Original Empirical Research

Implementation Research & Practice

Implementation Research and Practice
Volume 5: Jan-Dec 2024 1–20

DOI: 10.1177/26334895241279153
journals.sagepub.com/home/irp

Sage

Calculating power for multilevel implementation trials in mental health: Meaningful effect sizes, intraclass correlation coefficients, and proportions of variance explained by covariates

Nathaniel J. Williams[1,2], Nicholas C. Cardamone[3], Rinad S. Beidas[4] and Steven C. Marcus[5]

Abstract

Background

Despite the ubiquity of multilevel sampling, design, and analysis in mental health implementation trials, few resources are available that provide reference values of design parameters (e.g., effect size, intraclass correlation coefficient [ICC], and proportion of variance explained by covariates [covariate R^2]) needed to accurately determine sample size. The aim of this study was to provide empirical reference values for these parameters by aggregating data on implementation and clinical outcomes from multilevel implementation trials, including cluster randomized trials and individually randomized repeated measures trials, in mental health. The compendium of design parameters presented here represents plausible values that implementation scientists can use to guide sample size calculations for future trials.

Method

We searched NIH RePORTER for all federally funded, multilevel implementation trials addressing mental health populations and settings from 2010 to 2020. For all continuous and binary implementation and clinical outcomes included in eligible trials, we generated values of effect size, ICC, and covariate R^2 at each level via secondary analysis of trial data or via extraction of estimates from analyses in published research reports. Effect sizes were calculated as Cohen *d*; ICCs were generated via one-way random effects ANOVAs; covariate R^2 estimates were calculated using the reduction in variance approach.

Results

Seventeen trials were eligible, reporting on 53 implementation and clinical outcomes and 81 contrasts between implementation conditions. Tables of effect size, ICC, and covariate R^2 are provided to guide implementation researchers in power analyses for designing multilevel implementation trials in mental health settings, including two- and three-level cluster randomized designs and unit-randomized repeated-measures designs.

[1]Institute for the Study of Behavioral Health and Addiction, Boise State University, Boise, ID, USA
[2]School of Social Work, Boise State University, Boise, ID, USA
[3]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4]Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
[5]School of Social Policy and Practice, University of Pennsylvania, Philadelphia, PA, USA

Corresponding author:
Nathaniel J. Williams, Institute for the Study of Behavioral Health and Addiction, Boise State University, 1910 University Drive, Boise, ID 83725, USA.
Email: natewilliams@boisestate.edu

Conclusions

Researchers can use the empirical reference values reported in this study to develop meaningful sample size determinations for multilevel implementation trials in mental health. Discussion focuses on the application of the reference values reported in this study.

Plain Language Summary: To improve the planning and execution of implementation research in mental health settings, researchers need accurate estimates of several key metrics to help determine what sample size should be obtained at each level of a multi-level study (e.g., number of patients, doctors, and clinics). These metrics include the (1) effect size, which indicates how large of a difference in the primary outcome is expected between a treatment and control group, (2) intraclass correlation coefficient, which describes how similar two people in the same group might be, and (3) covariate R^2, which indicates how much of the variability in an outcome is explained by a background variable, such as level of health at the start of a study. We collected data from mental health implementation trials conducted between 2010 and 2020. We extracted information about each of these metrics and aggregated the results for researchers to use in planning their own studies. Seventeen trials were eligible, and we were able to obtain statistical information on 53 different outcome variables from these studies. We provide a set of values which will assist in sample size calculations for future mental health implementation trials.

Keywords

implementation research, multilevel power analysis, sample size, cluster randomized trial, hybrid effectiveness-implementation trial, mental health, intraclass correlation coefficient, effect size, covariate R^2

Background

Statistical power analysis is essential to designing implementation trials that achieve the dual goals of robustly testing scientific hypotheses while minimizing the use of finite scientific resources (Cohen, 2013; Faul et al., 2009; Kraemer & Blasey, 2016; Lipsey, 1995; Muthén & Muthén, 2002; Sedlmeier & Gigerenzer, 1989). Using power analysis, investigators determine the smallest sample size required to statistically detect the effect of interest (Hedges & Hedberg, 2007). If the sample is too small, statistical analyses will not detect an effect even if one exists (Cohen, 1992; Rosenthal et al., 1994). Conversely, enrolling a larger sample than necessary wastes resources and unnecessarily burdens participants. Balancing these priorities is especially important in implementation trials because the cost of recruiting and retaining sites and participants is often high.

Sample size calculation for implementation trials is often complicated by the multilevel context within which healthcare is delivered (Eccles & Mittman, 2006; Novins et al., 2013; Proctor et al., 2009). This multilevel context (e.g., patients nested within clinicians nested within clinics) generates dependencies within the data which result in incorrect statistical inferences if they are not accounted for by the statistical models (e.g., mixed effects; generalized estimating equations) (Goldstein, 2013; Snijders & Bosker, 1999). The use of these models requires specialized procedures for determining the necessary sample size at each level of the design.

Statistical theory for computing required sample sizes in multilevel designs is well-developed (Bell et al., 2008; Hox et al., 2018; Konstantopoulos, 2008b, 2008a; Raudenbush, 1997; Raudenbush & Liu, 2000; Scherbaum & Ferreter, 2009; Snijders, 2014) and software tools are available (Bhaumik et al., 2008; De Jong et al., 2010; Raudenbush & Liu, 2000; Zhang & Wang, 2009). However, these tools require investigators to input multiple design parameters which are often unknown, yet highly influential. Key design parameters for multilevel trials include (1) the anticipated effect size of the implementation strategy, which quantifies the expected standardized effect of the strategy relative to control at the study endpoint, (2) the intraclass correlation coefficient (ICC or ρ), which quantifies the degree of within-unit correlation or dependency among individuals nested within the same unit (e.g., patients nested within a provider); and (3) the proportion of variance explained by covariates (covariate R^2), which indicates how much of the variance in the outcome is explained by baseline control variables, such as participant demographic characteristics or pretest scores on the outcome (Snijders, 2014; Raudenbush & Liu, 2000). Obtaining accurate estimates of these design parameters is essential because relatively small changes in their values substantially influence the required sample size. For example, minor variations in effect size (Cohen's $d =$ .25, .35, .45) and ICC ($\rho = .02$, .08, .12) can alter the number of sites required for adequate power from 21 to more than 99.

For any trial, the choice of which design parameter estimates to use in power analysis should be guided by estimates obtained from similar prior research and knowledge of the minimum clinically important differences in the given context. For example, the effect size estimate would ideally reflect the minimum clinically

important benefit, with guidance from prior research about what effect sizes are reasonable given the substantive area. Unfortunately, design parameter estimates are rarely included in published research reports (Eldridge et al., 2004; Isaakidis, 2003), particularly in the still nascent field of implementation science (Pinnock et al., 2017; Wilson et al., 2017), despite clear guidance from the Consolidated Standards of Reporting Trials (CONSORT) extension for cluster randomized trials (Campbell et al., 2004). For example, in reviewing the effect of the CONSORT extension, Ivers et al. (2011) found that only 18% of 300 cluster randomized trials reported ICC values. Recent reviews suggest reporting rates are still low (Offorha et al., 2022). This leads investigators to develop power analyses based on crude rules of thumb (Killip, 2004) or on unreliable estimates from small pilot studies (Hedges & Hedberg, 2007; Killip, 2004; Kraemer et al., 2006; Leon et al., 2011). Many research proposals rely on Cohen's small, medium, and large effect size 'guidelines' even though Cohen strongly cautioned against using his guidelines as benchmarks (Cohen, 1988). Empirical studies confirm the importance of discipline-specific design parameter values by showing that these values vary greatly by discipline and outcome (Bosco et al., 2015; Dong et al., 2016; Hedges & Hedberg, 2007, 2013).

Discipline-specific plausible values of design parameters are needed to determine sample sizes for implementation trials in mental health but are currently not available. The goal of this study was to fill that gap by systematically identifying multilevel implementation trials focused on mental health evidence-based interventions (EBIs) and extracting from the analysis of those trials values of effect size, ICC, and covariate R^2 for all reported continuous and binary implementation and clinical outcomes. We sought to summarize the central tendency and range of these design parameters in order to provide plausible reference values for implementation scientists to determine sample size in their own trials in mental health settings. In addition, we examined the relationship between the magnitude of each of the design parameters and selected study characteristics (e.g., outcome type, measurement approach) with the goal of providing investigators with more nuanced guidance for designing trials.

Method

We identified potentially eligible mental health implementation trials through searches of NIH RePORTER, a congressionally mandated, web-based repository of NIH-funded research in the USA. RePORTER allows users to identify trials funded by specific Institutes/Centers (i.e., National Institute of Mental Health (NIMH); in response to specific funding opportunity announcements; by key words; by project type (e.g., research grants, centers); and for specific start and end dates. To ensure the capture of relevant trials, four searches were conducted using different functions: (1) key word search, (2) funding opportunity announcement search for dissemination- and implementation-related proposals, (3) study section search (e.g., Dissemination and Implementation Research in Health, now known as Science of Implementation in Health and Healthcare), and (4) grant mechanism search. With help from a reference librarian and RePORTER staff, we used key words to optimize trial discoverability. See Supplemental File 1 for search parameters.

Eligibility Criteria

We limited our search to trials funded by the US NIMH with the assumption that federally funded trials were more likely to be appropriately powered and rigorously designed. Trials were included if they (1) randomly assigned units (e.g., persons, teams, clinics) to implementation strategies, (2) collected quantitative data on one or more implementation outcome or clinical outcome, (3) included nested or clustered observations, and (4) ended from January 1, 2010 to December 30, 2020.

Trials were excluded if they (1) only examined feasibility, acceptability, or appropriateness of an EBI, (2) only examined feasibility or acceptability of an implementation strategy, (3) only examined technology-based delivery versus in-person delivery of an EBI, (4) focused on HIV/AIDS (in the US, funding for HIV/AIDS is provided under the umbrella of NIMH), (5) were administratively terminated by NIMH prior to study completion (and therefore lacked data), and (6) were clinical effectiveness trials or hybrid type I effectiveness-implementation trials that only tested a clinical intervention versus a no-treatment control group. The latter were excluded because they did not have comparative data on implementation strategy effects.

Study Identification

Title and Abstract Screen

Two reviewers (NW and a doctoral-level social work researcher) screened all project titles and abstracts to determine potential eligibility. Reviewers met weekly to compare inclusion/exclusion decisions; disagreements were resolved through discussion with a third author (SM). In cases where the abstract was unclear, trials were retained to avoid premature exclusion of eligible trials.

Research Report Identification and Selection

Each NIH RePORTER project webpage lists all published articles which acknowledge the project's support (https://report.nih.gov/faqs). We extracted all articles linked with potentially eligible projects. Duplicate articles, non-empirical, and review articles were excluded. Two members of the research team (NC and a doctoral-level researcher) screened titles and abstracts of the remaining

articles to identify those eligible for full-text retrieval and further screening. Disagreements on screening decisions were resolved through consensus. If no articles relevant to the study's primary aims were located from RePORTER, we searched for publications on ClinicalTrials.gov, PubMed, Google Scholar, and professional webpages of the study's principal investigator. Articles that met the criteria from these sources were downloaded as full texts to DistillerSR (*DistillerSR, 2023*), a web-based systematic review management software, for additional screening (see the study flow chart in Supplemental File 2).

We included all articles that (1) had a primary aim that tested the effects of one or more implementation strategies on implementation or clinical outcomes, or (2) reported quantitative data on implementation determinants in relation to implementation or clinical outcomes in mental health settings. Articles were excluded if there was no evidence of a clustered design, no randomization to conditions, or no quantitative outcomes. We also excluded articles that only used count-dependent variables. To minimize the burden on principal investigators, we only retained the article reporting on the first endpoint of trials with multiple eligible articles.

Data Collection and Analysis

Design parameter extraction and calculation

We aimed to extract three design parameters—effect size (Cohen's d), *ICC*, and covariate R^2 (R^2_{cov})—from all eligible trials for all continuous and binary implementation and clinical outcomes. We used two extraction procedures. First, two members of the research team (NW & NC) independently extracted design parameter values (d, *ICC*, R^2_{cov}) for all implementation and clinical outcomes from published study reports using a structured, web-based platform. Conflicts were resolved by re-reviewing the report to generate consensus. Second, if values were not reported in publications, we contacted principal investigators directly to calculate design parameter values from their primary data (i.e., secondary data analysis). Initial contacts with investigators occurred via email, followed by web-based virtual or telephone meetings. Investigators were offered two options to share design parameters from their study: (1) share de-identified data with our team and we would calculate the design parameters directly, or (2) receive code for the statistical package of their choice, written by our team, run it on their data locally, and share the resultant output with us (where the output included the targeted design parameters). We leveraged the substantial professional network of our authorship team to solicit participation. Investigators who did not respond were contacted twice within one month after the initial invitation and a third time approximately six months later. Participating investigators or their analysts were compensated $2500 for their time.

We followed well-established guidelines for calculating Cohen's d (Cohen, 1992; Rosenthal et al., 1994), ICC (Campbell et al., 2004; Hedges & Hedberg, 2013), and R^2_{cov} (Dong et al., 2016; Hedges & Hedberg, 2013) for each outcome variable. Effect sizes were calculated using Cohen's d (Cohen, 1992; Rosenthal et al., 1994) because it is the metric most often required by statistical power programs. We computed effect sizes based on raw data or values reported in articles when unable to obtain model-estimated effect size estimates. Calculation details are provided in Supplemental File 1. We were able to calculate the effect size unadjusted for covariates for most trials; however, one trial only provided covariate-adjusted effect sizes which we included in our analyses. To assess whether inclusion of this covariate-adjusted effect size had an important impact on the results, we compared the central tendency and variation with and without this value (i.e., sensitivity analysis).

Intraclass correlation coefficients for continuous outcomes were calculated using the one-way ANOVA with random effects method described by Raudenbush & Bryk (2010). This approach partitions the unadjusted variance in the outcome by the level of the design (e.g., within-cluster variance vs. between-cluster variance) and produces k-1 ICCs (where k is the number of levels in the design), each of which indicates the proportion of total variance between units at a given level (see Supplemental File 1 for further details). Consequently, we extracted k-1 ICCs per outcome, per study. The calculation of ICCs for dichotomous (binary) outcomes is an area of ongoing research (Chan, 2019; Eldridge et al., 2009). We used a formula that is employed by many power analysis programs (e.g., PASS) and which assumes a continuous underlying trait that follows a logistic distribution (Ahn et al., 2020, p. 20; Eldridge et al., 2009). Supplemental File 1 provides further details.

Values of R^2_{cov} were calculated using formulas provided by Hedges and Hedberg (2013) and Dong et al. (2016). This method, which is sometimes called the reduction in variance components approach (LaHuis et al., 2014), was selected because it is frequently used in power analysis programs. The method produces a measure of R^2_{cov} for each level of the design; these R^2_{cov} values describe the proportion of variance in the outcome *at that level* that is explained by the covariates.

Study characteristic coding

Two reviewers (NW and NC) independently coded study characteristics using a form based on the Standards for Reporting Implementation Studies checklist (Pinnock et al., 2017). Coded study characteristics included: participant population; provider population; study setting; EBI; trial type (e.g., hybrid type II); trial design (e.g., individual randomized repeated measures, cluster RCT, multisite cluster RCT, stepped wedge); level of randomization; sample size (by level); implementation strategy conditions;

implementation strategy targets (e.g., patient, provider, clinic); type of implementation outcome; type of clinical outcome; measurement approaches (e.g., observed, self-report); covariates included in the analysis (as applicable); and year of publication. Supplemental File 1 provides additional details on study characteristic coding.

Statistical Analysis

Data synthesis

We anticipated that the magnitude of effect sizes would depend on the implementation conditions compared. For example, a comparison of two active and equally potent implementation strategies might produce a smaller effect size than a comparison of a standard implementation strategy (e.g., training) vs. the same standard strategy plus an enhancement (e.g., training + audit and feedback). Accordingly, we categorized effect sizes into two types: (1) *standard vs. enhanced* effect sizes compared conditions in which both arms received one or more of the same strategies (i.e., standard) and one condition received an additional strategy (i.e., enhanced); (2) *comparative effectiveness* effect sizes compared conditions in which each arm received a distinct and potentially equally potent implementation strategy. Effect sizes for standard versus enhanced comparisons were calculated so that positive values indicated the superiority of the enhanced condition. Effect sizes from comparative effectiveness comparisons were calculated such that the condition with the more favorable effect was the referent, resulting in all positive effect size values. This permitted a standardized assessment of the overall magnitude of effects within these types of trials.

In order to generate meaningful descriptive analyses of the extracted ICCs, we divided ICCs into three conceptually distinct categories and produced separate descriptive analyses for each category. Category one included ICCs calculated from a repeated measures design (e.g., time within person) with only two levels, where time/observation = level 1 and person/unit = level 2. We labeled these *longitudinal* ICCs (ρ_L); they represent the within-unit correlation between observations on the same unit over time. The second category of ICCs included those calculated at level two of trials with cross-sectional designs (e.g., endpoint-only designs) as well as ICCs calculated at level three of trials with repeated measures designs (e.g., visit within person within clinic). We labeled these cross-sectional level 2 ICCs (ρ_{C2}); they represent the within-cluster correlation between participants within the same cluster. The third category included ICCs calculated at level three of trials with cross-sectional designs and at level four of trials with repeated measures designs. We labeled these cross-sectional level 3 ICCs (ρ_{C3}). Because of important differences in conceptualization and magnitude, we present separate summaries of each type of ICC.

No research reports provided values of R^2_{cov}; consequently, these were available only when research teams permitted secondary analysis of their data. We extracted R^2_{cov} values for each outcome at all study levels where investigators included covariates; models included all the covariates investigators used within the adjusted models. We did not calculate values of R^2_{cov} for longitudinal trials because most power analysis programs do not permit inclusion of covariates for these models (PASS 2024, 2024; Raudenbush et al., 2011). The lower bound for values of R^2_{cov} was constrained to 0.

Analysis. For each design parameter, we generated simple descriptive statistics (median, mean, minimum, maximum) and produced boxplots to visualize the range of observed values, including 25th and 75th percentiles, and outlying estimates (1.5 times the interquartile range). To test whether design parameter estimates varied by study characteristic, we conducted a series of bivariate linear mixed effects models in which a given design parameter (e.g., ICC) served as the dependent variable and a single study characteristic served as the predictor. These models included random intercepts that accounted for the clustering of design parameter estimates within studies and of effect sizes within outcomes within studies (López-López et al., 2018). For example, a study with three outcomes and three arms would produce three effect sizes for each outcome (one contrasting each pair of arms). These effect sizes would be nested within each outcome and the outcomes would be nested within the trial. Models analyzing effect sizes therefore included random intercepts for outcomes and trials. Models for ICC only included random intercepts for trials. These analyses were implemented in Stata (StataCorp, 2023) using the mixed command with restricted maximum likelihood estimation and the Kenward–Rogers correction to account for the small number of clusters (Bolin et al., 2019).

Results

Study Selection and Characteristics

In total, the search yielded 17 trials which met inclusion criteria, reported in 17 publications. Supplemental File 2 shows the flow of trials from initial identification into the pool of included trials. Table 1 presents descriptive information on the 17 trials included in our analyses. There were no stepped-wedge or multi-period cluster randomized trials; all trials employed parallel arms designs. Of the 17 trials, five were individually randomized with repeated measures, eight were cluster randomized, and four were cluster randomized with repeated measures. Among the latter two groups, the mean number of clusters at the highest level of clustering (e.g., level 2 in a two-level cluster randomized design and level 3 in a three-level cluster randomized with repeated measures design) was 29.4 (min = 5, max = 90).

From these trials, we extracted 53 outcome variables for which design parameters were calculated. The average

Table 1.
Characteristics of Included Implementation Trials.

Author	Clinical intervention	Sample size by level[a]	Randomization sequence and trial design	Conditions (duration)	Applicable ERIC strategy categories	Outcomes (description)	Measurement Approach
Kolko et al. (2012)	Alternatives for Families: A Cognitive-Mental Therapy	Child welfare or mental health agencies ($k = 10$) Community practitioners who treat children ($n = 182$)[b] Observations ($n = 548$)	Parallel arms, individually randomized with repeated measures (multisite)	AF-CBT Training w/ LCM vs. Training as usual (6 months)	Train and educate stakeholders	*Implementation:* • Fidelity (AF-CBT abuse-specific skills; AF-CBT abuse-specific techniques; AF-CBT teaching process)	Provider report
Williams et al. (2017)	Various EBPs for youth mental health	Children's mental health agencies ($k = 14$)[b] Clinicians providing outpatient youth mental health services ($n = 197$)	Parallel arms, cluster randomized	ARC vs. Service as usual (36 months)	Assess and redesign workflow; Use evaluative and iterative strategies	*Implementation:* • Adoption (EBP intentions; EBP adoption; EBP use)	Provider report
Wells et al. (2013)	Depression quality improvement (QI) toolkits	Primary care, mental health, substance use, social service, or other community programs ($k = 90$)[b] Clients with depression ($n = 1018$)	Parallel arms, cluster randomized	Community Engagement and Planning vs. Resources for Services (6 months)	Adapt and tailor to context; Train and educate stakeholders; Use evaluative and iterative strategies	*Clinical:* • Symptoms (MCS-12 greater than 40; PHQ-9 greater than 10)	Participant self-report
Waxmonsky et al. (2014)	Life Goals - Collaborative Care (LG-CC)	Community-based clinical practices ($k = 5$)[b] Individuals with bipolar disorder ($n = 384$)	Parallel arms, cluster randomized	Enhanced REP vs. Standard REP (6 months)	Adapt and tailor to context; Provide interactive assistance; Train and educate stakeholders; Use evaluative and iterative strategies	*Implementation:* • Fidelity (Number of group sessions; number of care management contacts; proportion that reached minimum fidelity; proportion that reached optimal fidelity)	Observer coded

(Continued)

Table 1.
(Continued)

Author	Clinical intervention	Sample size by level[a]	Randomization sequence and trial design	Conditions (duration)	Applicable ERIC strategy categories	Outcomes (description)	Measurement Approach
Brown et al. (2014)	Multidimensional Treatment Foster Care (MTFC)	County clusters[b] ($k = 6$) Counties ($n = 40$)	Parallel arms, cluster randomized	Community Development Team vs. Independent Strategy (3–6 years)	Provide interactive assistance; Train and educate stakeholders; Use evaluative and iterative strategies	*Implementation:* • Adoption (Proportion of counties that successfully started up MTFC) • Other (Stages of Implementation Composite; the proportion of counties achieving competence)	Research team rated based on data provided by sites
Lewis et al. (2015)	Measurement-based care (MBC)	Community mental health clinics ($k = 12$)[b] Clinicians ($n = 154$) Individuals with depression ($n = 2,059$) Sessions ($n = 5,522$)	Parallel arms, cluster randomized with repeated measures	Tailored MBC vs. Standardized MBC implementation (4-h training and 5 months)	Adapt and tailor to context; Assess and redesign workflow; Change infrastructure; Train and educate stakeholders; Use evaluative and iterative strategies	*Implementation:* • Fidelity (Completed PHQ-9 or not; discussed PHQ-9 or not) *Clinical:* • Symptoms (PHQ-9)	Fidelity outcomes were gathered through electronic metadata and clinical outcome was participant self-report
Cohen et al. (2016)	TF-CBT	Residential treatment facilities ($k = 18$) Therapists ($n = 129$)[b] Trauma-exposed children ($n = 339$)	Parallel arms, cluster randomized (multisite)	Web-based + Live vs. Web-based (12 months)	Provide interactive assistance; Train and educate stakeholders	*Implementation:* • Adoption (Percent of therapists conducting more than one screen) • Fidelity (Percent fidelity among engaged clients)	Provider report
Epstein et al. (2016)	ADHD quality improvement (QI)	Community-based pediatric primary care practices ($k = 50$)[b] Providers ($n = 213$) Children with ADHD ($n = 373$)	Parallel arms, cluster randomized with repeated measures	Technologically assisted QI intervention vs. Service as usual (four hour-long meetings; 12-months of follow up)	Assess and redesign workflow; Develop stakeholder interrelationships; Support clinicians; Train and educate stakeholders	*Implementation:* • Other (Proportion of days covered by medication prescriptions; number of contacts in first year) • Fidelity (Number of	Implementation outcomes collected through chart review; symptom data collected via caregiver report

(Continued)

Table 1.
(Continued)

Author	Clinical intervention	Sample size by level[a]	Randomization sequence and trial design	Conditions (duration)	Applicable ERIC strategy categories	Outcomes (description)	Measurement Approach
		Observations (n = 840)				teacher scales and parent scales to monitor treatment in first year) *Clinical:* • Symptoms (4-level scale of ADHD symptoms)	
Self-Brown et al. (2017)	SafeCare (SC)	Prevention or child welfare agencies (k = 17) Providers (n = 31)[b] Service participants (n = 169)	Parallel arms, cluster randomized (multisite)	SC-TA vs. SC-IU (Average: 7.74 months, Range: 4–10 months)	Engage consumers; Train and educate stakeholders; Use evaluative and iterative strategies	*Implementation:* • Fidelity (SafeCare Provider Fidelity Checklist)	Observer coded
Stice et al. (2017)	Body Project	Universities (k = 3) College-age women with body dissatisfaction (n = 680)[b] Observations (n = 1,915)	Parallel arms, individually randomized with repeated measures	Clinician-Led Body Project vs. Peer-Led Body Project vs. eBody Project vs. Educational video condition (4 weekly 1-h sessions)	Adapt and tailor to context; Provide interactive assistance; Train and educate stakeholders; Use evaluative and iterative strategies	*Clinical:* • Symptoms (Eating disorder symptoms; thin-ideal internalization; body dissatisfaction; negative affect)	Participant self-report
Locke et al. (2019)	Remaking Recess (RR)	Elementary schools (k = 12)[b] School personnel (n = 28) Children with autism (n = 31) Observations (n = 93)	Parallel arms, cluster randomized with repeated measures	RR + Implementation support vs. RR (6 weeks)	Adapt and tailor to context; Develop stakeholder interrelationships; Train and educate stakeholders	*Implementation:* • Fidelity (Self-rated percent of steps completed; accuracy; understanding; implementation; Coach rated percent of steps completed) *Clinical:* • Functioning (Social network inclusion;	All data was observer coded except for provider-report (self-rated) percent of steps completed

(Continued)

Table 1.
(Continued)

Author	Clinical intervention	Sample size by level[a]	Randomization sequence and trial design	Conditions (duration)	Applicable ERIC strategy categories	Outcomes (description)	Measurement Approach
						joint engagement; solitary play)	
Addington et al. (2019)	MARIGOLD	Individuals with depression ($n = 79$)[b] Observations ($n = 217$)	Multiple parallel arms, individually randomized with repeated measures	Online discussion board vs. Virtual badges vs. Facilitator contact (Self-paced, up to 5–7 weeks)	Develop stakeholder interrelationships; Engage consumers;	*Clinical:* • Symptoms (PHQ-8; CES-D)	Participant self-report
Bartels et al. (2022)	InSHAPE	Mental health provider organizations ($k = 55$)[b] Adults with serious mental illness who were overweight or obese ($n = 924$) Observations ($n = 2,555$)	Parallel arms, cluster randomized with repeated measures	Virtual learning collaborative (18 months) vs. Technical assistance (4 sessions at months 1, 2, 8, and 14)	Provide interactive assistance; Train and educate stakeholders; Use evaluative and iterative strategies	*Implementation:* • Reach (Number of participants who completed 6 or more sessions) • Fidelity (Program level InShape Fidelity Scale score)*Clinical:* • Functioning (Cardiovascular risk reduction)	Research team rated based on data provided by site
Wilfley et al. (2020)	Interpersonal psychotherapy	College counseling centers ($k = 24$)[b] Therapists for adults presenting with eating disorder symptoms ($n = 184$)	Parallel arms, cluster randomized	Trainer pretraining vs. Expert pretraining (One or two two-day workshops followed by 12 months of support)	Train and educate stakeholders	*Implementation:* • Fidelity (Adherence; competence; interpersonal psychotherapy knowledge)	Provider report
Smith et al. (2020)	Life Goals Collaborative Care Model Intervention (CCM)	Community practices treating patients with bipolar or other depressive disorders ($n = 43$)[b] Observations ($n = 555$)	Parallel arms, individually randomized with repeated measures	REP + EF/IF vs. REP + EF (6 months)	Adapt and tailor to context; Change infrastructure; Develop stakeholder interrelationships; Support clinicians; Train and educate stakeholders; Use evaluative and iterative	*Implementation:* • Reach (Number of patients receiving any Life Goal sessions) • Reach (Number of patients receiving three or more Life Goals sessions)	Provider report

(Continued)

Table 1.
(Continued)

Author	Clinical intervention	Sample size by level[a]	Randomization sequence and trial design	Conditions (duration)	Applicable ERIC strategy categories	Outcomes (description)	Measurement Approach
					strategies; Utilize financial strategies		
Jackson et al. (2021)	Parent-Child Interaction Therapy (PCIT)	Licensed psychiatric outpatient organizations ($k = 50$)[b] Clinicians treating children with disruptive behavior disorders ($n = 100$)	Parallel arms, cluster randomized	Learning Collaborative vs. Cascading Model vs. Distance Education (18 months)	Develop stakeholder interrelationships; Engage consumers; Provide interactive assistance; Train and educate stakeholders; Use evaluative and iterative strategies	*Implementation:* • Adoption (Approximate PCIT caseload; use of PCIT protocol) • Reach: (Number of families receiving PCIT)	Provider report
Parent et al. (2022)	Helping the Noncompliant Child (HNC)	Children with disruptive behavior disorders ($n = 101$)[b] Observations ($n = 285$)	Parallel arms, individually randomized with repeated measures	TE-HNC (Self-paced, 11.63 weeks) vs. HNC (Self-paced = 14.15)	Engage consumers; Provide interactive assistance; Train and educate stakeholders; Use evaluative and iterative strategies	*Clinical:* • Functioning (ECBI intensity; ECBI problems)	Provider report

Note. ADHD = Attention-Deficit/Hyperactivity Disorder; AF-CBT = Alternatives for Families: A Cognitive–Mental Therapy; ARC = Availability, Responsiveness, and Continuity; CES-D = Center for Epidemiologic Studies Depression Scale; ECBI = Eyberg Child Behavior Inventory; HNC = Helping the Noncompliant Child; LG-CC = Life Goals–Collaborative Care; LCM = Learning Collaborative Model; MARIGOLD = Mobile Affect Regulation Intervention with the Goal of Lowering Depression; MBC = Measurement-based care; MTFC = Multidimensional Treatment Foster Care; PCIT = Parent-Child Interaction Therapy; PHQ-8/9 = Patient Health Questionnaire-8/9; QI = Quality Improvement; REP = Replicating Effective Programs; RR = Remaking Recess; SC = SafeCare; SC-IU = SafeCare with In Vivo Utility; SC-TA = SafeCare with Technical Assistance; TE-HNC = Technology-Enhanced Helping the Noncompliant Child.
aCluster sample size at the highest level is indicated by *k*.
bIndicates the level of randomization.

number of outcomes per trial was 3.12 ($SD = 1.58$). For 31 of the outcomes, design parameters were calculated using primary data from the research team; for 22 outcomes, design parameters were extracted from published articles. Implementation outcomes ($n = 37$) were more common than clinical outcomes ($n = 16$). Supplemental File 3 provides a list of ERIC strategies used in eligible trials, as classified by the investigative team.

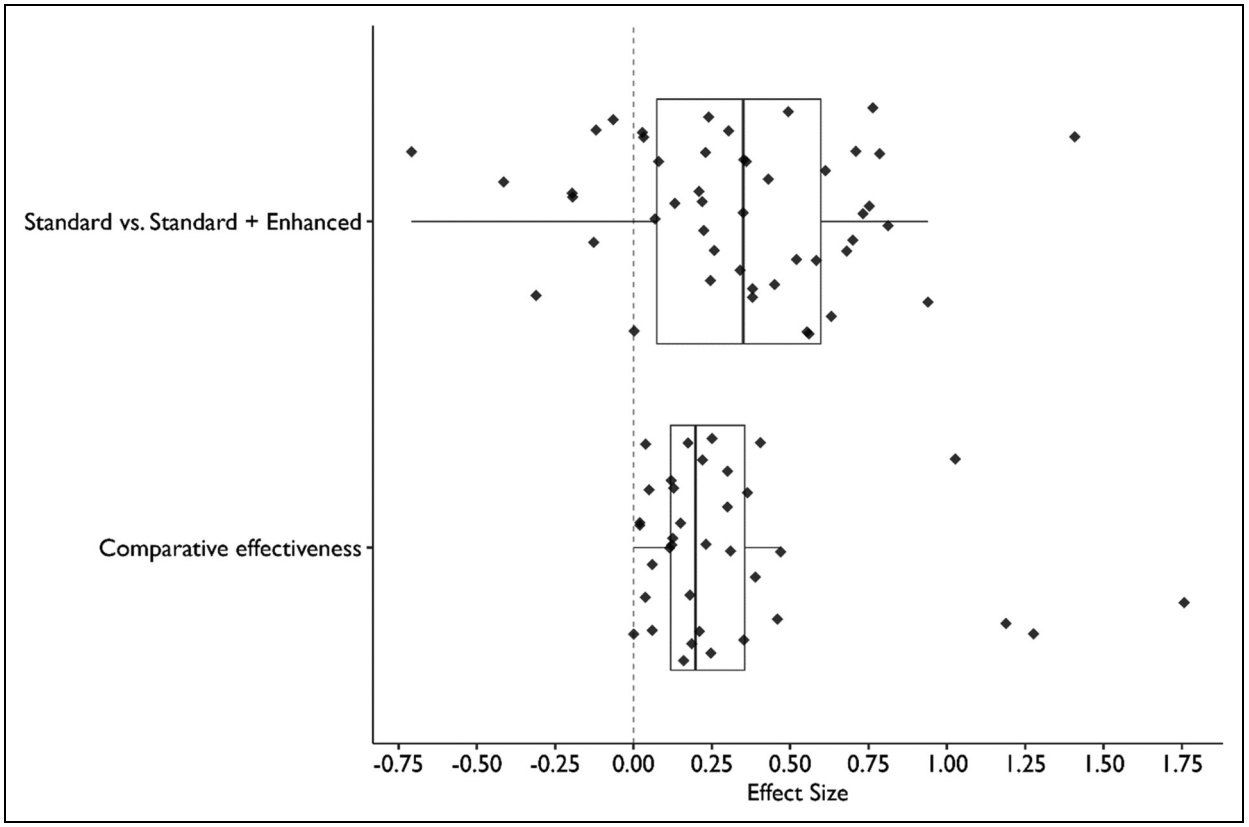
Effect Size

We extracted 83 effect sizes, representing 100% coverage of all 53 outcome variables in the 17 trials. The number of effect sizes is greater than the number of outcomes because trials with more than two conditions contributed multiple effect sizes per outcome. The average number of conditions per trial was 2.23 ($SD = 0.56$, min = 2, max = 4). Effect sizes for 35 contrasts (for 31 outcomes) were generated via secondary analysis of trial data; the remainder were generated from articles.

Figure 1 shows the distribution of effect sizes for each type of trial. The median effect size for *standard vs. enhanced* comparisons was $d = 0.35$ (IQR = 0.07–0.60, $n = 47$ effect sizes from n = 36 outcomes). The median effect size for comparative effectiveness trials was $d = 0.20$ (IQR = 0.12–0.35, $n = 36$ effect sizes from n = 24 outcomes from n = 8 studies).

Table 2 presents effect size values as a function of outcome type and measurement approach. Results of the linear mixed models failed to provide evidence of statistically significant differences in mean effect sizes by outcome type (*standard vs. enhanced*: $F[5, 28.06] = 0.91$, $p = .492$; *comparative effectiveness:* $F[4, 11.68] = 1.35$, $p = .308$) or measurement approach (*standard vs. enhanced*: $F[2, 24.21] = 1.46$, $p = .251$; *comparative effectiveness:* $F[2, 4.83] = 0.73$, $p = .527$), likely due to the small number of included studies. Sensitivity

Figure 1.
Distribution of Effect Sizes (Cohen's d*) from Implementation Trials in Mental Health Settings.*



Note. Sample size consisted of ($n = 47$) standard vs. Enhanced effect sizes and ($n = 36$) comparative effectiveness effect sizes. Standard vs. Enhanced designs consist of contrasts between a trial arm that receives one or more implementation strategies and a trial arm that receives all of the strategies in the other arm plus one or more additional strategies (e.g., distribute educational materials vs. Distribute educational materials + facilitation). Comparative effectiveness designs contrast trial arms with heterogeneous sets of implementation strategies (e.g., technical assistance vs. Facilitation).

Table 2.
Distribution of Effect Size by Study Characteristics.

	Standard vs. standard + enhanced strategy (*d*)					Comparative effectiveness (*d*)				
Factor	*M* (*SD*)	Min	Max	No. of effect sizes	No. of outcomes (No. of studies)	*M* (*SD*)	Min	Max	No. of effect sizes	No. of outcomes (No. of studies)
Overall	0.33 (0.39)	−0.71	1.41	47	36 (11)	0.32 (0.39)	0	1.76	36	24 (8)
Outcome type										
Implementation (Overall)	0.37 (0.45)	−0.71	1.41	26	23 (8)	0.49 (0.51)	0.04	1.76	17	17 (6)
Adoption	0.44 (0.28)	−0.06	0.79	9	7 (4)	0.16 (0.04)	0.13	0.19	2	2 (1)
Fidelity	0.21 (0.52)	−0.71	0.94	11	11 (5)	0.58 (0.52)	0.06	1.76	11	11 (4)
Reach	0.77 (0.90)	0.13	1.41	2	1 (1)	0.44 (0.58)	0.04	1.28	4	4 (3)
Other implementation	0.44 (0.28)	0.21	0.81	4	4 (2)	–	–	–	–	–
Clinical (Overall)	0.28 (0.31)	−0.41	0.71	21	13 (6)	0.16 (0.11)	0	0.41	19	7 (3)
Functioning	0.03 (0.37)	−0.41	0.61	5	5 (2)	0	0	0	1	1 (1)
Symptoms	0.36 (0.25)	−0.20	0.71	16	8 (4)	0.17 (0.10)	0.02	0.41	18	6 (2)
Measurement approach										
External observation	0.23 (0.39)	−0.41	0.81	17	17 (5)	0.57 (0.57)	0	1.76	10	10 (3)
Patient or caregiver report	0.36 (0.25)	−0.20	0.71	16	8 (4)	0.17 (0.10)	0.02	0.41	18	6 (2)
Provider report	0.42 (0.50)	−0.71	1.41	14	11 (5)	0.34 (0.41)	0.04	1.28	8	8 (3)

Note. Effect sizes (*d*) are reported as *M* and *SD*. Standard vs. standard + enhanced designs consist of contrasts between a trial arm that receives one or more implementation strategies and a trial arm that receives all of the strategies in the other arm plus one or more additional strategies (e.g., distribute educational materials vs. distribute educational materials + facilitation). Comparative effectiveness designs contrast trial arms with heterogeneous sets of implementation strategies (e.g., technical assistance vs. facilitation).

analyses indicated that inclusion of the covariate-adjusted effect size did not meaningfully alter the results.

Intraclass Correlation Coefficient

We extracted 47 ICC values from 31 (of 53, 58%) outcome variables across 10 (of 17; 59%) trials. Of these 47 ICC values, 18 were longitudinal (ρ_L), 22 were cross-sectional at level 2 (ρ_{C2}), and seven were cross-sectional at level 3 (ρ_{C3}). The median longitudinal ICC was 0.50 (IQR = 0.31–0.54, $n = 18$). The median cross-sectional level-2 ICC was 0.10 (IQR = 0.04–0.22, $n = 22$). The median cross-sectional level-3 ICC was 0.09 (IQR = 0.06–0.15, $n = 7$). Figure 2 shows the distribution of ICC values for each category. All ICC values were unadjusted for covariates.

Table 3 presents mean ICC values by outcome and study characteristic. Results of the linear mixed models failed to provide evidence of statistically significant differences of mean ICC values by outcome type (ρ_{C2}: $F[4, 9.87] = 2.39$, $p = .122$; ρ_{C3}: $F[2, 2.06] = 4.87$, $p = .166$; ρ_L: $F[3, 2.94] = 4.58$, $p = .124$), measurement approach (ρ_{C2}: $F[2, 17.01] = 1.76$, $p = .202$; ρ_{C3}: $F[1, 3.34] = 0.28$, $p = .628$; ρ_L: $F[2, 3.99] = 1.99$, $p = .252$), or population (ρ_{C2}: $F[1, 8.70] = 0.02$, $p = .898$; ρ_L: $F[2, 8.09] = 2.39$, $p = .153$).

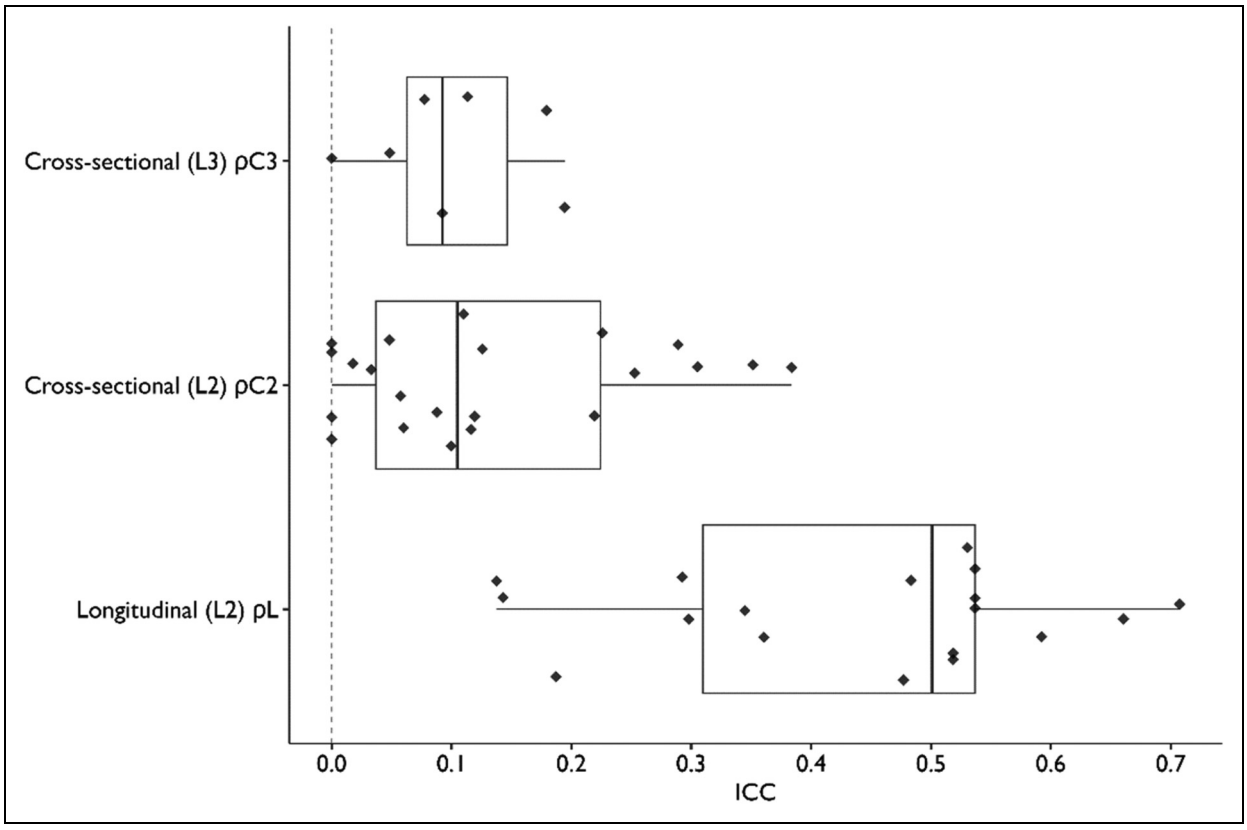
Covariate R^2

Only 10 trials, incorporating 27 outcomes, included covariates and there was substantial heterogeneity across trials in the covariates included. Given our goal of understanding how much variance in outcomes was explained by covariates that investigators believed were important, we computed R^2_{cov} values for all studies with available data (15 of 27 outcomes (60%) across six trials). The number of R^2_{cov} values extracted for each outcome varied depending on the number of levels in the design and the nature of the outcome variable (i.e., R^2_{cov} cannot be computed at level 1 for binary outcomes). From the 15 outcomes analyzed, we were able to extract 12 R^2_{cov} estimates at level 1 (R^2_1) (n = 4 trials), 15 R^2_{cov} estimates at level 2 (R^2_2) ($n = 5$ trials), and 4 R^2_{cov} estimates at level 3 (R^2_3) ($n = 1$ study). Table 4 presents descriptive statistics for R^2_{cov} at each level and indicates the specific covariates and number of covariates represented in each trial.

Discussion

This study bridges a major methodological gap in the design of implementation trials in mental health by presenting empirical reference values for effect size, ICC, and covariate R^2. The reference values presented in this paper represent plausible and meaningful estimates of

Figure 2.
Distribution of Intraclass Correlation Coefficients (ICC) from Implementation Trials in Mental Health settings.



Note. Sample size consisted of ($n = 18$) longitudinal ICC values, ($n = 22$) cross-sectional ICC values at level 2, and ($n = 7$) cross-sectional ICC values at level 3. We calculated ICC using methods from the study by Raudenbush and Bryk (2010).

design parameters to inform sample size determination for multilevel implementation trials in mental health. Given recent increases in NIH funding for implementation research (Neta et al., 2021; Purtle et al., 2015; Tinkle et al., 2013), we hope future research will add to the database of values presented here. In the meantime, these findings offer a crucial resource for investigators to design efficient and effective implementation trials. Below, we offer guidance on how investigators might use these values to inform their own trials.

One of the strengths (and weaknesses) of the nascent field of implementation science is the wide range of outcomes, populations, clinical interventions, implementation settings, implementation strategies, and trial design choices available to, and made by, implementation researchers. While this diversity likely increases the specificity and utility of knowledge generated by the field over the long term, in the short term it challenges efforts to produce evidence summaries. Despite this study's exclusive focus on implementation trials in mental health, examination of Tables 2 and 3 reveals several outcomes and study designs for which few or no trials were available. These sparse data signal that the reference values presented in this study offer a starting point for more robust compendiums in the future.

The distributions of design parameter values observed in this study are within the range of reference values reported in other substantive areas, yet are distinct enough to highlight the importance of field-specific reference values for determining sample size. For example, the mean effect sizes for clinical outcomes observed in these implementation trials were comparable to effects in trials comparing evidence-based psychotherapies to services as usual (e.g., Weisz et al., 2013). It was also not surprising that the mean effect sizes for implementation outcomes were descriptively higher than those for clinical outcomes, given implementation strategies' focus on changing practice. Studies of ICC for student achievement (e.g., math and reading) and social and mental health outcomes in schools have tended to generate higher ICC reference values than those observed for clinical outcomes in this study (Dong et al., 2016; Hedges & Hedberg, 2013). However, this study's clinical outcome ICCs of 0.08 and 0.05 at levels 2 and 3 are at the higher end of ICC values observed for clinical and quality of life outcomes in primary care settings (Elley et al., 2005;

Table 3.
Distribution of Intraclass Correlation Coefficients by Study Characteristics.

	Longitudinal (within-unit) ρ_L				Cross-sectional (Between-unit) $\rho_{C(2)}$				Cross-sectional (Between-unit) $\rho_{C(3)}$			
Factor	*M* (*SD*)	Min	Max	No. of outcomes (No. of studies)	*M* (*SD*)	Min	Max	No. of outcomes (No. of studies)	*M* (*SD*)	Min	Max	No. of outcomes (No. of studies)
Overall	0.44 (0.17)	0.14	0.71	18 (7)	0.13 (0.12)	0	0.38	22 (6)	0.10 (0.07)	0	0.19	7 (3)
Outcome type												
Implementation (overall)	0.48 (0.19)	0.14	0.71	7 (3)	0.17 (0.13)	0	0.38	15 (5)	0.10 (0.07)	0	0.19	6 (3)
Adoption	–	–	–	–	0.17 (0.11)	0.10	0.31	3 (1)	–	–	–	–
Fidelity	0.40 (0.16)	0.14	0.53	5 (2)	0.20 (0.13)	0.03	0.38	10 (4)	0.12 (0.09)	0	0.19	4 (3)
Reach	0.68 (0.03)	0.66	0.71	2 (1)	–	–	–	–	–	–	–	
Other implementation	–	–	–	–	0.03 (0.04)	0	0.06	2 (1)	0.08 (0.05)	0.05	0.11	2 (1)
Clinical (Overall)	0.41 (0.16)	0.14	0.59	11 (4)	0.05 (0.06)	0	0.12	7 (4)	0.08 (–)	0.08	0.08	1 (1)
Functioning	0.26 (0.09)	0.14	0.36	5 (2)	0.04 (0.06)	0	0.11	3 (1)	–	–	–	–
Symptoms	0.54 (0.03)	0.52	0.59	6 (2)	0.06 (0.06)	0	0.12	4 (3)	0.08 (–)	0.08	0.08	1 (1)
Measurement approach												
External observation	0.26 (0.11)	0.14	0.36	5 (2)	0.14 (0.14)	0	0.38	14 (4)	0.10 (0.07)	0	0.19	6 (3)
Patient or caregiver report	0.54 (0.03)	0.52	0.59	6 (2)	0.06 (0.06)	0	0.12	4 (3)	0.08 (–)	0.08	0.08	1 (1)
Provider report	0.48 (0.19)	0.19	0.71	7 (3)	0.19 (0.10)	0.10	0.31	4 (2)	–	–	–	–
Population												
Site (e.g., agency, school, etc.)	0.68 (0.03)	0.66	0.71	2 (1)	0.14 (0.12)	0	0.35	14 (4)	0.10 (0.07)	0	0.19	7 (3)
Provider (e.g., clinician, teacher)	0.38 (0.14)	0.14	0.53	7 (3)	0.12 (0.13)	0	0.38	8 (3)	–	–	–	–
Patient (e.g., client, student)	0.43 (0.17)	0.14	0.59	9 (4)	–	–	–	–	–	–	–	–

Note. *M* and *SD* are reported for longitudinal ICCs ρ_L, and cross-sectional ICCs at level 2 $\rho_{C(2)}$, and level 3 $\rho_{C(3)}$. All ICCs are unadjusted.

Table 4.
Distribution of R^2_{cov} *by Level*

Level	*M* (*SD*)	*Mdn*	Min	Max	No. of outcomes (No. of studies)	No. of covariates included per outcome, *M* (Min–Max)[a]
3	0.089 (0.163)	0.012	0	0.786	4 (1)	1 (1–1)
2	0.143 (0.262)	0.000	0	0.333	15 (5)	1.93 (1–8)
1	0.038 (0.063)	0.003	0	0.611	12 (4)	1 (1–1)

Note. All values included with negative values changed to zero. ($n = 14$) R^2_{cov} values are extracted from models with demographic covariates only, ($n = 13$) values are from models with pretest covariate only, and ($n = 5$) R^2_{cov} values are from models with both demographic and pretest covariates. *M* and *SD* are reported for R^2_{cov} at each level of analysis.
[a]Specific variables included in calculating R^2 for the five available studies were as follows: Study 1 (pretest and patient race), Study 2 (Organizational Social Climate profile), Study 3 (pretest, age, sex, more than three chronic conditions, education, race/ethnicity, income below federal poverty level, 12-month alcohol abuse or use of illicit drugs, and 12-month depressive disorder), Study 4 (pretest and new vs. existing patient), Study 5 (pretest-only).

Smeeth & Ng, 2002) and in implementation trials in primary care and hospitals (Campbell et al., 2005). In addition, the median cross-sectional ICC values for implementation outcomes observed in this study (i.e., 0.10 and 0.09 for levels 2 and 3, respectively) are higher than those reported for implementation outcomes in cluster-randomized implementation trials from primary care and hospital settings in the UK (median ICC of 0.06; Campbell et al., 2005), possibly due to greater heterogeneity of disease areas and medical specialities included in the study by Campbell et al. (2005).

Due to the heterogeneity of covariates included in these studies and the sparseness of data, we caution researchers against using the covariate R^2 values reported here as inputs for specific power analyses. However, these values do provide a window into the magnitude of variance in outcomes explained by the covariates that implementation researchers have used in their studies to date. Interestingly, the covariate R^2 values observed in these studies were considerably smaller than those observed in compendia of educational trials, where covariate R^2 routinely meets or exceeds 50% (Dong et al., 2016; Hedges & Hedberg, 2013). This is an important area for future implementation research in mental health because high values of covariate R^2 (e.g., from pretests of the outcome), can dramatically reduce the required cluster sample size (e.g., up to 50% or more, Bloom et al., 2007).

None of the studies located by our review used stepped-wedged, multiple-period, or sub-cluster designs, although, these are an important and growing part of implementation research (Davis-Plourde et al., 2023; Hemming et al., 2020; Ouyang et al., 2022). Special considerations are needed when calculating power for these designs (Korevaar et al., 2021) and we refer readers to relevant literature (Davis-Plourde et al., 2023; Hemming et al., 2020; Kasza et al., 2019; Kelcey et al., 2021; Ouyang et al., 2022). Caution may also be warranted when calculating ICCs using studies with few clusters.

Application

The selection of design parameter estimates for determining sample size in multilevel trials depends on the trial design (e.g., longitudinal vs. cross-sectional), variation between clusters, the feasibility of collecting covariate information, and what is considered a clinically meaningful difference in the given trial's context (Raudenbush, 1997). The design parameters presented in this study can be used to aid power analyses for multilevel implementation trials. For example, the effect size estimates presented in Table 2 can help ensure researchers are not overly optimistic; whereas values of ICC provide plausible estimates which may apply to a range of scenarios. In this section, we describe how our results might be applied to determine sample size for two hypothetical implementation trials.

A common design in our sample was a study in which patients were nested within providers within sites. As an example, consider a proposed study where the outcome of interest is a continuous score for intervention fidelity, measured at the patient level, with randomization of sites to implementation strategies and a two-tailed hypothesis that an enhanced implementation condition will have higher mean fidelity at the study endpoint compared to standard implementation. Using Table 2, an investigator may determine that a plausible minimum detectable effect size is $d = 0.4$. Drawing on Table 3, the investigator may determine that plausible values of ICC at the clinician (ρ_{C2}), and site levels (ρ_{C3}) are 0.21 and 0.10, respectively. Drawing on substantive knowledge of the implementation area, the investigator may estimate covariate R^2 values will be 0.05, 0.36., and 0.12 at levels 1, 2, and 3, respectively. Assuming the trial will enroll an average of four patients per clinician and seven clinicians per site, the investigator would need to enroll 28 sites to adequately power (>0.8) the trial at the 95% confidence level (a = .05), as calculated by the R Shiny App PowerUpR (Ataneka et al., 2023).

Another common design was a repeated-measures study with observations nested within sites and sites randomly assigned to conditions. For example, a study may randomly assign sites to two different implementation strategies and compare the difference in reach across sites, operationalized as the number of patients who received a target dose of an EBI (e.g., 3 + sessions) within the site, measured at baseline and quarterly thereafter for

12 months after the initial training (i.e., total of four measurement occasions). The contrast of interest would be differences in reach at 12 months. From Table 2, the investigator may select $d = 0.54$ as a reasonable value for the minimum detectable effect size at 12 months. Examining the ICCs for the reach outcome in Table 3, the investigator may select 0.68 as a reasonable estimate of ICC (i.e., ρ, representing the within-unit correlation) size. Assuming the model includes no covariates, and alpha is set at .05, the investigator would need to enroll 64 sites to adequately power (>0.8) the trial (Zhang et al., 2023), as calculated by the software PASS 2022 (which uses formulas from Ahn et al., 2020).

Future Implications

Results of our study highlight important directions for future research. First, the frequent omission of design parameters, such as ICC, from implementation research reports in mental health underscores the importance of changing scientific norms within the field. We consider this an opportune moment to open a conversation about the importance of ensuring that guidelines (Brown et al., 2015) to promote reporting of critical data elements are followed. Second, we believe normative change is needed within the discipline of implementation science with regard to data sharing. Despite our focus on federally-funded trials (for which data should presumably be available) and availability of resources to support researchers who shared data or conducted secondary analyses of their data, we were only able to obtain ICCs for 59% of trials. In some cases, investigators were amenable to sharing data or conducting analyses but were unable to do so due to resource limitations (e.g., the study statistician was no longer funded) or legal or policy restrictions (e.g., Institutional Review Board would not allow sharing of de-identified data). One practical way to address this is to change requirements for data sharing for federally-funded research. As of the writing of this article, it does not appear that the U.S. National Institute of Mental Health Data Archive (NDA; *NIMH Data Archive*, n.d.) requires trials with nested designs to report ICC for primary or secondary outcomes or to provide site identifiers for all observations (which represents the minimum necessary meta-data to calculate ICCs from within the archive). Modified reporting guidelines could address these limitations.

Limitations

This study represents the first step toward providing implementation researchers in mental health with meaningful reference values of design parameters to accurately determine sample size and calculate statistical power for trials. Given the preliminary stage of research in this area, our study has limitations. First, our focus on NIMH-funded trials, while intended to ensure studies were well-designed and adequately powered, may have resulted in the omission of important trials funded by other sources. For example, although NIMH funds research globally, none of the trials in our sample were conducted outside the United States. Second, the field was nascent during the period we collected data and in future years there will likely be more implementation trials from which larger samples of design parameters can be extracted. Future research can build on the database presented here. Third, the application of implementation science in mental health during this early period of the field's development is highly heterogeneous so we opted to analyze the data descriptively instead of using meta-analytic methods. Even when analyzed within subgroups that can be conceptualized as coming from the same sampling distribution (e.g., comparative effectiveness studies of strategies to improve fidelity), only a subset of trials provided the information necessary to compute the precision of the extracted design parameter estimates and some cell sizes were too small to robustly test the relationships between specific study characteristics and design parameter values. Also, even among studies where design parameters were available, the precision of these estimates might be questionable, especially in trials where there were a small number of clusters. We anticipate future reviews will address these weaknesses through the inclusion of larger samples of larger trials.

Conclusion

This work addresses a ubiquitous methodological barrier that undermines the advancement of implementation science in mental health by providing empirically based reference values for design parameters needed to determine sample sizes in multilevel implementation trials. The paper offers a model for future research that improves the field's ability to design efficient and adequately powered implementation trials.

Acknowledgments

We wish to thank Dr. Sara Cullen of the University of Pennsylvania School of Social Policy and Practice for her work screening titles and abstracts.

Contributions

NJW served as lead for conceptualization, methodology, writing–original draft, writing–review and editing, investigation, project administration, and funding acquisition. NCC served as lead for formal analysis, data curation, investigation, writing–original draft, and writing–review and editing. RSB served as lead for conceptualization and writing–review and editing. SCM served as lead for conceptualization, methodology, and writing–review and editing.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: NJW, NCC, and SCM have no conflicts of interest to declare. RSB is principal at Implementation Science & Practice, LLC. She is currently an appointed member of the National Advisory Mental Health Council and the NASEM study, "Blueprint for a national prevention infrastructure for behavioral health disorders," and serves on the scientific advisory board for AIM Youth Mental Health Foundation and the Klingenstein Third Generation Foundation. She has received consulting fees from United Behavioral Health and OptumLabs. She previously served on the scientific and advisory board for Optum Behavioral Health and has received royalties from Oxford University Press. All reported activities are outside of the submitted work.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was supported by the U.S. National Institute of Mental Health of the National Institutes of Health under Award Number R21MH126076 (PI: Williams). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ORCID iD

Nathaniel J. Williams https://orcid.org/0000-0002-3948-7480

Supplemental material

Supplemental material for this article is available online.

References

Addington, E. L., Cheung, E. O., Bassett, S. M., Kwok, I., Schuette, S. A., Shiu, E., Yang, D., Cohn, M. A., Leykin, Y., Saslow, L. R., & Moskowitz, J. T. (2019). The MARIGOLD study: Feasibility and enhancement of an online intervention to improve emotion regulation in people with elevated depressive symptoms. *Journal of Affective Disorders*, *257*, 352–364.

Ahn, C., Heo, M., & Zhang, S. (2020). *Sample size calculations for clustered and longitudinal outcomes in clinical research (First issued in paperback)*. CRC Press.

Ataneka, A., Kelcey, B., Dong, N., Bulus, M., & Bai, F. (2023). *PowerUp R Shiny App (Manual)* (Version 0.9) [Computer software]. https://www.causalevaluation.org/uploads/7/3/3/6/73366257/r_shinnyapp_manual_0.9.pdf

Bartels, S. J., Aschbrenner, K. A., Pratt, S. I., Zubkoff, L., Jue, K., Williams, G., Godfrey, M. M., Cohen, M. J., Banerjee, S., Xie, H., Wolfe, R., Naslund, J. A., & Bond, G. R. (2022). Virtual learning collaborative compared to technical assistance as a strategy for implementing health promotion in routine mental health settings: A hybrid type 3 cluster randomized trial. *Administration and Policy in Mental Health and Mental Health Services Research*, *49*(3), 1031–1046. https://doi.org/10.1007/s10488-022-01215-0

Bell, B., Ferron, J., & Kromrey, J. (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. In JSM Proceedings, Section on Survey Research Methods, 1122–1129.

Bhaumik, D. K., Roy, A., Aryal, S., Hur, K., Duan, N., Normand, S.-L. T., Brown, C. H., & Gibbons, R. D. (2008). Sample size determination for studies with repeated continuous outcomes. *Psychiatric Annals*, *38*(12). https://doi.org/10.3928/00485713-20081201-01

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59. https://doi.org/10.3102/0162373707299550

Bolin, J. H., Finch, W. H., & Stenger, R. (2019). Estimation of random coefficient multilevel models in the context of small numbers of level 2 clusters. *Educational and Psychological Measurement*, *79*(2), 217–248. https://doi.org/10.1177/0013164418773494

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

Brown, C. H., Chamberlain, P., Saldana, L., Padgett, C., Wang, W., & Cruden, G. (2014). Evaluation of two implementation strategies in 51 child county public service systems in two states: Results of a cluster randomized head-to-head implementation trial. *Implementation Science*, *9*, 1–15.

Brown, A. W., Li, P., Bohan Brown, M. M., Kaiser, K. A., Keith, S. W., Oakes, J. M., & Allison, D. B. (2015). Best (but oft-forgotten) practices: Designing, analyzing, and reporting cluster randomized controlled trials. *The American Journal of Clinical Nutrition*, *102*(2), 241–248. https://doi.org/10.3945/ajcn.114.105072

Campbell, M. K., Elbourne, D. R., & Altman, D. G. (2004). CONSORT statement: Extension to cluster randomised trials. *BMJ*, *328*(7441), 702–708. https://doi.org/10.1136/bmj.328.7441.702

Campbell, M. K., Fayers, P. M., & Grimshaw, J. M. (2005). Determinants of the intracluster correlation coefficient in cluster randomized trials: The case of implementation research. *Clinical Trials*, *2*(2), 99–107. https://doi.org/10.1191/1740774505cn071oa

Campbell, M. K., Grimshaw, J. M., & Elbourne, D. R. (2004). Intracluster correlation coefficients in cluster randomized trials: Empirical insights into how should they be reported. *BMC Medical Research Methodology*, *4*(1), 9. https://doi.org/10.1186/1471-2288-4-9

Chan, W. (2019). The relationship among design parameters for statistical power between continuous and binomial outcomes in cluster randomized trials. *Psychological Methods*, *24*(2), 179–195. https://doi.org/10.1037/met0000185

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. A., Mannarino, A. P., Jankowski, K., Rosenberg, S., Kodya, S., & Wolford, G. L. (2016). A randomized

implementation study of trauma-focused cognitive behavioral therapy for adjudicated teens in residential treatment facilities. *Child Maltreatment*, *21*(2), 156–167.

Cook, C. R., Lyon, A. R., Locke, J., Waltz, T., & Powell, B. J. (2019). Adapting a compilation of implementation strategies to advance school-based implementation research and practice. *Prevention Science*, *20*(6), 914–935. https://doi.org/10.1007/s11121-019-01017-1

Davis-Plourde, K., Taljaard, M., & Li, F. (2023). Sample size considerations for stepped wedge designs with subclusters. *Biometrics*, *79*(1), 98–112. https://doi.org/10.1111/biom.13596

De Jong, K., Moerbeek, M., & Van Der Leeden, R. (2010). A priori power analysis in longitudinal three-level multilevel models: An example with therapist effects. *Psychotherapy Research*, *20*(3), 273–284. https://doi.org/10.1080/10503300903376320

Distiller SR (2023). *DistillerSR* (Version 2.35) [Computer software]. DistillerSR Inc. https://www.distillersr.com/

Dong, N., Reinke, W. M., Herman, K. C., Bradshaw, C. P., & Murray, D. W. (2016). Meaningful effect sizes, intraclass correlations, and proportions of variance explained by covariates for planning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review*, *40*(4), 334–377. https://doi.org/10.1177/0193841X16671283

Eccles, M. P., & Mittman, B. S. (2006). Welcome to implementation science. *Implementation Science*, *1*(1), 1. https://doi.org/10.1186/1748-5908-1-1

Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R., & Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. *Clinical Trials*, *1*(1), 80–90. https://doi.org/10.1191/1740774504cn006rr

Eldridge, S. M., Ukoumunne, O. C., & Carlin, J. B. (2009). The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions. *International Statistical Review*, *77*(3), 378–394. https://doi.org/10.1111/j.1751-5823.2009.00092.x

Elley, C. R., Kerse, N., Chondros, P., & Robinson, E. (2005). Intraclass correlation coefficients from three cluster randomised controlled trials in primary and residential health care. *Australian and New Zealand Journal of Public Health*, *29*(5), 461–467. https://doi.org/10.1111/j.1467-842X.2005.tb00227.x

Epstein, J. N., Kelleher, K. J., Baum, R., Brinkman, W. B., Peugh, J., Gardner, W., Lichtenstein, P., & Langberg, J. M. (2016). Impact of a web-portal intervention on community ADHD care and outcomes. *Pediatrics*, *138*(2), e20154240. https://doi.org/10.1542/peds.2015-4240

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Goldstein, H. (2013). *Multilevel statistical models*. Wiley.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*(6), 445–489. https://doi.org/10.1177/0193841X14529126

Hemming, K., Kasza, J., Hooper, R., Forbes, A., & Taljaard, M. (2020). A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the shiny CRT calculator. *International Journal of Epidemiology*, *49*(3), 979–995. https://doi.org/10.1093/ije/dyz237

Hox, J. J., Moerbeek, M., & Schoot, R. v. d. (2018). *Multilevel analysis: Techniques and applications* (3rd ed). Routledge; Taylor & Francis Group.

Isaakidis, P. (2003). Evaluation of cluster randomized controlled trials in Sub-Saharan Africa. *American Journal of Epidemiology*, *158*(9), 921–926. https://doi.org/10.1093/aje/kwg232

Ivers, N. M, Taljaard, M., Dixon, S., Bennett, C., McRae, A., Taleban, J., Skea, Z., Brehaut, J. C., Boruch, R. F., Eccles, M. P., Grimshaw, J. M., Weijer, C., Zwarenstein, M., & Donner, A. (2011). Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: Review of random sample of 300 trials, 2000-8. *BMJ*, *343*(1), d5886. http://dx.doi.org/10.1136/bmj.d5886

Jackson, C. B., Herschell, A. D., Scudder, A. T., Hart, J., Schaffner, K. F., Kolko, D. J., & Mrozowski, S. (2021). Making implementation last: The impact of training design on the sustainability of an evidence-based treatment in a randomized controlled trial. *Administration and Policy in Mental Health and Mental Health Services Research*, *48*, 757–767.

Kasza, J., Hemming, K., Hooper, R., Matthews, J. N. S., & Forbes, A. B. (2019). Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Statistical Methods in Medical Research*, *28*(3), 703–716. https://doi.org/10.1177/0962280217734981

Kelcey, B., Xie, Y., Spybrook, J., & Dong, N. (2021). Power and sample size determination for multilevel mediation in three-level cluster-randomized trials. *Multivariate Behavioral Research*, *56*(3), 496–513. https://doi.org/10.1080/00273171.2020.1738910

Killip, S. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *The Annals of Family Medicine*, *2*(3), 204–208. https://doi.org/10.1370/afm.141

Kolko, D. J., Baumann, B. L., Herschell, A. D., Hart, J. A., Holden, E. A., & Wisniewski, S. R. (2012). Implementation of AF-CBT by community practitioners serving child welfare and mental health: A randomized trial. *Child Maltreatment*, *17*(1), 32–46.

Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, *1*(4), 265–288. https://doi.org/10.1080/19345740802328216

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, *1*(1), 66–88. https://doi.org/10.1080/19345740701692522

Korevaar, E., Kasza, J., Taljaard, M., Hemming, K., Haines, T., Turner, E. L., Thompson, J. A., Hughes, J. P., & Forbes, A. B. (2021). Intra-cluster correlations from the clustered outcome dataset bank to inform the design of longitudinal cluster trials. *Clinical Trials*, *18*(5), 529–540. https://doi.org/10.1177/17407745211020852

Kraemer, H. C., & Blasey, C. (2016). *How many subjects?: Statistical power analysis in research* (2nd edition). Sage Publicastions, Inc.

Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*, *63*(5), 484. https://doi.org/10.1001/archpsyc.63.5.484

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods*, *17*(4), 433–451. https://doi.org/10.1177/1094428114541701

Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The role and interpretation of pilot studies in clinical research. *Journal of Psychiatric Research*, *45*(5), 626–629. https://doi.org/10.1016/j.jpsychires.2010.10.008

Lewis, C. C., Fischer, S., Weiner, B. J., Stanick, C., Kim, M., & Martinez, R. G. (2015). Outcomes for implementation science: An enhanced systematic review of instruments using evidence-based rating criteria. *Implementation Science*, *10*(1), 155. https://doi.org/10.1186/s13012-015-0342-x

Lipsey, M. W. (1995). *Design sensitivity: Statistical power for experimental research (Nachdr.)*. Sage.

Locke, J., Shih, W., Kang-Yi, C. D., Caramanico, J., Shingledecker, T., Gibson, J., Frederick, L., & Mandell, D. S. (2019). The impact of implementation support on the use of a social engagement intervention for children with autism in public schools. *Autism: The International Journal of Research and Practice*, *23*(4), 834–845.

López-López, J. A., Page, M. J., Lipsey, M. W., & Higgins, J. P. T. (2018). Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Research Synthesis Methods*, *9*(3), 336–351. https://doi.org/10.1002/jrsm.1310

Muthén, L. K., & Muthén, B. O. (2002). How to use a monte carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Neta, G., Clyne, M., & Chambers, D. A. (2021). Dissemination and implementation research at the national cancer institute: A review of funded studies (2006–2019) and opportunities to advance the field. *Cancer Epidemiology, Biomarkers & Prevention*, *30*(2), 260–267. https://doi.org/10.1158/1055-9965.EPI-20-0795

NIMH Data Archive (n.d.). https://nda.nih.gov/.

Novins, D. K., Green, A. E., Legha, R. K., & Aarons, G. A. (2013). Dissemination and implementation of evidence-based practices for child and adolescent mental health: A systematic review. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(10), 1009–1025.e18. https://doi.org/10.1016/j.jaac.2013.07.012

Offorha, B. C., Walters, S. J., & Jacques, R. M. (2022). Statistical analysis of publicly funded cluster randomised controlled trials: A review of the national institute for health research journals library. *Trials*, *23*(1), 115. https://doi.org/10.1186/s13063-022-06025-1

Ouyang, Y., Li, F., Preisser, J. S., & Taljaard, M. (2022). Sample size calculators for planning stepped-wedge cluster randomized trials: A review and comparison. *International Journal of Epidemiology*, *51*(6), 2000–2013. https://doi.org/10.1093/ije/dyac123

Parent, J., Anton, M. T., Loiselle, R., Highlander, A., Breslend, N., Forehand, R., Hare, M., Youngstrom, J. K., & Jones, D. J. (2022). A randomized controlled trial of technology-enhanced behavioral parent training: Sustained parent skill use and child outcomes at follow-up. *Journal of Child Psychology and Psychiatry*, *63*(9), 992–1001.

PASS 2024 (2024). *Power analysis and sample size software* [Computer software]. NCSS Statistical Software. ncss.com/software/pass

Pinnock, H., Barwick, M., Carpenter, C. R., Eldridge, S., Grandes, G., Griffiths, C. J., Rycroft-Malone, J., Meissner, P., Murray, E., Patel, A., Sheikh, A., & Taylor, S. J. C. (2017). Standards for reporting implementation studies (StaRI) statement. *BMJ*, *356*, i6795. https://doi.org/10.1136/bmj.i6795

Proctor, E. K., Landsverk, J., Aarons, G., Chambers, D., Glisson, C., & Mittman, B. (2009). Implementation research in mental health services: An emerging science with conceptual, methodological, and training challenges. *Administration and Policy in Mental Health and Mental Health Services Research*, *36*(1), 24–34. https://doi.org/10.1007/s10488-008-0197-4

Purtle, J., Peters, R., & Brownson, R. C. (2015). A review of policy dissemination and implementation research funded by the national institutes of health, 2007–2014. *Implementation Science*, *11*(1), 1. https://doi.org/10.1186/s13012-015-0367-1

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2010). *Hierarchical linear models: Applications and data analysis methods*. Sage Publication.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. https://doi.org/10.1037/1082-989X.5.2.199

Raudenbush, S. W., Spybrook, J., Bloom, H., Congdon, R., Hill, C., & Martinez, A. (2011). *Optimal Design software for multi-level and longitudinal research (Version 3.01) [Computer* software]. Survey Research Center of the Institute of Social Research at University of Michigan.

Rosenthal, R., Cooper, H., & Hedges, L. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.) *The handbook of research synthesis* (2nd ed., Vol. 621, pp. 231–244). Russell Sage Foundation. https://www.jstor.org/stable/10.7758/9781610441377

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*(2), 347–367. https://doi.org/10.1177/1094428107308906

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316. https://doi.org/10.1037/0033-2909.105.2.309

Self-Brown, S. R, Osborne, C, Rostad, M, & Feil, E. (2017). A technology-mediated approach to the implementation of an evidence-based child maltreatment prevention program. *Child Maltreatment*, *22*(4), 344–353.

Smeeth, L., & Ng, E. S.-W. (2002). Intraclass correlation coefficients for cluster randomized trials in primary care. *Controlled*

Clinical Trials, *23*(4), 409–421. https://doi.org/10.1016/S0197-2456(02)00208-8

Smith, S. N., Liebrecht, C. M., Bauer, M. S., & Kilbourne, A. M. (2020). Comparative effectiveness of external vs blended facilitation on collaborative care model implementation in slow-implementer community practices. *Health Services Research*, *55*(6), 954–965.

Snijders (2014). Power and Sample Size in Multilevel Linear Models. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.), *Wiley StatsRef: Statistics Reference Online* (1st ed.). Wiley. https://doi.org/10.1002/9781118445112.stat06584

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications.

StataCorp (2023). *Stata statistical software: Release 18 [Computer software]*. StataCorp LLC.

Stice, E., Rohde, P., Shaw, H., & Gau, J. M. (2017). Clinician-led, peer-led, and internet-delivered dissonance-based eating disorder prevention programs: Acute effectiveness of these delivery modalities. *Journal of Consulting and Clinical Psychology*, *85*(9), 883–895.

Tinkle, M., Kimball, R., Haozous, E. A., Shuster, G., & Meize-Grochowski, R. (2013). Dissemination and implementation research funded by the US national institutes of health, 2005–2012. *Nursing Research and Practice*, *2013*, 1–15. https://doi.org/10.1155/2013/909606

Waxmonsky, J., Kilbourne, A. M., Goodrich, D. E., Nord, K. M., Lai, Z., Laird, C., Clogston, J., Kim, H. M., Miller, C., & Bauer, M. S. (2014). Enhanced fidelity to treatment for bipolar disorder: Results from a randomized controlled implementation trial. *Psychiatric Services*, *65*(1), 81–90.

Weisz, J. R., Kuppens, S., Eckshtain, D., Ugueto, A. M., Hawley, K. M., & Jensen-Doss, A. (2013). Performance of evidence-based youth psychotherapies compared with usual clinical care. *JAMA Psychiatry*, *70*(7), 750. https://doi.org/10.1001/jamapsychiatry.2013.1176

Wells, K. B., Jones, L., Chung, B., Dixon, E. L., Tang, L., Gilmore, J., Sherbourne, C., Ngo, V. K., Ong, M. K., Stockdale, S., Ramos, E., Belin, T. R., & Miranda, J. (2013). Community-partnered cluster-randomized comparative effectiveness trial of community engagement and planning or resources for services to address depression disparities. *Journal of General Internal Medicine*, *28*(10), 1268–1278. https://doi.org/10.1007/s11606-013-2484-3

Wilfley, D. E., Agras, W. S., Fitzsimmons-Craft, E. E., Bohon, C., Eichen, D. M., Welch, R. R., Jo, B., Raghavan, R., Proctor, E. K., & Wilson, G. T. (2020). Training models for implementing evidence-based psychological treatment: A cluster-randomized trial in college counseling centers. *JAMA Psychiatry*, *77*(2), 139–147.

Williams, N. J., Glisson, C., Hemmelgarn, A., & Green, P. (2017). Mechanisms of change in the ARC organizational strategy: Increasing mental health clinicians' EBP adoption through improved organizational culture and capacity. *Administration and Policy in Mental Health and Mental Health Services Research*, *44*, 269–283.

Wilson, P. M., Sales, A., Wensing, M., Aarons, G. A., Flottorp, S., Glidewell, L., Hutchinson, A., Presseau, J., Rogers, A., Sevdalis, N., Squires, J., & Straus, S. (2017). Enhancing the reporting of implementation research. *Implementation Science*, *12*(1), 13. https://doi.org/10.1186/s13012-017-0546-3

Zhang, Z., Mai, Y., Yang, M., Xu, Z., & McNamara, C. (2023). *Package 'WebPower'* (Version 0.9.4) [Computer software]. https://webpower.psychstat.org/wiki/

Zhang, Z., & Wang, L. (2009). Statistical power analysis for growth curve models using SAS. *Behavior Research Methods*, *41*(4), 1083–1094. https://doi.org/10.3758/BRM.41.4.1083