02]	4.11 [3.70, 4.53]
Pepper robot [f]	7	9	182	3.01 [2.68, 3.34]	4.06 [3.75, 4.37]

Figure 3. Forest plots for average anthropomorphism and likeability scores by robot model. k_1 = Number of samples, k_2 = number of ratings, N = total sample size. [a]Foster et al. (2012), Giuliani et al. (2013), Keizer et al. (2014); [b]Ghiglino et al. (2020), Lehmann et al. (2016), Mazzola et al. (2020), Willemse & Wykowska (2019); [c]Hoegen (2013), Lohse et al. (2013); [d]Barlas (2019), Cuijpers et al. (2011), Ham et al. (2015), van der Hout (2017), Lehmann et al. (2020), Mirnig, Stollnberger, Giuliani, et al. (2017), Mirnig, Stollnberger, Miksch et al. (2017), Rosenberg-Kima et al. (2020), Rosenthal-von der Pütten et al. (2017, 2018), Schneider (2019), Zanatto, Patacchiola, Goslin, & Cangelosi (2019), Zanatto, Patacchiola, Goslin, Thill, & Cangelosi (2020); [e]Churamani et al. (2017), Kerzel et al. (2020); [f]Iwashita & Katagami (2020), Rhim et al. (2019), Straßmann et al. (2020).

Table 2. Polynomial meta-regression tests for the Uncanny Valley Hypothesis

	Model 1	Model 2	Model 3	Model 4
Intercept	8.94*** (2.10)	8.72*** (2.18)	8.66* (3.53)	6.95*** (2.07)
Anthropomorphism				
1. Linear term	−6.19** (2.32)	−5.57** (2.41)	−5.93$^+$ (3.75)	−3.53 (2.31)
2. Quadratic term	2.28** (0.85)	2.03* (0.88)	2.20 (1.30)	1.23 (0.84)
3. Cubic term	−0.25* (0.10)	−0.22* (0.11)	−0.24 (0.15)	−0.12 (0.10)
Control variables				
4. Average age[a]		0.00 (0.00)		
5. Share of women[b]		0.58* (0.26)		
Country[c]				
6. United Kingdom		−0.17 (0.16)		
7. Other country		−0.23* (0.09)		
8. Publication year[d]		0.00 (0.02)		
9. Movement[e]		−0.14$^+$ (0.07)	7.23 (5.37)	
10. Communication[e]		−0.17* (0.08)		5.89 (8.80)
11. Interaction with real robot[e]		0.02 (0.09)		
12. Statistics reported[e]		−0.16 (0.11)		
13. Risk of study bias[f]		−0.06 (0.04)		
Moderating effects				
14. 1. × 8.			−10.60 (6.75)	
15. 2. × 8.			4.87$^+$ (2.84)	
16. 3. × 8.			−0.72$^+$ (0.40)	
17. 1. × 9.				−7.98 (9.67)
18. 2. × 9.				3.16 (3.50)
19. 3. × 9.				−0.39 (0.42)
Random effects (τ_s/τ_e)	0.39/0.08	0.35/0.04	0.40/0.04	0.37/0.04
I^2	96%	95%	96%	95%
R^2	5%	23%***	3%***	17%***

Note. Dependent variable are likeability ratings. Presented are meta-regression coefficients with standard errors in parentheses. τ_s/τ_e = Standard deviations of random effects for samples and evaluations; R^2 = Explained random variance. [a]Centered at 25 years, [b]Centered at .50, [c]Dummy coded with Germany as reference category, [d]Centered at year 2020, [e]0 = yes, 1 = no, [f]Centered at 4. ***$p < .001$; ** $p < .01$; * $p < .05$; $^+p < .10$.

the first quantitative, meta-analytical review of the main assumptions underlying the uncanny valley effect. We focused on the characteristic relationship between user assessments of human likeness (the *x*-axis) and likeability (the *y*-axis) that was proposed by Mori (1970, Figure 1), based on the Godspeed scales (Bartneck, Kulić, et al., 2009), a standard measure in the field (see Weiss & Bartneck, 2015). To this end, state-of-the-art meta-analytic methods that acknowledged dependencies between samples using a random-effects structure (see Cheung, 2019; Van den Noortgate et al., 2013) were used to study the nonlinear hypothesis with polynomial meta-regression analyses. From our quantitative assessment of the 93 independent samples that comprised our meta-analytic database, a main insight is the limited range of anthropomorphism and likeability scores in the examined primary studies (Figure 2). In the large majority of studies, the focal robot was experienced as being not quite human-like with means ranging below the scale's midpoint. Means above 3.5 on a 5-point scale were almost entirely missing. Likewise, and even more pronounced, the focal robots were experienced as highly likable on average in the primary studies. The large majority of studies reported mean likeability scores above the midpoint of the scale. The limited range of the primary study scores is highly relevant for our main meta-analytic aim, gathering quantitative evidence for or against the uncanny valley hypothesis. According to Mori (1970) and contemporary interpretations of his ideas (e.g., Diel & MacDorman, 2021; Wang et al., 2015), the characteristic drop in likeability is experienced at the higher end of the human likeness continuum. Based on the studies underlying our meta-analysis, this higher end of the human likeness continuum is unchartered territory.

We deduced several functional properties from the curvilinear explication of the uncanny valley hypothesis. Despite the identified limitations in scale range, likeability scores supported the first assumption (a) derived from Mori's uncanny valley hypothesis in that increasing human likeness was found to be associated with increased positive user responses within the spectrum of low to medium

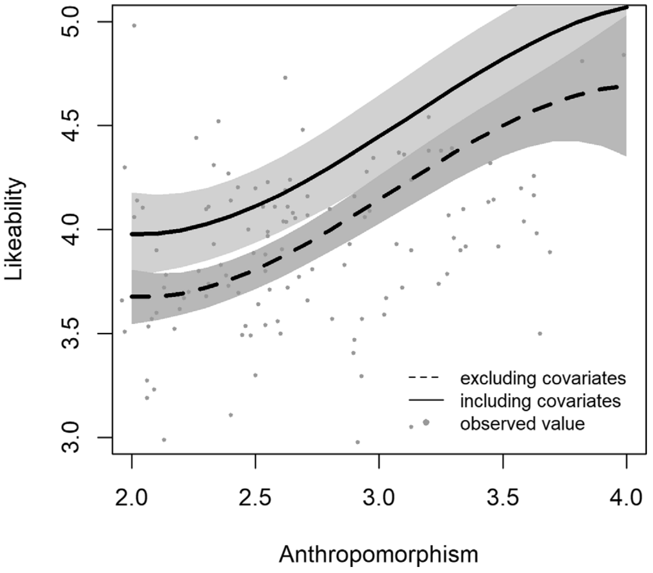


Figure 4. Predicted likeability ratings with 95% confidence intervals.

anthropomorphism. Important against the backdrop of the uncanny valley literature and in line with the assumption (b), our results also suggest a nonlinear effect, leading to a flattening of the likeability curve at about 75% of the anthropomorphism scale range (*x*-axis). However, because hardly any robots had been rated as highly human-like in the available primary studies, neither assumption (c) that such robots would lead to a pronounced drop in acceptance nor assumption (d) that near-to-perfect copies of humans at the end of the continuum would lead to an ultimate grow in acceptance could be evaluated. Mori's core proposition about the adverse effects of android robots can therefore neither be rejected nor confirmed at this stage.

We further examined several potential moderating variables. A comparison between static and moving robots was of particular relevance to the original uncanny valley hypothesis. Static robots were evaluated as less human-like than moving robots, but the movement had no impact on likeability ratings. Importantly, the linear, quadratic, and cubic associations between human likeness and likeability did not differ significantly between statically presented robots and such that were moving. Assumption (e), based on Mori's description of a potentially intensifying role of robotic motion, therefore must be rejected in view of the current data.

Limitations and Implications

As outlined above, our quantitative test of the uncanny valley hypothesis is preliminary, as primary studies that captured high degrees of human likeness were missing. The low human likeness scores observed could be a function of several factors. First, the robotic platforms examined in the primary studies do not stipulate high human likeness (e.g., NAO and similar designs, see ESM 1). Second, participants naïve to robotics may use expectations derived from science-fiction as a point of comparison (Appel et al., 2016; Mara & Appel, 2015). Due to the fact that the state of today's technological advancement rarely matches sci-fi worlds, robots examined in human-robot interaction research have to fall short compared to fictional robots. The original movie Blade Runner (Scott, 1982), for example, showed a world in the year 2019 in which humans and human-like robotic replicants mingled. Participants with high technological knowledge or even a study emphasis in computer science, in turn, may be aware of technological glitches or wizard-of-oz simulated interactions.

We deliberately restricted our study pool to primary studies that reported data on the Godspeed Scales (Bartneck, Kulić, et al., 2009) to achieve high comparability and to prevent an influx of data with low reliability or validity, which has been described as a substantial problem in the field (Wang et al., 2015). The Godspeed Scales are in particular widespread use, constituting one of the standard measures in the field. Despite their popularity, it should not be dismissed that the Godspeed Scales themselves have also faced some criticism in the past (Carpinella et al., 2017; Ho & MacDorman, 2010). For example, an exploratory factor analysis conducted by Carpinella and colleagues (2017) indicated low eigenvalues and low reliabilities for some of the five Godspeed components. However, this was mainly true for the animacy and safety scales, but not for anthropomorphism and likeability. Consistent with this and in support of our decision to use the Godspeed Sales, our database showed high reliabilities for both the anthropomorphism scale and the likeability scale. That said, future meta-analyses could apply more liberal inclusion criteria. Promising alternative measures include the scales by Ho and MacDorman (2010, 2017), which were developed specifically for research on the uncanny valley hypothesis, or the Robotic Social Attributes Scale (Carpinella et al., 2017), which assesses warmth and competence as components of social perception and discomfort as a potential measure for the uncanny experience.

We further restricted our meta-analysis to genuine implementations of robotic systems. Studies that relied on verbal descriptions, drawings of robots, or morphed pictures (e.g., Lischetzke et al., 2017; MacDorman & Ishiguro, 2006) were excluded. Whereas these stimuli could arguably increase human likeness (e.g., morphs between robots and humans, Lischetzke et al., 2017), such stimuli have been criticized for lacking external validity, for example, morphs may show ghosting artifacts by the computer graphics software (Kätsyri et al., 2019).

Several measures were taken to secure a standard of sufficient data quality in the primary study pool and, therefore, our meta-analysis as a summary of the quantitative results. This includes the restriction to the experience of genuine technical implementations and the Godspeed Scales as operationalizations of the key variables. We further implemented a risk of study bias assessment (Nudelman & Otto, 2020) and controlled our meta-analytic results for the respective scores. These scores revealed remarkable weaknesses regarding design or reporting. We need to acknowledge these shortcomings of the primary study data, and we emphasize two implications for human-robot interaction research:

First, our review of studies revealed that a substantial number of publications failed to report basic information on the sample and descriptive results. Authors of quantitative results sections should make sure to report (subgroup-) sample sizes and results on variance (e.g., the standard deviation) along with mean values (or any other measure of central tendency). Zero-order correlations and raw descriptive statistics are particularly helpful for (meta-analytic) summaries and comparisons within a field of research. Second, sample sizes were remarkably small, *Mdn*(*N*) = 21, from a general psychological perspective. They arguably reflect the studied topic in human-robot interaction research for which the technological requirements complicate or impede larger sample sizes. Nevertheless, minimal sample size recommendations should be adhered to (Simmons et al., 2011). Note that 20 participants per cell, for example, is insufficient to "detect in a representative sample that men are heavier than women" (Simmons et al., 2018, p. 256). The problem of low sample size is even more serious for complex between-subjects designs (e.g., a focal moderation effect based on a 2 × 2 experimental design). The authors of several other recent meta-analyses and reviews in the field of human-robot interaction also identified similar problems in data reporting and statistical power of primary studies and made similar recommendations to the interdisciplinary research community (Leichtmann & Nitsch, 2020; Oliveira et al., 2021; Stower et al., 2021). We are therefore optimistic that future empirical work will benefit from the lessons learned and, through larger sample sizes and greater transparency, will make important contributions to our understanding of user responses to robots.

Conclusion

The uncanny valley hypothesis is a major perspective to explaining and predicting negative responses to humanoid and android robots. The available research covers user experiences of low to moderate human likeness, whereas robots with high human likeness are largely unchartered territory. Within these low to moderate levels of human likeness, our findings follow the assumptions derived from the uncanny valley hypothesis insofar as likeability ratings initially increase but then level off to a plateau as a result of a nonlinear function. Movement appears to be no factor that intensifies the characteristic nonlinear association between human likeness and likeability.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/2151-2604/a000486

ESM 1. Coded variables and data, diagrams, Meta-Analyses of Godspeed Scale Scores by Robot, and further analyses. Table E2: Meta-Analyses of Godspeed Scale Scores by Robot Model.

References

References marked with * were included in the meta-analysis.

Appel, M., Izydorczyk, D., Weber, S., Mara, M., & Lischetzke, T. (2020). The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior, 102*, 274–286. https://doi.org/10.1016/j.chb.2019.07.031

Appel, M., Krause, S., Gleich, U., & Mara, M. (2016). Meaning through fiction: Science fiction and innovative technologies. *Psychology of Aesthetics, Creativity, and the Arts, 10*, 472–480. https://doi.org/10.1037/aca0000052

*Avelino, J., Correia, F., Catarino, J., Ribeiro, P., Moreno, P., Bernardino, A., & Paiva, A. (2018). The power of a hand-shake in human-robot interactions. In *Proceedings of the 2018 International Conference on Intelligent Robots and Systems* (pp. 1864–1869). IEEE. https://doi.org/10.1109/IROS.2018.8593980

*Barlas, Z. (2019). When robots tell you what to do: Sense of agency in human-and robot-guided actions. *Consciousness and Cognition, 75*, Article 102819. https://doi.org/10.1016/j.concog.2019.102819

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2007). Is the uncanny valley an uncanny cliff? In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 368–373). IEEE. https://doi.org/10.1109/ROMAN.2007.4415111

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009). My robotic doppelgänger – A critical look at the uncanny valley. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 269–276). IEEE. https://doi.org/10.1109/ROMAN.2009.5326351

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*, 71–81. https://doi.org/10.1007/s12369-008-0001-3

*Busch, B., Cotugno, G., Khoramshahi, M., Skaltsas, G., Turchi, D., Urbano, L., Wächter, M., Zhou, Y., Asfour, T., Deacon, G., Russell, D., & Billard, A. (2019). Evaluation of an industrial robotic assistant in an ecological environment. In *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication* (pp. 1–8). IEEE. https://doi.org/10.1109/RO-MAN46459.2019.8956399

Čapek, K. (1920/2001). *R.U.R. (Rossum's Universal Robots)*. Dover.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The Robotic Social Attributes Scale (RoSAS): Development and validation. In *Proceedings of the International Conference on Human-Robot Interaction* (pp. 254–262). IEEE. https://doi.org/10.1145/2909824.3020208

Cave, S., Dihal, K., & Dillon, S. (2020). Introduction: Imagining AI. In S. Cave, K. Dihal, & S. Dillon (Eds.), *AI Narratives: A history of imaginative thinking about intelligent machines* (pp. 1–21). . Oxford University Press. https://doi.org/10.1093/oso/9780198846666.001.0001

Cheung, M. W. L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review, 29*(4), 387–396. https://doi.org/10.1007/s11065-019-09415-6

*Churamani, N., Anton, P., Brügger, M., Fließwasser, E., Hummel, T., Mayer, J., Mustafa, W., Ng, H. G., Nguyen, T. L. C., Nguyen, Q., Soll, M., Springenberg, S., Griffiths, S., Heinrich, S., Navarro-Guerrero, N., Strahl, E., Twiefel, J., Weber, C., & Wermter, S. (2017). The impact of personalisation on human-robot interaction in learning scenarios. In B. Wrede, Y. Nagai, T. Komatsu, M. Hanheide, & L. Natale (Eds.), *Proceedings of the 5th International Conference on Human Agent Interaction* (pp. 171–180). Association for Computing Machinery. https://doi.org/10.1145/3125739.3125756

*Cuijpers, R. H., Bruna, M. T., Ham, J. R., & Torta, E. (2011). Attitude towards robots depends on interaction but not on anticipatory behaviour. In B. Mutlu, C. Bartneck, J. Ham, V. Evers, & T. Kanda (Eds.), *Proceedings of the 2011 International Conference on Social Robotics* (pp. 163–172). . Springer. https://doi.org/10.1007/978-3-642-25504-5_17

Diel, A., & MacDorman, K. F. (2021). Creepy cats and strange high houses: Support for configural processing in testing predictions of nine uncanny valley theories. *Journal of Vision, 21*(4), 1–20. https://doi.org/10.1167/jov.21.4.1

Döring, N., Mohseni, M. R., & Walter, R. (2020). Design, use, and effects of sex dolls and sex robots: Scoping review. *Journal of Medical Internet Research, 22*(7), Article e18551. https://doi.org/10.2196/18551

*Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., & Petrick, R. P. (2012). Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 3–10). ACM. https://doi.org/10.1145/2388676.2388680

*Fu, C., Yoshikawa, Y., Iio, T., & Ishiguro, H. (2021). Sharing experiences to help a robot present its mind and sociability. *International Journal of Social Robotics, 13*, 341–352. https://doi.org/10.1007/s12369-020-00643-y

*Ghiglino, D., De Tommaso, D., Willemse, C., Marchesi, S., & Wykowska, A. (2020). Can I get your (robot) attention? Human sensitivity to subtle hints of human-likeness in a humanoid robot's behavior. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Virtual Meeting of the Cognitive Science Society* (pp. 952–958). Cognitive Science Society. https://doi.org/10.31234/osf.io/kfy4g

*Giuliani, M., Petrick, R. P., Foster, M. E., Gaschler, A., Isard, A., Pateraki, M., & Sigalas, M. (2013). Comparing task-based and socially intelligent behaviour in a robot bartender. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (pp. 263–270). ACM. https://doi.org/10.1145/2522848.2522869

Gnambs, T. (2019). Attitudes towards emergent autonomous robots in Austria and Germany. *Elektrotechnik und Informationstechnik, 136*, 296–300. https://doi.org/10.1007/s00502-019-00742-3

Gnambs, T., & Appel, M. (2019). Are robots becoming unpopular? Changes in attitudes towards autonomous robotic systems in Europe. *Computers in Human Behavior, 93*, 53–61. https://doi.org/10.1016/j.chb.2018.11.045

*Ham, J., Cuijpers, R. H., & Cabibihan, J. J. (2015). Combining robotic persuasive strategies: The persuasive power of a storytelling robot that uses gazing and gestures. *International Journal of Social Robotics, 7*(4), 479–487. https://doi.org/10.1007/s12369-015-0280-4

*Haring, K. S., Silvera-Tawil, D., Takahashi, T., Velonaki, M., & Watanabe, K. (2015). Perception of a humanoid robot: a cross-cultural comparison. In *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 821–826). IEEE. https://doi.org/10.1109/ROMAN.2015.7333613

*Haring, K. S., Silvera-Tawil, D., Takahashi, T., Watanabe, K., & Velonaki, M. (2016). How people perceive different robot types: A direct comparison of an android, humanoid, and non-biomimetic robot. In *Proceedings of the 8th International Conference on Knowledge and Smart Technology* (pp. 265–270). IEEE. https://doi.org/10.1109/KST.2016.7440504

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39–65. https://doi.org/10.1002/jrsm.5

Ho, C. C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior, 26*(6), 1508–1518. https://doi.org/10.1016/j.chb.2010.05.015

Ho, C. C., & MacDorman, K. F. (2017). Measuring the uncanny valley effect. *International Journal of Social Robotics, 9*(1), 129–139. https://doi.org/10.1007/s12369-016-0380-9

*Hoegen, R. (2013). *The influence of a robot's voice on proxemics in human-robot interaction* [Unpublished manuscript]. University of Twente.

Hoffmann, M., & Pfeifer, R. (2018). Robots as powerful allies for the study of embodied cognition from the bottom up. In A. Newen, L. de Bruin, & S. Gallagher (Eds.), *The Oxford handbook of 4E cognition* (pp. 841–862). . Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.013.45

Ishiguro, H. (2016). Android science. In M. Kasaki, H. Ishiguro, M. Asada, M. Osaka, & T. Fujikado (Eds.), *Cognitive neuroscience robotics A: Synthetic approaches to human understanding* (pp. 193–234). Springer. https://doi.org/10.1007/978-4-431-54595-8

*Iwashita, M., & Katagami, D. (2020). Psychological effects of compliment expressions by communication robots on humans. In *Proceedings of the 2020 International Joint Conference on Neural Networks* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN48605.2020.9206898

*Johanson, D. L., Ahn, H. S., Lim, J., Lee, C., Sebaratnam, G., MacDonald, B. A., & Broadbent, E. (2020). Use of humor by a healthcare robot positively affects user perceptions and behavior. *Technology, Mind, and Behavior, 1*(2), 1–33. https://doi.org/10.1037/tmb0000021

Johnson, D. O., Cuijpers, R. H., Pollmann, K., & van de Ven, A. A. (2016). Exploring the entertainment value of playing games with a humanoid robot. *International Journal of Social Robotics, 8*(2), 247–269.

Kätsyri, J., de Gelder, B., & Takala, T. (2019). Virtual faces evoke only a weak uncanny valley effect: An empirical investigation with controlled virtual face images. *Perception, 48*(10), 968–991. https://doi.org/10.1177/0301006619869134

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology, 6*, Article 390. https://doi.org/10.3389/fpsyg.2015.00390

*Keizer, S., Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., & Lemon, O. (2014). Handling uncertain input in multi-user human-robot interaction. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication* (pp. 312–317). IEEE. https://doi.org/10.1109/ROMAN.2014.6926271

*Kerzel, M., Pekarek-Rosin, T., Strahl, E., Heinrich, S., & Wermter, S. (2020). Teaching NICO how to grasp: An empirical study on crossmodal social interaction as a key factor for robots learning from humans. *Frontiers in Neurorobotics, 14*, Article 28. https://doi.org/10.3389/fnbot.2020.00028

Kim, B., Bruce, M., Brown, L., de Visser, E., & Phillips, E. (2020). A comprehensive approach to validating the uncanny valley using the Anthropomorphic RoBOT (ABOT) database. In *Proceedings of 2020 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 1–6). IEEE. https://doi.org/10.1109/SIEDS49339.2020.9106675

*Kühnlenz, B. (2013). *Alignment strategies for information retrieval in prosocial human-robot interaction* [Unpublished doctoral dissertation]. Technical University Munich.

*Kühnlenz, B., Sosnowski, S., Buß, M., Wollherr, D., Kühnlenz, K., & Buss, M. (2013). Increasing helpfulness towards a robot by emotional adaption to the user. *International Journal of Social Robotics, 5*(4), 457–476. https://doi.org/10.1007/s12369-013-0182-2

*Lehmann, H., Rojik, A., & Hoffmann, M. (2020, September). *Should a small robot have a small personal space? Investigating personal spatial zones and proxemic behavior in human-robot interaction*. Paper presented at the CognitIve RobotiCs for intEraction (CIRCE) Workshop at the 2020 IEEE International Conference on Robot and Human Interactive Communication. https://arxiv.org/abs/2009.01818

*Lehmann, H., Roncone, A., Pattacini, U., & Metta, G. (2016). Physiologically inspired blinking behavior for a humanoid robot. In A. Agah, J. J. Cabibihan, A. Howard, M. Salichs, & H. He (Eds.), *Proceedings of the 2016 International Conference on Social Robotics* (pp. 83–93). . Springer. https://doi.org/10.1007/978-3-319-47437-3_9

Leichtmann, B., & Nitsch, V. (2020). How much distance do humans keep toward robots? Literature review, meta-analysis, and theoretical considerations on personal space in human-robot interaction. *Journal of Environmental Psychology, 68*, 101386. https://doi.org/10.1016/j.jenvp.2019.101386

Lischetzke, T., Izydorczyk, D., Hüller, C., & Appel, M. (2017). The topography of the uncanny valley and individuals' need for structure: A nonlinear mixed effects analysis. *Journal of Research in Personality, 68*, 96–113. https://doi.org/10.1016/j.jrp.2017.02.001

*Löffler, D., Hurtienne, J., & Nord, I. (2019). Blessing robot blessU2: A discursive design study to understand the implications of social robots in religious contexts. *International Journal of Social Robotics*. Advance online publication. doi: https://doi.org/10.1007/s12369-019-00558-3

*Lohse, M., van Berkel, N., Van Dijk, E. M., Joosse, M. P., Karreman, D. E., & Evers, V. (2013). The influence of approach speed and functional noise on users' perception of a robot. In *Proceedings of the 2013 International Conference on Intelligent Robots and Systems* (pp. 1670–1675). IEEE. https://doi.org/10.1109/IROS.2013.6696573

*Lugrin, B., Dippold, J., & Bergmann, K. (2018). Social robots as a means of integration? An explorative acceptance study considering gender and non-verbal behaviour. In *Proceedings of the 2018 International Conference on Intelligent Robots and Systems* (pp. 2026–2032). IEEE. https://doi.org/10.1109/IROS.2018.8593818

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies, 7*(3), 297–337. https://doi.org/10.1075/is.7.3.03mac

Mara, M., & Appel, M. (2015). Science fiction reduces the eeriness of android robots: A field experiment. *Computers in Human Behavior, 48*, 156–162. https://doi.org/10.1016/j.chb.2015.01.007

Mara, M., Schreibelmayr, S., & Berger, F. (2020). Hearing a nose? User expectations of robot appearance induced by different robot voices. In *Proceedings of the Companion of the 2020 International Conference on Human-Robot Interaction* (pp. 355–356). ACM/IEEE. https://doi.org/10.1145/3371382.3378285

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition, 146*, 22–32. https://doi.org/10.1016/j.cognition.2015.09.008

*Mazzola, C., Aroyo, A. M., Rea, F., & Sciutti, A. (2020). Interacting with a social robot affects visual perception of space. In *Proceedings of the 2020 ACM/IEEE International Conference on Human Robot Interaction* (pp. 549–557). ACM. https://doi.org/10.1145/3319502.3374819

*Meghdari, A., Shariati, A., Alemi, M., Vossoughi, G. R., Eydi, A., Ahmadi, E., Mozafari, B., Nobaveh, A. A., & Tahami, R. (2018). Arash: A social robot buddy to support children with cancer in a hospital environment. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 232*(6), 605–618. https://doi.org/10.1177/0954411918777520

*Mirnig, N., Stollnberger, G., Giuliani, M., & Tscheligi, M. (2017). Elements of humor: How humans perceive verbal and non-verbal aspects of humorous robot behavior. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 211–212). Association for Computing Machinery. https://doi.org/10.1145/3029798.3038337

*Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI, 4*, Article 21. https://doi.org/10.3389/frobt.2017.00021

*Moon, A., Parker, C. A., Croft, E. A., & Van der Loos, H. M. (2013). Design and impact of hesitation gestures during human-robot resource conflicts. *Journal of Human-Robot Interaction, 2*(3), 18–40. https://doi.org/10.5898/JHRI.2.3.Moon

Mori, M. (1970). Bukimi no tani [The uncanny valley]. *Energy, 7*, 33–35.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley. *IEEE Robotics & Automation Magazine, 19*, 98–100. https://doi.org/10.1109/MRA.2012.2192811

*Müller, S. L., Schröder, S., Jeschke, S., & Richert, A. (2017). Design of a robotic workmate. In V. Duffy (Ed.), *International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management* (pp. 447–456). Springer. https://doi.org/10.1007/978-3-319-58463-8_37

Nudelman, G., & Otto, K. (2020). The development of a new generic risk-of-bias measure for systematic reviews of surveys. *Methodology, 16*, 278–298. https://doi.org/10.5964/meth.4329

Ogawa, K., Nishio, S., Koda, K., Taura, K., Minato, T., Ishii, C. T., & Ishiguro, H. (2011). Telenoid: Tele-presence android for communication. In *Proceedings of the SIGGRAPH 2011 Emerging Technologies* (pp. 1). ACM. https://doi.org/10.1145/2048259.2048274

Oliveira, R., Arriaga, P., Santos, F. P., Mascarenhas, S., & Paiva, A. (2021). Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior, 114*, Article 106547. https://doi.org/10.1016/j.chb.2020.106547

*Paetzel, M., Perugia, G., & Castellano, G. (2020). The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 73–82). ACM. https://doi.org/10.1145/3319502.3374786

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

Pandey, A. K., & Gelin, R. (2018). A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine, 25*(3), 40–48. https://doi.org/10.1109/MRA.2018.2833157

*Petrak, B., Weitz, K., Aslan, I., & André, E. (2019). Let me show you your new home: Studying the effect of proxemic-awareness of robots on users' first impressions. In *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication* (pp. 1–7). IEEE. https://doi.org/10.1109/RO-MAN46459.2019.8956463

Pick, J. L., Nakagawa, S., & Noble, D. W. (2019). Reproducible, flexible and high-throughput data extraction from primary literature: The *metaDigitise* R package. *Methods in Ecology and Evolution, 10*, 426–431. https://doi.org/10.1111/2041-210X.13118

Piwek, L., McKay, L. S., & Pollick, F. E. (2014). Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition, 130*(3), 271–277. https://doi.org/10.1016/j.cognition.2013.11.001

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

*Rhim, J., Cheung, A., Pham, D., Bae, S., Zhang, Z., Townsend, T., & Lim, A. (2019). Investigating positive psychology principles in affective robotics. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction* (pp. 1–7). IEEE. https://doi.org/10.1109/ACII.2019.8925475

*Rosenberg-Kima, R. B., Koren, Y., & Gordon, G. (2020). Robot-supported collaborative learning (RSCL): Social robots as teaching assistants for higher education small group facilitation. *Frontiers in Robotics and AI, 6*, Article 148. https://doi.org/10.3389/frobt.2019.00148

*Rosenthal-von der Pütten, A. M., Bock, N., & Brockmann, K. (2017). Not your cup of tea? How interacting with a robot can increase perceived self-efficacy in HRI and evaluation. In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 483–492). IEEE. https://doi.org/10.1145/2909824.3020251

*Rosenthal-von der Pütten, A. M., Krämer, N. C., & Herrmann, J. (2018). The effects of humanlike and robot-specific affective nonverbal behavior on perception, emotion, and behavior. *International Journal of Social Robotics, 10*(5), 569–582. https://doi.org/10.1007/s12369-018-0466-7

*Ruijten, P. A., & Cuijpers, R. H. (2018). If drones could see: Investigating evaluations of a drone with eyes. In S. S. Ge, J.-J. Cabibihan, M. A. Salichs, E. Broadbent, H. He, A. R. Wagner, & Á. Castro-González (Eds.), *Proceedings of the 10th International Conference on Social Robotics* (pp. 65–74). Springer. https://doi.org/10.1007/978-3-030-05204-1_7

*Schneider, S. (2019). *Socially assistive robots for exercise scenarios* [Unpublished dissertation]. Bielefeld University. https://doi.org/10.4119/unibi/2934006

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Scott, R. (. Director). (1982). *Blade runner* [movie]. Warner Bros.

*Shariati, A., Shahab, M., Meghdari, A., Nobaveh, A. A., Rafatnejad, R., & Mozafari, B. (2018). Virtual reality social robot platform: A case study on Arash social robot. In S. S. Ge, J.-J. Cabibihan, M. A. Salichs, E. Broadbent, H. He, A. R. Wagner, & Á. Castro-González (Eds.), *Proceedings of the 10th International Conference on Social Robotics* (pp. 551–560). Springer. https://doi.org/10.1007/978-3-030-05204-1_54

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science, 13*(2), 255–259. https://doi.org/10.1177/1745691617698146

Stein, J. P., & Ohler, P. (2017). Venturing into the uncanny valley of mind – The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition, 160*, 43–50. https://doi.org/10.1016/j.cognition.2016.12.010

Stower, R., Calvo-Barajas, N., Castellano, G., & Kappas, A. (2021). A meta-analysis on children's trust in social robots. *International Journal of Social Robotics*. Advance online publication. https://doi.org/10.1007/s12369-020-00736-8

*Straßmann, C., Grewe, A., Kowalczyk, C., Arntz, A., & Eimler, S. C. (2020). Moral robots? How uncertainty and presence affect humans' moral decision making. In C. Stephanidis & M. Antona (Eds.), *Proceedings of the 2020 International Conference on Human-Computer Interaction* (pp. 488–495). Springer. https://doi.org/10.1007/978-3-030-50726-8_64

*Syrdal, D. S., Dautenhahn, K., Koay, K. L., Walters, M. L., & Ho, W. C. (2013). Sharing spaces, sharing lives – the impact of robot mobility on user perception of a home companion robot. In G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *Proceedings of the 2013 International Conference on Social Robotics* (pp. 321–330) . Springer. https://doi.org/10.1007/978-3-319-02675-6_32

Thompson, J. C., Trafton, J. G., & McKnight, P. (2011). The perception of humanness from the movements of synthetic agents. *Perception, 40*(6), 695–704. https://doi.org/10.1068/p6900

*Trovato, G., Ramos, J. G., Azevedo, H., Moroni, A., Magossi, S., Ishii, H., Simmons, R., & Takanishi, A. (2015a). Designing a receptionist robot: Effect of voice and appearance on anthropomorphism. In *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 235–240). IEEE. https://doi.org/10.1109/ROMAN.2015.7333573

*Trovato, G., Ramos, J. G., Azevedo, H., Moroni, A., Magossi, S., Ishii, H., Simmons, R., & Takanishi, A. (2015b). "Olá, my name is Ana": A study on Brazilians interacting with a receptionist robot. In *Proceedings for the 2015 International Conference on Advanced Robotics* (pp. 66–71). IEEE. https://doi.org/10.1109/ICAR.2015.7251435

*Ueno, A., Hlaváč, V., Mizuuchi, I., & Hoffmann, M. (2020). Touching a human or a robot? Investigating human-likeness of a soft warm artificial hand. In *Proceedings of the 29th IEEE International Conference on Robot and Human Interactive Communication* (pp. 14–20). IEEE. https://doi.org/10.1109/RO-MAN47096.2020.9223523

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*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

*Van der Hout, V. M. (2017). *The touch of a robotic friend* [Unpublished master's thesis]. University of Twente. http://purl.utwente.nl/essays/73221

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

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which data to meta-analyze, and how? A specification-curve and multiverse-analysis approach to meta-analysis. *Zeitschrift für Psychologie, 227*(1), 64–82. https://doi.org/10.1027/2151-604/a000357

Voskuhl, A. (2013). *Androids in the enlightenment: Mechanics, artisans, and cultures of the self*. University of Chicago Press.

Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology, 19*(4), 393–407. https://doi.org/10.1037/gpr0000056

Weiss, A., & Bartneck, C. (2015). Meta analysis of the usage of the Godspeed Questionnaire series. In *Proceedings of the 2015 International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 381–388). IEEE. https://doi.org/10.1109/ROMAN.2015.7333568

*Wieser, I., Toprak, S., Grenzing, A., Hinz, T., Auddy, S., Karaoğuz, E. C., Chandran, A., Remmels, M., Shinawi, A. E., Josifovski, J., Vankadara, L. C., Wahab, F. U., Bahnemiri, A. M., Sahu, D., Heinrich, S., Navarro-Guerrero, N., Strahl, E., Twiefel, J., & Wermter, S. (2016). A robotic home assistant with memory aid functionality. In G. Friedrich, M. Helmert, & F. Wotawa (Eds.), *Joint German/Austrian Conference on Artificial Intelligence* (pp. 102–115). Springer. https://doi.org/10.1007/978-3-319-46073-4_8

*Willemse, C., & Wykowska, A. (2019). In natural interaction with embodied robots, we prefer it when they follow our gaze: A gaze-contingent mobile eyetracking study. *Philosophical Transactions of the Royal Society B, 374*(1771), Article 20180036. https://doi.org/10.1098/rstb.2018.0036

Yoshikawa, M., Matsumoto, Y., Sumitani, M., & Ishiguro, H. (2011). Development of an android robot for psychological support in medical and welfare fields. In *Proceedings of the 2011 International Conference on Robotics and Biomimetics* (pp. 2378–2383). IEEE. https://doi.org/10.1109/ROBIO.2011.6181654

*Zanatto, D., Patacchiola, M., Goslin, J., & Cangelosi, A. (2019). Investigating cooperation with robotic peers. *PLoS One, 14*(11), Article e0225028. https://doi.org/10.1371/journal.pone.0225028

*Zanatto, D., Patacchiola, M., Goslin, J., Thill, S., & Cangelosi, A. (2020). Do humans imitate robots? An investigation of strategic social learning in human-robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 449–457). Association for Computing Machinery. https://doi.org/10.1145/3319502.3374776

Zhang, J., Li, S., Zhang, J. Y., Du, F., Qi, Y., & Liu, X. (2020). A literature review of the research on the uncanny valley. In P.-L. P. Rau (Ed.), *Cross-cultural design. User experience of products, services, and intelligent environments* (pp. 255–268). Springer.

History

Received February 24, 2021
Revision received November 15, 2021
Accepted November 18, 2021
Published online February 2, 2022

Acknowledgments

The authors are grateful to the research assistants Christine Busch, Lina Curth, Laura Moradbakhti, Simon Schreibelmayr, and Sandra Siedl, who contributed to the coding of the primary studies included in this meta-analysis.

Conflict of Interest

We have no conflicts of interest to disclose.

Open Data

The meta-analysis was not preregistered. The data, material, computer code, and analysis results are publicly available at the Open Science Framework (OSF; https://osf.io/t9rdk).

ORCID

Martina Mara
https://orcid.org/0000-0003-3447-0556

Timo Gnambs
https://orcid.org/0000-0002-6984-1276

Martina Mara
LIT Robopsychology Lab
Johannes Kepler University Linz
Altenberger Straße 69
4040 Linz
Austria
martina.mara@jku.at

Original Article

Block-Wise Model Fit for Structural Equation Models With Experience Sampling Data

Julia Norget and Axel Mayer

Faculty of Psychology and Sport Science, Bielefeld University, Germany

Abstract: Common model fit indices behave poorly in structural equation models for experience sampling data which typically contain many manifest variables. In this article, we propose a block-wise fit assessment for large models as an alternative. The entire model is estimated jointly, and block-wise versions of common fit indices are then determined from smaller blocks of the variance-covariance matrix using simulated degrees of freedom. In a first simulation study, we show that block-wise fit indices, contrary to global fit indices, correctly identify correctly specified latent state-trait models with 49 occasions and N = 200. In a second simulation, we find that block-wise fit indices cannot identify misspecification purely between days but correctly rejects other misspecified models. In some cases, the block-wise fit is superior in judging the strength of the misspecification. Lastly, we discuss the practical use of block-wise fit evaluation and its limitations.

Keywords: structural equation modeling, fit indices, latent state-trait theory, experience sampling

In psychological research, we often measure people's affect, behavior or cognition in different situations. Changes in measures from one occasion to another may reflect a change of the attribute in question, the different situations in which it was assessed, or be due to measurement error. With several measurement occasions, latent state-trait theory and its revised version (LST-R theory; Steyer et al., 1999, 2015) allows researchers to distinguish between occasion-specific (state residual) and stable (trait) influences on the observed attribute. State residuals reflect the influence of a specific situation and the person-situation interaction on the observed variable. A trait is an attribute of the person at the time of measurement (Steyer et al., 2015).

When we research states that fluctuate over short periods, experience sampling (ES) studies can be useful. In ES studies, participants respond about their behavior or thoughts several times a day during one or more weeks (Mehl et al., 2011), leading to large datasets. LST-R theory can also be applied to ES datasets. Eid et al. (2012) give an overview of models for ES data. These models include autoregressive effects to account for short time lags and can be defined in the LST-R framework (Eid et al., 2017).

There are multiple other approaches to assessing the (in)stability of constructs with structural equation models, for example, the single indicator STARTS model (Kenny & Zautra, 1995, 2001), the integrated state-trait model (Hamaker et al., 2007), the random intercept cross-lagged panel model (RI-CLPM; Hamaker et al., 2015), or multilevel approaches such as dynamic structural equation models (DSEM; e.g., Asparouhov et al., 2018; Zhang et al., 2008).

LST-R Theory

In this article, we focus on LST-R models for ES data. LST-R theory is an extension of classical test theory (CTT) for longitudinal data. While CTT can differentiate between person ("trait") effects and measurement error, LST-R theory also considers the influences of the situation and person-situation interaction. A revised version (LST-R theory; Steyer et al., 2015) recognizes that a person changes with experience and thus that traits can change over time.

Each observed variable (indicator) is denoted as Y_{it}, where i (i = 1, 2, 3,. . .) stands for the indicator and t (t = 1, 2, 3,. . .) for the time point. Each indicator can be decomposed into a latent state variable (τ_{it}) and measurement error. The latent state variables are defined as the expected value of Y_{it} given the person-at-time-t and the situation-at-time-t. The measurement error variable (ϵ_{it}) is the difference between Y_{it} and τ_{it}. The latent state variable is further decomposed into the latent trait variable (ξ_{it}) and the state residual variable (ζ_{it}). The latent trait variable is defined as the expected value of Y_{it} given the person-at-time-t. The state residual variable is the difference between the latent

Zeitschrift für Psychologie (2022), *230*(1), 47–59
https://doi.org/10.1027/2151-2604/a000482

state and the latent trait variables. Overall, we obtain the following equation:

$$Y_{it} = \xi_{it} + \zeta_{it} + \epsilon_{it}. \quad (1)$$

The latent trait variable represents the person-at-time-t-specific influence on the measurement. Since the person can change with experience, we could also call the trait variable an occasion-specific disposition. The state residual variable represents the influences of the situation and person-situation interaction.

Based on this decomposition, LST-R theory defines three important coefficients for each indicator. Consistency is the proportion of variance due to the trait variable: $\text{Con}(Y_{it}) = \text{Var}(\xi_{it})/\text{Var}(Y_{it})$. Occasion-specificity is the proportion of variance due to the state residual variable: $\text{Spe}(Y_{it}) = \text{Var}(\zeta_{it})/\text{Var}(Y_{it})$. Reliability is the sum of both, or in other words, it is the proportion of variance not due to unsystematic measurement error: $\text{Rel}(Y_{it}) = 1 - \text{Var}(\epsilon_{it})/\text{Var}(Y_{it})$.

With these definitions alone, it is not yet possible to estimate an LST-R model. Additional assumptions about the equivalence of latent state and trait variables need to be made to obtain an identified model. For a model with a single trait variable, the most restrictive equivalence assumptions (state- and trait-equivalence), assume that the state and trait variables are measured on the same scale with the same intercept, meaning that the intercept is zero and factor loadings are fixed to one. The model equation with these assumptions is $Y_{it} = \theta + \zeta_t + \epsilon_{it}$. There is one single trait variable θ for all occasions and several occasion-specific state residual variables ζ_t, as can be seen in the path model in Figure 1A. For a detailed overview of the definitions and additional assumptions in LST-R theory, see Steyer and colleagues (2015).

Models With Autoregressive Effects

In ES studies, time intervals between measurements are very short. Measures taken close together in time are more similar than measures taken further apart, and autoregressive effects are common in ES data (Bolger & Laurenceau, 2013). Eid and colleagues (2012) therefore propose different LST models with autoregression, which can be defined in the framework of LST-R theory (Eid et al., 2017). Autoregressive paths are added at the level of occasion-specific residual variables. The latent state variable is decomposed into the latent trait variable and an occasion factor (OCC_{ij}). Autoregressive paths are added between these occasion factors. The OCC_{ij} variables have a residual, which is the state residual variable ζ_{ij}. This means that the occasion factors are the current state residual plus a linear combination of all previous state residuals. The first occasion factor is identical to the first state residual. A model with autoregression is depicted in Figure 1B. It is also possible to add autoregressive paths between the latent states, but for most short-term longitudinal studies, autoregression between occasion-specific residual variables seems more suitable (Stadtbäumer et al., 2021).

Indicator- and Day-Specific Traits

While the models described above included a single trait, the constructs explored in ES research are often more dynamic and a single trait across the entire measurement period is not always realistic. Eid and colleagues (2012) describe models with indicator- and day-specific traits. If the indicators in the model are not homogeneous (e.g., positive and negative valence) indicator-specific traits can capture the specific components which are not shared. Given that the indicators are supposed to measure the same construct, indicator-specific traits should correlate highly. Indicator-specific LST-R models can also include indicator-specific equivalence assumptions, meaning that we assume state- and trait-equivalence separately for the manifest variables of each indicator. When the construct in question is stable within days but less stable across the entire measurement period, it is also possible to include day-specific traits. The day-specific trait variables can capture within-day stability, while the correlation between traits gives an indication of between-day stability. Day-specific models can also have day-specific state- and trait-equivalence assumptions, meaning that equivalence is assumed within each day. Day-specific and indicator-specific traits can also be combined. Some path models of indicator-specific and day-specific models can be found in the Electronic Supplementary Material 1 (ESM 1, Figure E1), which illustrates the design of the simulation study.

Model Fit Evaluation

In LST-R models for ES data, it is difficult to estimate model fit. Fit indices are less reliable for models with many manifest variables: they show inflated χ^2-values with rejection rates of up to 100% for correctly specified models (Moshagen, 2012). This so-called model size effect is largely influenced by the number of manifest variables (p) and the sample size (N) (Shi et al., 2019). The number of free parameters (q) has a smaller influence. Moshagen (2012) found no influence of q on inflated Type I error rates, but Shi and colleagues (2019) found such an effect. However, with a large number of manifest variables ($p \geq 60$) Type I error rates are dramatically inflated, independent of q and even with very large sample sizes (N = 2,000) (Shi et al., 2019). The model size effect disappears asymptotically (i.e., when N approaches infinity).

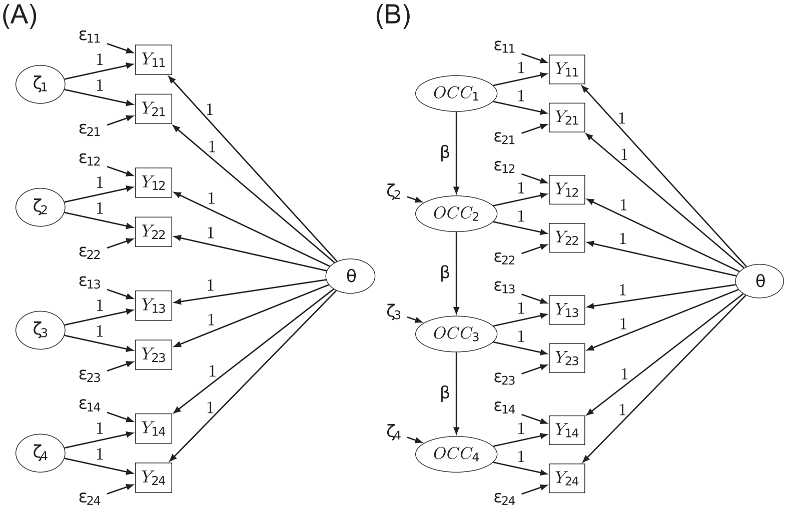


Figure 1. Path models of two LST-R single-trait models. (A) LST-R model with state residuals on the left side and a single trait on the right side. (B) LST-R model with autoregression.

There are different χ^2-corrections to counteract the model size effect, such as the ones by Bartlett (1950), Swain (1975), and Yuan and colleagues (2015). A comparison by Shi and colleagues (2018) showed that the correction by Yuan et al. (2015) performs best and results in acceptable Type I error rates, except with very large p (≥ 90) and small N (= 200). Yuan and colleagues (2015) multiply the empirical maximum likelihood χ^2 test statistic with the correction factor $e = [N - (2.381 + 0.361p + 0.006q)]/(N - 1)$.

Common fit indices such as Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) are based on the χ^2-value and are also biased in larger models (Kenny & McCoach, 2003; Shi et al., 2019). In the case of misspecified dimensionality, CFI and TLI worsen with more manifest variables but improve when the data-generating model includes residual correlations which are omitted in the analysis. RMSEA values decrease with more manifest variables (Kenny & McCoach, 2003; Savalei, 2012; Shi et al., 2019). Guidelines for interpreting these values are based on studies with smaller models (e.g., 15 manifest variables in the study by Hu & Bentler, 1999). For models with ES data, relying on these fit indices may lead to the incorrect rejection of acceptable models.

Local Fit Evaluation

The bias of fit indices is associated with model size, so a more local evaluation of smaller model elements seems intuitive. Maydeu-Olivares and Shi (2017) suggest that the local source of misspecification can be visually detected through areas (for misspecified trait dimensionality) or rows (for misspecified secondary loadings) of high residual correlations. However, with two items measured at only 14 time points, a residual correlation matrix has 784 entries, making it difficult to detect meaningful patterns. The number of large residual correlations also increases with matrix size. If there are no obvious patterns, this approach may not tell us if model rejection is due to model size or legitimate model misfit and may not be helpful for judging the fit of ES models.

Another approach to local fit evaluation is testing individual implications of the proposed model. Thoemmes and colleagues (2018) suggest conditional independence test for implications of the model structure and tetrad tests if latent variables are involved. The number of conditional independence constraints equals the degrees of freedom (*df*), and the number of tetrad constraints is large even with few latent variables. For models with ES data, there will be thousands of tests, making it difficult to derive what they imply for the model structure.

Recently, Rosseel and Loh (2021) presented the Structural After Measurement (SAM) framework, where parameters of the measurement part are estimated first, followed by the parameters of the structural part. The measurement part can be estimated as (1) a single measurement block containing all latent variables, (2) separate measurement blocks for each latent variable, or (3) several measurement blocks which can contain more than one latent variable. There may not be equality constraints, cross-loadings, or correlated residuals between indicators in different blocks. Fit indices are derived for each measurement block and the structural part. A special case of SAM is step-wise factor score regression, where the measurement models of each latent variable are estimated independently, and their relationships are modeled with factor scores (Devlieger, 2019). For ES LST-R models, measurement models with two or three indicators are too small for factor score regression or option 2 of the SAM framework, and option 3 does not work for models with a single trait, indicator-specific traits, measurement invariance over time or other equivalence assumptions. The first option, however, is not recommended and offers little benefit over SEM.

While local fit assessment has typically been recommended as a follow-up analysis, Rosseel and Loh (2021) show that it can also be useful as an alternative to global fit evaluation. Unfortunately, for the evaluation of most ES LST-R models, the SAM framework provides little added benefit. In this article, we will thus show that a new approach to local fit assessment can be a viable alternative to global evaluation for ES LST-R models.

We propose an approach where the full variance-covariance matrix is first estimated for the global model based on all postulated relationships. Then, local versions of fit indices are determined for each day (or other blocks) based on the global variance-covariance matrix using simulated block-wise *df*. This approach can take all kinds of relationships across measurement models and days into account and provides familiar fit indices. We will first show how local block-wise fit indices can be estimated. We then show in two simulation studies under which conditions they provide a more reliable fit assessment than global fit measures and discuss the implication of our results for the evaluation of large SEMs.

Block-Wise Model Fit Indices

In the past decades, a variety of fit indices have been developed to examine how well a theoretical model is supported by empirical data. Some of the most common indices are the χ^2, CFI, TLI, and RMSEA. In this section, we explain how these indices can be computed for blocks (e.g., days) of LST-R models for ES data in three steps: (1) estimating the overall model, (2) extracting blocks from $\hat{\Sigma}$ and S, and (3) calculating fit indices from these blocks.

First, the model including all latent constructs and postulated relationships is specified and estimated with maximum likelihood, yielding a model implied variance-covariance matrix $\hat{\Sigma}$. In the second step, a number of substantively meaningful blocks is chosen, such as the days in an ES study. All manifest variables are uniquely associated with one block, and all blocks contain the same number of manifest variables $p_k = \frac{p}{K}$, where p is the total number of manifest variables and K is the number of blocks. We use the subscript k for all block-specific parameters, with $k = 1, \ldots, K$. Then, the (co)variances of the manifest variables of each block are extracted from $\hat{\Sigma}$ and S. This results in K $p_k \times p_k$ model-implied ($\hat{\Sigma}_k$) and observed (S_k) (co)variance matrices (i.e., $\hat{\Sigma}_1$ for block 1, etc.). The block-wise matrices $\hat{\Sigma}_k$ and S_k must be invertible so that block-wise χ^2_k-values can be determined. In step three, model fit indices are determined for each block based on $\hat{\Sigma}_k$ and S_k with the regular formulas adapted for block-wise use.

This block-wise approach can be applied to large LST-R or other longitudinal models, where we can identify a substantively meaningful number of blocks. The block-wise approach allows for a day-specific evaluation of models that include restrictions across days, such as a single trait or measurement invariance over time.

Block-Wise χ^2

In a structural equation model, the χ^2-test evaluates the discrepancy between $\hat{\Sigma}$ and S, with the null hypothesis that $\hat{\Sigma}$ is identical to the (co)variances in the population from which the sample is drawn. An insignificant χ^2 value at an α-level of .05 is often used as an indicator of good fit. The χ^2-value is the product of the fitting function and sample size. The most common estimator to minimize the fitting function is the maximum likelihood (Bollen, 1989). Adapted for block-wise use, we get the formula:

$$\begin{aligned}\chi^2_k = {} & \big(\log|\hat{\Sigma}_k| + \mathrm{tr}(\hat{\Sigma}_k^{-1} S_k) - \log|S_k| - p_k \\ & + (\bar{x}_k - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (\bar{x}_k - \hat{\mu}_k)\big) \cdot (N-1),\end{aligned} \tag{2}$$

where p_k is the number of observed variables per block, $\bar{x}_k$ the vector of sample means, and $\hat{\mu}_k$ the vector of model-implied means, both for the items in block k. N is the sample size. Although the sample estimates for χ^2 and other fit indices include $N - 1$ in the formulas, both lavaan and MPlus use N instead. For the sake of consistency, we therefore used N in the computations for the simulation study. With large sample sizes, the χ^2-test yields significant p-values even for models with a minor misfit. This is one of the reasons which have inspired the development of different fit indicators including RMSEA, CFI, and TLI.

Block-Wise Degrees of Freedom

In order to test the null hypothesis and to calculate other fit indices, we need the *df*. In SEM, *df* are the difference between the number of empirical parameters (means, variances, and covariances of the manifest variables) and estimated parameters. All estimated parameters are involved in computing the implied (co)variances in $\hat{\Sigma}$. However, not all estimated parameters are uniquely associated with only one block. Time-invariant factor loadings, or the variance and mean of a single trait, affect the calculation of co(variances) in more than one $\hat{\Sigma}_k$ matrix.

Since it is unclear how estimated parameters can be split among blocks, we suggest simulating block-wise df_k. Under the null hypothesis, empirical χ^2 values follow a χ^2-distribution with $df = E(\chi^2)$. Thus, we can approximate the df by simulating many datasets from the true model and computing the mean of the block-wise χ^2-values $M(\chi^2_k)$.

Simulation of Block-Wise df_k

To test this behavior, we simulated data for 4 ES LST-R models (single-trait and day-specific model with and without autoregressive effect, with 7 occasions per day and 2 indicators per occasion), differing numbers of days (1, 2, 7), and sample sizes (200, 10,000). For each condition, 1,000 datasets were created and analyzed with the data-generating model. Block-wise χ^2_k values were calculated for 2 or 7 blocks, conferring to the number of days. We examined global df, $M(\chi^2)$, and the distribution of the χ^2 values, as well as block-wise $M(\chi^2_k)$ and the distribution of all χ^2_1, that is, the χ^2 values of the first block.

Simulation results show that global $M(\chi^2)$ for 2 or 7 days and $N = 200$ are overestimated, for example, the autoregressive model with day-specific traits for 7 days has $df =$ 4,852, but $M(\chi^2) = 5{,}965.70$. With $N = 10{,}000$, the χ^2 inflation almost disappeared (for the same model: $M(\chi^2) =$ 4,867.79). For models with 1 day, $M(\chi^2)$ closely resembles df (e.g., for autoregressive model with day-specific traits, $N = 200$: $df = 109$, $M(\chi^2) = 112.62$). We also computed Kolmogorov-Smirnov (KS) distances, which express the maximum difference on a scale from 0 to 1 between the observed distribution of the χ^2-values and their theoretical distribution with df degrees of freedom and checked which proportion of simulated χ^2-values fall within each decile of a χ^2-distribution with $df = M(\chi^2)$. Both approaches indicate that the simulated values are χ^2-distributed with $df = M(\chi^2)$. A table with all results is included in ESM 2.

Overall, the block-wise χ^2_k values are approximately χ^2-distributed with $df = M(\chi^2_k)$. Based on these results, we recommend simulating datasets based on $\hat{\Sigma}$ and the model-implied means with the actual sample size, computing block-wise χ^2 values for all datasets with Formula 2 and using $M(\chi^2_k)$ as an approximation of block-wise df_k. This approximation can then be used for the calculation of other fit indices. As an alternative, one can directly simulate the distribution of the test statistic and use its empirical distribution to test for significance. However, in this case, the other fit indices cannot be computed.

Block-Wise Absolute Fit Indices

Absolute fit indices such as the RMSEA can better be understood as measures of misfit, where small values indicate little misfit. The RMSEA is based on the χ^2 statistic but corrects for model complexity. The block-wise version for each block k can be calculated as follows:

$$\text{RMSEA}_k = \sqrt{\max\left(0, \frac{\chi^2_k - df_k}{df_k \cdot (N-1)}\right)}. \tag{3}$$

Although fit indices were developed to judge the extent of model (mis)fit, they are also affected by other influences such as the strength of factor loadings (Heene et al., 2011) or, as discussed before, the number of manifest variables. Common rules of thump should thus be used with great caution, and several indices need to be taken into consideration to judge model fit. According to Hu and Bentler (1999), RMSEA values < .06 indicate good fit. Another common rule of thumb is that RMSEA values $\leq$ .05 indicate close fit, and values $\leq$ .08 indicate reasonable fit (Browne & Cudeck, 1992). Another absolute fit index is the SRMR. This fit index does not depend on the χ^2-test statistic, but we provide information on block-wise SRMR in ESM 1.

Block-Wise Incremental Fit Indices

Incremental fit indices (e.g., CFI and TLI; Bentler, 1990; Tucker & Lewis, 1973) do not use the χ^2-statistic directly but compare the proposed model to the worst possible (null) model. The null model only includes variances for the observed variables, but no relationships are modeled. For block-wise CFI and TLI, the null model is computed for each block based on the manifest variables from the block in question, p_k. The block-wise null model has $p_k(p_k - 1)/2$ df, and a χ^2-value is estimated according to Formula 2. The block-wise Bentler Comparative Fit Index (CFI$_k$) can then be calculated as follows (Shi et al., 2019):

$$\text{CFI}_k = \frac{\max\left(d_k(\text{Null Model}), 0\right) - \max\left(d_k(\text{Proposed Model}), 0\right)}{\max\left(d_k(\text{Null Model}), 0\right)}, \tag{4}$$

where $d = \chi^2_k - df_k$ for the null and proposed model. CFI$_k$ can range between 0 and 1. The block-wise TLI$_k$ is calculated as follows:

$$\text{TLI}_k = \frac{\chi^2_k/df_k(\text{Null Model}) - \chi^2_k/df_k(\text{Proposed Model})}{\chi^2_k/df_k(\text{Null Model}) - 1}. \tag{5}$$

Since the TLI is not normed, TLI > 1 or negative values are possible. For both CFI and TLI, values $\geq$ .97 indicate a good fit between model and data, but values between .95 and .97 are considered acceptable (Hu & Bentler, 1999; Schermelleh-Engel et al., 2003).

Simulation Studies

We have shown that common advice for judging model fit is not suitable for models with many manifest variables (e.g., Kenny & McCoach, 2003; Moshagen, 2012; Savalei, 2012; Shi et al., 2019) and have proposed block-wise evaluation. In order to demonstrate that block-wise evaluation is a viable alternative for models with many manifest variables, we conduct two simulation studies. We first simulate correctly specified data for ES LST-R models and evaluate the effect of model size and sample size on global and block-wise fit indices. Here, we expect that global fit indices will incorrectly reject models with more days and a smaller sample size, which is common in ES studies. We expect that block-wise fit indices can correctly identify these models. In a second simulation study, we generate data for the same ES LST-R models but with different misspecifications and evaluate the effects of model size, sample size, and misspecification on global and block-wise fit. Block-wise fit evaluation is based on (co)variances within each day, so we expect that block-wise fit indices will correctly reject models which are misspecified within days, but fail to identify models which are misspecified purely between days.

Study 1: Correctly Specified Models

Method

In the first simulation study, model fit is evaluated for two different ES LST-R models, with varying model and sample size. Overall, we have a 2 (models) × 2 (model size) × 2 (sample size) design. The analysis models are (1) an autoregressive multistate–singletrait model, where a single trait is assumed across all measurements, and (2) an autoregressive multistate–multitrait model with day-specific traits. We included both models because the multistate–singletrait model is most common in applications of LST(-R) theory, but the model with day-specific traits is suitable for many applications with ES data. We did not have different hypotheses for these two models. The models include 2 ndicators for each occasion and 7 occasions for each day. LST-R models can include more indicators, but ES studies typically include as few questions as possible to keep the strain on participants low. Models with 2–3 indicators thus seem realistic for ES LST-R models. Both models have η-equivalent and θ-equivalent measures within each day. This implies that all factor loadings are set to 1, all intercepts are 0, and all (state) residual variances are equal within each day ($\text{Var}(\epsilon_t) = \text{Var}(\epsilon_u)$ and $\text{Var}(\zeta_t) = \text{Var}(\zeta_u)$, $t \neq u$). Autoregressive effects are restricted to be equal between all occasion-specific factors. Parameters in the data-generating population models were $\text{Var}(\zeta_t) = .3$, $\text{Var}(\theta_{(u)}) = .3$ for the trait in the single-trait model and all day-specific trait variables θ_u in the day-specific traits model, $\text{Var}(\epsilon_{it}) = .4$, $\text{Cov}(\theta_u, \theta_V, u \neq v) = 0.21$ (corresponding to a correlation of $r = .7$), $M(\theta_{(u)}) = 2.2$, and autoregressive effects $\beta = .1$. This implies equal occasion-specificity and consistency, with item reliabilities between .60 and .61. These values are approximately based on an empirical application with ratings of perceived conflict of interest in social situations (Norget et al., 2021). The trait and state residual variances are adjusted to be equal because many constructs assessed in longitudinal studies have both stable and occasion-specific aspects (Geiser, 2021). Please refer to Figure E1 (ESM 1) for path models of the single-trait and the day-specific traits model. Data was generated for models with 2 or 7 days (i.e., 28 or 98 manifest variables) and sample sizes of 200 or 1,000. Typical data situations in ES studies include sample sizes around or smaller than $N = 200$ and data collection on several days, often one or two weeks. For each condition of the study, we estimate global fit indices as well as block-wise fit indices for each day. For comparison, we also simulated global *df* in the same way as we described for the block-wise *df* and computed all global fit indices using these simulated *df*. Additionally, we computed the Yuan et al. (2015) corrected χ^2-estimates to compare rejection rates and χ^2/df-ratios.

For each of the 8 conditions, 500 datasets were generated and analyzed. Block-wise (and global) *df* were simulated for the first dataset in each condition. These estimates were then used for all 500 datasets in the same condition. In a test phase, we simulated the block-wise *df* several times for the same condition and found very small deviations between the estimates. The simulation study was conducted in R (R Core Team, 2020; RStudio Team, 2019) using the packages SimDesign (Chalmers & Adkins, 2020), lavaan (Rosseel, 2012), lsttheory (Mayer, 2020), and MASS (Venables & Ripley, 2002). We discuss χ^2-rejection rates at $\alpha = .05$, KS distances and mean CFI, TLI, and RMSEA values for the different conditions and point out the most important aspects of this visual analysis. Results are shown in Figure 2 and Figures E2 and E3 (ESM 1).

Results

There were (almost) no differences between the two models. We present the results for the day-specific model here and provide results for the other model in Figure E2 (ESM 1). We will refer to the global fit indices as implemented in common SEM software as "global" χ^2, CFI, TLI, and RMSEA. Estimates based on simulated global *df* are "simulated global" values, and Yuan and colleagues (2015) corrected values are "Yuan-corrected".

χ^2-Rejection and Kolmogorov-Smirnov Distance

For correctly specified models, χ^2-rejection rates at $\alpha = .05$ should be around 5%. As shown in Figure 2A, χ^2-rejection

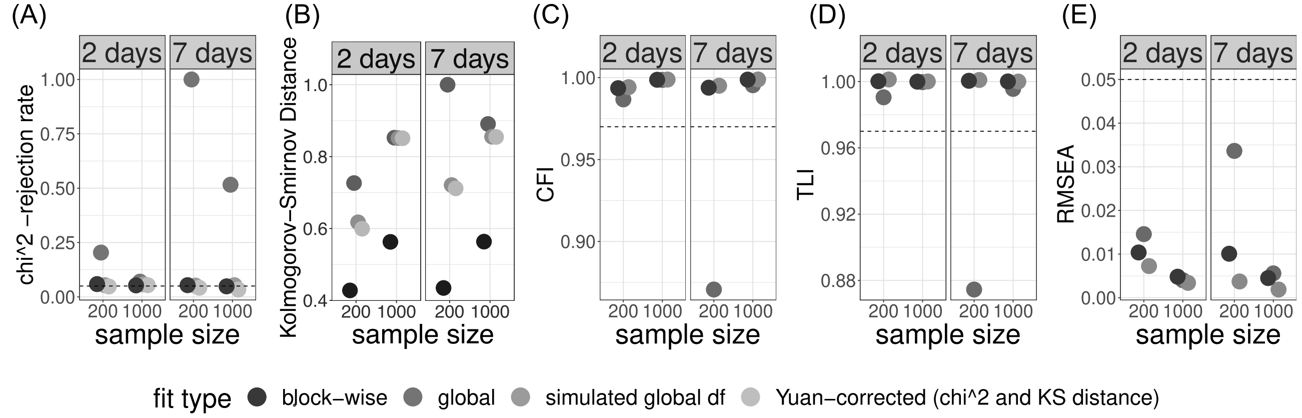


Figure 2. Results of Study 1 for the day-specific model (single-trait figures are included in Figure E2 (ESM 1). (A) χ^2-rejection rates at α = .05; (B) KS distance single-trait model; (C) CFI values; (D) TLI values; (E) RMSEA values.

rates for global evaluation, with globally simulated *df*, and Yuan-corrected χ^2 for models with 2 days and N = 1,000 are close to the expected rejection rate. With smaller sample sizes and more days, global rejection rates increase up to 100% for models with 7 days and N = 200. The KS distances (Figure 2B) show that the distribution of global, simulated global, and Yuan-corrected χ^2 values differs more strongly from their theoretical distribution than block-wise χ^2_k. Global χ^2 values are most strongly overestimated. Again, this difference is especially large for models with 7 days and N = 200. Overall, global χ^2, as implemented in most software, highly overestimates the test statistic and too often rejects correctly specified models, especially in the most likely data scenario in ES studies, while block-wise χ^2_k performs much better.

Comparative Fit Index and Tucker-Lewis Index

For CFI and TLI most models yield estimates ≥ .97, indicating a good model fit. However, we can see in Figures 2C and 2D that global CFI and TLI indicate worse fit for smaller samples (N = 200 vs. N = 1,000) and for larger models (7 vs. 2 days). In the most likely data scenario in ES studies (7 days and N = 200) global indices reject the correctly specified model (CFI = .88, TLI = .88). However, simulated global CFI and TLI, and block-wise CFI$_k$ and TLI$_k$ correctly indicate a good fit in all cases.

Root Mean Square Error of Approximation

All types of RMSEA$_{(k)}$ correctly identify a good fit in all four scenarios (see Figure 2E). For models with N = 1,000, global, simulated global, and block-wise RMSEA$_{(k)}$-values are very small (.003–.006). For models with N = 200, RMSEA$_{(k)}$ values are slightly higher, and global values indicate worse fit than simulated global or block-wise values. RMSEA indicates good fit in all scenarios, but global RMSEA is noticeably worse for N = 200 and 7 days (.034) compared to all other conditions. Block-wise RMSEA$_{(k)}$ clearly indicates a better fit in this case. Simulated global values indicate better fit in all scenarios. Since all RMSEA$_{(k)}$ correctly indicate a good fit, block-wise evaluation may offer less benefit over global evaluation in the case of RMSEA compared to other fit indices. However, block-wise and simulated global RMSEA$_{(k)}$ still correctly indicate a better fit than global RMSEA.

Discussion

Overall, correctly specified models were correctly identified by block-wise χ^2_k, CFI$_k$, TLI$_k$, and RMSEA$_k$, but not always by their global counterparts. Simulated global indices behave similarly to block-wise indices. The biases in global fit are in line with previous simulation studies (Kenny & McCoach, 2003; Moshagen, 2012). Especially in the most likely data scenario with experience sampling data, models for 7 days (49 occasions), and sample sizes of N = 200, block-wise fit evaluation seems to offer a good alternative to global evaluation.

Study 2: Misspecified Models

Method

While Study 1 showed that block-wise fit correctly identifies correctly specified models, it is also important to consider under which conditions block-wise fit can correctly reject misspecified models. Since block-wise fit is based on the (co)variances of each block, we expect that misspecifications within blocks should be identified correctly, while misspecifications purely between blocks should be undetectable for block-wise fit indices. In Study 2, we generated data with different misspecifications in a 2 (models) × 2 (model size) × 2 (sample size) × 6 (misspecification) design. Analyzing ES LST-R models were the same as described in Study 1: a single-trait and a day-specific model. Again, we generated data for either 2 or 7 days, with sample sizes of 200 or 1,000. Global CFI and TLI worsen with more manifest variables and misspecified dimensionality, but improve with omitted residual correlations (Shi et al., 2019), so we included models with omitted residual correlations between and within days, as well as structural misspecifications similar to those in Shi et al. (2019). Pathmodels are provided in Figure E1 (ESM 1). The misspecified models include (1) small or (2) large residual correlations between days, that is, both items measured on occasion 1, 2, and so forth on each day are correlated with the same item measured on the same occasion on other days. Residual correlations are small (r = .15) or large (r = .40); (3) small (r = .15) or (4) large (r = .40) residual correlations within days, that is, the residuals of item 1 on all occasions within the same day are correlated, and likewise for item 2; (5) small or (6) large structural error, that is, each trait is split into two indicator-specific traits in the population model, with correlations of r = .90 (small error) or r = .60 (larger error). Other population values are identical to Study 1. We expected that block-wise fit would detect the structural error and the omitted residual correlations within days but not between days.

For the χ^2, we discuss rejection rates at α = .05 and provide further analysis for the ratio between χ^2 and df. χ^2/df = 1 indicates perfect fit. For CFI, TLI, and RMSEA, we analyze their global, simulated global, and block-wise values using analyses of variance (ANOVAs) with the respective fit index as the outcome and the four predictors: (1) model (single-trait/day-specific traits), (2) number of days (2/7), (3) sample size (200/1,000), and (4) the type of the fit index (global/simulated global/block-wise). All predictors are coded as factors, and we use Type III sum of squares and sum to zero contrasts. We used the R package car (Fox & Weisberg, 2019) to fit the ANOVAs and the package effectsize (Ben-Shachar et al., 2020) for effect sizes. Normal distribution of the residuals and variance homogeneity were visually checked, and the assumptions were met sufficiently.

Results

χ^2-*Rejection Rates and* χ^2/df Ratios

The χ^2-rejection rates at α = .05 are displayed in Figure 3A. Colored figures can be found in Figures E4a–E4e (ESM 1). Most models have rejection rates of around 100%. Block-wise χ^2 cannot detect the omitted residual correlations between days and incorrectly indicates perfect fit (i.e., rejection rates around 5%). For small misspecifications and N = 200, block-wise, and to a lesser degree also simulated global and Yuan-corrected $\chi^2_{(k)}$ sometimes have rejection rates notably lower than 100%; global χ^2 for N = 200 and 2 days as well, but with higher rejection rates than the other types.

The ANOVA for the χ^2/df ratio revealed substantial main effects of the misspecification, F(5, 179,876) = 412,102, p < .001, η^2 = .26, and sample size, F(1, 179,876) = 952,053, p < .001, η^2 = .12, meaning that all χ^2/df ratios are for a large part similarly affected by these two influences. Since we are more interested in the differences between the types of fit, we will focus on interaction effects with the type of fit measure. A table with the complete ANOVA results is included in Table E1 (ESM 1).

First, there is considerable two-way interaction between the type of fit measure and the misspecification, F(15, 179,876) = 105,156.1, p < .001, η^2 = .20. Block-wise χ^2_k/df_k ratios are higher (i.e., indicate worse fit) than global, simulated global, or Yuan-corrected ratios for models with large structural misspecification and omitted residual correlations within days. However, block-wise χ^2_k/df_k ratios indicate a perfect fit for the models with omitted residual correlations between days.

Second, there is an interaction effect between type of fit and sample size, F(3, 179,876) = 578,95.2, p < .001, η^2 = .02). Looking at the types of fit separately, the effect of sample size remains substantial for all, with lower ratios for N = 200 than N = 1,000. The difference between sample sizes is larger for block-wise ratios ($M_{1000} - M_{200}$ = 3.48) than for global ($M_{1000} - M_{200}$ = 1.72), simulated global ($M_{1000} - M_{200}$ = 1.84) and Yuan-corrected ($M_{1000} - M_{200}$ = 1.84) ratios.

Furthermore, there is a 3-way interaction between type of fit, misspecification, and sample size, F(15, 179,876) = 58,386.4, p < .001, η^2 = 0.11. Figure 3B shows that for block-wise χ^2_k/df_k, and to a lesser extend for global, simulated global, and Yuan-corrected ratios, the difference between N = 200 and N = 1,000 is larger with strongly misspecified models compared to their less strongly misspecified counterparts.

Contrary to our expectations, there was no noteworthy interaction between the type of fit and the number of days, F(3, 179,876) = 20,434.2, p < .001, η^2 = .008, or main effect of the number of days, F(1, 179,876) = 107,115.8, p < .001, η^2 = .01.

Comparative Fit Index and Tucker-Lewis Index

The results for CFI and TLI barely differ, and results are reported together. A figure with the TLI results is included in Figure E4d (ESM 1). Most effects are significant in the ANOVAs, and we focus on those with notable effect sizes. All main effects, except for the effect of the model (single-trait vs. day-specific), are noteworthy and interact with the type of fit. We will focus on these interactions here since we are mostly interested in how global and block-wise fit are differently affected by other influences. Full results are included in Tables E2 and E3 (ESM 1).

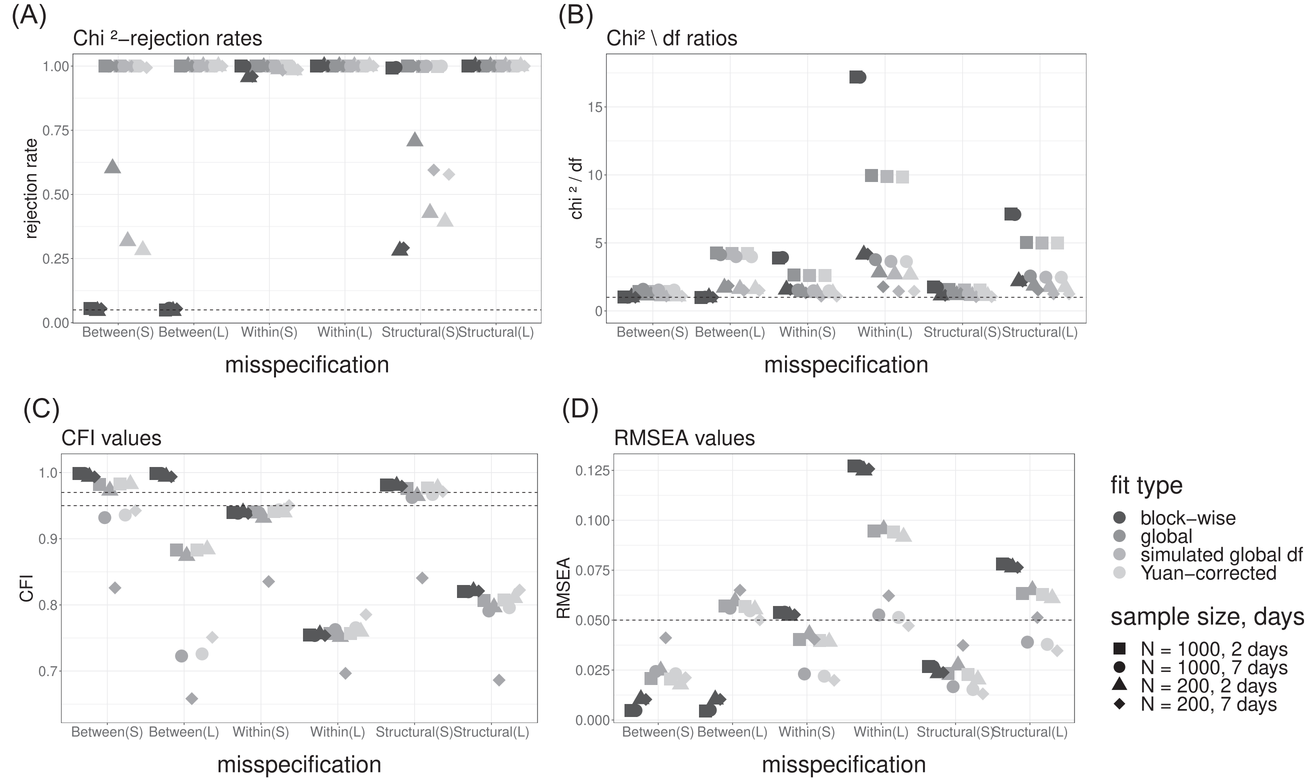


Figure 3. Overview of results of Study 2 for each fit index. There are six misspecifications on the *x*-axis. Between(S): omitted residuals correlations between days (*r* = .15); Between(L): likewise but with *r* = .40; Within(S): omitted residual correlations within days (*r* = .15); Within(L): likewise but with *r* = .40, Structural(S): data is generated with correlated indicatorspecific traits (*r* = .90); Structural(L): likewise but with *r* = .60. (A) χ^2-rejection rates at α = .05; (B) χ^2/df ratios; (C) CFI values (TLI values look almost identical); (D) RMSEA.

Most notably, there is a substantial interaction between the misspecification and type of fit (CFI: $F(10, 155{,}903) = 36{,}904.1$, $p < .001$, $\eta^2 = .13$; TLI: $F(10, 155{,}903) = 38{,}277.6$, $p < .001$, $\eta^2 = .11$). This effect is largely due to the models with omitted residual correlations between days. Here, block-wise CFI_k and TLI_k indicate perfect fit, while global and simulated global CFI and TLI can identify the misspecification.

The number of days also interact with the type of fit (CFI: $F(2, 155{,}903) = 42{,}835.7$, $p < .001$, $\eta^2 = .03$; TLI: $F(2, 155{,}903) = 72{,}257.6$, $p < .001$, $\eta^2 = .04$). Global CFI and TLI values are lower for models with 7 than 2 days (CFI: $t(23{,}394) = 67.04$, $p < .001$, $d = .87$; TLI: $t(22{,}675) = 77.21$, $p < .001$, $d = 1.00$), to a lesser extend this is also true for simulated global indices (CFI: $t(23{,}783) = 24.44$, $p < .001$, $d = .32$; TLI: $t(23{,}666) = 24.58$, $p < .001$, $d = .32$) but for block-wise CFI_k and TLI_k there is no notable difference between 2 and 7 days (CFI: $t(38{,}858) = 0.60$, $p = .55$; TLI: $t(38{,}741) = 0.03$, $p = .98$).

There is a smaller interaction between the sample size and type of fit (CFI: $F(2, 155{,}903) = 19{,}282.2$, $p < .001$, $\eta^2 = .01$; TLI: $F(2, 155{,}903) = 33{,}428.6$, $p < .001$, $\eta^2 = .02$). Block-wise CFI_k and TLI_k are barely affected by sample size (CFI_k: $t(107{,}980) = -2.66$, $p = .008$, $d = -0.02$; TLI_k: $t(107{,}980) = 4.07$, $p < .001$, $d = 0.02$), and the effect on simulated global indices is also small (CFI: $t(23{,}939) = 6.63$, $p < .001$, $d = 0.09$; TLI: $t(23{,}326) = 18.45$, $p < .001$, $d = 0.24$). Here, models with $N = 200$ fit better than with $N = 1{,}000$. Global CFI and TLI are generally worse for smaller samples (CFI: $t(23{,}739) = -40.2$, $p < .001$, $d = -0.52$) TLI: $t(23{,}611) = -40.2$, $p < .001$, $d = -0.52$).

Type of fit also interacts with number of days and misspecification (CFI: $F(10, 155{,}903) = 6{,}071.9$, $p < .001$, $\eta^2 = .02$; TLI: $F(10, 155{,}903) = 7{,}123.1$, $p < .001$, $\eta^2 = .02$). The interaction between days and misspecification remains noteworthy for global (CFI: $F(5, 23{,}988) = 1{,}965.8$, $p < .001$, $\eta^2 = .06$), and simulated global (CFI: $F(5, 23{,}988) = 6{,}697.9$, $p < .001$, $\eta^2 = .09$) but not for block-wise fit (CFI: $F(5, 107{,}988) = 3.8175$, $p = 002$, $\eta^2 < .001$). Figure 3C shows that especially for models with omitted residual correlations between days, global and simulated global CFI and TLI for 7 days (circle and diamond shape) is smaller than for 2 days (square and triangle). This difference is smaller for other misspecifications.

Another interesting three-way interaction is between the type of fit, number of days, and sample size (CFI:

$F(1,\ 155{,}903) = 13{,}693.5$, $p < .001$, $\eta^2 = .01$; TLI: $F(2,\ 155{,}903) = 21{,}879.2$, $p < .001$, $\eta^2 = .01$). This effect can easily be understood when we look at Figure 3C: For $N = 200$ and 7 days the global values (i.e., medium-gray diamond) are systematically lower than global and block-wise values for other combinations of the three predictors.

Root Mean Square Error of Approximation

Most effects on the RMSEA are statistically significant, and we only discuss those with notable effect sizes. Complete results are included in Table E4 (ESM 1). In terms of main effects, the misspecification accounts for the majority of variance in all RMSEA$_{(k)}$ values, $F(5,\ 155{,}903) = 192{,}939.9$, $p < .001$, $\eta^2 = .53$, and there is a small main effect of the number of days, $F(1,\ 155{,}903) = 27{,}321.8$, $p < .001$, $\eta^2 = .01$.

Again, the strongest interaction is between the type of fit and misspecification, $F(10,\ 155{,}903) = 53{,}426.9$, $p < .001$, $\eta^2 = .29$. Figure 3D shows that block-wise RMSEA$_k$ generally indicates worse fit than global and simulated global RMSEA, except in the case of omitted residual correlation between days, which cannot be detected by block-wise fit.

Additionally, there is an interaction between type of fit, misspecification, and the number of days, $F(10,\ 155{,}903) = 31{,}80.7$, $p < .001$, $\eta^2 = .02$. The interaction between misspecification and number of days remains notable for global RMSEA, $F(5,\ 23{,}988) = 6{,}241$, $p < .001$, $\eta^2 = .14$, and simulated global RMSEA, $F(5,\ 23{,}988) = 11{,}129$, $p < .001$, $\eta^2 = .12$, but there is no interaction for block-wise RMSEA$_k$, $F(5,\ 107{,}988) = 2.06$, $p = .67$. A look at the medium- and light-gray shapes in Figure 3D reveals that (simulated) global RMSEA indicates better fit for models with 7 than 2 days in the case of omitted residual correlations within days or large structural misspecification. The figure also shows that global and simulated RMSEA tend to assess strongly misspecified models with 7 days as acceptable, while block-wise fit can identify them as fitting badly.

There is also a small interaction effect of the type of fit with number of days, $F(1,\ 155{,}903) = 11{,}240.8$, $p < .001$, $\eta^2 = .01$. While the RMSEA values with 7 days are lower than with 2 day for global, $t(19{,}600) = 32.44$, $p < .001$, $d = 0.42$, and simulated global RMSEA, $t(19{,}234) = 58.15$, $p < .001$, $d = 0.75$, there is no difference between the number of days for block-wise RMSEA$_k$, $t(38{,}839) = -0.32$, $p = .75$.

Discussion

In general, all fit types indicate a less-than-perfect fit for misspecified models and stronger misfit for more strongly misspecified models. As expected, block-wise fit cannot identify the misspecification between blocks because block-wise fit indices are based on the (implied and observed) (co)-variances of items associated with the same block.

In line with previous research (Kenny & McCoach, 2003; Moshagen, 2012), global χ^2 is strongly affected by sample size. The same remains true for block-wise χ^2_k, χ^2 evaluation with simulated *df* and Yuan et al. (2015) corrected χ^2. Contrary to the correctly specified models, number of days (and thus number of manifest variables) does not affect the χ^2-tests for misspecified models. This could be due to the fact that the number of misspecified covariances also increases with model size.

CFI and TLI behave practically identical in our simulation study. Globally, they are sensitive to the number of days in the model, to a lesser extend also when they are estimated globally with simulated *df*. Their block-wise counterparts are not affected by the number of days. Regular global CFI and TLI indicate worse fit for all models with 7 days and $N = 200$, which is a likely ES data scenario. Block-wise CFI$_k$ and TLI$_k$, and to a lesser extend global CFI and TLI with simulated *df*, generally indicate better fit than regular global indices. Contrary to Kenny and McCoach (2003) and Shi and colleagues (2019), global CFI and TLI also worsened with more days for models with omitted residual correlations. In previous studies, the number of misspecified covariances remained stable with more manifest variables in the model, and the proportion of misspecified covariances decreased with model size. In our study, the number of misspecified covariances also increased with model size, explaining our different results. In fact, for the model with omitted residual correlations within days, the proportion of misspecified covariances is larger for 2 days than 7 days, but CFI and TLI indicate a worse fit for 7 days. This demonstrates that these global indices are indeed strongly affected by the number of manifest variables.

Global RMSEA and global RMSEA based on simulated *df* indicate a slightly better fit for models with more days (i.e., more manifest variables), while the number of days does not affect block-wise RMSEA$_k$. Especially in the case of strongly misspecified models for 7 days, global RMSEA would still let us erroneously conclude that these models are acceptable, while block-wise evaluation can identify them as fitting badly. Block-wise RMSEA$_k$ generally indicates a worse fit than both global indices, which is desirable in misspecified models. The behavior of global RMSEA is largely in line with previous research (Kenny & McCoach, 2003; Savalei, 2012; Shi et al., 2019).

Global Discussion

In this article, we introduced block-wise model fit evaluation for LST-R models with experience sampling data. We performed two simulation studies to compare block-wise fit evaluation to traditional global evaluation. We also included Yuan and colleagues (2015) corrected χ^2 estimates and global fit indices derived with simulated degrees of freedom for comparison. In Study 1, we investigated if the different fit indices properly identify correctly specified models. Results show that traditional global fit evaluation too often leads to the rejection of correctly specified models, especially with realistic sample and model sizes in ES studies. Block-wise fit evaluation, global fit derived from simulated *df*, and Yuan et al. (2015) corrected χ^2 correctly identified that these models fit well.

In the second study, we investigated under which conditions block-wise fit indices correctly reject misspecified models. As expected, if models are misspecified purely between days, block-wise fit cannot identify the misfit. Furthermore, traditional global CFI and TLI generally indicate a worse fit for the most realistic ES data scenario (7 days and N = 200) compared to other models and sample sizes. Block-wise and global indices with simulated *df* do not share this bias. Block-wise RMSEA$_k$ can more often identify (strongly) misspecified models than both types of global RMSEA.

Practical Usage

Based on the simulation results, we recommend using block-wise fit indices for LST models with many measurement occasions (e.g., several occasions on each day for one or more weeks) and sample sizes around 200 or smaller. With large sample sizes of around 1,000, there is less benefit in using block-wise fit evaluation. However, sample sizes around or under 200 are much more common in empirical research, so block-wise fit is useful for data from a typical ES study.

When using block-wise fit, researchers need to find logical blocks in their data. This decision should be based on the (LST) model and the study design. For example, blocks can correspond to days. When data was collected over several weeks with fewer occasions per day, blocks may better correspond to weeks.

Compared with existing corrections for the model size effect, such as Yuan et al. (2015), the block-wise evaluation provides valuable additional information about each block. In empirical data, it may happen that some blocks indicate acceptable fit, while others do not. This information about the source of misfit can be used to reflect on the model and data collection. For example, were there any structural differences between the days of assessment, such as weekdays and weekends being on the same days of the study for all participants? The R-function to determine block-wise fit indices is available in ESM 3.

Limitations and Future Research

The main limitation of the block-wise fit approach is evident from Study 2: misspecifications between blocks cannot be detected. We have proposed a block-wise fit for each block, but it is usually not possible to determine a block-wise fit between two blocks. To calculate block-wise fit indices, we extract blocks from $\hat{\Sigma}$ which only contain the (implied) (co)variances between items of the same block. Theoretically, it would be possible to extract the sections containing only covariances of items from two different blocks. However, if we assume any kind of measurement invariance between the blocks, the section of $\hat{\Sigma}$ which contains only the implied covariances between two blocks i and j, $i \neq$ j contains identical (and thus linearly dependent) vectors. The determinant of such a matrix is zero, log(0) is not defined, and a block-wise χ^2_{ij} cannot be determined. As a consequence, the block-wise fit is not informative about misfit between blocks. In future studies, the block-wise approach could be extended to include information between blocks. It should be possible to extract (co)variances of two consecutive or non-consecutive days and estimate a block-wise indices from these blocks. Blocks of different sizes could also be an option, for example, if a researcher is interested in morning- and evening-blocks with different numbers of measurements.

Also, the influence of differing numbers of indicators per block on block-wise fit indices was not assessed, and we cannot give advice on the number of manifest variables per block. Studies on which common advice for interpreting fit indices are based might serve as an orientation. For example, Hu and Bentler (1999) used 15 indicators.

Furthermore, missing data is common in ES studies. In the article, we have not yet discussed how Full Information Maximum Likelihood (FIML), a common missing data strategy, could be applied for block-wise fit indices. To date, the approach we introduced works with multiply imputed datasets. For practical reasons, it will be helpful to extend this approach to FIML. We have also focused on the χ^2-test, CFI, TLI and RMSEA, but the block-wise approach can be extended to other fit indices.

Electronic Supplementary Material

The electronic supplementary materials are available with the online version of the article at https://doi.org/10.1027/2151-2604/a000482

ESM 1. Formula of the Standardized Root Mean Square Residual (SRMR), design, and results of Studies 1 and 2. Figure E1: Design of Studies 1 and 2. Figure E2: Results of Study 1 for the single-trait model. Figure E3: Results of Study 1 for the model with day-specific traits. Figure E4: Results of Study 2 for each fit index, (a)–(e). Tables E1–E4: ANOVA tables showing the effects on χ^2/df ratios, CFI, TLI, and RMSEA.

ESM 2. Results of the block-wise degrees of freedom pilot simulation study.

ESM 3. The R-function to compute block-wise fit indices based on a fitted lavaan object.

References

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology, 3*(2), 77–85.

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software, 5*(56), Article 2815. https://doi.org/10.21105/joss.02815

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology, 16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Devlieger, I. (2019). *Factor score regression* (Doctoral dissertation). Ghent University.

Eid, M., Courvoisier, D. S., & Lischetzke, T. (2012). Structural equation modeling of ambulatory assessment data. In M. R. Mehl & T. S. Connor (Eds.), *Handbook of research methods for studying daily life* (pp. 384–406). Guilford Press.

Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (2017). On the definition of latent-state-trait models with autoregressive effects. *European Journal of Psychological Assessment, 33*(4), 285–295. https://doi.org/10.1027/1015-5759/a000435

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Geiser, C. (2021, June). *Are psychological constructs traits or states? Preliminary findings from a review of applied latent state-trait studies*. Paper presented at the EAPA Digital Event 2021 Conference. https://www.eapa.science/services/ability-assessment-1-1-1

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated state-trait model. *Journal of Research in Personality, 41*, 295–315. https://doi.org/10.1016/j.jrp.2006.04.003

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20*(1), 102–116. https://doi.org/10.1037/a0038889

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods, 16*(3), 319–336.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10*(3), 333–351. https://doi.org/10.1207/S15328007SEM1003_1

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology, 63*(1), 52–59. https://doi.org/10.1037/0022-006X.63.1.52

Kenny, D. A., & Zautra, A. (2001). Trait-state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *Decade of behavior. New methods for the analysis of change* (pp. 243–263). American Psychological Association. https://doi.org/10.1037/10409-008

Maydeu-Olivares, A., & Shi, D. (2017). Effect sizes of model misfit in structural equation models. *Methodology, 13*(S1), 23–30. https://doi.org/10.1027/1614-2241/a000129

Mayer, A. (2020). *Lsttheory: Latent state-trait theory*. [R package version 0.2-1.002]. https://github.com/amayer2010/lsttheory

Mehl, M. R., Conner, T. S., & Csikszentmihalyi, M. (2011). *Handbook of research methods for studying daily life* (1st ed.). Guilford Press.

Moshagen, M. (2012). The model size effect in sem: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling, 19*(1), 86–98. https://doi.org/10.1080/10705511.2012.634724

Norget, J., Columbus, S., Mayer, A., & Balliet, D. (2021, June). *Latent state-trait models of subjective interdependence*. Paper presented at the EAPA Digital Event 2021 Conference. https://www.eapa.science/services/ability-assessment-1-1-1

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Rosseel, Y., & Loh, W. W. (2021). *A structural after measurement (SAM) approach to SEM*. https://osf.io/pekbm/

RStudio Team. (2019). *Rstudio: Integrated development environment for R*. RStudio. http://www.rstudio.com/

Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement, 72*(6), 910–932. https://doi.org/10.1177/0013164412452564

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23–74.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement, 79*(2), 310–334. https://doi.org/10.1177/0013164418783530

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling, 25*(1), 21–40. https://doi.org/10.1080/10705511.2017.1369088

Stadtbäumer, N., Kreissl, S., & Mayer, A. (2021). *Comparing reformulated latent state-trait models with autoregressive effects*. Manuscript submitted for publication.

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits-revised. *Annual Review of Clinical Psychology, 11*, 71–98. https://doi.org/10.1146/annurev-clinpsy-032813-153719

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13*(5), 389–408. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0CO;2-A

Swain, A. J. (1975). *Analysis of parametric structures for variance matrices* (Doctoral dissertation). Adelaide.

Thoemmes, F., Rosseel, Y., & Textor, J. (2018). Local fit evaluation of structural equation models using graphical criteria. *Psychological Methods, 23*(1), 27–41. https://doi.org/10.1037/met0000147

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1–10. https://doi.org/10.1007/BF02291170

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. http://www.stats.ox.ac.uk/pub/MASS4/

Yuan, K.-H., Tian, Y., & Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika, 80*(2), 379–405. https://doi.org/10.1007/s11336-013-9386-5

Zhang, Z., Hamaker, E. L., & Nesselroade, J. R. (2008). Comparisons of four methods for estimating a dynamic factor model. *Structural Equation Modeling, 15*(3), 377–402. https://doi.org/10.1080/10705510802154281

History
Received April 30, 2021
Revision received October 9, 2021
Accepted November 2, 2021
Published online February 2, 2022

Funding
Open access publication enabled by Bielefeld University.

ORCID
Julia Norget
https://orcid.org/0000-0002-3388-8873

Julia Norget
Faculty of Psychology and Sport Science
Bielefeld University
Universitätsstraße 25
33615 Bielefeld
Germany
julia.norget@uni-bielefeld.de

Original Article

Analyzing Data of a Multilab Replication Project With Individual Participant Data Meta-Analysis

A Tutorial

Robbie C. M. van Aert

Department of Methodology and Statistics, Tilburg University, The Netherlands

Abstract: Multilab replication projects such as Registered Replication Reports (RRR) and Many Labs projects are used to replicate an effect in different labs. Data of these projects are usually analyzed using conventional meta-analysis methods. This is certainly not the best approach because it does not make optimal use of the available data as a summary rather than participant data are analyzed. I propose to analyze data of multilab replication projects with individual participant data (IPD) meta-analysis where the participant data are analyzed directly. The prominent advantages of IPD meta-analysis are that it generally has larger statistical power to detect moderator effects and allows drawing conclusions at the participant and lab level. However, a disadvantage is that IPD meta-analysis is more complex than conventional meta-analysis. In this tutorial, I illustrate IPD meta-analysis using the RRR by McCarthy and colleagues, and I provide R code and recommendations to facilitate researchers to apply these methods.

Keywords: meta-analysis, registered replication report, replication, multilevel analysis, individual participant data meta-analysis

Multilab replication projects are exemplary for the increased attention for replication research in psychology. Prominent effects in the psychological literature are replicated in these multilab replication projects in different labs across the world. These projects yield highly relevant insights about whether an effect can actually be replicated and also whether the effect depends on contextual factors such as the location where a study was conducted. Multiple registered replication reports (RRRs; Simons et al., 2014) have been conducted where a single effect is replicated in different labs as well as Many Labs projects (Ebersole et al., 2016, 2020; Klein et al., 2014, 2018, 2021) where multiple effects are replicated in a large collaborative project.

The main publication outlet for multilab replication projects within psychology was the journal *Perspectives on Psychological Science*, but *Advances in Methods and Practices in Psychological Science* has taken over this role since its launch in 2018. Twelve RRRs were published in these journals since the introduction of RRRs and until September 6, 2021. Moreover, the Many Labs projects replicated 12, 28, 10, 1, and 10 effects in Many Labs 1, 2, 3, 4, and 5, respectively. These published RRRs and Many Labs projects show that multilab replication projects are not uncommon, and these projects are expected to become more popular due to the increased attention for replications and the desire to study the credibility of psychological science.

The usual analysis strategy for analyzing the data of a single effect in multilab replication projects is equivalent to how a conventional meta-analysis is conducted. That is, a summary effect size (e.g., [standardized] mean difference or correlation) and corresponding sampling variance (i.e., squared standard error) is computed for each lab and these summary effect sizes are then usually synthesized by means of a random-effects meta-analysis. The meta-analytic average effect size is of interest as well as whether the true effect size of the labs is heterogeneous and whether this heterogeneity can be explained by moderator variables in a so-called meta-regression model (e.g., Thompson & Sharp, 1999; Van Houwelingen et al., 2002). This is a valid but certainly also suboptimal approach because the differences of participants within a lab are lost by aggregating the data to summary effect sizes. I propose analyzing data of multilab replication projects through an individual participant data (IPD) meta-analysis where the participant data are analyzed rather than summary effect sizes (e.g., L. A. Stewart & Tierney, 2002). Multilab replication projects are ideal for applying IPD meta-analysis as the participants' data is, in contrast to traditional studies, readily available.

IPD meta-analysis is popular among medical researchers, and it is commonly referred to as individual *patient* data meta-analysis. In contrast to research in psychology,

https://doi.org/10.1027/2151-2604/a000483

medical research has a long history with respect to sharing data that enables researchers to conduct IPD meta-analysis. For example, the prominent medical journal BMJ required authors to agree on sharing the IPD data of clinical trials of drugs or devices on request in 2013, and this policy was extended to all trials in 2015 (Godlee, 2012; Loder & Groves, 2015). Medical research also frequently uses binary data (e.g., dead vs. alive and treatment vs. placebo group), and these data can easily be reported in a 2 × 2 frequency table, making reporting of IPD data less cumbersome compared to fields like psychology that mainly use continuous data. These developments together with the call for more personalized treatments (Hingorani et al., 2013) made that IPD meta-analysis is nowadays seen as the gold standard for synthesizing studies in medical research (Riley et al., 2008; Rogozińska et al., 2017; Simmonds et al., 2005).

IPD meta-analysis has many advantages over conventional meta-analysis (Riley et al., 2010; L. A. Stewart & Tierney, 2002). Two advantages are especially valuable for analyzing data of multilab replication projects. First, participant-level moderators can be included to explain heterogeneity in true effect size, which is one of the main aims of multilab replication projects. Heterogeneity in the conventional meta-analysis can only be attributed to study level characteristics and not to characteristics of the participants within a lab because summary statistics of the primary studies are analyzed rather than the underlying participant data. Researchers who draw conclusions at the participant level using summary effect sizes may introduce aggregation bias and commit an ecological fallacy (e.g., Berlin et al., 2002; Borenstein et al., 2009), which will be illustrated below. Second, statistical power to test moderating effects is usually larger than of conventional meta-regression. Simmonds and Higgins (2007) analytically showed that the statistical power of testing a moderator variable in IPD meta-analysis is always larger than of conventional meta-regression in a fixed-effect meta-analysis (aka equal-effect) model. The only exception is when all participant scores on the moderator variable within primary studies are the same because the statistical power of conventional meta-regression and IPD meta-analysis is equivalent in this situation. Lambert and colleagues (2002) compared statistical power of IPD meta-analysis with conventional meta-regression in a fixed-effect meta-analysis model using simulations and showed that statistical power of IPD meta-analysis was especially larger when the effect size, number of primary studies, and sample size in the primary studies was small.

The goal of this paper is to illustrate how data of a multilab replication project can be analyzed through an IPD meta-analysis. The focus of this paper will be on the estimation of the average effect size as well as on quantifying the heterogeneity in true effect size and explaining this heterogeneity with moderator variables because both aspects are generally studied in multilab replication projects (e.g., Ebersole et al., 2016; Klein et al., 2014, 2018). Two different approaches to IPD meta-analysis are a one-stage and two-stage approach that I will both explain and illustrate. Before turning to IPD meta-analysis, I will first provide an example of aggregation bias in a meta-regression model. Subsequently, I will introduce the RRR by McCarthy and colleagues (2018) that will illustrate the methods and explain how these data are commonly analyzed using conventional random-effects meta-analysis. The paper ends with a conclusion section that contains recommendations for analyzing data of a multilab replication project.

Illustration of Aggregation Bias in Meta-Regression

Aggregation bias or an ecological fallacy refers to a situation where conclusions are drawn for individuals based on aggregated data (Robinson, 1950). Meta-analysts can easily fall into the trap of introducing aggregation bias if they do not realize that differences between labs in a meta-regression analysis can only be attributed to lab level characteristics (e.g., Berlin et al., 2002; Borenstein et al., 2009). Figure 1A shows data of three labs using a two-independent groups design where scores of participants in the experimental and control group are denoted by E and C, respectively. The main interest in this analysis is to study whether age has a moderating effect on the grouping variable, so whether the effect of the manipulation is strengthened (or weakened) by the participant's age.

The model underlying the data of all three labs is a linear regression model. That is, for lab 1: $51 - 18x + x \times \text{age}$, for lab 2: $46 - 30x + x \times \text{age}$, and for lab 3: $41 - 42x + x \times \text{age}$, where x denotes whether a participant belongs to the experimental ($x = 1$) or control ($x = 0$) group and *age* is the participant's age. Within each lab, the age of participants in the experimental group is larger than that of the participants in the control group. This may occur in practice if participants are not randomly assigned to one of the two groups. The regression equations show that the only differences between the labs are the intercept and the effect of the manipulation. These data indicate that there is a positive interaction effect between the grouping variable and age at the participant level, so the effect of the manipulation is strengthened by the participant's age.

Table 1 shows the summary statistics that are used as input for the meta-regression analysis. The focus in the meta-regression analysis is on the relationship between the raw *mean* difference of the experimental and control group and the lab's *mean* age. This implies that we are no longer allowed to draw conclusions at the participant level

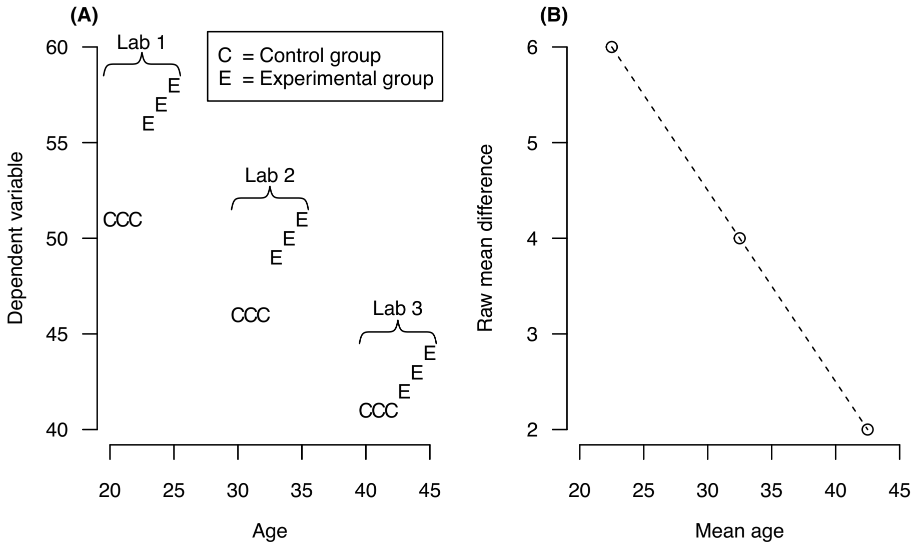


Figure 1. Artificial example to illustrate aggregation bias in the context of meta-regression analysis. (A) Individual participant data; (B) Data analyzed in the meta-regression analysis.

as we are analyzing summary statistics of the labs. Figure 1B shows the raw mean difference and mean age per lab. The relationship between the raw mean difference and mean age is negative (dashed line in Figure 1B) and contradicts the finding of the analysis based on the participant data.

This example illustrates that the interaction effect may be substantially different at the lab compared to the participant level. The effect at a higher level can be in the opposite direction compared to the lower level (Aitkin & Longford, 1986; Snijders & Bosker, 1999). Although this example was created in a way to illustrate aggregation bias, it may also occur in practice and can only be studied if participant data are available. Hence, this example also shows that a meta-regression cannot be used to draw conclusions at the participant level as it is prone to committing an ecological fallacy. A meta-regression is, however, suitable to draw conclusions about moderating effects measured at the level of the lab. This implies that the results of the meta-regression in this example can be used to draw conclusions about the lab's mean age on the raw mean difference.

Table 1. Sample means of the dependent variable in the experimental and control group and the moderator age. Raw mean difference is the raw mean difference of the sample means in the experimental and control group

	Sample means			Raw mean
	Experimental	Control	Age	difference
Lab 1	57	51	22.5	6
Lab 2	50	46	32.5	4
Lab 3	43	41	42.5	2

Example of a Registered Replication Report

The RRR by McCarthy and colleagues (2018) replicated the study by Srull and Wyer (1979) on assimilative priming. Assimilative priming refers to the idea that "exposure to priming stimuli causes subsequent judgments to incorporate more of the qualities of the primed construct" (McCarthy et al., 2018, p. 322). In the replicated experiment, participants were first asked to perform a sentence construction task where either 20% or 80% of the sentences described hostile behavior. Participants were then asked to read a vignette about a man called Donald who behaved in an ambiguously hostile way and rated Donald's behavior on 12 traits to get a score of the extent he was perceived as hostile. All 12 traits were measured on a scale ranging from 0 (= *not at all*) to 10 (= *extremely*), and six of these traits were averaged to create a hostility rating. The tested hypothesis was that participants who were exposed to a larger number of sentences describing hostile behavior would rate Donald's behavior as more hostile.

The RRR by McCarthy and colleagues (2018) was selected for illustrating the different meta-analysis models because the data are well-documented, it was possible to reproduce the reported results, variables were reported that could be included in the models as moderator, and two-independent groups design was used, which is common in psychology. The effect size measure of interest was, as by McCarthy and colleagues (2018), the raw mean difference. The raw mean difference is a common effect size measure in multilab replication projects because the dependent variable is measured in the same way in each lab. Hence, computing standardized mean differences is not necessary and even undesired if the data can be analyzed on its original (unstandardized) scale (e.g., Baguley, 2009; Bond et al., 2003; Wilkinson, 1999). The study was replicated in 22 labs and the total sample size was 7,373 (see McCarthy et al., 2018 for more details). All analyses were conducted in the statistical software R (Version 4.1.0, R Core Team, 2021), the R package papaja (Aust & Barth, 2020) was used for writing the article, and annotated R code to analyze the RRR is available in the supplemental materials at the Open Science Framework (OSF; Van Aert, 2019a: https://osf.io/c9zep/).

Random-Effects Model

The conventional random-effects model is usually fitted to data of multilab replication projects, and this is also how the data of the RRR by McCarthy and colleagues (2018) were analyzed. A requirement for applying the random-effects model is that summary effect sizes and corresponding sampling variances per lab are computed. Formulas for computing these summary effect sizes and sampling variances are available in Borenstein and Hedges (2019). I will continue by describing the random-effects model before applying this model to the RRR.

Statistical Model

The random-effects model assumes that the effect size y_i is observed for each ith lab. The statistical model can be written as (e.g., Borenstein et al., 2009)

$$y_i = \mu + \mu_i + \varepsilon_i, \qquad (1)$$

where μ is the average true effect size, μ_i is the random effect denoting the difference between the average true effect size μ and a lab's true effect size θ_i, and e_i reflects the sampling error. The random effect μ_i is commonly assumed to follow a normal distribution with mean zero and variance τ^2, and the sampling error e_i is assumed to follow a normal distribution with mean zero and variance σ_i^2. The μ_i and e_i are assumed to be mutually independent of each other, and it is common practice to estimate σ_i^2 and then assume that its value is known.

The most interesting outcomes in a multilab replication project are the parameters μ and τ^2. The parameter μ denotes the meta-analytic average effect size estimate yielding insight into the true effect size of the replicated study and can also be used to assess whether the original study can be deemed to be successfully replicated. The parameter τ^2 reflects the between-study variance in true effect size and indicates whether the lab's true effect sizes θ_i are all the same (homogeneous) or different from each other (heterogeneous). Heterogeneity in true effect size can be explained by extending the statistical model in (1) to a random-effects meta-regression model where study characteristics are included as moderators (e.g., Thompson & Sharp, 1999; Van Houwelingen et al., 2002). That is, a lab's true effect size becomes a regression equation in a random-effects meta-regression model (e.g., $\beta_0 + \beta_1 x$ where x is a moderator variable).

Fitting the Random-Effects Model to the Data

Before fitting the random-effects model to the RRR, I first computed the raw mean differences and corresponding sampling variances for each lab (see Van Aert, 2019a: https://osf.io/c9zep/). I used the R package `metafor` (Version 3.0.2, Viechtbauer, 2010) for fitting the random-effects model. The random-effects model can be fitted using the `rma()` function of the `metafor` package by providing the lab's raw mean differences (argument `yi`) and the corresponding sampling variances (argument `vi`). R code for fitting the random-effects model is[1]

```
rma(yi = yi, vi = vi, data = ma_dat)
```

where `ma_dat` is a data frame containing the `yi` and `vi`.

The results of fitting the random-effects model are presented in the first row of Table 2. These results exactly match those of Figure 1 in McCarthy and colleagues (2018). The average true effect size estimate is equal to $\hat{\mu}$ = 0.083 (95% confidence interval (CI) [0.004; 0.161]), and the null-hypothesis of no effect was rejected (z = 2.058, two-tailed p = .040). These results imply that the average raw mean difference between the mean hostility rating of participants in the 80%-hostile priming condition and those in the 20%-hostile priming condition was 0.083. Hence, the mean hostility rating of participants in the 80%-hostile priming conditions was larger than those in the 20%-hostile priming condition. There was a small amount of heterogeneity observed in the true effect sizes. The estimate of the between-study variance $\hat{\tau}^2$ = 0.006 (95% CI [0; 0.043]),[2] Cochran's Q-test (Cochran, 1954) for testing

[1] The restricted maximum likelihood estimator (Raudenbush, 2009) was used for estimating the between-study variance τ^2. This is the default estimator of `metafor` and also allows direct comparison with the results of IPD meta-analysis as these also rely on restricted maximum likelihood estimation.

[2] The 95% CI for the between-study variance τ^2 is not in the output of fitting the random-effects model. Such a CI can, for instance, be obtained using the Q-profile method (Viechtbauer, 2007) via the function `confint()` where the only argument of the function is the object obtained by running the function `rma()`. See the supplemental materials for the actual code and output at https://osf.io/c9zep/ (Van Aert, 2019a).

Table 2. Results of fitting a random-effects model (RE MA) and two-stage and one-stage individual participant data meta-analysis to the registered replication report by McCarthy and colleagues (2018)

	$\hat{\mu}$ (*SE*)	(95% CI)	Test H_0: $\mu = 0$	$\hat{\tau}^2$	(95% CI)	Test H_0: $\tau^2 = 0$
RE MA	0.083 (0.040)	(0.004; 0.161)	$z = 2.058$, $p = .040$	0.006	(0; 0.043)	$Q(21) = 25.313$, $p = .234$
Two-stage	0.082 (0.040)	(0.004; 0.161)	$z = 2.055$, $p = .040$	0.006	(0; 0.043)	$Q(21) = 25.266$, $p = .236$
One-stage	0.090 (0.038)	(0.017; 0.164)	$t(18.6) = 2.356$, $p = .030$	0.002	(0; 0.012)	$\chi^2(2) = 0.554$, $p = .758$[a]

Note. $\hat{\mu}$ = estimate of the average true effect size; *SE* = standard error; CI = confidence interval; $\hat{\tau}^2$ is the estimate of the between-study variance obtained with restricted maximum likelihood estimation. [a]The `anova()` function conducts the likelihood-ratio test by first fitting the models to be compared with full maximum likelihood estimation.

the null-hypothesis of no between-study variance was not statistically significant, $Q(21) = 25.313$, $p = .234$.

The null-hypothesis of no heterogeneity could not be rejected, which is common for multilab replication projects that consist of direct replications (Olsson-Collentine et al., 2020). However, the estimated small between-study variance suggested that a small amount of heterogeneity in the true effect size was present in the meta-analysis. This heterogeneity can be explained by including moderators measured at the lab level in a random-effects meta-regression analysis. The moderator variable mean age of participants per lab is included in this paper for illustrating the methods, but the procedure is similar for any moderator variable. After computing this mean age per lab, the random-effects meta-regression model can be fitted to the data using the following code

```
rma(yi = yi, vi = vi, mods = ~ m_age,
         data = ma_dat)
```

where `mods = ~ m_age` indicates that mean age of participants per lab is included as moderator.

The results of fitting the random-effects meta-regression model are shown in the first two rows of Table 3.[3] The coefficient of the variable mean age is 0.050 ($z = 1.237$, two-tailed $p = .216$, 95% CI [−0.029; 0.128]) implying that a one unit increase in *mean* age leads to a predicted increase of 0.050 in the average raw mean difference. The estimate of the residual between-study variance was $\hat{\tau}^2 = 0.005$ (95% CI [0; 0.043], $Q(20) = 23.456$, $p = .267$). These results of fitting the random-effects model and random-effects meta-regression model will be contrasted with the results of IPD meta-analysis when describing those results.

Individual Participant Data Meta-Analysis

Meta-analysis models can be seen as a special case of multilevel models (also known as mixed-effects models) with at level 1 the participants within studies and at level 2 the studies. This is also the reason why meta-analysis models are discussed in books on multilevel models (e.g., Hox et al., 2018). This equivalence between meta-analysis and multilevel models becomes even more apparent when we move from the conventional random-effects model analyzing summary effect sizes to IPD meta-analysis analyzing the participants' data directly because IPD meta-analysis models are actually multilevel models applied to participants who are nested in studies.

Two different approaches to IPD meta-analysis are common: the one-stage and two-stage approach. In the two-stage approach, effect sizes are first computed for each lab and these are subsequently meta-analyzed. The one-stage approach does not require the computation of effect sizes per lab because the data are modeled directly using a multilevel model. Both approaches allow drawing inferences regarding moderator variables at the participant level in contrast to the meta-regression model. Moreover, both approaches generally yield similar (average) effect size estimates (e.g., Koopman et al., 2008; G. B. Stewart et al., 2012; Tierney et al., 2020; Tudur Smith & Williamson, 2007), but larger practically relevant differences can also be observed (Tudur Smith et al., 2016).

The two-stage approach appeals to researchers familiar with conventional meta-analysis models due to the close similarities between the two. One of the conventional meta-analysis models (i.e., the fixed-effect or random-effects model) is fitted in the second step of the two-stage approach. However, the differences between the conventional and two-stage IPD meta-analysis model also offers opportunities to gain better insights. Additional variables can be included in the first step of the two-stage approach to control for these variables, which is impossible in the conventional meta-analysis model. The most important difference is that analyzing the participant data in the first step of the two-step approach allows drawing inferences at the *participant* level. The conventional meta-analysis model uses summary statistics per lab for studying the effect of

[3] The intercept of this random-effects meta-regression model refers to the average true effect size estimate conditional on a mean age of zero. If the intercept is of interest to the meta-analyst, it is advised to center the variable mean age at, for instance, the grand mean (i.e., the overall mean of age) to increase the interpretability. The intercept can then be interpreted as the average true effect size estimate conditional on a mean age equal to the grand mean of age.

moderators and therefore only allows for drawing inferences at the *lab* level.

Despite these appealing properties of two-stage IPD meta-analysis, there are reasons for applying a one-stage rather than a two-stage IPD meta-analysis approach. For example, the two-stage approach has lower statistical power except for situations where all labs have the same mean on the moderator variable (Fisher et al., 2011; Simmonds & Higgins, 2007). Furthermore, the one-stage approach is also more flexible and does not require the assumption of known sampling variances σ_i^2 (Papadimitropoulou et al., 2019). This approach is, however, also more complicated to implement as convergence problems may arise in the one-stage approach, whereas these are less common in the two-stage approach (Kontopantelis, 2018).

I generally recommend applying one-stage IPD meta-analysis, but the two-stage approach is a useful "stepping stone" to move from the random-effects meta-analysis model to a one-stage IPD meta-analysis. Hence, I continue with describing two-stage IPD meta-analysis before illustrating one-stage IPD meta-analysis.

Statistical Model Two-Stage Approach

The first step of the two-stage approach consists of fitting a linear regression model to the participant data of each *i*th lab. In case of raw mean differences, the linear regression model is (e.g., Riley et al., 2008)

$$y_{ij} = \phi_i + \theta_i x_{ij} + \varepsilon_{ij}, \tag{2}$$

where y_{ij} denotes the score on the dependent variable of participant j in lab i, ϕ_i is a fixed lab effect, x_{ij} is a dummy variable indicating whether participant j in lab i belongs to the experimental or control group, and ε_{ij} is the sampling error of participant j in lab i. The same assumptions as for the random-effects model apply, so $\theta_i \sim N(\mu, \tau^2)$, $\varepsilon_{ij} \sim N(0, \sigma_i^2)$, and θ_i and ε_i are assumed to be mutually independent. There is no heterogeneity between labs if all θ_i are equal, and the parameters μ and τ^2 are again the main parameters of interest as these indicate the average treatment effect and the between-study variance in true effect size.

The linear regression model in (2) is fitted to the data of each *i*th lab in order to get an estimate of the raw mean difference ($\hat{\theta}_i$) and corresponding sampling variance. In the second step of the two-stage approach, these mean differences $\hat{\theta}_i$ are combined using the random-effects model in statistical model (1). That is, a conventional random-effects model is fitted using as input $\hat{\theta}_i$ as effect size estimate and $\text{Var}[\hat{\theta}_i]$ as sampling variance for each study.

The effect of moderator variables in a two-stage IPD meta-analysis is studied by adding interactions between the moderators and the grouping variable x_{ij} to the linear regression model described in (2). In case of one moderator variable, the linear regression model fitted to the data of each *i*th lab is (e.g., Riley et al., 2008)

$$y_{ij} = \phi_i + \alpha_i w_{ij} + \theta_i x_{ij} + \gamma_i w_{ij} x_{ij} + \varepsilon_{ij}, \tag{3}$$

where α_i is the predicted change in the dependent variable for participants in the control group if the moderator variable w_{ij} increases with one unit and γ_i denotes the interaction effect of moderator w_{ij} with the grouping variable x_{ij}. Inclusion of the main effect of the moderator variable is especially beneficial if participants were not randomly assigned to either the experimental or control group because it controls for differences between these groups.

Estimates of γ_i and the corresponding sampling variances have to be stored for each *i*th lab if moderator effects are studied in the two-stage approach. The second step when estimating moderator effects is equivalent to the second step when estimating the average true effect except that now the random-effects model in (1) is fitted to the γ_i. This two-stage approach is also called a "meta-analysis of interactions" since moderator effects are now meta-analyzed (Simmonds & Higgins, 2007).

Applying the Two-Stage Approach to the Data

A linear regression model can be fitted to the participant data of each *i*th lab by using the function `lm()` in the preloaded R package `stats` (R Core Team, 2021). The `lm()` function requires as argument the regression equation in so-called formula notation. The linear regression model in (2) can be fitted using the code

```
lm(y ~ x)
```

where `y ~ x` denotes that a linear regression model is fitted with dependent variable `y` and independent variable `x`. The variables `y` and `x` refer to y_{ij} and x_{ij} of the *i*th lab in the linear regression model (2). This R code has to be executed per lab, and the regression coefficient of variable x_{ij} and its sampling variance has to be stored for each lab. The supplemental materials at https://osf.io/c9zep/ provide code for extracting this information from the output in R (Van Aert, 2019a).

R code of the second step is highly similar to the code for fitting the random-effects model,

```
rma(yi = thetai_hat, vi = vi_thetai_hat,
             data = ma_dat)
```

where `thetai_hat` is the regression coefficient of variable x_{ij} and `vi_thetai_hat` is the corresponding sampling variance.

The results of the two-stage IPD meta-analysis are presented in the second row of Table 2. These results were highly similar to the ones of the random-effects model

Table 3. Results of fitting a random-effects meta-regression model (RE MR) and two-stage and one-stage individual participant data meta-analysis where age is included as a moderator variable to data of the registered replication report by McCarthy and colleagues (2018)

	Estimate (*SE*)	(95% CI)	Test of no effect	$\hat{\tau}^2$	(95% CI)	Test H_0: $\tau^2 = 0$
RE MR				0.005	(0; 0.043)	$Q(20) = 23.456$, $p = .267$
Intercept	−0.921 (0.812)	(−2.512; 0.671)	$z = -1.134$, $p = .257$			
Mean age	0.050 (0.040)	(−0.029; 0.128)	$z = 1.237$, $p = .216$			
Two-stage				0.000	(0; 0.011)	$Q(21) = 18.006$, $p = .649$
Age	0.053 (0.024)	(0.007; 0.100)	$z = 2.238$, $p = .025$			
One-stage				0.003	(0; 0.011)	$\chi^2(2) = 0.355$, $p = .837$[a]
Intercept	8.264 (0.353)	(7.570; 8.951)	$t(1{,}701.0) = 23.420$, $p < .001$			
x	−0.791 (0.814)	(−2.308; 0.820)	$t(18.8) = -0.972$, $p = .343$			
Age	−0.064 (0.017)	(−0.096; −0.030)	$t(4{,}477.1) = -3.780$, $p < .001$			
Age within	0.050 (0.024)	(0.003; 0.096)	$t(5{,}331.4) = 2.074$, $p = .038$			
Age between	0.044 (0.040)	(−0.036; 0.118)	$t(18.8) = 1.087$, $p = .291$			

Note. SE = standard error; CI = confidence interval; $\hat{\tau}^2$ = estimate of the between-study variance obtained with restricted maximum likelihood estimation. "x" is a dummy variable that determines whether a participant is in the control (= reference category) or experimental group, "Age within" is the within-lab interaction between age and "x," and "Age between" is the between-lab interaction between age and "x." [a]The `anova()` function conducts the likelihood-ratio test by first fitting the models to be compared with full maximum likelihood estimation.

fitted to the summary effect sizes. The average true effect size estimate slightly decreased ($\hat{\mu}$ = 0.082, 95% CI [0.004; 0.161]), but was still statistically significant (z = 2.055, two-tailed p = .040). The estimate of the between-study variance remained the same ($\hat{\tau}^2$ = 0.006, 95% CI [0; 0.043]) and was not statistically significant, $Q(21)$ = 25.266 with p = .236.

The linear regression model in (3) has to be fitted in the first step of a two-stage IPD meta-analysis in order to study whether age has a moderating effect on the dependent variable. This can be done by using the `lm()` function,

```
lm(y ~ x + age + x:age)
```

where `age` is the age of participant j in lab i and `x:age` denotes the interaction effect between the grouping variable and the moderating variable age. After storing the estimated coefficient of the interaction effect and its sampling variance, the random-effects model can be fitted analogous to how we fitted this model for the two-stage IPD meta-analysis for the lab's estimated treatment effect $\hat{\theta}_i$,

```
rma(yi = gammai, vi = vi_gammai,
    data = ma_dat)
```

where `gammai` and `vi_gammai` are the estimated coefficient of the interaction effect and corresponding sampling variance, respectively.

The results of the two-stage IPD meta-analysis are presented in the third row of Table 3. The coefficient of the variable age was slightly larger than the coefficient of the variable mean age obtained with the random-effects meta-regression model (0.050 vs. 0.053), which suggested that the effects at the participant and lab level were comparable. The variable age was statistically significant in the two-stage IPD meta-analysis (z = 2.238, two-tailed p = .025). This indicates that the effect of assimilative priming on the hostility rating was moderated by age. The between-study variance of the true effects of the interaction was estimated as $\hat{\tau}^2$ = 0, and the null-hypothesis of no heterogeneity was not rejected, $Q(21)$ = 18.006 with p = .649.

Statistical Model One-Stage Approach

The linear regression model in (2) is fitted in a single analysis using a multilevel model in one-stage IPD meta-analysis. A controversial modeling decision is whether the effects of the labs (parameter ϕ_i in linear regression model (2)) have to be treated as fixed or random effects (Brown & Prescott, 2015; Higgins et al., 2001). Fixed effects imply that separate intercepts are estimated for each lab, so the number of parameters increases if the number of labs increase. This makes the model not parsimonious, and its results can be difficult to interpret. Treating the effects as fixed implies that inferences can only be drawn for the included effects. Treating the effects as random implies the assumption that the effects are a random sample from a population of effects. Random effects allow, in contrast to fixed effects, researchers to generalize the results to the population effects. This is the reason why including the lab's effects as random effects has been argued as more appropriate than fixed lab's effects (Schmid et al., 2004). However, estimation of the variance of the population of effects may be difficult in the case of a small number of labs (Brown & Prescott, 2015), so random effects may still be incorporated as fixed parameters in the model to avoid imprecise estimation of this variance. Another solution is to fit this model in a Bayesian framework where prior information about the variance of the population effects can be incorporated (e.g., Chung et al., 2013).

The linear regression model in (3) can be fitted in a single analysis to include moderator variables in a one-stage IPD meta-analysis approach. However, the within and between-lab interaction between the grouping and moderating variable are not disentangled by fitting this model. A better approach that disentangles the within and between lab interaction is to fit the linear regression model (Riley et al., 2008)

$$y_{ij} = \phi_i + \alpha_i w_{ij} + \theta_i x_{ij} + \gamma_{\mathrm{W}} x_{ij}\left(w_{ij} - m_i\right) + \gamma_{\mathrm{B}} x_{ij} m_i + \varepsilon_{ij}, \tag{4}$$

where m_i is the mean of the moderator of the ith lab and γ_{W} and γ_{B} is the within and between-lab interaction between the moderating and grouping variable. The term $\gamma_{\mathrm{W}} x_{ij}(w_{ij} - m_i)$ is the interaction effect of the grouping variable and the moderator variable minus the ith lab's mean of the moderating variable. This is known as group-mean centering in the literature on multilevel modeling (e.g., Enders & Tofighi, 2007). Also including the interaction between the grouping variable and the lab mean in the model (i.e., $\gamma_{\mathrm{B}} x_{ij} m_i$) allows for disentangling the within and between-lab interaction of the grouping and moderator variable.

Applying the One-Stage Approach to the Data

The one-stage IPD meta-analysis model can be fitted to the data by using the R package `lme4` (Version 1.1.27.1, Bates et al., 2015) and the R package `lmerTest` (Version 3.1.3, Kuznetsova et al., 2017) has to be loaded to get p-values for hypothesis tests of fixed effects.[4] I show how to fit the one-stage IPD meta-analysis model with random effects for lab's effects in the paper, but R code for fitting the model with fixed effects as lab's effects is available in the supplemental material at https://osf.io/c9zep/ (Van Aert, 2019a).[5]

The statistical model in (2) can be fitted with random lab effects using the R code

```
lmer(y ~ x + (x | lab), data = ipd_dat)
```

where `ipd_dat` is a data frame containing the variables that are included in this model. Random effects are specified in the `lmer()` function by including terms between brackets. Here `(x | lab)` indicates that a model is fitted with a random intercept for lab and a random slope for the treatment effect that is allowed to be correlated.

The results of fitting one-stage IPD meta-analysis to the data are shown in the last row of Table 2. The results are similar to the ones obtained with the random-effects model and two-stage IPD meta-analysis. The average effect size estimate is $\hat{\mu}$ = 0.090 (95% CI [0.017; 0.164]), and this effect size is significantly different from zero, t(18.6) = 2.356, two-tailed p = .030. The estimate of the between-study variance was close to zero ($\hat{\tau}^2$ = 0.002) and not statistically significant, $\chi^2(2)$ = 0.554, p = .758. The correlation between the intercepts and slopes of the labs was equal to 0.591, so labs with a larger hostility rating in the control group also showed a larger effect of assimilative priming.

The statistical model in (4) to study the interaction effect between age and the grouping variable can also be fitted with the `lmer()` function. The following R code fits the model

```
lmer(y ~ x + (x | lab) + age + I(age-
age_gm):x + age_gm:x, data = ipd_dat)
```

where `I(age-age_gm):x` is the interaction effect between the grouping variable and the group-mean centered age variable, and `age_gm:x` is the interaction effect between the mean age per lab and the grouping variable.

The results of one-stage IPD meta-analysis with age as moderating variable are included in the last rows of Table 3. Estimates of the intercept and the "x" are controlled for other variables in the model and reflect the estimated average score of participants in the control group and the estimated treatment effect. Estimates of the variables "Age within" and "Age between" are of particular interest as these indicate the interaction effect between the grouping variable and age within and between labs. There was a small positive interaction effect within labs $\hat{\gamma}_{\mathrm{W}}$ = 0.050

[4] There is debate about whether p-values should be reported in the context of multilevel models because it is currently unknown how the denominator degrees of freedom should be computed. I decided to explain how to obtain p-values and report those for the one-stage IPD meta-analysis as researchers have a strong desire to interpret and report p-values. However, it is important to realize that these p-values are based on approximate rather than exact denominator degrees of freedom. Luke (2017) showed by means of simulations that the default Satterthwaite approximation implemented in the R package `lmerTest` (Kuznetsova et al., 2017) adequately controlled Type-I error and had the comparable statistical power to other methods.

[5] I conducted a small Monte-Carlo simulation study to examine whether the estimate of the treatment effect, its standard error, and the estimate of the between-study variance were different for models with random and fixed effects as lab's effects. Data were generated using a procedure to stay as close as possible to the data of the RRR by McCarthy and colleagues (2018). That is, parameter estimates of the one-stage IPD meta-analysis with random effects for lab's effects were used for generating data, and the same number of labs as in the RRR was used. Sample sizes were based on the observed sample sizes in the labs, but these were also systematically varied as small sample sizes were expected to be favorable for fixed effects as lab's effects. Results were highly similar for the two different one-stage IPD meta-analysis models. Non-convergence occurred in approximately 50% of the iterations. For more details about this Monte-Carlo simulation study, R code, and all results see Van Aert, 2019b: https://osf.io/r5kqy/.

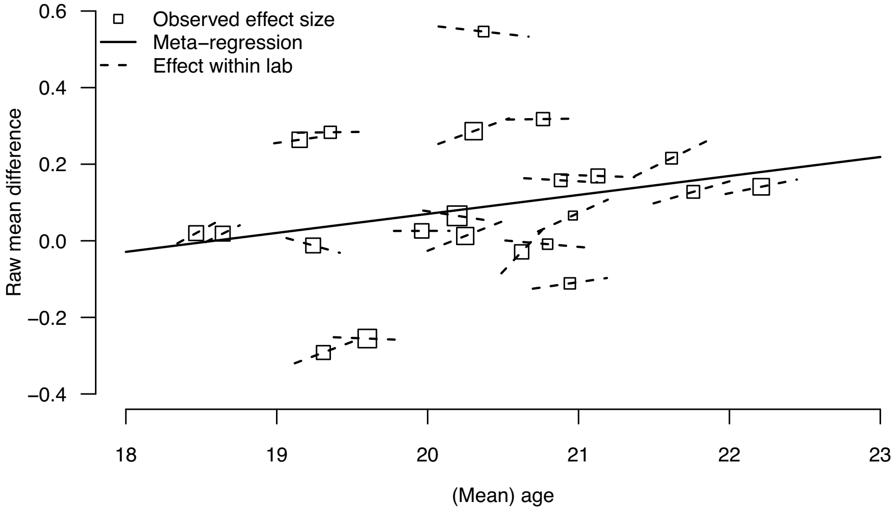


Figure 2. The effect of participant's age and mean age per lab on the raw mean difference in the RRR by McCarthy and colleagues (2018). Squares denote the observed effect sizes and mean age in the labs. The size of squares is proportional to the inverse of the standard error of the effect sizes. The solid line shows the estimated effect between labs based on the meta-regression. The dashed lines show the effect of age within lab obtained in the two-stage IPD meta-analysis (i.e., $\hat{\gamma}_i$ in model (3)). The length of the dashed lines is proportional to the standard deviation of age per lab.

(95% CI [0.003; 0.096], $t(5{,}331.4) = 2.074$, two-tailed $p = .038$), but not between labs $\hat{\gamma}_B = 0.044$ (95% CI [−0.036; 0.118], $t(18.8) = 1.087$, two-tailed $p = .291$). However, $\hat{\gamma}_W$ and $\hat{\gamma}_B$ were highly comparable, so there were no clear indications that the interaction effect was different between and within labs. Also, note the difference in degrees of freedom for testing these interaction effects that may cause a statistically significant effect within but not between labs. The between-study variance in lab's true effect size was negligible ($\hat{\tau}^2 = 0.003$) and not statistically significant, $\chi^2(2) = 0.355$, $p = .837$. The correlation between the intercepts and slopes of the labs was equal to 0.371.

Figure 2 provides an overview of the effect of (mean) age within and between labs. The solid line represents the relationship between labs that was estimated by the meta-regression model. Squares denote the observed effect size and mean age per lab, with the dashed line reflecting the effect of age within each lab that was obtained in the first step of the two-stage IPD meta-analysis. The slope of a dashed line illustrates to what extent the treatment effect within a lab is moderated by age. Although the slopes of the within lab effect differs across labs, this figure corroborates the results in Table 3 showing that the effect of (mean) age was not substantially different between and within labs.

Conclusion

Multilab replication projects are becoming more popular to examine whether an effect can be replicated and to what extent it depends on contextual factors. Data of these projects are commonly analyzed using lab's summary statistics by means of conventional meta-analysis methods. This is certainly a suboptimal approach because differences within a lab are lost. This paper illustrated a better approach for analyzing data of multilab replication projects using IPD meta-analysis.

IPD meta-analysis allows for distinguishing the effect at the participant and lab level in contrast to conventional meta-analysis models. An artificial example illustrated that drawing conclusions at the participant level using the conventional meta-regression model is not allowed and that it could lead to committing an ecological fallacy if it is done. Other advantages of IPD meta-analysis are larger statistical power for testing moderator effects than conventional meta-analysis (Lambert et al., 2002; Simmonds & Higgins, 2007) and more modeling flexibility. Applying one-stage and two-stage IPD meta-analysis to the RRR by McCarthy and colleagues (2018) did not alter the main conclusion that assimilative priming had a small but statistically significant effect on hostility ratings. An interesting finding obtained with IPD meta-analysis was that the moderating effect of age was present within but not between labs.

IPD meta-analysis was illustrated by using raw mean difference as effect size measure because this is a common effect size measure for multilab replication projects and it was used in the RRR of McCarthy and colleagues (2018). However, these models can also be applied for other effect size measures as, for example, the correlation coefficient and binary data (see for illustrations Pigott et al., 2012; Turner et al., 2000; Whitehead, 2002). In the case of the Pearson correlation coefficient, the independent and dependent variables need to be standardized before being included in a one-stage IPD meta-analysis. The one-stage IPD meta-analysis then returns an estimate of the average correlation because the regression coefficient of a standardized dependent variable regressed on a standardized independent variable equals a Pearson correlation coefficient. An IPD meta-analysis based on binary data is generally less cumbersome than for other effect size measures since participant data can be extracted from cell frequencies of contingency tables in a study.

I recommend analyzing data of any multilab replication project using one-stage IPD meta-analysis. One-stage IPD meta-analysis is preferred over two-stage IPD meta-analysis because it generally has larger statistical power (Fisher et al., 2011; Simmonds & Higgins, 2007) and has more modeling flexibility. For example, moderators at the first level (participant) and second level (lab) can be added as well as interaction effects between these moderators or an extra random effect can be added to take into account that labs are located in different countries. The model flexibility of a one-stage IPD meta-analysis can also be used to make different assumptions about the within-study residual variance. This residual variance was assumed to be the same in all control and experimental groups of the labs in the used one-stage IPD meta-analysis, but researchers may have theoretical reasons to impose a weaker assumption on the within-study residual variance. Another advantage of one-stage IPD meta-analysis is that it does not require specialized meta-analysis software in contrast to two-stage IPD meta-analysis and also conventional meta-analysis. Popular statistical software packages such as R, SPSS, Stata, and SAS all include functionality to fit multilevel models that can also be used for one-stage IPD meta-analysis.

A drawback of one-stage IPD meta-analysis is that it is more complex to implement compared to two-stage IPD and conventional meta-analysis. This increased complexity is caused by the modeling flexibility that requires researchers to carefully think about how to specify their model. This complexity of one-stage IPD meta-analysis is illustrated by Jackson and colleagues (2018), who identified six one-stage IPD meta-analysis models for synthesizing studies with odds ratio as effect size measure, and five of these models showed acceptable statistical properties. Hence, there is currently not a single one-stage IPD meta-analysis model, and future research is needed to assess what the best one-stage IPD meta-analysis models are. Another drawback of one-stage IPD meta-analysis is that convergence problems may arise. These problems may be solved by simplifying the random part of the model. For example, researchers may opt for one-stage IPD meta-analysis with fixed rather than random lab effects. Researchers may use two-stage IPD meta-analysis to analyze their data as a last resort if convergence problems of one-stage IPD meta-analysis cannot be resolved.

This paper and the proposed recommendations are in line with a recent article (McShane & Böckenholt, 2020) that advocated meta-analysts by means of a thought experiment to think about how they would analyze their data if they would possess the participant data rather than only the summary data. This thought experiment will motivate researchers to apply more advanced and appropriate meta-analysis models such as a three-level meta-analysis model (e.g., Konstantopoulos, 2011; Van den Noortgate & Onghena, 2003) when the nesting of studies in labs is, for instance, taken into account or multivariate meta-analysis where multiple outcomes are analyzed simultaneously (e.g., Hedges, 2019; Van Houwelingen et al., 2002). One-stage IPD meta-analysis is also ideally suited for fitting these more advanced meta-analysis models due to its modeling flexibility if the participant data are available.

Fitting IPD meta-analysis models to data in psychology and this tutorial paper, in particular, may become more relevant in the distant future when publishing participant data hopefully becomes the norm. However, IPD meta-analysis models can already be applied within psychology in other situations than multilab replication projects. For instance, meta-analyzing studies in a multistudy paper in a so-called internal meta-analysis (e.g., Cumming, 2008, 2012; Maner, 2014; McShane & Böckenholt, 2017) has increased in popularity (Ueno et al., 2016). The usual approach of an internal meta-analysis is to meta-analyze summary data, whereas analyzing the participant data by means of an IPD meta-analysis is a better alternative. There are, however, also rare cases where computing summary statistics based on IPD data is beneficial. In the case of Big Data, it may be unfeasible to analyze the IPD data directly because the data are too large to handle with a computer. A solution could be to analyze the data using a split/analyze/meta-analyze (SAM) approach where the data are (1) split into smaller chunks, (2) each chunk is analyzed separately, and (3) the results of the analysis of each chunk are combined using a meta-analysis (Cheung & Jak, 2016; Zhang et al., 2018). This approach is comparable to a two-stage IPD meta-analysis.

To conclude, the application of IPD meta-analysis methods to multilab replication projects has the potential to yield relevant insights that could not have been obtained by conventional meta-analysis methods. I hope that this paper creates awareness for IPD meta-analysis methods within the research field of psychology and enables researchers to apply these methods to their own data.

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: General, 149*(1), 1–43.

Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://github.com/crsh/papaja

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 48. https://doi.org/10.18637/jss.v067.i01

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., & Feldman, H. I. (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Statistics in Medicine, 21*(3), 371–387. https://doi.org/10.1002/sim.1023

Bond, C. F. Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8*(4), 406–418. https://doi.org/10.1037/1082-989X.8.4.406

Borenstein, M., & Hedges, L. V. (2019). Effect sizes for meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 207–244). Rusell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Brown, H., & Prescott, R. (2015). *Applied mixed models in medicine*. Wiley.

Cheung, M. W.-L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*(738), 1–13. https://doi.org/10.3389/fpsyg.2016.00738

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika, 78*(4), 685–709. https://doi.org/10.1007/s11336-013-9328-2

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101–129.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Change, R., & Nosek, B. A. (2016). Many Labs 3: Evaluating participant poor quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Change, R., & Nosek, B. A. (2020). Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science, 3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Fisher, D. J., Copas, A. J., Tierney, J. F., & Parmar, M. K. B. (2011). A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *Journal of Clinical Epidemiology, 64*(9), 949–967. https://doi.org/10.1016/j.jclinepi.2010.11.016

Godlee, F. (2012). Clinical trial data for all drugs in current use. *British Medical Jorunal, 345*, Article e7304. https://doi.org/10.1136/bmj.e7304

Hedges, L. V. (2019). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 281–297). Russell Sage Foundation.

Higgins, J. P. T., Whitehead, A., Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine, 20*(15), 2219–2241. https://doi.org/10.1002/sim.918

Hingorani, A. D., Van der Windt, D. A., Riley, R. D., Abrams, K. R., Moons, K. G. M., Steyerberg, E. W., Change, R., & Hemingway, H. (2013). Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *British Medical Journal, 346*, Article e5793. https://doi.org/10.1136/bmj.e5793

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications*. Routledge.

Jackson, D., Law, M., Stijnen, T., Viechtbauer, W., & White, I. R. (2018). A comparison of 7 random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in Medicine, 37*, 1059–1085. https://doi.org/10.1002/sim.7588

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., Change, R., & Ratliff, K. A. (2021). *Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement*. https://doi.org/10.31234/osf.io/vef2c

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., Change, R., & Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology, 45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Change, R., & Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2*(1), 61–76. https://doi.org/10.1002/jrsm.35

Kontopantelis, E. (2018). A comparison of one-stage vs two-stage individual patient data meta-analysis methods: A simulation study. *Research Synthesis Methods, 9*(3), 417–430. https://doi.org/10.1002/jrsm.1303

Koopman, L., Van der Heijden, G. J. M. G., Hoes, A. W., Grobbee, D. E., & Rovers, M. M. (2008). Empirical comparison of subgroup effects in conventional and individual patient data meta-analyses. *International Journal of Technology Assessment in Health Care, 24*(3), 358–361. https://doi.org/10.1017/S0266462308080471

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

Lambert, P. C., Sutton, A. J., Abrams, K. R., & Jones, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology, 55*(1), 86–94. https://doi.org/10.1016/S0895-4356(01)00414-0

Loder, E., & Groves, T. (2015). The *BMJ* requires data sharing on request for all trials. *British Medical Journal, 350*, Article h2373. https://doi.org/10.1136/bmj.h2373

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Maner, J. K. (2014). Let’s put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science, 9*(3), 343–351. https://doi.org/10.1177/1745691614528215

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Change, R., & Yıldız, E. (2018). Registered replication report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science, 1*(3), 321–336. https://doi.org/10.1177/2515245918777487

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research, 43*(6), 1048–1063. https://doi.org/10.1093/jcr/ucw085

McShane, B. B., & Böckenholt, U. (2020). Enriching meta-analytic models of summary data: A thought experiment and case study. *Advances in Methods and Practices in Psychological Science, 3*(1), 81–93. https://doi.org/10.1177/2515245919884304

Meijer, E., Simons, D. J., McCarthy, R. J., Verschuere, B., Jim, A., & Hoogesteyn, K. (2018). *Data, analysis scripts, and results for McCarthy et al., 2018*. https://osf.io/mcvt7/

Olsson-Collentine, A., Wicherts, J. M., & Van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin, 146*(10), 922–940. https://doi.org/10.1037/bul0000294

Papadimitropoulou, K., Stijnen, T., Dekkers, O. M., & Le Cessie, S. (2019). One-stage random effects meta-analysis using linear mixed models for aggregate continuous outcome data. *Research Synthesis Methods, 10*(3), 360–375. https://doi.org/10.1002/jrsm.1331

Pigott, T. D., Williams, R., & Polanin, J. (2012). Combining individual participant and aggregated data in a meta-analysis with correlational studies. *Research Synthesis Methods, 3*(4), 257–268. https://doi.org/10.1002/jrsm.1051

R Core Team. (2021). *R: A language and environment for statistical computing*. http://www.r-project.org/

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–315). Russell Sage Foundation.

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *British Medical Journal, 340*, Article c221. https://doi.org/10.1136/bmj.c221

Riley, R. D., Lambert, P. C., Staessen, J. A., Wang, J., Gueyffier, F., Thijs, L., & Boutitie, F. (2008). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine, 27*(11), 1870–1893. https://doi.org/10.1002/sim.3165

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351–357. https://doi.org/10.2307/2087176

Rogozińska, E., Marlin, N., Thangaratinam, S., Khan, K. S., & Zamora, J. (2017). Meta-analysis using individual participant data from randomised trials: Opportunities and limitations created by access to raw data. *Evidence Based Medicine, 22*(5), 157–162. https://doi.org/10.1136/ebmed-2017-110775

Schmid, C. H., Stark, P. C., Berlin, J. A., Landais, P., & Lau, J. (2004). Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *Journal of Clinical Epidemiology, 57*(7), 683–697. https://doi.org/10.1016/j.jclinepi.2003.12.001

Simmonds, M. C., & Higgins, J. P. T. (2007). Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Statistics in Medicine, 26*(15), 2982–2999. https://doi.org/10.1002/sim.2768

Simmonds, M. C., Higgins, J. P. T., Stewart, L. A., Tierney, J. F., Clarke, M. J., & Thompson, S. G. (2005). Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. *Clinical Trials, 2*(3), 209–217. https://doi.org/10.1191/1740774505cn087oa

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at Perspectives on Psychological Science. *Perspectives on Psychological Science, 9*(5), 552–555. https://doi.org/10.1177/1745691614543974

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications.

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology, 37*(10), 1660–1672. https://doi.org/10.1037/0022-3514.37.10.1660

Stewart, G. B., Altman, D. G., Askie, L. M., Duley, L., Simmonds, M. C., & Stewart, L. A. (2012). Statistical analysis of individual participant data meta-analyses: A comparison of methods and recommendations for practice. *PLoS One, 7*(10), Article e46042. https://doi.org/10.1371/journal.pone.0046042

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions, 25*(1), 76–97. https://doi.org/10.1177/0163278702025001006

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine, 18*(20), 2693–2708. https://doi.org/10.1002/(SICI)1097-0258(19991030)18:20%3C2693::AIDSIM235%3E3.0.CO;2-V

Tierney, J. F., Fisher, D. J., Burdett, S., Stewart, L. A., & Parmar, M. K. B. (2020). Comparison of aggregate and individual participant data approaches to meta-analysis of randomised trials: An observational study. *PLoS Medicine, 17*(1), Article e1003019. https://doi.org/10.1371/journal.pmed.1003019

Tudur Smith, C., Marcucci, M., Nolan, S. J., Iorio, A., Sudell, M., Riley, R., Change, R., & Williamson, P. R. (2016). Individual participant data meta-analyses compared with meta-analyses based on aggregate data. *Cochrane Database of Systematic Reviews*, (9), 1–56. https://doi.org/10.1002/14651858.MR000007.pub3

Tudur Smith, C., & Williamson, P. R. (2007). A comparison of methods for fixed effects meta-analysis of individual patient data with time to event outcomes. *Clinical Trials, 4*(6), 621–630. https://doi.org/10.1177/1740774507085276

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., & Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine, 19*(24), 3417–3432.

Ueno, T., Fastrich, G. M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology: General, 145*(5), 643–654. https://doi.org/10.1037/xge0000159

Van den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement, 63*(5), 765–790.

Van Aert, R. C. M. (2019a). *Supplemental materials to "Analyzing data of a multi-lab replication project with individual participant data meta-analysis: A tutorial"*. https://osf.io/c9zep/

Van Aert, R. C. M. (2019b). *Supplemental materials to "Analyzing data of a multi-lab replication project with individual participant data meta-analysis: A tutorial"*. https://osf.io/r5kqy/

Van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine, 21*(4), 589–624. https://doi.org/10.1002/sim.1040

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine, 26*(1), 37–52. https://doi.org/10.1002/sim.2514

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Whitehead, A. (2002). *Meta-analysis of controlled clinical trials*. Wiley.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Zhang, Y. E., Liu, S., Xu, S., Yang, M. M., & Zhang, J. (2018). Integrating the split/analyze/meta-analyze (SAM) approach and a multilevel framework to advance big data research in psychology: Guidelines and an empirical illustration via the human Resource management investment-firm performance relationship. *Zeitschrift für Psychologie, 226*(4), 274–283. https://doi.org/10.1027/2151-2604/a000345

History

Received January 18, 2021
Revision received September 13, 2021
Accepted November 9, 2021
Published online February 2, 2022

Acknowledgments

I thank Marcel van Assen, Marjan Bakker, and Jelte Wicherts for their valuable comments on an earlier draft of this paper.

Conflict of Interest

The author declares that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Data

The datasets analyzed in this paper are available on the Open Science Framework (OSF) on the project website of McCarthy and colleagues (2018) at https://osf.io/qegfd/ and also in Meijer and colleagues (2018) at https://osf.io/mcvt7/. Annotated R code used to analyze the data is available in the online supplemental materials available at https://osf.io/c9zep/ (Van Aert, 2019a). Details about the Monte-Carlo simulation study, R code, and all results are available at https://osf.io/r5kqy/ (Van Aert, 2019b).

Funding

This work was supported by the European Research Council under the grant number: 726361 (IMPROVE).

ORCID

Robbie C. M. van Aert
https://orcid.org/0000-0001-6187-0665

Robbie C. M. van Aert
Department of Methodology and Statistics
Tilburg University
PO Box 90153
5000 LE Tilburg
The Netherlands
r.c.m.vanaert@tilburguniversity.edu

Call for Papers

"COVID-19 and Coping With Future Crises: Perspectives of Educational and Developmental Psychology"

A Topical Issue of the *Zeitschrift für Psychologie*

Guest Editors: Marko Lüftenegger[1], Martin Daumiller[2], and Ingrid Schoon[3,4]

[1]Department of Developmental and Educational Psychology, University of Vienna, Austria
[2]Department of Psychology, University of Augsburg, Germany
[3]Berlin Social Science Center, Germany
[4]Institute of Education, University College London, University of London, UK

Focus of the Topical Issue, Aims, and Scope

COVID-19 has challenged societies and our educational systems in particular, with dramatic cuts into established practices and imposing challenging new requirements. As a consequence, vast differences emerged in how individual students, teachers, and parents, different schools, and different educational systems managed to cope with this unprecedented crisis. Through understanding these differences, we can learn from this health and economic crisis and support educational policies in steering and mitigating such crises in the future. In this call for papers for the *Zeitschrift für Psychologie*, we invite contributions that use the lens of educational and developmental psychology to derive what we have can learned from this pandemic. From a macro, meso, and micro perspective, we aim at improving future pandemic preparedness by tackling three key questions: (1) How can educational systems be prepared for future health and associated economic crises to mitigate impacts of learning and development as well as psychosocial functioning of students? (2) Which competencies need to be fostered in students or teachers and how can this be achieved? (3) How can social inequalities be alleviated and an increase in educational gaps during crises be counteracted? We seek original research contributions that use survey or experimental data, social media data or objective data, or combinations thereof. The main focus will be on presenting original data to derive and justify implications and recommendations.

How to Submit

There is a two-stage submissions process. Initially, interested authors are requested to submit extended abstracts of their proposed papers. Authors of the selected abstracts will then be invited to submit full papers. All papers will undergo blind peer review.

Stage 1: Structured Abstract Submission

Authors interested in this special issue must submit a structured abstract of the planned manuscript before submitting a full paper. The goal is to provide authors with prompt feedback regarding the suitability and relevance of the planned manuscript to the special issue.

The deadline for submitting structured abstracts is June 15, 2022

Feedback on whether or not the editors encourage authors to submit a full paper will be given by July 15, 2022.

Submission Guidelines for Structured Abstracts

Structured abstracts should be no more than 1,500 words and may encompass information on each of the following headings: (a) Background, (b) Objectives, Research question(s) and/or hypothesis/es, (c) Method/Approach,

https://doi.org/10.1027/2151-2604/a000490

(d) Results/Findings, (e) Conclusions and implications (expected).

Structured abstracts should be submitted via email to Marko Lüftenegger (marko.lueftenegger@univie.ac.at).

Stage 2: Full Paper Submission

For those who have been encouraged to submit a full paper,

the deadline for submission of manuscripts is October 15, 2022

The full papers must be submitted through the online submission system of the journal, Editorial Manager. Full manuscripts will undergo a blind peer-review process.

Submission Guidelines for Full Papers

- Only English-language submissions can be considered.
- Contributions must be original (not published previously or currently under review for publication elsewhere).
- Original research articles should not exceed 45,000 characters and spaces in length, including references, figures, and tables (allowances for figures and tables should be deducted on the basis of size: approximately 1,250 characters for a quarter-page figure/table).
- Other submission formats (short reports, research summaries, opinion pieces, etc.) are also considered, please contact the editors for details.
- Reference citations in the text and in the reference list should be in accordance with the principles set out in the *Publication Manual of the American Psychological Association* (7th ed.).
- Supplementary material must be made available through digital open access repositories such as PsychArchives: https://www.psycharchives.org/
- See also any recent issue of the journal.

For detailed author guidelines, please see the journal's website at www.hogrefe.com/j/zfp/

For additional information, please contact: marko.lueftenegger@univie.ac.at, martin.daumiller@phil.uni-augsburg.de, i.schoon@ucl.ac.uk

Timeline

- **June 15, 2022:** Abstract submissions due
- **July 15, 2022:** Feedback to authors
- **October 15, 2022:** Full paper submissions due
- **December 15, 2023:** Feedback to authors of full paper submissions due
- **February 15, 2023:** Revised manuscripts due
- **February 28, 2023:** Editorial decision about acceptance/refusal of revised papers due
- **Issue 3 (2023):** Publication of topical issue

About the Journal

The *Zeitschrift für Psychologie*, founded in 1890, is the oldest psychology journal in Europe and the second oldest in the world. One of the founding editors was Hermann Ebbinghaus. Since 2007 it is published in English and devoted to publishing topical issues that provide state-of-the-art reviews of current research in psychology. For more detailed information about the journal please visit the official website at http://www.hgf.io/zfp

Call for Papers

"Are All Conspiracy Theories Created Equal? The Form, Functions, and Consequences of General Conspiracy Mindsets Versus Specific Beliefs"

A Topical Issue of the *Zeitschrift für Psychologie*

Guest Editors: Roland Imhoff[1], Aleksandra Cichocka[2], Biljana Gjoneska[3], and Olivier Klein[4]

[1]Social and Legal Psychology, Johannes Gutenberg University Mainz, Germany
[2]School of Psychology, University of Kent, UK
[3]Macedonian Academy of Sciences and Arts, Skopje, North Macedonia
[4]Center for Social and Cultural Psychology, Université libre de Bruxelles, Belgium

Focus of the Topical Issue, Aims, and Scope

Events like the 2021 storming of the US capitol or protests against protective measures during the global COVID-19 pandemic have brought conspiracy beliefs and their consequences to the center stage of political discourse. Although each of these actions is motivated by very concrete conspiracy allegations, psychological research has long suspected a coherent mindset behind concrete conspiracy beliefs, labelled a conspiracy mentality, conspiracist ideation, conspiratorial worldview or mindset. In this special issue we welcome submissions that help elucidate the psychology behind this general propensity to endorse conspiracy theories as well as scholarly critique and empirical arguments against this view. We invite original scholarly contributions that aim to tackle the different functions and manifestations of different conspiracy beliefs as well as a generalized conspiracy mentality.

Both original papers as well as systematic reviews and meta-analyses are welcome in this special issue. Studies spanning across longer periods of time and wider geographical regions, would be of special interest to editors. In particular, the editors may prioritize longitudinal over cross-sectional data, experimental over correlational evidence, data from underrepresented samples over convenience or crowdsourced samples, or pre-registered research over non-pre-registered research.

The topics covered may include (but are not limited to):

- What are different antecedents and consequences of generalized conspiracy mentality vs. specific conspiracy beliefs?
- To what extent are conspiracy beliefs just an expression of an underlying worldview or highly specific beliefs that serve intergroup as well as personal purposes?
- What is conspiracy mentality and what are the (dis) similarities with political ideology?
- As a counterpoint to the notion of a uniform worldview: Are there meaningful differences between different conspiracy beliefs that are associated with differential antecedents, functions, and consequences (e.g., intergroup vs. intragroup; personally affected vs. not; targeting elites vs. marginalized groups)?
- How does the concept of a uniform mentality help or hinder interventions aimed at mitigating the potential adverse effects of conspiracy beliefs?

How to Submit

There is a two-stage submission process. Initially, interested authors are requested to submit extended abstracts of their proposed papers. Authors of the selected abstracts will then be invited to submit full papers. All manuscripts will undergo blind peer review.

https://doi.org/10.1027/2151-2604/a000489

Stage 1: Structured Abstract Submission

Authors interested in this special issue must submit a structured abstract of the planned manuscript before submitting a full paper. The goal is to provide authors with prompt feedback regarding the suitability and relevance of the planned manuscript to the special issue.

The deadline for submitting structured abstracts is June 15, 2022.

Feedback on whether or not the editors encourage authors to submit a full paper will be given by August 1, 2022.

Structured abstracts should be within 500 to 1,000 words maximum and may encompass information on each of the following headings: (a) Background and Objectives, (b) Methods, (c) Results, (d) Conclusions. For empirical papers, please specify whether the study was pre-registered or not.

Structured abstracts should be submitted by email to the guest editor Roland Imhoff (roland.imhoff@uni-mainz.de).

Stage 2: Full Paper Submission

For those who have been encouraged to submit a full paper,

the deadline for submission of manuscripts is January 1, 2023.

The full papers must be submitted through the online submission system of the journal, Editorial Manager. Full manuscripts will undergo a blind peer-review process.

Submission Guidelines for Full Papers

- Only English-language submissions can be considered.
- Contributions must be original (not published previously or currently under review for publication elsewhere).
- Review and original articles should not exceed 45,000 characters and spaces in length (roughly 6,500 words), including references, figures, and tables (allowances for figures and tables should be deducted on the basis of size: approximately 1,250 characters for a quarter-page figure/table).
- All research syntheses should adhere to the meta-analytic reporting standards (MARS) proposed by the APA (http://www.apa.org/pubs/authors/jars.pdf). Additionally, authors should include a statement in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (most current version available at http://www.prisma-statement.org) as a supplemental file for review.
- Other submission formats (short reports, opinion pieces, etc.) will also be considered – please contact the editors for details. Research spotlights reporting single studies will be reserved for studies including hard-to-reach-populations, advanced methods, exceptionally large datasets, or any other feature that justifies the single study format.
- Reference citations in the text and in the reference list should be in accordance with the principles set out in the Publication Manual of the American Psychological Association (7th ed.).
- Supplementary material must be made available through PsychArchives: https://www.psycharchives.org/.
- For exemplary articles, please see any recent issue of the journal.

For more detailed instructions for authors, please visit the following link https://tinyurl.com/2mbzfjth.

Timeline

- **June 15, 2022:** Abstract submissions due
- **August 1, 2022:** Deadline for abstract selection/call for full papers
- **January 1, 2023:** Full paper submissions due
- **March 15, 2023:** Guest editor feedback to authors
- **June 1, 2023:** Deadline for revised papers
- **September 30, 2023:** Guest editor feedback to revised papers
- **Issue 1 (2024):** Publication topical issue

About the Journal

The *Zeitschrift für Psychologie*, founded in 1890, is the oldest psychology journal in Europe and the second oldest in the world. One of the founding editors was Hermann Ebbinghaus. Since 2007 it is published in English and devoted to publishing topical issues that provide state-of-the-art reviews of current research in psychology. For more detailed information about the journal please visit the official website at http://www.hgf.io/zfp.

Instructions to Authors

The *Zeitschrift für Psychologie* publishes high-quality research from all branches of empirical psychology that is clearly of international interest and relevance, and does so in four topical issues per year. Each topical issue is carefully compiled by guest editors. The subjects being covered are determined by the editorial team after consultation within the scientific community, thus ensuring topicality. The *Zeitschrift für Psychologie* thus brings convenient, cutting-edge compilations of the best of modern psychological science, each covering an area of current interest.

***Zeitschrift für Psychologie* publishes the following types of articles:** Review Articles, Original Articles, Research Spotlights, Horizons, and Opinions.

Manuscript submission: A call for papers is issued for each topical issue. Current calls are available on the journal's website at http://www.hgf.io/zfp. Manuscripts should be submitted as Word or RTF documents by e-mail to the responsible guest editor(s). An article can only be considered for publication in the *Zeitschrift für Psychologie* if it can be assigned to one of the topical issues that have been announced. The journal does not accept general submissions.

Detailed instructions to authors are provided at **http://www.hgf.io/zfp**

Copyright Agreement: By submitting an article, the author confirms and guarantees on behalf of themselves and any coauthors that he or she holds all copyright in and titles to the submitted contribution, including any figures, photographs, line drawings, plans, maps, sketches and tables, and that the article and its contents do not infringe in any way on the rights of third parties. The author indemnifies and holds harmless the publisher from any third-party claims. The author agrees, upon acceptance of the article for publication, to transfer to the publisher on behalf of themselves and any coauthors the exclusive right to reproduce and distribute the article and its contents, both physically and in nonphysical, electronic, and other form, in the journal to which it has been submitted and in other independent publications, with no limits on the number of copies or on the form or the extent of the distribution. These rights are transferred for the duration of copyright as defined by international law. Furthermore, the author transfers to the publisher the following exclusive rights to the article and its contents:

1. The rights to produce advance copies, reprints, or offprints of the article, in full or in part, to undertake or allow translations into other languages, to distribute other forms or modified versions of the article, and to produce and distribute summaries or abstracts.
2. The rights to microfilm and microfiche editions or similar, to the use of the article and its contents in videotext, teletext, and similar systems, to recordings or reproduction using other media, digital or analog, including electronic, magnetic, and optical media, and in multimedia form, as well as for public broadcasting in radio, television, or other forms of broadcast.
3. The rights to store the article and its content in machine-readable or electronic form on all media (such as computer disks, compact disks, magnetic tape), to store the article and its contents in online databases belonging to the publisher or third parties for viewing or downloading by third parties, and to present or reproduce the article or its contents on visual display screens, monitors, and similar devices, either directly or via data transmission.
4. The rights to reproduce and distribute the article and its contents by all other means, including photomechanical and similar processes (such as photocopying or facsimile), and as part of so-called document delivery services.
5. The right to transfer any or all rights mentioned in this agreement, as well as rights retained by the relevant copyright clearing centers, including royalty rights to third parties.

Online Rights for Journal Articles: Guidelines on authors' rights to archive electronic versions of their manuscripts online are given in the document "Guidelines on sharing and use of articles in Hogrefe journals" on the journal's web page at http://www.hgf.io/zfp

August 2021

The joys and pains of using big data and research synthesis methods in the field of psychology

This sixth collection of "Hotspots in Psychology" goes beyond presenting state-of-the-art systematic reviews and meta-analyses in research fields to also discuss the fruitfulness and challenges of using big data in psychological research. For instance, topics such as intensive longitudinal data (e.g., time series and experience sampling), nonobtrusive methods of data gathering (e.g., sensors and log data), and how big data can be handled and analyzed.

Five contributions explore the application of individual participants meta-analyses as a way to replicate studies, the role of the degree of anthropomorphism ("human-likeness") in human–robot interactions, the challenge of multiple dependent effect sizes when conducting a meta-analytical structural equation model, the value of using log data from online platforms as a way to predict learning outcomes, and the utility of a block-wise fit evaluation in structural equation models with many longitudinally measured variables. To promote open science, supplemental material is available in a repository.

Contents include:

Hogrefe Publishing Group

Göttingen · Berne · Vienna · Oxford · Paris
Boston · Amsterdam · Prague · Florence
Copenhagen · Stockholm · Helsinki · Oslo
Madrid · Barcelona · Seville · Bilbao
Zaragoza · São Paulo · Lisbon

www.hogrefe.com

ISBN 978-0-88937-622-9