Required sample size to detect mediation in 3-level implementation studies

Nathaniel J. Williams, PhD[1,2]*

Kristopher J. Preacher, PhD[3]

Paul D. Allison, PhD[4]

David Mandell[5,6]

Steven C. Marcus, PhD[5,7]

[1] Institute for the Study of Behavioral Health and Addiction, Boise State University, 1910 University Drive, Boise, ID 83725-1940, natewilliams@boisestate.edu

[2] School of Social Work, Boise State University, Boise, ID

[3] Department of Psychology & Human Development, Vanderbilt University, 230 Appleton Place, Nashville, TN, 37203-5721, kris.preacher@vanderbilt.edu

[4] Statistical Horizons LLC, P.O. Box 282, Ardmore, PA, 19003, allison@statisticalhorizons.com

[5] Penn Center for Mental Health, University of Pennsylvania School of Medicine, 3535 Market Street, Philadelphia, PA 19104, David.Mandell@pennmedicine.upenn.edu#

[6] Department of Psychiatry, University of Pennsylvania School of Medicine, 3535 Market Street, Philadelphia, PA

[7] School of Social Policy & Practice, University of Pennsylvania, 3701 Locust Walk, Philadelphia, PA, 19104-6214, marcuss@upenn.edu

*Corresponding author (NJW natewilliams@boisestate.edu)

Abstract

Background: Statistical tests of mediation are important for advancing implementation science; however, little research has examined the sample sizes needed to detect mediation in 3-level designs (e.g., organization, provider, patient) that are common in implementation research. Using a generalizable Monte Carlo simulation method, this paper examines the sample sizes required to detect mediation in 3-level designs under a range of conditions plausible for implementation studies.

Method: Statistical power was estimated for 17,496 3-level mediation designs in which the independent variable (X) resided at the highest cluster level (e.g., organization), the mediator (M) resided at the intermediate nested level (e.g., provider), and the outcome (Y) resided at the lowest nested level (e.g., patient). Designs varied by sample size per level, intraclass correlation coefficients of M and Y, effect sizes of the two paths constituting the indirect (mediation) effect (i.e., X→M and M→Y), and size of the direct effect. Power estimates were generated for all designs using two statistical models—conventional linear multilevel modeling of manifest variables (MVM) and multilevel structural equation modeling (MSEM)—for both 1- and 2-sided hypothesis tests.

Results: For 2-sided tests, statistical power to detect mediation was sufficient (≥0.8) in only 463 designs (2.6%) estimated using MVM and 228 designs (1.3%) estimated using MSEM; the minimum number of highest-level units needed to achieve adequate power was 40; the minimum total sample size was 900 observations. For 1-sided tests, 808 designs (4.6%) estimated using MVM and 369 designs (2.1%) estimated using MSEM had adequate power; the minimum number of highest-level units was 20; minimum total sample was 600. At least one large effect size for either the X→M or M→Y path was necessary to achieve adequate power across all conditions.

Conclusions: While our analysis has important limitations, results suggest many of the 3-level mediation designs that can realistically be conducted in implementation research lack statistical power to detect mediation of highest-level independent variables unless effect sizes are large and 40 or more highest-level units are enrolled. We suggest strategies to increase statistical power for multilevel mediation designs and innovations to improve the feasibility of mediation tests in implementation research.

Key words: statistical power, indirect effects, mediation, multilevel, Mplus

Contributions to the Literature

- Multilevel mediation analysis is an important tool for testing mechanisms in implementation science; however, little is known about the sample sizes required to adequately power these studies particularly within the range of sample sizes that are feasible for implementation research
- We calculated statistical power to detect mediation in 3-level designs (e.g., organization, provider, patient) using a range of plausible input values and sample sizes for implementation research
- Less than 5% of designs had adequate statistical power to detect mediation; large effect sizes and samples of 40 or more clusters (e.g., organizations) were typically required
- Results indicate changes are needed in how mechanisms are studied in implementation science and in the expectations of research funders

Background

The goal of implementation science is to improve the quality and effectiveness of health services by developing strategies that promote the adoption, implementation, and sustainment of empirically supported interventions in routine care (1). Understanding the causal processes that influence healthcare professionals' and participants' behavior greatly facilitates this aim (2,3); however, knowledge regarding these processes is in its infancy (4,5). One popular approach to understanding causal processes is to conduct mediation studies in which the relationship between an independent variable (X) and a dependent variable (Y) is decomposed into two relationships—an indirect effect that occurs through an intervening or mediator variable (M), and a direct effect that does not occur through an intervening variable (6,7). Figure 1 shows a mediation model in which the effect of X on Y is decomposed into direct (*c'*) and indirect effects (the product of the *a* and *b* paths). Estimates of the *a*, *b*, and *c'* paths shown in Figure 1 can be obtained from regression analyses or structural equation modeling. Under certain assumptions, these estimates allow for inference regarding the extent to which the effect of X on Y is mediated, or transmitted, through the intervening variable M (8-10). Interpreted appropriately, mediation analysis enables investigators to test hypotheses about *how* X contributes to change in Y and thereby to elucidate the mechanisms of change that influence implementation (5,9,10). Recently, several major research funders, including the National Institutes of Health in the United States, have emphasized the importance of an experimental therapeutics approach to translational and implementation research in which mechanisms of action are clearly specified and tested (11-13). Mediation analysis offers an important method for such tests.

Mediation analysis has long been of importance in implementation science, with recent studies emphasizing the need to increase the frequency and rigor with which this method is used (5,69). Guided by theoretical work on implementation mechanisms (70,20), emerging methodsfocused guidance for implementation research calls for the use of mediation analyses in randomized implementation trials to better understand how implementation strategies influence healthcare processes and outcomes (5,22). A systematic review of studies examining implementation mechanisms indicated mediation analysis was the dominant method for testing mechanisms in the field, used by 30 of 46 studies (4). Other systematic reviews highlight deficits in the quality of published mediation analyses in implementation science to date and have called for increased and improved use of the method (5,75). Reflecting its growing importance within the field, mediation analyses feature prominently in several implementation research protocols published in the field's leading journal, *Implementation Science*, during the last year (71-74). Chashin et al. (76) recently published guidance for reporting mediation analyses in implementation studies, including the importance of determining required sample sizes for mediation tests *a priori*.

Designing mediation studies requires estimates of the sample size needed to detect the indirect effect. This seemingly simple issue takes on special nuance and heightened importance in implementation research because of the complexity of statistical power analysis for multilevel research designs—which are the norm in implementation research (21,22)—and the constraints on sample size posed by the practical realities of conducting implementation research in healthcare systems. While statistical power analysis methods and tools for single-level mediation are well-developed and widely available (8,14-18), these approaches are inappropriate for testing mediation in studies with two or more hierarchical levels, such as patients nested within providers nested within organizations (9,29,53). Generating correct inferences about mediation from multilevel research designs requires multilevel analytic approaches and associated power analyses to determine the required sample size (19,23-26).

While some tools have begun to emerge to estimate required sample sizes for 2- and 3-level mediation designs (27,28), findings from this preliminary research indicate that calculation of statistical power for multilevel mediation is complex and depends on the anticipated range and configuration of study design input values—such as effect sizes and sample sizes—at each level (e.g., organization, clinician, patient). As a result, the feasibility of obtaining adequate sample sizes to test multilevel mediation is highly field-dependent; which mediation hypotheses can be realistically tested in implementation science depends on the anticipated range and configuration of realistic study design input values for the field. In implementation research, resource and practical constraints often limit the sample sizes that are feasible to recruit and enroll at the highest level of the design—for example, the number of geographical areas, organizations, or clinics that can be studied—thus potentially restricting the mediation hypotheses that can be realistically tested. Further, the structure of healthcare systems and natural constraints on healthcare processes (e.g., patient flow) often limit the number of providers available within higher-level units over a project period as well as the number of patients each provider serves. These field-specific constraints on sample sizes at each level create a more specific and high-stakes question for implementation scientists interested in using mediation analysis: what are the minimum sample sizes required—at each level—to detect mediation in 3-level designs, *given what is realistic for implementation settings*?

Mediation Analysis in Multilevel Studies

Krull and MacKinnon describe multilevel mediation designs by the level of each variable in the X→M→Y chain (23). Each level in the design represents a different level of sampling (e.g., organization, clinician, provider) and units at lower levels (e.g., patients) are assumed to be nested within units at higher levels (e.g., clinicians). For example, organizations may be at the highest level (level 3), clinicians may be nested within organizations (level 2), and patients may be nested within clinicians (level 1).

Figure 2 presents a conceptual model of a 3-level mediation design and the parameter values the investigator must supply to estimate statistical power or the required sample size. Similar to a protocol by Aarons et al. (78), in this example, an organization-level implementation strategy (X) at level 3, is designed to influence a patient-level implementation outcome at level 1 (Y) through its effects on a level 2 clinician mediator (M). The X variable is random assignment to an organizational implementation strategy versus a control condition. Aarons et al. (78) describe a strategy that trains organizational leaders in skills and strategies that improve clinicians' implementation citizenship behaviors. Increases in clinicians' implementation citizenship behavior (level 2 M) is hypothesized to increase patients' experience of high-fidelity care (level1 Y). In the figure, these relationships correspond to the a_3 and b_3 paths, respectively, which make up the indirect effect at level 3. The c'_3 path represents the direct effect.

To estimate statistical power for this example, the investigator must supply: (1) alpha level (typically set at $\alpha=0.05$); (2) 1- vs. 2-sided hypothesis test; (3) sample size for each level; (4) standardized effect sizes for the a_3, b_3, and c'_3 paths at level 3; (5) a standardized effect size for the b_2 path at level 2; and (6) values of the intraclass (or intracluster) correlation coefficient (ICC) for the mediator M at level 3 (ICC_{m3}) and, for the outcome Y, at levels 2 (ICC_{y2}) and 3 (ICC_{y3}). The ICC is a ratio describing the proportion of variance in a variable that resides at each level of the design (77); it can be interpreted as the extent to which observations within a cluster are correlated with one another (29). In this example, ICC_{y3} represents the variance of the outcome Y that occurs between organizations (e.g., the variance in the means of Y across organizations), and ICC_{y2} represents the variance of the outcome that occurs between clinicians *within* organizations (77). ICC_{m3} represents the variance of the mediator M that occurs between organizations.

In multilevel designs, one can test mediation hypotheses using two different statistical approaches: traditional multilevel modeling based on manifest (i.e., observed) variables (MVM) or multilevel structural equation modeling (MSEM). MVM approaches test mediation based on observed data using traditional multilevel models (19), which are sometimes referred to as hierarchical linear models (29) or mixed effects models (53). Many software programs provide routines to analyze data using these models (24). MSEM uses structural equation modeling to partition observed variables into latent components at different levels of the design and subsequently tests mediation using these latent components (25, 26). Analogous to the relationship between linear regression and single-level structural equation modeling (55), MSEM represents a large-sample approach to multilevel mediation analysis that engenders greater modeling flexibility and produces more accurate effect estimates relative to MVM at the cost of higher standard errors and lower statistical power (25, 52).

Study Aims

In this study, we address the issue of statistical power and minimum sample sizes required to test mediation in 3-level implementation studies using a generalizable method for calculating statistical power based on Monte Carlo simulations. We examined statistical power for mediation in 17,496 3-level designs that varied across a range of design parameter input values deemed plausible for implementation research in healthcare settings. As is shown in Figure 3, power was estimated for all designs using two statistical models: MVM (cells A & C) and MSEM (cells B & D) for both 2-sided (cells A & B) and 1-sided (cells C & D) hypothesis tests.

Our study makes four contributions to implementation science. First, our power analyses address a specific range of realistic design parameter input values for implementation studies in healthcare. As such, our results represent a useful resource and potentially cautionary note for implementation scientists planning multilevel mediation studies. Second, our simulation-based approach to determining statistical power overcomes the limitations of prior formula-based work (28) that does not address power for MSEM designs. While some tools are available to estimate statistical power for multilevel mediation in 2- (27) and 3-level trials (28), these approaches do not accommodate MSEM designs. Often, they accommodate cluster randomized trials but not observational studies. By providing our simulation code to investigators, we offer a power analysis template for multilevel mediation that addresses MSEM for 3-level observational or cluster-randomized designs that can be easily modified for 2-level designs. Third, our approach overcomes the limitations of formula-based tools for 3-level mediation designs which make the restrictive and unrealistic assumption that the direct effect is zero (e.g., 28). This is important because direct effects are rarely equal to zero in implementation studies (see [5]) and because non-zero direct effects meaningfully influence statistical power and sample size in 3-level designs (as is shown below). Fourth, our simulation-based approach incorporates sufficient flexibility to allow investigators to revise the code to address hypotheses regarding moderated mediation (i.e., effect modifiers) and other design variations which are not possible with the limited formula-based tools currently available for 2- or 3-level mediation designs (80).

Focusing on design parameters that are realistic for implementation studies in healthcare, the research questions were: (1) How many of the plausible designs studied had adequate statistical power to detect mediation? (2) What study characteristics were associated with increased statistical power to detect mediation? (3) What was the range of minimum required sample sizes to detect mediation within this set of plausible designs? We provide our code in Additional File 1 as a resource for investigators to estimate statistical power for designs not examined here.

Method

Our method for estimating statistical power was based on empirical Monte Carlo simulations (30,31). Under this approach, many samples of a specified size are generated from a hypothetical population and the model of interest is estimated in each sample. Statistical power is computed as the proportion of samples (e.g., 400 out of 500) in which the parameter of interest is statistically significant. Monte Carlo simulation methods are well-established as a general approach to determining statistical power; they make similar assumptions as formula-based approaches but have greater flexibility for estimating power in complex models derived from hierarchically selected samples (30,31). We followed guidelines for reporting Monte Carlo simulation studies as suggested by Boomsma (68).

We used simulations to estimate statistical power for designs that incorporated a continuous outcome and mediator and varied systematically with regard to the population design parameters shown in Figure 2. Values for each of nine design parameters were fully crossed, resulting in 17,496 designs (3^7*4*2). Following prior work (18,32), values of the two standardized paths that make up the indirect effect (i.e., a_3 and b_3) were set at 0.14, 0.39, and 0.59, which represent small (~2% of the variance), medium (~13% of the variance), and large (~26% of the variance) effect sizes, respectively[1], as suggested by Cohen (33). Based on the same logic, values of the standardized c'_3 path, which represents the direct effect, were set at 0.14 (small) and 0.39 (medium). Values of the standardized b_2 path, which is not typically of

[1] Percentages in parentheses are approximate for the b_3 and c'_3 paths because they are partial coefficients.

substantive interest in implementation studies, were fixed at 0.39 (medium). Values of ICC for the mediator and outcome were set at 0.05, 0.10, and 0.20 at each relevant level of the design. These correspond to small, medium, and large ICCs based on research describing ranges of ICC for process and endpoint variables in implementation research and healthcare settings (34,35,79).

We studied a range of sample sizes relevant to implementation research. As is shown in Table 1, the level-3 sample size (N_3) represents the number of highest-level clusters (e.g., organizations), the level-2 sample size (N_2) represents the number of intermediate-level units per cluster (e.g., providers), and level-1 sample size (N_1) represents the number of lowest-level units per intermediate unit (e.g., patients). Guided by the range of sample sizes observed in systematic reviews of implementation studies (5,36-39), level-3 sample sizes were set at 10, 20, 40, and 60. We chose 10 because it was the expected lower limit on the number of level-3 units (e.g., organizations) necessary to achieve adequate power and 60 because reviews of implementation studies suggest 60 is often the largest feasible sample size. Level-2 sample sizes were set at 5, 10, and 20, reflecting a minimum number of intermediate-level units (e.g., providers) expected to achieve adequate power and an upper limit expected to reflect larger samples in healthcare settings. Level-1 sample sizes were set at 3, 6, and 12, reflecting a minimum number of lowest-level units (e.g., patients) to justify clustering and an anticipated upper limit feasible to recruit during a time-limited period. The code in Additional File 1 can be modified to calculate power for designs not studied here.

For each design, 500 simulated datasets were generated using the MONTECARLO command in Mplus 8 (40). These were analyzed using the TYPE=THREELEVEL option of the ANALYSIS command with the default maximum likelihood estimator with robust standard errors (MLR). Simulations were conducted on multi-processor computing platforms which allowed for simultaneous estimation of models.

We generated statistical power estimates for each of the 17,496 designs under four different conditions shown in Figure 3. Cells A and C in Figure 3 represent statistical power estimates generated for traditional multilevel models with manifest variables (MVM). Cells B and D represent statistical power estimates generated for MSEM. Indirect effects for MVM models were calculated using the "centered within context with means reintroduced" approach described by Zhang et al. (19). MSEM indirect effects were based on latent partitioning of variables (25).

Cells A and B in Figure 3 represent statistical power estimates for both MVM and MSEM using a 2-sided null hypothesis test (H_0: $a_3*b_3 = 0$) with alpha set at α=0.05. For these tests, we used the first-order delta method which is sometimes called the Sobel test (41). The Sobel test is widely used for mediation analyses across multiple disciplines (6) and is slightly more conservative than computationally intensive bootstrapping methods (57) or the Monte Carlo confidence interval approach (15,58).

Cells C and D in Figure 3 represent statistical power estimates for MVM and MSEM using a 1-sided hypothesis test. Many mediation hypotheses could reasonably be specified as directional (i.e., 1-sided) because the implementation strategy is anticipated to have a positive (or negative) effect on the mediator and outcome. Use of a 1-sided test should reduce the sample size needed to detect mediation. Estimates of statistical power for 1-sided tests were generated using an algebraic transformation of the results from the 2-sided simulations and thus did not require additional computational time (details available upon request).

Results

Completion of the simulations required 591 days of computational time. Completion rates, defined as the number of replications within a simulation that successfully converged (e.g., 500 out of 500), were high: 97.8% (*n*=17,114) of the MVM simulations exhibited complete convergence (i.e., 500 of 500 replications were successfully estimated) and 79.4% (*n*=13,889) of the MSEM simulations exhibited complete convergence. The lowest number of completions for any design was 493 (out of 500). The high rate at which the replications were completed increases confidence in the resulting simulation-based estimates of statistical power.

How many of the designs studied had adequate statistical power to detect mediation?

Table 1 shows the frequency and percent of designs studied that had adequate statistical power ($\geq$ 0.8) to detect mediation by study characteristic based on a conventional MVM model, using a 2-sided test (cell A in Figure 3). Only 463 of the 17,496 (2.6%) designs had adequate statistical power to detect mediation. As expected, statistical power was higher for the designs in cell C of Figure 3 which were estimated using MVM and a 1-sided hypothesis test: 808 of these designs (4.6%) had adequate power to detect mediation.

As an alternative to MVM, investigators may use MSEM. Focusing on cell B of Figure 3 (MSEM, 2-sided test), results indicated that 228 of the 17,496 designs (1.3%) studied had adequate statistical power to detect mediation. Shifting to cell D of Figure 3 (MSEM, 1-sided test): 369 of the designs (2.1%) had adequate statistical power.

In summary, less than 5% of the 3-level mediation designs studied had adequate statistical power to detect mediation regardless of the statistical model employed (i.e., MVM vs. MSEM) or whether tests were 1- vs. 2-sided.

What study characteristics were associated with increased statistical power to detect mediation?

Table 1 presents the frequency and percent of designs with adequate statistical power to detect mediation by study characteristic for the 17,496 designs in cell A of Figure 3 (MVM, 2-sided test). Because results were similar for all four cells in Figure 3, we focus on the results from cell A and describe variations for the other cells as appropriate. Additional File 2 presents the frequency and percent of study designs with adequate statistical power to test mediation by study characteristic for all four cells shown in Figure 3.

First, consistent with expectations, statistical power to detect mediation increased as the magnitude of effect sizes increased for the two paths that constitute the indirect effect (i.e., a_3 and b_3). Notably, none of the designs in Table 1 had adequate power when either the a_3 or b_3 paths were small; less than 1% of designs had adequate power when the a_3 or b_3 paths were medium.

Second, the number of adequately powered designs increased as sample sizes increased at each level, with the level-3 sample size having the largest effect on power. In Table 1, no designs with fewer than 40 level-3 clusters (e.g., organizations) had adequate power to detect mediation. This finding also held for the MSEM designs (cells B and D in Figure 3; see Additional File 2). However, for cell C in Figure 3 (MVM, 1-sided test), 11 designs (0.1%) had adequate power to detect mediation with level-3 sample sizes of 20 (see Additional File 2).

Third, larger total sample sizes were associated with increased power, although this relationship was not monotonic because the total sample size consisted of the sum of the sample sizes at each level. In Table 1, the minimum total required sample size to detect mediation was N=900 level-1 units. The minimum total sample for cell C in Figure 3 (MVM, 1-sided test) was N=600. The minimum total sample for cell B in Figure 3 (MSEM, 2-sided test) was N=1,800 and the minimum total sample for cell D in Figure 3 (MSEM, 1-sided test) was N=1,200.

What was the range of minimum sample sizes required to detect mediation?

Table 2 presents the minimum sample sizes required to achieve statistical power $\geq$ 0.8 to detect mediation by values of effect size for the a_3 and b_3 paths that constitute the indirect effect, the size of the direct effect, and the level-3 ICCs of the mediator and outcome. Results in Table 2 are based on cell A of Figure 3 (MVM, 2-sided). In each cell of Table 2, two sample sizes are provided, one assuming a small direct effect (c_s) and the other assuming a medium direct effect (c_m). Sample sizes are presented as N_3 [N_2 [N_1]] where N_3 = number of level-3 units (e.g., organizations), N_2 = number of level-2 units (e.g., providers) per cluster, and N_1 = number of level-1 units (e.g., patients) per level-2 unit. Because the N_3 sample size is typically the most resource intensive to recruit in implementation studies, and because multiple combinations of N_1, N_2, and N_3 can achieve the same total sample size in a given cell, the minimum sample sizes shown in Table 2 were selected based on the sample combination with adequate power and the smallest N_3, followed by the smallest N_2, followed by the smallest N_1. Blank cells (-) are informative in that they indicate there were no sample sizes that achieved adequate statistical power to detect mediation for that design; for these cells, it is not possible to design a study with adequate statistical power to test mediation within the range of sample sizes and input values we tested. Additional File 3 provides a similar table for cell C of Figure 3 (MVM, 1-sided test).

Table 2 provides additional insights into the design features necessary to test mediation in 3-level designs under conditions that are plausible for implementation research. First, most of the cells in Table 2 are empty, indicating no design in that cell had adequate power to detect mediation. This underscores the limited circumstances under which one can obtain a sample large enough to test mediation in 3-level implementation designs. Second, no designs with combinations of small or medium effects for the a_3 and b_3 paths had adequate statistical power.

This indicates at least one large effect size for either the a_3 or b_3 path is needed to achieve adequate statistical power to test mediation. Third, the size of the level-3 ICC of the mediator (ICC_{m3}) is extremely important. When ICC_{m3} is small, there are no designs with adequate power except those that have large effect sizes for both a_3 and b_3 paths.

Discussion

Thought leaders and funders in the field of implementation science have increasingly called for a stronger focus on understanding implementation mechanisms (13,20,69,70), with methodologists pointing to mediation analysis as a recommended tool in this effort (5,22). Because statistical power to test mediation in multilevel designs depends on the specific range of input values that are feasible within a given research area, we estimated what sample sizes, effect sizes, and ICCs are required to detect mediation in 3-level implementation research designs. We estimated statistical power and sample size required to detect mediation using a range of input values feasible for implementation research. Designs were tested under four different conditions representing two statistical models (MVM vs. MSEM) and 1- versus 2-sided hypothesis tests (see Figure 3). Fewer than 5% of the designs studied had adequate statistical power to detect mediation. In almost all cases, the smallest number of level-3 clusters necessary to achieve adequate power was 40, the upper limit of what is possible in many implementation studies. This raises important questions about the feasibility of mediation analyses in implementation research as it is currently practiced. Enrolling 40 organizations usually requires substantial resources and may not be feasible within a limited geographic area or timeframe (21,37). In many settings, it also may not be possible to enroll enough *level-2* units per setting (e.g., nurses on a ward, primary care physicians in a practice, specialty mental health clinicians in a clinic) or level-1 units (e.g., patients per provider). Below, we discuss the implications of these findings for researchers, funders of research, and for the field.

Implications for Researchers

Implementation research commonly randomizes highest-level units to implementation strategies and measures characteristics of these units that may predict implementation, such as organizational climate or culture, organizational or team leadership, or prevailing policies or norms within geopolitical units. If researchers wish to study multilevel mediation, they must either obtain a large number of highest-level units or choose potential mediating variables that are likely to have large effects. While it is not known how often such level-3 independent variables have large effects on putative lower-level mediators, there are some encouraging data on the potential for large associations between lower-level mediators and lowest-level outcomes. For example, in a meta-analysis of 79 studies, Godin et al. found variables from social cognitive theories explained up to 81% of the variance in providers' intention to execute healthcare behaviors and 28% of the variance in physicians' behaviors, 24% of the variance in nurses' behavior, and 55% of the variance in other healthcare professionals' behavior (46). These effect sizes are comparable to or larger than the effect size for the b_3 path used in this study, suggesting that the variables proposed as antecedents to behavior in these theoretical models may serve as effective mediators linking level-3 independent variables to level-1 implementation outcomes.

Researchers can take steps to increase statistical power. One approach is to include a baseline covariate that is highly correlated with the outcome, ideally a pretest measure of the outcome itself, which can significantly increase statistical power, in some cases reducing the required sample size by 50% (28,29,47,48). The higher the correlation between the pretest covariate and the outcome, the lower the required sample size. Including a pretest of the mediator or outcome also increases the likelihood that the design meets the assumptions required to make causal inferences (49,50). However, whereas some settings like schools often have readily available pretests (e.g., academic achievement measures), pretests of implementation outcomes are not always available or may not make conceptual sense. For example, in implementation studies examining fidelity to a new practice, collecting pretest fidelity data may confuse participants because they have not yet learned the practice. Other approaches to increasing statistical power for indirect effects include using 1-sided hypothesis tests when appropriate (33), optimizing the reliability of measurement instruments (33), and using significance tests that likely to engender higher statistical power, such as the distribution of the product method or Monte Carlo confidence intervals (15). The chronic underuse of 1-sided hypothesis tests for indirect effects is puzzling considering they have significantly more power and are often justified by theory. Our results strongly support the use of 1-sided hypothesis tests for theory-informed multilevel mediation hypotheses.

Implications for Funders

Over the last decade, funding agencies like the US National Institutes of Health have made understanding the mechanisms by which interventions work part of funding announcements and the review process for implementation research (59). The implications of this requirement, combined with other requirements that call for tests of mediation and moderation (i.e., sex as a biological variable; the role of treatment fidelity on outcome [60,61]), place considerable demands on recruitment and measurement, even as the maximum budget for an R01 has not increased in almost 20 years. Funders may wish to change expectations for implementation research or emphasize trials that measure implementation outcomes but not clinical outcomes, which may allow for larger sample sizes at higher levels. Funders also may wish to develop funding mechanisms that provide additional funds to address the need for substantially larger sample sizes to test theories about mechanisms in multilevel contexts.

Implications for the Field

Our results are sobering, and cause for reflection about how implementation science as a field approaches research designs that elucidate how our implementation strategies result in change. First, our results suggest the need for immediate studies to help researchers make sample size decisions. Because implementation science is a relatively new discipline, little data are available for estimating ICCs for outcomes at different levels. The field needs studies that summarize a wide range of ICCs for many implementation and clinical outcomes and for mediation targets across settings, populations, and interventions. The field also needs research that clarifies how different formal tests for mediation influence power in multilevel models. Although some studies have tested the performance of mediation tests in multilevel models (24,28), much more work is needed. This line of research is especially important considering research from single-level models showing that some mediation tests display better balance between Type I error rates and statistical power (15).

Second, the field needs accurate measures of putative mediating variables. Increasing accuracy of measurement will increase our ability to observe effects (33). At present, the field does not have standardized ways to measure, for example, the constructs from cognitive theories often used as putative mediators (62). The field could benefit from close collaboration with experts in those areas to develop agreed upon (and then tested) measurement strategies.

Third, the field should consider implementation strategies that are less expensive to implement. The expense of many implementation strategies has been documented in the literature, raising questions about scalability (63-65). Less expensive strategies would increase our ability to test mechanism, but more importantly, increase the resources available to recruit more organizations into studies. Similarly, we should consider pragmatic trials that reduce measurement burden and allow us to enroll larger samples. Pragmatic trials differ from more traditional RCTs in that they can have more inclusive eligibility standards, the comparison condition, practitioner expertise and use of the intervention, primary outcome, and how these components are measured (66). The focus of pragmatic trials is highly consistent with the goal of implementation science in understanding strategies to increase use of evidence-based care in community practice and researchers have developed tools to describe the level of pragmatism in implementation trials (67).

Study Caveats and Limitations

Our results indicate that investigators are unlikely to detect mediation in 3-level studies with samples of less than 40 highest-level units under conditions that are feasible in implementation science, although examples of positive studies may occur. In those cases, our results provide important context for interpreting the exceptional study's results. First, low power to detect an effect does not mean it is impossible. Second, 3-level studies with samples of fewer than 40 highest-level units that do not detect mediation are likely never published, making the few published examples appear more common and representative than they are. Third, in some multilevel studies, indirect effects may be improperly specified and therefore statistically significant but not theoretically justified (19,26,51). Fourth, studies may compensate for low N_3 by having very large samples at other levels or higher effect sizes than those tested in our study.

The design parameters investigated in this study reflect a broad range of plausible values for 3-level designs in implementation research; however, there are undoubtedly important additional parameter values not studied here. We provide our code so investigators can study designs with other parameter values. The computational demands of bootstrapping and Monte Carlo confidence interval approaches led us to use the Sobel test for our study; consequently, power is likely to be slightly higher if investigators use these more powerful methods. Our study assessed mediation only in 3-2-1 designs that are broadly applicable to implementation science. Additional research should evaluate required sample sizes for power in other designs (e.g., 3-3-1, 3-1-1). To optimize potential generalizability and parsimony, our study did not include covariates in the mediation model; most notably, we did not include a pretest of the outcome. Covariates can reduce the required sample size to detect indirect effects (28) and future research is needed to characterize the types of pretest covariates that are available in implementation research as well as the strength of the relationship between these covariates and pertinent implementation and clinical outcomes as these will be important for study planning. Future research should also examine how unbalanced clusters influence power in multilevel mediation.

Conclusions

This study assesses the sample sizes needed to test mediation in 3-level designs that are typical and plausible in implementation science in healthcare. Results suggest large effect sizes coupled with 40 or more highest-level units are needed to test mediation. Innovations in research design are likely needed to increase the feasibility of studying mediation within the multilevel contexts common to implementation science.

List of abbreviations

EBP = evidence-based practice

MSEM = multilevel structural equation modeling

MVM = conventional linear multilevel regression analysis using manifest (observed) variables

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Additional File 1 provides Mplus code for running the simulations so investigators can recreate the results or produce new simulations for designs not studied here. Full results of the simulations are presented Additional Files 2 and 3.

Competing interests

The authors report no competing interests.

Funding

This work was funded in part through a grant from the U. S. National Institute of Mental Health (P50MH113840) to the University of Pennsylvania P50 ALACRITY Center for Transforming Mental Health Care Delivery through Behavioral Economics and Implementation Science (PIs: Mandell, Beidas, Volpp). The views expressed herein are those of the authors and do not necessarily reflect those of NIMH.

Authors' contributions

All authors (NJW, KJP, PDA, DM, SCM) contributed to conceptualization of the research questions and design of the study. KJP generated the simulation code for all analyses. SCM supervised completion of the simulations and tabulation of the results. NJW, KJP, PDA, DM, and SCM aided in interpretation of the simulation data. NJW drafted the initial manuscript. All authors (NJW, KJP, PDA, DM, SCM) revised the manuscript for essential intellectual content and approved the final manuscript.

Acknowledgments

The authors would like to thank Eliza Macneal for her assistance programming the simulations for this study.

Authors' Information

- Nathaniel J. Williams, Ph.D. is Associate Professor in the School of Social Work and Research and Evaluation Coordinator in the Institute for the Study of Behavioral Health and Addiction at Boise State University, Boise, ID, natewilliams@boisestate.edu
- Kristopher J. Preacher, Ph.D. is Professor of Quantitative Methods and Associate Chair of the Department of Psychology & Human Development at Vanderbilt University, Nashville, TN, kris.preacher@vanderbilt.edu
- Paul D. Allison, Ph.D. is President of Statistical Horizons LLC, Ardmore, PA, allison@statisticalhorizons.com
- David Mandell, Sc.D. is the Kenneth E. Appel Professor of Psychiatry and the Director of the Penn Center for Mental Health at the University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, mandell@upenn.edu
- Steven C. Marcus, Ph.D. is Research Associate Professor in the School of Social Policy & Practice at the University of Pennsylvania, Philadelphia, PA, marcuss@upenn.edu

References

1. Eccles MP, Mittman BS. Welcome to Implementation Science. Implement. Sci. 2006;1:1. doi:10.1186/1748-5908-1-1.
2. Nilsen P. Making sense of implementation theories, models and frameworks. Implement. Sci. 2015;10:53. doi:10.1186/s13012-015-0242-0.
3. Williams NJ, Beidas RS. Annual Research Review: The state of implementation science in child psychology and psychiatry: a review and suggestions to advance the field. J Child Psychol Psychiatry. 2019;60:430-450. doi:10.1111/jcpp.12960.

4. Lewis CC, Boyd MR, Walsh-Bailey C, Lyon AR, Beidas R, Mittman B, Aarons GA, Weiner BJ, Chambers DA. A systematic review of empirical studies examining mechanisms of implementation in health. Implement. Sci. 2020;15:21. doi:10.1186/s13012-020-00983-3.
5. Williams NJ. Multilevel mechanisms of implementation strategies in mental health: integrating theory, research, and practice. Adm Policy Ment Health. 2016;43:783-798. doi:10.1007/s10488-015-0693-2.
6. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J Pers Soc Psychol. 1986;51:1173-1182. doi:10.1037/0022-3514.51.6.1173.
7. MacKinnon DP. Introduction to statistical mediation analysis. Routledge; 2007.
8. Hayes AF. Introduction to mediation, moderation, and conditional process analysis: a regression-based approach. 1st ed. Guilford Publications; 2017.
9. Preacher KJ. Advances in mediation analysis: A survey and synthesis of new developments. Annu Rev Psychol. 2015;66:825-852. doi:10.1146/annurev-psych-010814-015258.
10. VanderWeele T. Explanation in causal inference: Methods for mediation and interaction. Oxford University Press; 2015.
11. Insel TR. The NIMH experimental medicine initiative. World Psychiatry. 2015;14:151.
12. Lewandowski KE, Ongur D, Keshavan MS. Development of novel behavioral interventions in an experimental therapeutics world: challenges, and directions for the future. Schizophr Res. 2018;192:6-8.
13. Nielsen L, Riddle M, King JW, Aklin WM, Chen W, Clark D, Collier E, Czajkowski S, Esposito L, Ferrer R. The NIH science of behavior change program: transforming the science through a focus on mechanisms of change. Behav Res Ther. 2018;101:3-11.
14. Fritz MS, Mackinnon DP. Required sample size to detect the mediated effect. Psychol Sci. 2007;18:233-239. doi:10.1111/j.1467-9280.2007.01882.x.
15. Hayes AF, Scharkow M. The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: does method really matter? Psychol Sci. 2013;24:1918-1927.
16. MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. Psychol Methods. 2002;7:83-104. doi:10.1037/1082-989X.7.1.83.
17. Schoemann AM, Boulton AJ, Short SD. Determining power and sample size for simple and complex mediation models. Social Psychol Personality Sci. 2017;8:379-386. doi.org/10.1177/1948550617715068.
18. Thoemmes F, Mackinnon DP, Reiser MR. Power analysis for complex mediational designs using monte carlo methods. Struct Equ Modeling. 2010;17:510-534. doi:10.1080/10705511.2010.489379.
19. Zhang Z, Zyphur MJ, Preacher KJ. Testing multilevel mediation using hierarchical linear models: problems and solutions. Organ Res Methods. 2009;12:695-719.
20. Weiner BJ, Lewis MA, Clauser SB, Stitzenberg KB. In search of synergy: strategies for combining interventions at multiple levels. J Natl Cancer Inst Monogr. 2012;34-41.
21. Mazzucca S, Tabak RG, Pilar M, Ramsey AT, Baumann AA, Kryzer E, Lewis EM, Padek M, Powell BJ, Brownson RC. Variation in research designs used to test the effectiveness of dissemination and implementation strategies: a review. Front Public Health. 2018;6:32. doi:10.3389/fpubh.2018.00032.
22. Wolfenden L, Foy R, Presseau J, Grimshaw JM, Ivers NM, Powell BJ, Taljaard M, Wiggers J, Sutherland R, Nathan N, Williams CM, Kingsland M, Milat A, Hodder RK, Yoong SL.

Designing and undertaking randomised implementation trials: guide for researchers. BMJ. 2021;372:m3721. doi:10.1136/bmj.m3721.

23. Krull JL, MacKinnon DP. Multilevel modeling of individual and group level mediated effects. Multivariate Behav Res. 2001;36:249-277. doi: 10.1207/S15327906MBR3602_06.
24. Pituch KA, Murphy DL, Tate RL. Three-level models for indirect effects in school-and class-randomized experiments in education. Journal of Experimental Education. 2009;78:60-95.
25. Preacher KJ. Multilevel SEM strategies for evaluating mediation in three-level data. Psychol Methods. 2011;46:691-731. doi.org/10.1080/00273171.2011.589280.
26. Preacher KJ, Zyphur MJ, Zhang Z. A general multilevel SEM framework for assessing multilevel mediation. Psychol Methods. 2010;15:209. doi.org/10.1037/a0020141.
27. Kelcey B, Spybrook J, Dong N. Sample size planning for cluster-randomized interventions probing multilevel mediation. Prev Sci. 2019;20:407-418. doi: 10.1007/s11121-018-0921-6.
28. Kelcey B, Xie Y, Spybrook J, Dong N. Power and sample size determination for multilevel mediation in three-level cluster-randomized trials. Multivariate Behav Res. 2021;56:496-513. doi: 10.1080/00273171.2020.1738910.
29. Raudenbush SW, Bryk AS. Hierarchical linear models: Applications and data analysis methods. Vol. 1. Sage; 2002.
30. Muthén LK, Muthén BO. How to use a Monte Carlo study to decide on sample size and determine power. Struct Equ Modeling. 2002;9:599-620. doi.org/10.1207/S15328007SEM0904_8.
31. Skrondal A. Design and Analysis of Monte Carlo Experiments: Attacking the Conventional Wisdom. Multivariate Behav Res. 2000;35:137-67. doi:10.1207/s15327906mbr3502_1.
32. Zhang Z. Monte Carlo based statistical power analysis for mediation models: Methods and software. Behav res methods. 2014;46:1184-1198.
33. Cohen J. A power primer. Psychol Bull. 1992;112:155-9. doi:10.1037//0033-2909.112.1.155.
34. Ben Charif A, Croteau J, Adekpedjou R, Zomahoun HTV, Adisso EL, Légaré F. Implementation research on shared decision making in primary care: inventory of intracluster correlation coefficients. Med Decis Making. 2019;39:661-672. doi:10.1177/0272989x19866296.
35. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. Clin Trials. 2005;2:99-107. doi:10.1191/1740774505cn071oa.
36. Forman-Hoffman VL, Middleton JC, McKeeman JL, Stambaugh LF, Christian RB, Gaynes BN, Kane HL, Kahwati LC, Lohr KN, Viswanathan M. Quality improvement, implementation, and dissemination strategies to improve mental health care for children and adolescents: a systematic review. Implement Sci. 2017;12:93. doi:10.1186/s13012-017-0626-4.
37. Novins DK, Green AE, Legha RK, Aarons GA. Dissemination and implementation of evidence-based practices for child and adolescent mental health: a systematic review. J Am Acad Child Adolesc Psychiatry. 2013;52:1009-1025.e18. doi:10.1016/j.jaac.2013.07.012.
38. Powell BJ, Proctor EK, Glass JE. A systematic review of strategies for implementing empirically supported mental health interventions. Res Soc Work Pract. 2014;24:192-212. doi:10.1177/1049731513505778.
39. Rabin BA, Glasgow RE, Kerner JF, Klump MP, Brownson RC. Dissemination and implementation research on community-based cancer prevention: a Systematic Review. Am J Prev Med. 2010;38:443-456. doi: 10.1016/j.amepre.2009.12.035.

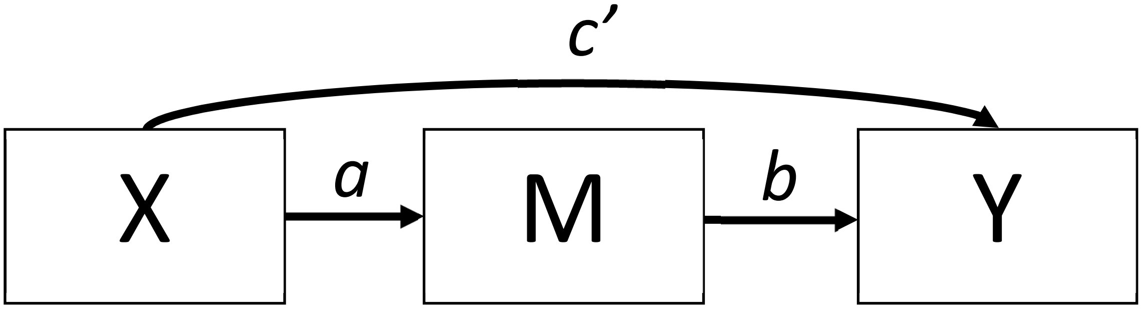
40. Muthén LK, Muthén BO. Mplus user's guide: Statistical analysis with latent variables. 8 ed. Muthén & Muthén; 2017.
41. Sobel ME. Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology. 1982;13:290-312. doi.org/10.2307/270723.
42. Flory JH, Mushlin AI, Goodman ZI. Proposals to conduct randomized controlled trials without informed consent: a narrative review. J Gen Intern Med. 2016;31:1511-1518. doi:10.1007/s11606-016-3780-5.
43. Sim J, Dawson A. Informed consent and cluster-randomized trials. Am J Public Health. 2012;102:480-485. doi:10.2105/AJPH.2011.300389.
44. Zelen M. Randomized consent designs for clinical trials: an update. Stat Med. 1990;9:645-56. doi:10.1002/sim.4780090611.
45. Relton C, Torgerson D, O'Cathain A, Nicholl J. Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design. BMJ. 2010;340:c1066. doi:10.1136/bmj.c1066.
46. Godin G, Bélanger-Gravel A, Eccles M, Grimshaw J. Healthcare professionals' intentions and behaviours: a systematic review of studies based on social cognitive theories. Implement Sci. 2008;3:36. doi:10.1186/1748-5908-3-36.
47. Bloom HS, Richburg-Hayes L, Black AR. Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. Educ Eval and Policy Anal. 2007;29:30-59. doi.org/10.3102/0162373707299550.
48. Konstantopoulos S. The impact of covariates on statistical power in cluster randomized designs: which level matters more? Multivariate Behav Res. 2012;47:392-420. doi: 10.1080/00273171.2012.673898.
49. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychol Methods. 2010;15:309-334. doi:10.1037/a0020761.
50. VanderWeele TJ. Mediation analysis: a practitioner's guide. Annu Rev Public Health. 2016;37:17-32.
51. Cronbach LJ. Research on classrooms and schools: Formulation of questions, design, and analysis. Stanford University Evaluation Consortium; 1976.
52. Gonzalez-Roma V, Hernandez A. Conducting and evaluating multilevel studies: Recommendations, resources, and a checklist. Org Res Methods. 2022. doi: 10.1177/10944281211060712.
53. Snijders TA, Bosker RJ. Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sage; 2011.
54. Preacher KJ, Zhang Z, Zyphur MJ. Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. Structural Equation Modeling. 2011;18:161-82. doi.org/10.1080/10705511.2011.557329.
55. Bollen KA. Structural equations with latent variables. John Wiley & Sons; 1989.
56. Lüdtke O, Marsh HW, Robitzsch A, Trautwein U, Asparouhov T, Muthén B. The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. Psychological Methods. 2008;13:203-229.
57. Efron B, Tibshirani TJ. An introduction to the bootstrap. Chapman & Hall; 1993.
58. Preacher KJ, Selig JP. Advantages of Monte Carlo confidence intervals for indirect effects. Communication Methods and Measures. 2012;6:77-98.

59. Insel T. NIMH's new focus in clinical trials. www.nimh.nih.gov/funding/grant-writing-and-application-process/concept-clearances/2013/nimhs-new-focus-in-clinical-trials. Accessed 6 June 2022.
60. US National Institute of Mental Health. Consideration of sex as a biological variable in NIH-funded research. https://grants.nih.gov/grants/guide/notice-files/not-od-15-102.html. Accessed 6 June 2022.
61. US National Institute of Mental Health. Dissemination and implementation research in health (R01 clinical trial optional) PAR-22-105. https://grants.nih.gov/grants/guide/pa-files/PAR-22-105.html. Accessed 6 June 2022.
62. Fishman J, Yang C, Mandell D. Attitude theory and measurement in implementation science: a secondary review of empirical studies and opportunities for advancement. Implement Sci. 2021;16(1):87. doi: 10.1186/s13012-021-01153-9.
63. Cidav Z, Mandell D, Pyne J, Beidas R, Curran G, Marcus S. A pragmatic method for costing implementation strategies using time-driven activity-based costing. Implement Sci. 2020;15(1):28. doi:10.1186/s13012-020-00993-1.
64. Saldana L, Ritzwoller DP, Campbell M, Block EP. Using economic evaluations in implementation science to increase trans-parency in costs and outcomes for organizational decision-makers. Implement Sci Commun. 2022;3(1):40. doi: 10.1186/s43058-022-00295-1.
65. Dopp AR, Kerns SEU, Panattoni L, Ringel JS, Eisenberg D, Powell BJ, Low R, Raghavan R. Translating economic evaluations into financing strategies for implementing evidence-based practices. Implement Sci. 2021;16(1):66. doi: 10.1186/s13012-021-01137-9.
66. Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. Trials. 2009;10:37. doi.org/10.1186/1745-6215-10-37.
67. Norton, W.E., Loudon, K., Chambers, D.A. et al. Designing provider-focused implementation trials with purpose and intent: introducing the PRECIS-2-PS tool. Implementation Sci. 2021;16:7. doi.org/10.1186/s13012-020-01075-y.
68. Boomsma A. Reporting Monte Carlo simulation studies in structural equation modeling. Structural Equation Modeling. 2013;20:518-540. doi: 10.1080/10705511.2013.797839.
69. Lewis CC, Powell BJ, Brewer SK, Nguyen AM, Schriger SH, Vejnoska SF, Walsh-Bailey C, Aarons GA, Beidas RS, Lyon AR, Weiner B. Advancing mechanisms of implementation to accelerate sustainable evidence-based practice integration: protocol for generating a research agenda. BMJ Open. 2021 Oct 1;11(10):e053474.
70. Grol RP, Bosch MC, Hulscher ME, Eccles MP, Wensing M. Planning and studying improvement in patient care: the use of theoretical perspectives. The Milbank Quarterly. 2007 Mar;85(1):93-138.
71. Beidas, R.S., Ahmedani, B., Linn, K.A. *et al.* Study protocol for a type III hybrid effectiveness-implementation trial of strategies to implement firearm safety promotion as a universal suicide prevention strategy in pediatric primary care. Implementation Sci. 2021; 16**(**89). https://doi.org/10.1186/s13012-021-01154-8
72. Kohrt, B.A., Turner, E.L., Gurung, D. *et al.* Implementation strategy in collaboration with people with lived experience of mental illness to reduce stigma among primary care providers in Nepal (RESHAPE): protocol for a type 3 hybrid implementation effectiveness cluster randomized controlled trial. Implementation Sci. 2022;17(39). https://doi.org/10.1186/s13012-022-01202-x
73. Cumbe, V.F.J., Muanido, A.G., Turner, M. *et al.* Systems analysis and improvement approach to optimize outpatient mental health treatment cascades in Mozambique (SAIA-

MH): study protocol for a cluster randomized trial. Implementation Sci. 2022; 17(37). https://doi.org/10.1186/s13012-022-01213-8

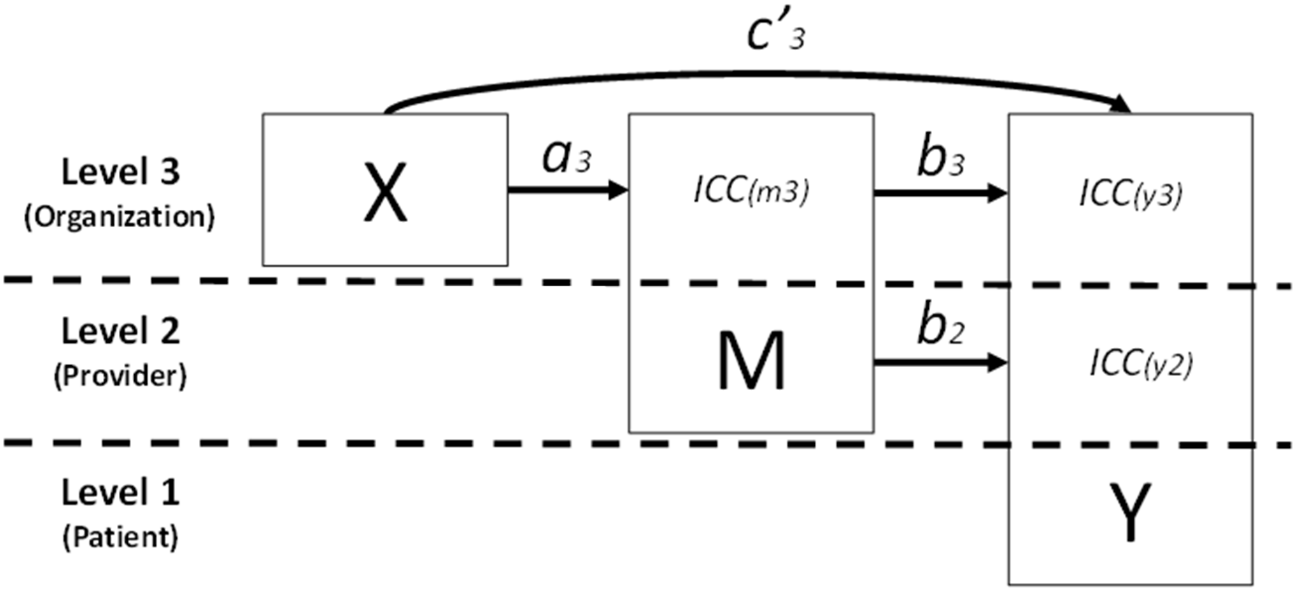
74. Swindle, T., Rutledge, J.M., Selig, J.P. *et al.* Obesity prevention practices in early care and education settings: an adaptive implementation trial. Implementation Sci. 2022;17(25). https://doi.org/10.1186/s13012-021-01185-1
75. McIntyre SA, Francis JJ, Gould NJ, Lorencatto F. The use of theory in process evaluations conducted alongside randomized trials of implementation interventions: a systematic review. Transl Behav Med. 2020;10:168-78.
76. Cashin AG, McAuley JH, Lee H. Advancing the reporting of mechanisms in implementation science: A guideline for reporting mediation analyses (AGReMA). Implementation Research and Practice. 2022;3:26334895221105568.
77. Wang X, Turner EL, Preisser JS, Li F. Power considerations for generalized estimating equations analyses of four-level cluster randomized trials. Biometrical Journal. 2022;64(4):663-80.
78. Aarons, G.A., Ehrhart, M.G., Moullin, J.C. *et al.* Testing the leadership and organizational change for implementation (LOCI) intervention in substance abuse treatment: a cluster randomized trial study protocol. Implementation Sci. 2017; 12(29). https://doi.org/10.1186/s13012-017-0562-3
79. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. Evaluation Review. 2003;27(1):79-103.
80. Muthén BO, Muthén LK, Asparouhov T. *Regression and mediation analysis using Mplus*. 2017. Los Angeles, CA: Muthén & Muthén.

Figure 1. Single-level mediation model.



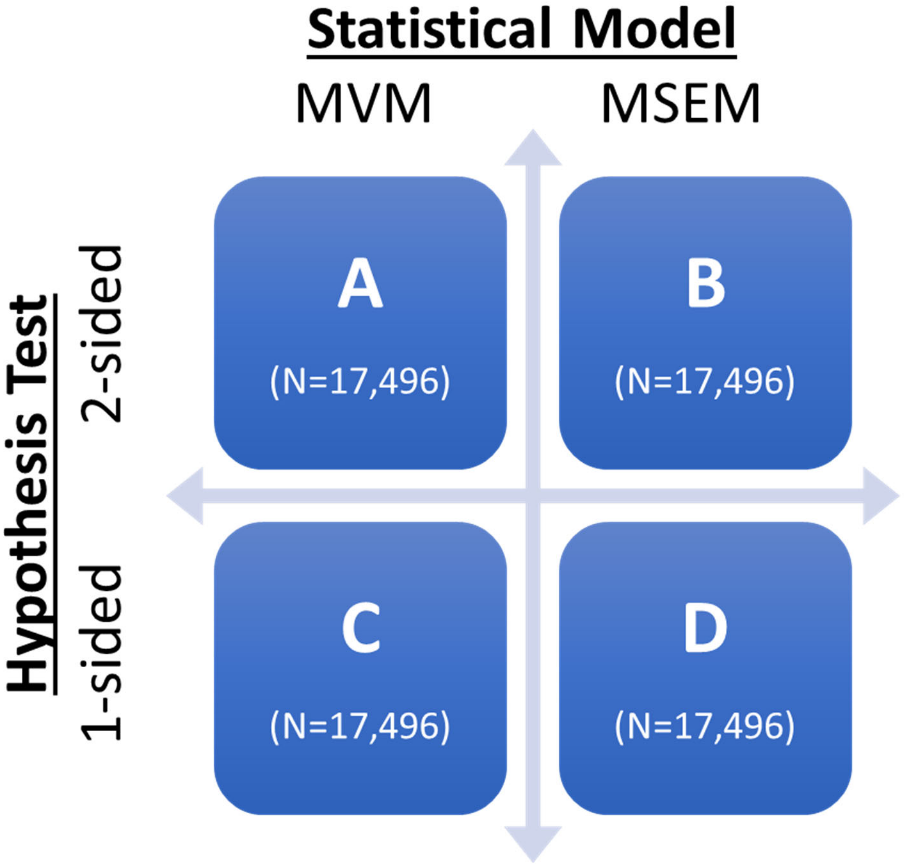
Note: X = independent variable; M = mediator, Y = outcome. The indirect effect is estimated as the product of the *a* and *b* paths (i.e., *a***b*). The *c'* path represents the direct effect of X on Y (i.e., the effect of X on Y that is not transmitted through the mediator, M).

Figure 2. Multilevel mediation model (3-2-1).



Note: The diagram presents the 3-2-1 mediation design for which statistical power was calculated in this study. The boxes signify each construct in the design and show the levels at which the construct exhibits variance: X = independent variable which varies only at level 3; M = mediator which resides at level 2 but exhibits variance at levels 2 and 3 (due to clustering); Y = outcome which resides at level 1 but exhibits variance at levels 1, 2, and 3 (due to clustering). The variance of M and Y at the higher levels of analysis are represented by ICC values. Arrows indicate effects that can be estimated through conventional multilevel regression (MVM) (19) or through multilevel structure equation modeling (MSEM) (26). The paths that make up the indirect effect (i.e., mediation at level-3) are a_3*b_3. The c'_3 path represents the direct effect. The b_2 path is typically not of substantive interest; it represents the relationship between the within-organization component of M and within-organization component of Y.

Figure 3. Statistical models and types of hypothesis tests studied.



Note: We conducted statistical power simulations for 17,496 implementation research designs under four different conditions. The conditions represent a fully crossed matrix of two different statistical models (traditional multilevel modeling of manifest variables [MVM] vs. multilevel structural equation modeling [MSEM]) and two different hypothesis tests for the mediation effect (1- and 2-sided).

Table 1. Frequency of study designs with statistical power ≥ 0.8 by study characteristic (*N* = 17,496 designs).

Study Characteristic	**Parameter value**	**Total *N* of designs**	***N* of adequately powered designs (>= .8)**	**Proportion of adequately powered designs (% >=.8)**
Total		17,496	463	2.6%
a_3 (Standardized X→M coefficient)	0.14	5,832	0	0.0%
	0.39	5,832	46	0.8%
	0.59	5,832	417	7.2%
b_3 (Standardized M→Y coefficient)	0.14	5,832	0	0.0%
	0.39	5,832	32	0.5%
	0.59	5,832	431	7.4%
c'_3 (Standardized direct effect)	0.14	8,748	161	1.8%
	0.39	8,748	302	3.5%
ICC_{M3} (Level-3 intraclass correlation coefficient for mediator M)	0.05	5,832	36	0.6%
	0.10	5,832	125	2.1%
	0.20	5,832	302	5.2%
ICC_{Y2} (Level-2 intraclass correlation coefficient for outcome Y)	0.05	5,832	136	2.3%
	0.10	5,832	148	2.5%
	0.20	5,832	179	3.1%
ICC_{Y3} (Level-3 intraclass correlation coefficient for outcome Y)	0.05	5,832	140	2.4%
	0.10	5,832	161	2.8%
	0.20	5,832	162	2.8%
N_3 (Level-3 sample size; *N* of highest-level units/ clusters, e.g., organizations)	10	4,374	0	0.0%
	20	4,374	0	0.0%
	40	4,374	110	2.5%
	60	4,374	353	8.1%
N_2 (Level-2 sample size; *N* of nested intermediate-level units per cluster, e.g., providers)	5	5,832	42	0.7%
	10	5,832	129	2.2%
	20	5,832	292	5.0%
N_1 (Level-1 sample size; *N* of nested lowest-level units per intermediate unit, e.g., patients)	3	5,832	131	2.2%
	6	5,832	159	2.7%
	12	5,832	173	3.0%
Total Sample Size (N_3* N_2* N_1)	150	486	0	0.0%
	300	1,458	0	0.0%
	600	2,916	0	0.0%
	900	486	9	1.9%
	1,200	3,402	8	0.2%

1,800	972	43	4.4%
2,400	2,916	33	1.1%
3,600	1,458	112	7.7%
4,800	1,458	39	2.7%
7,200	972	116	11.9%
9,600	486	30	6.2%
14,400	486	73	15.0%

Note: Power was calculated for N = 17,496 designs based on Monte Carlo simulations (500 replications per design) conducted in Mplus 8. All models represent 3-2-1 mediation designs estimated using maximum likelihood with robust standard errors based on a linear multilevel model with manifest variables (MVM). For each design, power was calculated as the proportion of replications (out of 500) for which the null hypothesis, H_0: $a_3 * b_3 = 0$, was rejected based on the Sobel test, assuming $\alpha = 0.05$ (two-tailed).

Table 2. Minimum sample sizes required for adequate statistical power to detect mediation.

		Standardized Effect Sizes for a_3 path (X→M) and b_3 path (M→Y)								
ICC_{m3}	ICC_{y3}	SS	SM	SL	MS	MM	ML	LS	LM	LL
S	S	-	-	-	-	-	-	-	-	c_s: 60[20[6]] c_m: 60[20[3]]
S	M	-	-	-	-	-	-	-	-	c_s: 60[20[6]] c_m: 60[20[3]]
S	L	-	-	-	-	-	-	-	-	c_s: 60[20[6]] c_m: 60[20[3]]
M	S	-	-	-	-	-	-	-	c_s: - c_m: 60[20[6]]	c_s: 40[20[12]] c_m: 40[20[3]]
M	M	-	-	-	-	-	-	-	c_s: - c_m: 60[20[6]]	c_s: 40[20[6]] c_m: 40[20[3]]
M	L	-	-	-	-	-	-	-	-	c_s: 40[20[12]] c_m: 40[20[3]]
L	S	-	-	-	-	-	c_s: 60[20[3]] c_m: 60[20[3]]	-	c_s: - c_m: 60[10[12]]	c_s: 40[20[3]] c_m: 40[10[3]]
L	M	-	-	-	-	-	c_s: 60[20[3]] c_m: 60[10[12]]	-	c_s: - c_m: 60[10[6]]	c_s: 40[10[12]] c_m: 40[10[3]]
L	L	-	-	-	-	-	c_s: 60[20[3]] c_m: 60[10[12]]	-	c_s: - c_m: 60[10[6]]	c_s: 40[10[12]] c_m: 40[5[6]]

Note: Sample sizes shown are the smallest sample size required to achieve statistical power $\geq$ 0.8 to reject the null hypothesis $a_3*b_3 = 0$ given the design parameters shown. Within each cell, two sample sizes are provided, one assuming a small direct effect (c_s) and the other assuming a medium direct effect (c_m). Sample sizes are presented as $N_3[N_2[N_1]]$ where N_3 = number of highest-level clusters (level-3), N_2 = number of intermediate nested units (level-2) per cluster, and N_1 = number of lowest-level nested observations (level-1) per level-2 unit Blank cells (-) indicate there were no sample sizes that achieved adequate power for that design. Required samples sizes were generated using linear multilevel modeling with manifest variables assuming α=0.05 (2-tailed). ICC_{m3} = level-3 intraclass correlation coefficient of the mediator, ICC_{y3} = level-3 intraclass correlation coefficient of the outcome. ICCs were evaluated at S=0.05, M=.1, L=.2. Standardized effect sizes indicate the size of the a_3 path followed by the size of the b_3 path, where S=.14, M=.39, and L=.59.