Is Replication Possible without Fidelity?

Michelle R. Ellefson

University of Cambridge

and

Daniel M. Oppenheimer

Carnegie Mellon University

Author Notes

This manuscript was accepted for publication in *Psychological Methods* on 08 November 2021. This is the accepted version.

We have not presented this data in any forum. R scripts and generated data for this manuscript are openly available from https://osf.io/7y4v5 and are linked to a preprint (Ellefson & Oppenheimer, 2021, https://psyarxiv.com/bqnk4). Thanks to Richard Parkin for advice on programming.

Correspondence concerning this article should be addressed to Michelle R. Ellefson, University of Cambridge Faculty of Education, 184 Hills Road, Cambridge, CB2 8PQ, United Kingdom. Email: mre33@cam.ac.uk.

Abstract

Failure of replication attempts in experimental psychology might extend beyond *p*-hacking, publication bias or hidden moderators; reductions in experimental power can be caused by violations of fidelity to a set of experimental protocols. In this paper, we run a series of simulations to systematically explore how manipulating fidelity influences effect size. We find statistical patterns that mimic those found in ManyLabs style replications and meta-analyses, suggesting that fidelity violations are present in many replications attempts in psychology. Scholars in intervention science, medicine, and education have developed methods of improving and measuring fidelity, and as replication becomes more mainstream in psychology, the field would benefit from adopting such approaches as well.

Keywords: replication; fidelity; effect size; heterogeneity; simulations

Is Replication Possible without Fidelity?

About a decade ago, results on improbable topics with implausible findings generated discussion about whether experimental psychology methods were sufficiently robust (e.g., Chambers, 2017; Spellman, 2013; 2015). Around the same time, scholars uncovered a number of research and statistical practices that massively increased the likelihood of false positives (Simmons et al., 2011), and which were commonly used in psychological science (John et al., 2012).

These discussions led to important initiatives to improve scholarship in psychology such as large replication projects (e.g., ManyLabs, Psychological Science Accelerator, SCORE: Systematizing Confidence in Open Research and Evidence) and the birth of metascience as a viable subfield in psychology. One common result from these initiatives is that the effect sizes of replications are frequently lower than in the study being replicated, even for successfully replicated studies (e.g., Camerer, et al., 2018; Klein et al., 2018).

Research in meta-analysis has consistently shown that even direct replications exhibit surprising amounts of statistical heterogeneity. Linden & Hönekopp (2021) define heterogeneity as the amount by which effect sizes differ between studies that cannot be explained by sampling error. The authors found heterogeneity was smaller for close replications than for conceptual replications, reflecting the fact that the more similar two studies are, the more similar the effect sizes are likely to be. However, they, and other scholars find heterogeneity to be large even for direct replications.

ManyLabs-2 (Klein et al., 2018) found statistically significant heterogeneity in nearly 40% of the replication attempts, with other massive pre-registered replication efforts replicating this heterogeneity (e.g., 36% in Hagger et al., 2016; nearly 45% in Eerland et al., 2016). van Erp et al. (2017) explored heterogeneity across over 700 meta-analyses published in *Psychological Bulletin* over the past 30 years and calculated a median heterogeneity of over 70% (although many of these meta-analyses examined conceptual replications rather than direct replications,

with differences between paradigms explaining some of the observed heterogeneity). Stanley et al. (2018) find that heterogeneity accounts for three times as much variance between studies as what would be expected from sampling error and argue that this heterogeneity could underlie failed replication attempts.

Possible reasons for this trend could include *p*-hacking, publication bias, or hidden moderators (e.g., Klein et al., 2018). However, one additional explanation could include subtle variations from the original protocol. Indeed, authors frequently respond to replication attempts with details about differences between the replication and the original experimental paradigm (e.g., Monin et al., 2014). Even well-intentioned scholars working in good faith can make errors when attempting to follow an experimental protocol. The larger the research team (e.g., the more research assistants involved in running studies) or the farther design is from a scholar's area of expertise, the more likely a study is to incorporate inadvertent deviations from the original protocol (see Table 1 for concrete examples of how fidelity violations could occur)

The *fidelity* to experimental protocols is an important element of grappling with the challenges of designing fair replications. Fidelity can be conceptualized as consistency of a method between different implementations (e.g., Santacroce et al., 2004). For example, measures of how much each teacher is adhering to a curriculum or how much a community-based intervention varies across sites. Variation in how well research teams adhere to instructions or manuals and how similar they are to each other can undercut experimental power and lead to Type II errors. Fortunately, there is a large literature in education, medicine, and intervention science specifically on the issue of fidelity that can be imported into experimental psychology to help improve the quality of replications (and original research as well!).

Typically, four core areas of fidelity checks are important for research (e.g., Bellg et al., 2004; Gearing et al., 2011; Mowbray et al., 2003): (1) *Design* - including a clear theoretical framework; (2) *Training* - including the initial training and follow up training protocols; (3)

Delivery - including clear guidelines about how implementers run the program, plans for monitoring drift from the protocol and feedback to the implementer; and (4) *Participation* - including measuring how much of the intervention each user completed (dosage) and followed. In addition, lists of threats and measures for each of these fidelity checks can increase fidelity. Various checklists (e.g., Gearing et al., 2011; Humphrey et al., 2016; Pines, 2020) and metrics (e.g., Aboud & Prado, 2018) are available for a variety of research settings to help researchers better document fidelity.

In educational settings, fidelity has been linked to how much students benefit from an intervention (e.g., Dusenbury et al., 2003; Zvoch, 2009) and effect sizes (Hulleman & Cordray, 2009; Lendrum & Humphrey, 2012). For example, Scammacca et al. (2015) conducted a meta-analysis of reading interventions between 1980 and 2011. They found that effect sizes were larger when researchers, rather than teachers, administered the intervention, and this was particularly true prior to when practices to increase fidelity became standard.

Fidelity is key to intervention work, but even in fields where fidelity is emphasized only a minority of studies measure or report fidelity, and even fewer measure fidelity quantitatively. For example, Maynard et al. (2013) found that only 29% of after-school intervention programs measured fidelity and only 4% included fidelity metrics in the analyses of the intervention (see Capin et al., 2018; Kechter et al., 2019; for similar patterns across reading intervention and mindfulness programs). Naleppa & Cagle (2010) found that only 15% of social work intervention studies collected fidelity data. Obtaining fidelity metrics involves detailed oversight over how well each researcher or implementer keeps to the protocol (e.g., Bellg et al., 2004; Gersten et al. 2004). Specific fidelity metrics include but are not limited to: (1) keeping checklists on implementation or researcher tasks; (2) measuring how well each person implementing the task does after a training event, but before running the experiment or intervention; (3) trained observers coding adherence to the protocol during live intervention events or from video/audio recordings; (4) interviews with those administering the intervention/task; and (5) participant

surveys. When qualitative or observational data are collected, then it is best to use multiple observers who are blind to experimental condition and to calculate inter-rater reliability metrics

There is growing concern that the substantial amount of null findings of educational interventions in the What Works Clearinghouse (https://ies.ed.gov/ncee/wwc/) and the Education Endowment Foundation (https://educationendowmentfoundation.org.uk/evidence-summaries/) could be due to a lack of fidelity, resulting in potentially sound programs being unnecessarily scrapped (Stockard, 2010).

Given the link between fidelity and effect size in intervention work, fidelity should play an important role in experimental psychology replication studies because they typically involve teams of researchers. Individuals on those teams might unintentionally introduce variance into the study based on individual differences in their research skills and/or their overall delivery of the study (e.g., differences in prosody, pacing, etc.). Moreover, researchers rarely explicate fully every element of a data collection protocol, meaning that good faith replication attempts may include subtle differences in how faithfully the replication adheres to the conditions of the original study (see Table 1). These minor variations will decrease fidelity. However, few experimental studies in psychology, whether they be replications or not, report any fidelity measures.

Even though there is widespread acceptance that fidelity is important for intervention studies, there is not yet a well-accepted quantified metric to show just how much it matters. In this paper, we run a series of simulations to systematically explore how manipulating fidelity influences effect size. In addition, we argue that statistical patterns characteristic of fidelity challenges are evident in large scale replication attempts such as ManyLabs.

Table 1.

Examples of Research Team and Environment Differences that Could Affect Fidelity

	Research Team	**Research Environment**
Qualities Potentially Reducing Fidelity	***Differences across the team for individual researcher characteristics, including:*** ***Comprehensibility:*** Researchers who speak quietly, or quickly, or with an accent, may make it harder for participants to hear, understand, or otherwise be influenced by instructions or manipulations. ***Demeanor*****:** Researchers who are more friendly (i.e., smiling) or have more enthusiasm/charisma may differentially motivate participants to take the task seriously. ***Diligence*****:** Researchers who are less conscientious may forget to include some elements of a manipulation, be less precise in timing, or otherwise inadvertently modify experimental protocols.	***Differences across research settings, including:*** ***Timing:*** Research conducted at different times could yield participants who are more fatigued or stressed and thus less attentive to experimental interventions. For example, studies run in the early morning, late evening, or just after eating or those conducted with students during examinations. ***Temperature:*** Studies in overly air-conditioned or overheated rooms may cause participants to be uncomfortable and less diligent /responsive to interventions. ***Distractions/Information:*** Research laboratories could differ in terms of the information posted inside or outside of the building/room. Furthermore, it is often difficult to control the environment in which studies are run, especially online or field studies. Unobserved online participants might be watching television or quietly working, they could be alone or surrounded by others, and it is difficult to know whether they are surrounded by cues, anchors, or primes that could influence the effectiveness of an intervention.
Real World Examples	An undergraduate research assistant in Oppenheimer's lab was found to have run participants in the dark because the protocol did not specify that the lights should be turned on at the beginning of the study. This added unexpected difficulty and reduced the effectiveness of the intervention. Those participants' data were discarded. A former teacher recruited to help with data collection during one round of data collection within a 3-year school-based intervention project was seen coaching children to get the correct answer, deviating from a detailed administration script (see Ellefson et al 2018), including standardized tests of cognitive ability. The data were not published.	Ellefson's lab ran a study on young children's chemistry reasoning. Before each instance of seeing various solids mixed with water, they were asked to predict what they thought would happen. In one school, they used a small side room a food pyramid poster on the wall. Some children from that school (and only that school) referred to that information when responding to the prompts. Oppenheimer's lab used to be located near the field where the university marching band practiced. From 3-5pm, three days a week, the band was loudly audible from the lab, making for a very distracting environment (they quickly decided to avoid running participants during that time).
Take Home Message	In general, the larger the research team, and the less automated the research the more opportunities there are for fidelity violations.	Research teams needs to be mindful of every detail in the research environment; and take careful stock of all fieldwork settings.

Notes. Some of these examples could be classified as moderators in some experimental designs For example, a kind vs. mean researcher would be a fidelity violation for designs involving feeling rapport and/or motivation but might not be for designs where researcher demeanor matters less.

Method

We generated data for our simulations using *tidyverse* and *data.table* packages in R (Dowle & Srinivasan, 2020; R core team, 2019; Wickam et al., 2019). An R script (openly available - https://osf.io/7y4v5) was created to generate simulated data that could be used to test how fidelity influences group differences (means, SDs, effect sizes). We generated data for three different models (see Table 2). Model One tests fidelity influences on effect sizes within a classroom setting without incorporating how other factors like site characteristics influences the findings. Model Two extends Model One by evaluating how these other factors influences the influence of fidelity on effect size. Model Three extends Model Two to a structure more like psychology laboratory studies.

Input Parameters

Sample Size

In Models One and Two, the total number of participants and classrooms was selected to represent a large study (N = 1000) that included a decent number of classrooms (n = 40) randomly assigned to each condition, and the number of students within a classroom being a fair representation of educational practice (n = 25). Model Three uses the same overall number of participants (N = 1000), but they are distributed into fewer labs (n = 100 participants each for 10 labs) representing a sampling approach that is more like ManyLabs.

Ability Score

Ability score represents individual participants' baseline ability and corresponds to individual differences in baseline characteristics within the sampled population. The mean base-ability parameter was set at 100 and the *SD* at 10% of the mean. Delta-ability represents the true effectiveness of the intervention (how much the intervention actually influences ability). It was sampled across a range representing 0 to 20% of the base-ability mean and 0 to 200% of the base-ability standard deviation.

Fidelity Parameters

Conceptually, fidelity represents the percentage of an intervention to which each classroom adhered. Thus, fidelity should not be smaller than 0.00 or larger than 1.00. Fidelity *Ms* and *SDs* were sampled across the range from 0 to 1.00 at intervals of 0.20 for individual classrooms/labs. We refrained from using absolute 0.00 or 1.00 because of potential errors in the underlying mathematics (0.01 represented zero, 0.99 represented 1.00). Fidelity *M* and *SD* combinations where the sum exceeded 1.00 (e.g., *M* = 0.70, *SD* = 0.40) or where the difference was less than 0.00 (e.g., *M* = 0.20, *SD* = .50) were excluded.

Classroom/Lab Bias Parameters

Classroom/Lab Bias represents any factor that affects performance in a specific classroom or lab that is not related to the intervention of interest (e.g., a better teacher could improve student performance). Classroom/lab bias parameters were applied at the classroom/lab level for Models Two and Three; the specific *M* values tested represented 0 to 20% of the base-ability mean; the *SD* ranged between 0 and 200% of the base-ability score.

Table 2.

Input Parameter Descriptions and Values for Models One, Two and Three

Input Parameter	***Brief Description***	***Level Simulated***	***Model One***	***Model Two***	***Model Three***
Classrooms/Labs (n)	Number of classrooms/labs in each simulation		40	40	10
Students/Participants (n)	Number of students/participants in each simulation		1000	1000	1000
Students/Participants per Classroom/Lab (n)	Number of student/participants per classroom/lab in each simulation		25	25	100
Base-Ability (M)	Gaussian distribution of base ability score values for each simulation based on these M and SD values	participant	100	100	100
Base-Ability (SD)		participant	10	10	10
Delta-Ability	Amount of change in ability score produced by the intervention	experimental condition	0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20	0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20	0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20
Fidelity (M)	Gaussian distribution of fidelity values for each simulation based on these M and SD values	classroom/lab	0.01, 0.20, 0.40, 0.60, 0.80, 0.99	0.01, 0.20, 0.40, 0.60, 0.80, 0.99	0.01, 0.20, 0.40, 0.60, 0.80, 0.99
Fidelity (SD)		participant	0.01, 0.20, 0.40, 0.60, 0.80, 0.99	0.01, 0.20, 0.40, 0.60, 0.80, 0.99	0.01, 0.20, 0.40, 0.60, 0.80, 0.99
Classroom/Lab Bias (M)	Gaussian distribution of classroom/lab bias values for each simulation based on these M and SD values	classroom/lab	0	-20, -15, -10, -5, 0, 5, 10, 15, 20	-20, -15, -10, -5, 0, 5, 10, 15, 20
Classroom/Lab Bias (SD)		classroom/lab	0	0.01, 5, 10, 15, 20, 25, 30, 35, 40	0.01, 5, 10, 15, 20, 25, 30, 35, 40
Random Assignment to Condition (p = .50)	Odds of being assigned to experimental vs control group	varies by model	Classroom	Classroom	Participant

Generating Data

Means and standard deviations for the input parameters were used to generate a Gaussian distribution of base-ability scores for 1000 simulated participants, as well as Fidelity and classroom/lab bias scores. We selected a Gaussian distribution because the most common statistical tests in psychology research are parametric (e.g., t-test, ANOVA, regression, etc.) and assume a normal distribution of data. We ran additional models using uniform, Poisson, exponential, binomial, and random distributions. Those models produced similar results and are openly available (see https://osf.io/7y4v5).

The probability of being assigned to the intervention vs control conditions was 50%. This assignment was by classroom for Models One and Two, and by participant for Model Three. A post-hoc audit confirmed roughly equal sample sizes for the conditions.

For each participant, we calculated final scores based on their base-ability score, the classroom-bias, and fidelity (see Equation 1).

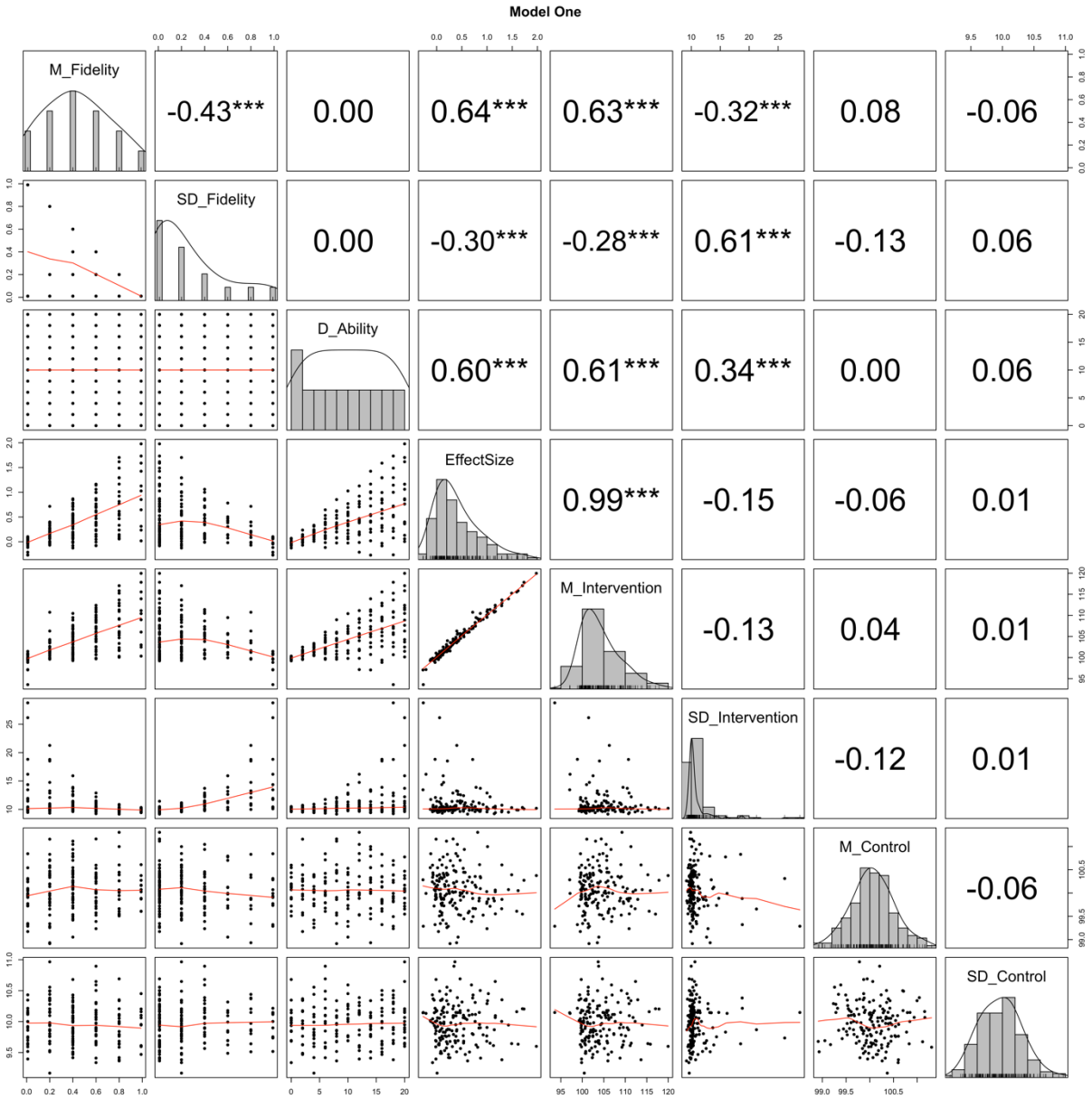
$$\textit{Final Score} = \textit{Base-Ability} + \textit{Classroom/Lab Bias} + (\textit{Fidelity} * \textit{Delta-Ability}) \quad \text{Eq. 1}$$

The input parameter for the delta-ability score was used for participants in the experimental group but set to zero for the control group. Mean scores for the experimental vs. control group were used to generate effect sizes (Cohen's *d*) for each simulation.

For each model, we ran a simulation for each viable combination of input parameters (165 simulations for Model One, 5,445 simulations each for Models Two and Three). Each simulation represented 1,000 participants and the number of classroom/labs outlined in Table 2. For each model, we examined the relationships between the various input parameter values to the key output values of effect size, Ms and SDs using *tidyverse*, *psych* and *ppcor* R packages (Dowle & Srinivasan, 2020; Kim, 2015; Revelle, 2020). This procedure is analogous to meta-analyses comparing results across different studies. These simulation-level comparisons provide a way of measuring the influence of fidelity on effect size while controlling for unobserved variables/moderators in a way not possible in real-world research.

Figure 1.

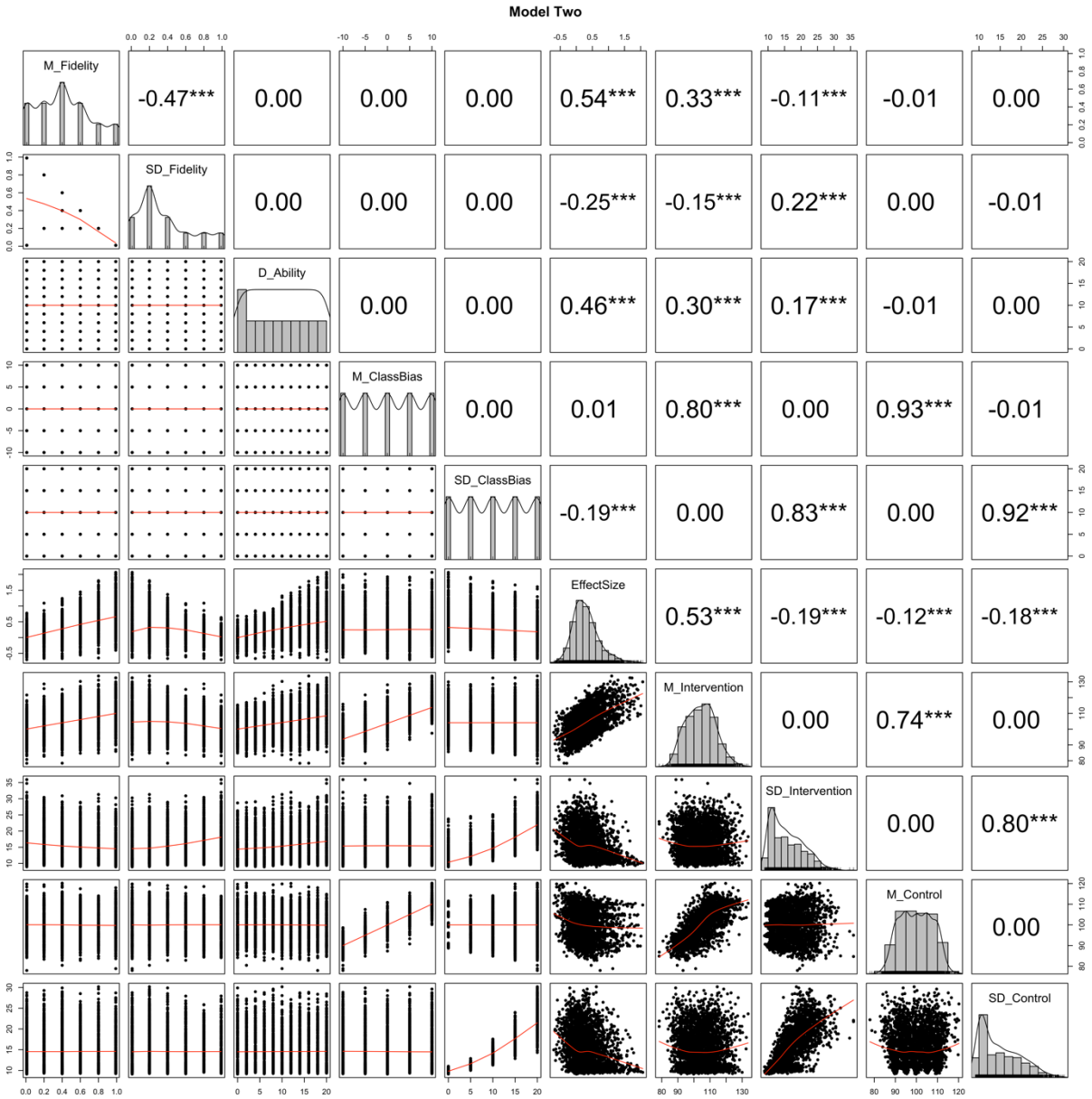
Model One Histograms, Scatterplots and Pairwise Correlations (N = 165 simulations)



Notes. M_Fidelity, SD_Fidelity, and D_Ability are input parameters for the simulated models; M_Intervention, SD_Intervention, M_Control, and SD_Control are data generated by the model. M = mean, SD = standard deviation, D = difference / change.

Figure 2.

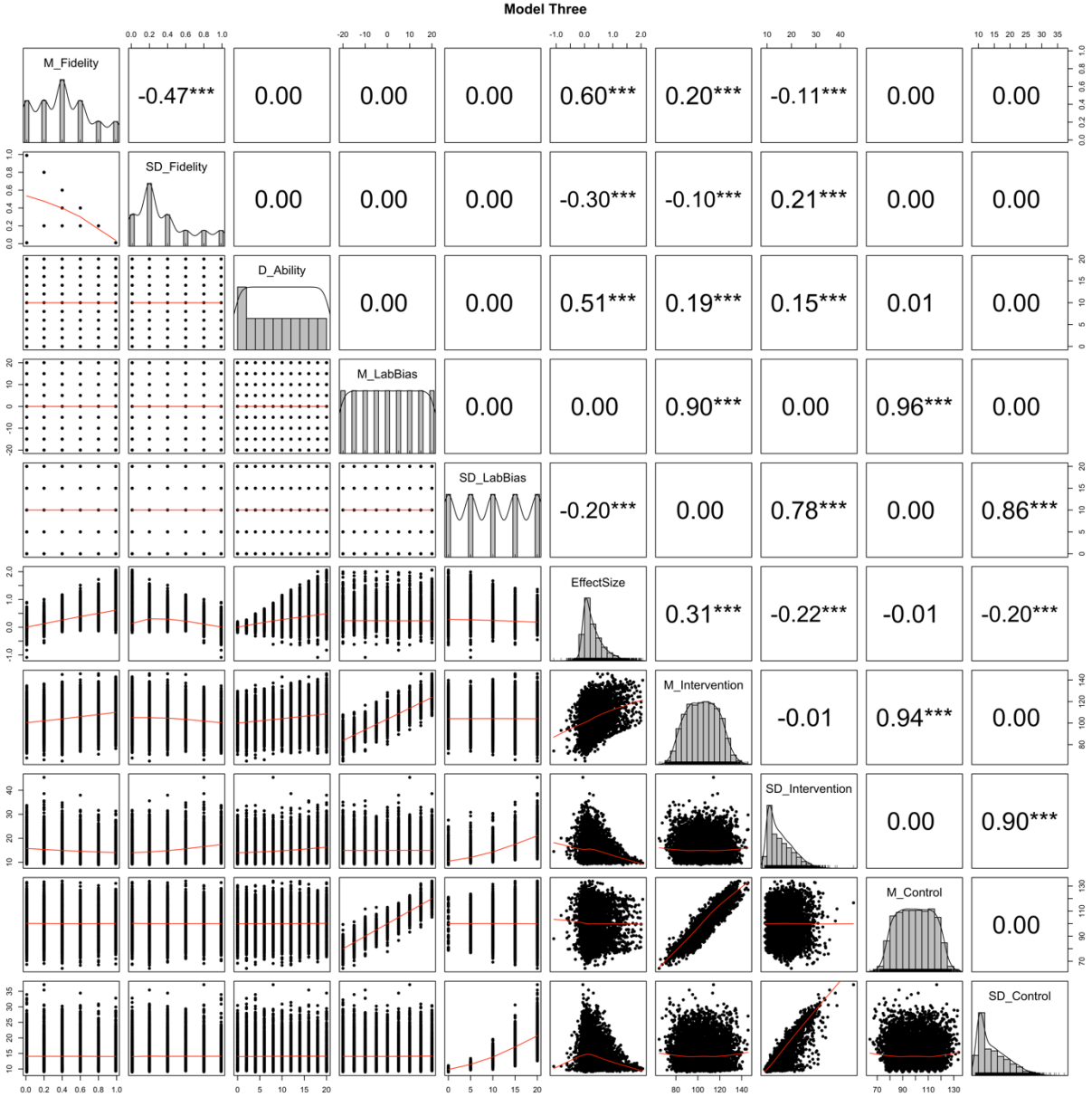
Model Two Histograms, Scatterplots and Pairwise Correlations (N = 5,445 simulations)



Notes. M_Fidelity, SD_Fidelity, and D_Ability are input parameters for the simulated models; M_Intervention, SD_Intervention, M_Control, and SD_Control are data generated by the model. M = mean, SD = standard deviation, D = difference / change.

Figure 3.

Model Three Histograms, Scatterplots and Pairwise Correlations (N = 5,445 simulations)



Notes. M_Fidelity, SD_Fidelity, and D_Ability are input parameters for the simulated models; M_Intervention, SD_Intervention, M_Control, and SD_Control are data generated by the model. M = mean, SD = standard deviation, D = difference / change.

Results and Discussion

Fidelity Influence Effect Sizes

Unsurprisingly, as fidelity decreased, so did effect sizes (see Figures 1, 2 and 3). This pattern replicated through all of the models (r = .64, .54, .60, for Models One, Two and Three, respectively, ps < .001), suggesting that results were not reliant on particular arbitrary parameterization decisions. Importantly, these simulations quantify exactly how much seemingly small reductions in fidelity influenced effect size, with every 5% fidelity reduction producing a 5% reduction in effect size (F (1,163) = 116.1, p < .001, R^2 = .42, B = 1.01, *Intercept* = -0.04).

While lowering fidelity increases variance, one surprising finding is that it does not do so uniformly. The increase in variance is a function of the base effect size; fidelity matters more for experiments with larger effect sizes. This finding does align with large scale replication projects and meta-analyses. For example, Linden and Hönekopp, (2021) found that research areas with larger effect sizes have more heterogeneity (r = .70; see also Kenny & Judd, 2019). Similarly, ManyLabs-2 (Klein et al., 2018) studies with the largest heterogeneity tended to be those with the largest effect sizes.

In the absence of fidelity challenges, one would expect that larger effect sizes indicate higher signal to noise ratios, which would yield low variance between classrooms/labs because the signal would swamp the noise. Instead, the opposite pattern exists for large scale replications, meta-analyses, and our own simulations. Given the controls that we were able to apply to the simulations, these findings suggest that low fidelity could be an issue for many replication studies.

Classroom Bias Indirectly Influences Effect Sizes

Models Two and Three (incorporating classroom/lab bias), had similar, albeit weaker correlations between fidelity and effect size. Classroom/lab bias means correlated strongly with the means of the intervention and control groups (rs > .80, ps < .001). Similar strong correlations existed for SDs (rs > .78, ps < .001). Although classroom/lab bias did not correlate

directly with effect sizes, a partial correlation indicted a strong negative correlation after controlling for the mean of the intervention group (r = -.82, -.69 for Model Two and Three, respectively, ps < .001). This finding indicates that heterogeneity in classroom/lab factors has less influence on effect sizes than fidelity.

Strengths and Weaknesses of Our Approach

There are a number of advantages to using simulations to explore the relationship between fidelity and effect sizes. For one, we can cheaply and easily run sample sizes well beyond what can plausibly be done even in a ManyLabs type study or meta-analysis. Moreover, we have full control over our input parameters and can rule out common interpretations of heterogeneity observed in meta-analysis (e.g., unmeasured moderators, *p*-hacking) and observe what patterns arise from reductions in fidelity in isolation. How those patterns align with the findings from meta-analyses of non-simulated data provides a sense of fidelity's importance in replication attempts. Our patterns correspond quite well to findings from meta-analyses (e.g., Stanley et al., 2018) and ManyLabs (Klein et al., 2018), suggesting that those patterns emerge without *p*-hacking or unmeasured moderators.

Given that *p*-hacking and unmeasured moderators are not necessary to produce these results does not preclude the possibility that they exist in the published literature. Our findings demonstrate that fidelity violations lead to heterogeneity, but do not indicate that we can safely conclude that observed heterogeneity in the literature is always caused by fidelity violations alone. Still, this finding can highlight that a failure to consider fidelity in replication attempts can undermine the inferential value of those studies.

In our simulations, classroom/lab bias was held orthogonal to fidelity. In the real world, there is reason to believe that better teachers/researchers also more diligently follow protocols (Phillips et al., 2017), thus ensuring fidelity. It is plausible that our simulations underestimate the extent of this issue. Further research should investigate real world sources of bias and fidelity reduction, so as to better understand their interactions.

Challenges of Democratizing Psychology

As psychology grapples with safeguarding the integrity of our science, some authors have suggested that students enrolled in research methods classes be assigned to replicate studies as part of course requirements (Frank & Saxe, 2012; Gernsbacher, 2018). While we do not quibble with the formative value of doing so as a learning opportunity, it is worth noting that delegating replications to students who are not fully trained researchers, and who may not exhibit the attention to detail, rigor, and diligence of professional scholars may yield studies with low fidelity (see Table 1). Indeed, even published scholars may struggle to perfectly recreate the conditions of a study run outside their area of expertise. Additionally, the more researchers involved in a project, the more likely that fidelity is breached during the course of the study.

Given the reputational harms to individual scholars and public confidence in the field as a whole that can arise from failed replications, it is important to limit preventable type II errors. While we encourage practices that promote replication attempts, we need to be careful to do so in ways that promote fidelity. This likely requires that we be thoughtful about who we trust to engage in replications, and that we establish norms of ensuring, measuring and reporting replication fidelity. This includes having original authors sharing stimulus presentation code, RA training procedures, and videos of the experimental being conducted. Replicators should adopt the methods from education, medicine, and intervention sciences to preserve fidelity, such as fidelity checklists (c.f. Gearing et al., 2011), fidelity metrics (c.f. About & Prado, 2018), and thorough documentation regarding how they ensured adherence to experimental protocols. Recent multisite psychology studies publications do include details about training programs and strict protocols that can improve fidelity (e.g., Vohs et al., 2021) or documenting heterogeneity across sites (e.g., ManyLabs2 – Klein et al., 2018, Vohs et al. 2021) albeit without using the term fidelity or presenting fidelity metrics. These studies illustrate the level of detail and care to fidelity that all psychology studies should adopt. However, we suggest that quantitative metrics of fidelity should be added to further improve reproducibility and reliability of findings.

Conclusions

Replications are difficult. Even when researchers engage in good faith attempts to perfectly reproduce an experimental protocol, it is possible to introduce inadvertent deviations. As shown here, these violations of fidelity can influence effect sizes, and lead to heterogeneity of findings in the literature. Indeed, the patterns that emerged in our simulations mimic those that are found in large scale replication attempts and meta-analyses, suggesting that psychology, as a field, may still have work to do in ensuring the fidelity of our replications. We need to recognize that replication success is based not just on participant noise, but instead on that plus fidelity. Scholars in intervention science, medicine, and education have developed methods of improving and measuring fidelity, and as replication becomes more mainstream in psychology, the field would benefit from adopting such approaches as well.

References

Aboud, F. E., & Prado, E. L. (2018). Measuring the implementation of early childhood development programs. *Annals of the New York Academy of Sciences*, *1419*, 249-263. https://doi.org/10.1111/nyas.13642

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., et al. (2004). Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology*, *23*, 443. https://doi.apa.org/doi/10.1037/0278-6133.23.5.443

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637-644. https://doi.org/10.1038/s41562-018-0399-z

Capin, P., Walker, M.A., Vaughn, S., & Wanzek, J. (2018). Examining how treatment fidelity is supported, measured, and reported in K–3 reading intervention research. *Educational Psychology Review*, *30*, 885-919. https://doi.org/10.1007/s10648-017-9429-z

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Dowle, M., & Srinivasan, A. (2020). *data.table: Extension of `data.frame`* (R package version 1.13.6). https://CRAN.R-project.org/package=data.table

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, *18*, 237-256. https://doi.org/10.1093/her/18.2.237

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M., Crocker, C., Ferretti, T. R., Kibbe, M. R., Knepp, M. M., Kurby, C. A., Melcher, J. M., Michael, S. W., Poirier, C., & Prenoveau, J. M. (2016).

Registered Replication Report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*, 158-171. https://doi.org/10.1177/1745691615605826

Ellefson, M. R., Baker, S. T., & Gibson, J. L. (2019). Lessons for successful cognitive developmental science in educational settings: The case of executive functions. *Journal of Cognition and Development*, *20*, 253-277. https://doi.org/10.1080/15248372.2018.1551219

Ellefson, M. R., & Oppenheimer, D. (2021, October 20). *Is replication possible without fidelity?* Preprint from: https://doi.org/10.31234/osf.io/bqnk4 Data available from https://osf.io/7y4v5

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*, 600-604. https://doi.org/10.1177/1745691612460686

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review*, *31*, 79-88. https://doi.org/10.1016/j.cpr.2010.09.007

Gernsbacher, M. A. (2018). Three ways to make replication mainstream. *Behavioral and Brain Sciences*, *41*, e129. https://dx.doi.org/10.1017%2FS0140525X1800064X

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*, 149-164. https://doi.org/10.1177%2F001440290507100202

Hagger, M. S., Chatzisarantis, N., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D., Dewitte, S., Elson, M., … & Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-

Depletion Effect. *Perspectives on Psychological Science*, 11, 546-573. https://doi.org/10.1177/1745691616652873

Hulleman, C. S., & Cordray, D. S. (2009) Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*, 88-110, https://doi.org/10.1080/19345740802539325

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). *Implementation and process evaluation (IPE) for interventions in educational settings: An introductory handbook*. Education Endowment Foundation.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524-532. https://doi.org/10.1177/0956797611430953

Kim, S. (2015). *ppcor: Partial and semi-partial (part) correlation* (R package version 1.1). https://CRAN.R-project.org/package=ppcor

Kechter, A., Amaro, H., & Black, D. S. (2019). Reporting of treatment fidelity in mindfulness-based intervention trials: A review and new tool using NIH Behavior Change Consortium guidelines. *Mindfulness 10*, 215-233. https://doi.org/10.1007/s12671-018-0974-4

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*, 578. https://doi.org/10.1037/met0000209

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, 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*, 443-490. https://doi.org/10.1177/2515245918810225

Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, *38*, 635-652, https://doi.org/10.1080/03054985.2012.734800

Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*. https://doi.org/10.1177%2F1745691620964193

Maynard, B. R., Peters, K. E., Vaughn, M. G., & Sarteschi, C. M. (2013). Fidelity in after-school program intervention research: A systematic review. *Research on Social Work Practice*, *23*, 613-623. https://doi.org/10.1177/1049731513491150

Monin, B., Oppenheimer, D. M., Ferguson, M. J., Carter, T. J., Hassin, R. R., Crisp, R. J., Miles, E., Husnu, S., Schwarz, N., Strack, F., Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., . . . & Kahneman, D. (2014). Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*, *45*, 299-311. https://doi.org/10.1027/1864-9335/a000202

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement and validation. *American Journal of Evaluation*, *24*, 315−340. https://doi.org/10.1177%2F109821400302400303

Naleppa, M. J., & Cagle, J. G. (2010). Treatment fidelity in social work intervention research: A review of published studies. *Research on Social Work Practice*, *20*, 674-681. https://doi.org/10.1177/1049731509352088

Phillips, B. M., Ingrole, S. A., Burris, P. W., & Tabulda, G. (2017). Investigating predictors of fidelity of implementation for a preschool vocabulary and language curriculum, *Early Child Development and Care*, *187*, 542-553. https://doi.org/10.1080/03004430.2016.1251428

Pines, J. M. (2020). Narrowing the gap between efficacy and effectiveness using the TIDieR checklist. *American Journal of Emergency Medicine*, *38*, 1178-1179. https://doi.org/10.1016/j.ajem.2020.03.023

R core team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2020). *psych: Procedures for personality and psychological research* (R package version 2.0.12). Northwestern University. https://CRAN.R-project.org/package=psych

Santacroce, S. J., Maccarelli, L. M., & Grey, M. (2004). Intervention fidelity. *Nursing Research*, *53*, 63-66.

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A Meta-Analysis of Interventions for Struggling Readers in Grades 4–12: 1980–2011. *Journal of Learning Disabilities*, *48*, 369-390. https://doi.org/10.1177/0022219413504995

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%2F0956797611417632

Spellman, B. A. (2013). Introduction to the special section on advancing science. *Perspectives on Psychological Science*, *8*, 412-413. https://doi.org/10.1177/1745691613493245

Spellman, B. A. (2015). Introduction to the special section on methods: Odds and end. *Perspectives on Psychological Science*, *10*, 359-360. https://doi.org/10.1177/1745691615582201

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

Stockard, J. (2010). An analysis of the fidelity implementation policies of the What Works Clearinghouse. *Current Issues in Education*, *13*. Retrieved from http://cie.asu.edu/

van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990-2013. *Journal of Open Psychology Data*, *5*, 4. http://doi.org/10.5334/jopd.33

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., Brizi, A., Bunyi, A., Butschek, G. J., Campbell, C., Capaldi, J., Cau, C., Chambers, H., Chatzisarantis, N. L. D., Christensen, W. J., Clay, S. L., Curtis, J., … Albarracín, D. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*. https://doi.org/10.1177/0956797621989733

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin Pedersen, T., Miller, E., Milton Bache, S., Müller, K., Ooms, J., Robinson, D., Paige Seidel, D., Spinu, V., . . . & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*, 1686, https://doi.org/10.21105/joss.01686

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. *American Journal of Evaluation*, *30*, 44-61. https://doi.org/10.1177/1098214008329523