This is an Accepted Manuscript of an article published by Taylor & Francis in Disability and Rehabilitation on 2018-08-19, available online: https://www.tandfonline.com/10.1080/09638288.2018.1500647.

Balance of Group Sizes in Randomized Controlled Trials Published in APA Journals

Mara Cañedo-Ayala, BSc[a]; Danielle B. Rice, MSc[a,b]; Alexander W. Levis, MSc[c]; Matthew Chiovitti, MSc[a]; Brett D. Thombs, PhD[a,b,d,e,f,g]

[a]Lady Davis Institute of the Jewish General Hospital, Montreal, Quebec, Canada; [b]Department of Psychology, McGill University, Montreal, Canada; [c]Department of Biostatistics, Harvard University, Cambridge, USA; Departments of [d]Epidemiology, Biostatistics and Occupational Health; [e]Psychiatry, [f]Medicine, and [g]Educational and Counselling Psychology, McGill University, Montreal, Canada.

Dr. Brett D. Thombs, Jewish General Hospital; 4333 Côte-Sainte-Catherine Road; Montréal, Québec, Canada; H3T 1E4; Telephone: (514) 340-8222 ext. 26811; Email address: brett.thombs@mcgill.ca.

ABSTRACT

Objective: We evaluated whether sample size differences between arms of two-arm parallel group randomized controlled trials (RCTs) published in American Psychological Association (APA)-affiliated journals were consistently smaller than expected by chance with simple randomization.

Methods: We searched PsycINFO for two-arm parallel group RCTs in APA-affiliated journals published January 2007 to September 2017 that used individual randomization (1:1 allocation ratio), reported the number of participants randomized, and did not describe employing restrictive randomization (e.g., blocking). We queried authors because randomization processes were often not described in articles, and we conducted a post-hoc logistic regression analysis to attempt to identify factors associated with overly balanced groups.

Results: We identified 203 eligible trials, but after the author query, it was determined that only 115 used simple randomization. Among those 115 trials, there was a significantly greater number of trials with smaller sample size differences between trial arms than would be expected by chance ($p < .001$); 89 of 115 (77%) had differences in trial arm sample sizes smaller than the 50% prediction interval threshold for these differences. Greater proportionate imbalance may be associated with larger trial size (odds ratio of 0.27, 95% CI 0.08 to 0.94 for $N > 200$ versus $N \leq 100$); greater balance may be more common in higher impact journals, though this was not statistically significant.

Conclusions: Education is needed to ensure that randomization procedures are implemented as intended and fully and accurately reported and that balanced group sample sizes are not understood as an indicator of trial quality.

Key Words: randomized controlled trial, American Psychological Association, sample size

INTRODUCTION

In randomized controlled trials (RCTs), participants are randomly assigned to a study condition in order to test the benefits or harms of an intervention. The APA Presidential Task Force on Evidence Based Practice has emphasized that RCTs represent the standard for inferring causation and the effectiveness of psychological interventions (American Psychological Association Task Force, 2006). Establishment of empirically supported treatments also includes a review of evidence from high-quality RCTs (Tolin, McKay, Forman, Klonsky, & Thombs, 2015).

In order for findings of RCTs to reflect the true effects of an intervention, they must be conducted properly, including adequate randomization. Randomization has a fundamental role in reducing risk of bias in trial results through balancing participants in different trial arms on potential confounding variables (Jadad & Enkin, 2007; Simon, 2001). In simple randomization, all trial participants are randomized individually and have the same chance of being assigned to different trial arms (Schulz & Grimes, 2002). This ensures unpredictability of assignment and reduces the possibility of selection bias, which can occur if allocation to trial arms is influenced by personal characteristics or if future assignment can be predicted based on randomization restrictions (Berger & Bears, 2003). Simple randomization sometimes leads to differences in the number of participants and characteristics of participants in different trial arms, particularly when the sample size is small (Lachin, 1988; Schulz & Grimes, 2002). Loss of power due to group imbalances, however, is negligible, except in extreme cases, and covariate imbalances are typically of little consequence in large, adequately powered trials (Lachin, 1988). Restricted randomization methods (e.g., blocking, urn randomization) are sometimes used to balance participant characteristics across groups and can facilitate subgroup analyses. A potential disadvantage is that some of these methods, such as permuted block designs (Matts & Lachin, 1988), may give rise to predictability of patient assignment and increase the likelihood of selection bias. Urn designs are an option that provide adequate balancing properties while protecting against selection bias (Schulz & Grimes, 2002; Wei & Lachin, 1988).

Discrepancies in the numbers of study participants assigned to different trial arms are expected with simple randomization. Nonetheless, there is a common misconception that successful randomization must yield equal sample sizes in comparison groups and that unequal group sizes reflect poorly on the credibility of the trial (Friedman, Furberg, DeMets, Reboussin, & Granger, 2015; Schulz & Grimes, 2002). Because of this, it is possible that reviewers and editors may favor trials with balanced group sizes or that some researchers may use non-randomized methods to force balance between group sizes when randomization generates imbalances (Altman & Doré, 1990; Schulz & Grimes, 2002; Schulz et al., 1994). Previous research has found that reporting of randomization procedures in healthcare intervention trials is sometimes suboptimal and that some trials described as RCTs may not be truly randomized (Altman & Doré, 1990; Chan & Altman, 2005; DerSimonian, Charette, McPeek, & Mosteller, 1982; Grimes & Schulz, 1996; Mosteller, Gilbert, & McPeek, 1980; Schulz, Chalmers, Grimes, & Altman, 1994; Schulz, Chalmers, Hayes, & Altman, 1995; Smith, Moffatt, Gelskey, Hudson, & Kaita, 1997).

The degree of expected divergence between sample sizes of different trial arms in individual trials can be calculated based on the total number of participants randomized. Patterns in sets of trials that reflect significantly less divergence than expected due to chance can be an indication that simple randomization may not have occurred or been carried out as described in some of the trials. Two studies, both published in the early 1990s, evaluated this. One study evaluated 80 trials published in four high-impact general medicine journals (*Annals of Internal Medicine*, *British Medical Journal*, *Lancet*, *New England Journal of Medicine*) and found that 60% of trials did not report information about the type of randomization used and that, among studies that did not report using restrictive randomization, group sizes were significantly more similar than would be expected by chance if simple randomization had occurred (Altman & Doré, 1990). A second study, which included 206 trials published in four obstetrics and gynecology journals, found that only 32% of studies described an adequate method of randomizing participants, only 23% described steps taken to conceal the randomization sequence, and just 9% did both. There were 96 trials that did not report using restricted randomization; of these, only 8 had differences in the number of participants assigned to the two trial arms that were greater than the threshold that 50% of trials of the same sample size would be expected to exceed, far too few to be plausible if simple randomization had been used and properly conducted. (Schulz et al., 1994). Finding group sizes across trials that are too similar to be statistically plausible could occur if non-randomized assignment was used in some trials described as RCTs (e.g., alternation, days of week), if trialists did not always adhere to randomization results when randomization generated imbalances in group sizes, or if trials were poorly reported and restrictive randomization procedures were used but not mentioned in published trial reports (Schulz et al., 1994).

Both of these studies were conducted over 20 years ago and prior to the development of reporting guidelines, including the Consolidated Standards of Reporting Trials Statement (CONSORT; (Moher, Schulz, Altman, & Group, 2001; Schulz, Altman, & Moher, 2010) and the APA's Journal Article Reporting Standards (JARS; Appelbaum, Cooper, Kline, Mayo-Wilson, Nezu & Rao, 2018). They did not examine trials of psychological treatments. Recently, members of our team evaluated whether specialty health care journals that publish trials of primarily non-regulated interventions (e.g., psychology, surgery, rehabilitation, nutrition) require prospective trial registration and whether registration policies are associated with higher rates of trial registration (Azar et al., 2019). In the process, we noticed that many trials that did not describe using restrictive randomization procedures reported group sample sizes that were surprisingly similar. Among more than 300 trials reviewed, we found that < 5% had differences that were outside a 50% prediction interval, taking trial sample size into consideration; this did not change when we only evaluated 148 trials whose authors clarified that simple randomization had been used (Thombs et al., 2020). However, fewer than 15% of the trials were behavioral or psychological interventions, and only 2% were from APA-affiliated journals.

The APA has taken significant steps to improve the conduct and reporting of psychological research, including the publication of its JARS criteria (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). The objectives of the present study were to determine whether differences in sample sizes between trial arms in two-arm parallel-group RCTs with 1:1 allocation that did not describe using restrictive randomization processes were consistently smaller than would be expected by chance if randomization had truly occurred (1) among all trials published in APA-affiliated journals; (2) among a subset for which trial authors verified that they used simple randomization based on an email query; and (3) for trials with confirmed simple randomization that were published in *Health Psychology* or the *Journal of Consulting and Clinical Psychology*, since these journals publish far more psychological intervention trials than any other APA journal (e.g., Azar, Riehm, McKay, & Thombs, 2015). Additionally, we evaluated year of publication and, in a post-hoc analysis, trial and journal factors that may be associated with greater likelihood of reporting balanced trial arm sizes.

METHODS

This study used a cross-sectional design and adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) reporting guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009).

Included Journals and Identification of Eligible RCTs

Journals on the "APA and Affiliated Journals" website that were listed under the subject "Clinical Psychology" and had publications listed in PsycINFO were included in the search strategy. Of 37 journals listed under "Clinical Psychology" on the website (American Psychological Association, 2017), two journals (*Clinician's Research Digest: Adult Populations*; *Clinician's Research Digest: Child and Adolescent Populations*) did not have any citations in PsycINFO. Thus, 35 APA or APA-affiliated journals were included in the search strategy (see full list in Appendix 1).

We searched PsycINFO on September 26, 2017 for examples of RCTs published in the 35 included journals between January 2007 and September 2017. The search strategy combined the journal titles with (Treatment Effectiveness Evaluation/ OR exp Treatment Outcomes/ OR Psychotherapeutic Outcomes/ OR PLACEBO/ OR Followup Studies/ OR placebo*.tw. OR random*.tw. OR comparative stud*.tw. OR randomi#ed controlled trial*.tw. OR (clinical adj3 trial$).tw. OR (research adj3 design).tw. OR (evaluat* adj3 stud*).tw. OR (prospectiv* adj3 stud*).tw. OR ((singl* or doubl* or trebl* or tripl*) adj3 (blind* or mask*)).tw.) (see full search strategy in Appendix 2).

Eligible publications had to report results from a two-arm parallel-group RCT that randomized individual patients, not clusters or groups; that used 1:1 randomization assignment; that did not report the use of any form of restrictive randomization (e.g., blocking, urn randomization, randomized allocation) (Altman & Bland, 1999, 2005; Hewitt & Torgerson, 2006); and that reported the number of participants randomized to each trial arm, not just the number analyzed. We restricted to 1:1 randomization and two-arm trials for simplicity. If a single publication reported findings from more than one RCT, only the first RCT described in the publication was evaluated. If more than one publication from the same RCT was identified, the RCT was counted only once. In these cases, we included the trial publication with the largest total sample size at randomization, if there were differences, or the earliest trial publication if reported sample sizes did not differ. Randomization information reported in all identified publications associated with the RCT, including trial protocols, was examined and used to code randomization procedure. Publications that analyzed data from trial participants but did not compare the effects of assignment to different trial arms on outcomes were excluded (e.g., cross-sectional analyses of baseline data).

Search results were uploaded into the systematic review software *DistillerSR*, which was used to code and track results. First, two investigators independently evaluated titles and abstracts for potential eligibility. Articles deemed potentially eligible by either investigator were included for full-text review. Then, two investigators independently conducted full-text reviews. Disagreement between investigators was resolved through consensus, with a third investigator consulted as necessary.

Of the 203 eligible trials in our initial sample that did not describe using restrictive randomization, 181 (89%) did not provide enough information to determine how randomization had been conducted. To verify if simple randomization had been used, we emailed the corresponding authors of all 203 studies, once per week, up to five weeks (see Appendix 3). If the email for the corresponding author from the trial publication did not function or if we did not receive a response, we searched for alternate email address and emailed co-authors. Based on these queries, we determined which trials were verified to have used simple randomization.

Data Extraction

For all eligible RCTs, one member of the research team initially extracted data from the identified trial report plus any other published reported associated with the RCT, including first author last name, first author country, year of publication, journal name, study objectives, intervention description, whether there was explicit reporting of simple randomization, method of randomization (e.g., coin toss, computer-generated random number), and the number of participants in each trial arm. A second reviewer validated all extracted data using the *DistillerSR* Quality Control function. Discrepancies were resolved by consensus and consultation with a third investigator, if needed.

Statistical Analyses

The number of participants assigned to the treatment and control arms in each trial were compared by subtracting the number of participants in the control group from the number of participants in the intervention group. Based on the assumption that under simple randomization, the number of participants randomized to the treatment arm should follow a binomial distribution, prediction intervals were constructed for these differences. Prediction intervals, from the binomial distribution are well-approximated using the normal distribution, particularly for $p = .50$ (Agresti & Coull, 1998). Thus, as was done in previous studies (Altman & Doré, 1990; Schulz et al., 1994), for each included trial, based on the total number randomized, we calculated the differences between group sample sizes that corresponded to 50% and 95% prediction intervals. The width of the interval depends on the total sample size of the trial. Thus, for each included trial, separately, we can calculate the expected distribution of differences between group sample sizes and determine if the difference in the trial was within the interval where we would expect to find 50% of the differences, between the 50% and 95% intervals, or beyond the 95% interval. If simple randomization had been performed, we would expect 50% of these differences to lie within (-0.6745*sqrt(total trial sample size) and +0.6745*sqrt(total trial sample size)), and 95% of the differences in sample sizes to lie within (-1.96*sqrt(total trial sample size) and +1.96*sqrt(total trial sample size)), since 0.6745 and 1.96 reflect, approximately, the 75th and 97.5th percentiles of the standard normal distribution, respectively. We compared the numbers of RCTs with observed differences in sample size that fell within the 50% interval, between the 50% and 95% intervals, and beyond the 95% interval to the expected numbers assuming simple randomization using a chi-squared test ($\alpha = 0.05$) programmed in an Excel worksheet. Analyses were performed for all included trials “as published”, trials verified to have used simple randomization (“author verified”), and trials verified to have used simple randomization that were published in *Health Psychology* or the *Journal of Consulting and Clinical Psychology*.

Because the most recent version of the CONSORT statement was published in 2010 (Schulz, et al., 2010), we evaluated trials in our sample published from 2007-2010 and 2011-2017, separately. We additionally generated a Pearson’s correlation of publication year and proportion of trials outside of the 50% prediction interval.

Additionally, in a post-hoc analysis, we used logistic regression to evaluate trial and publication factors that may be associated with the likelihood of trial arm sample sizes that were balanced (within 50% prediction interval), including number of participants randomized, if the randomization procedure was reported adequately in the publication, and the journal impact factor.

We evaluated sample size for two reasons. First, it is possible that improper balancing via non-random methods to correct for imbalances from randomization could be more common in small trials, since this would require manipulation of the status of only a few participants rather than large-scale manipulation. Second, akin to publication bias (e.g., Franco, Malhotra, & Simonovits, 2014), it is possible that some reviewers and journal editors may evaluate small trials with imbalanced trial arm sample sizes negatively, which would lead to a greater proportion of balanced trials among smaller trials. Journal impact factor was evaluated because, similarly, if reviewers and editors favor trials with balanced sample sizes, those trials would be more likely to be published in journals with higher impact factors than trials with imbalances, all else equal. We used 2017 journal impact factor for all included studies to avoid conflating year of publication, since journal impact factors tend to increase over time. We assessed whether the simple randomization method was clearly articulated in the published article versus only determined via author query to evaluate whether reporting may have a role.

Power and Sample Size

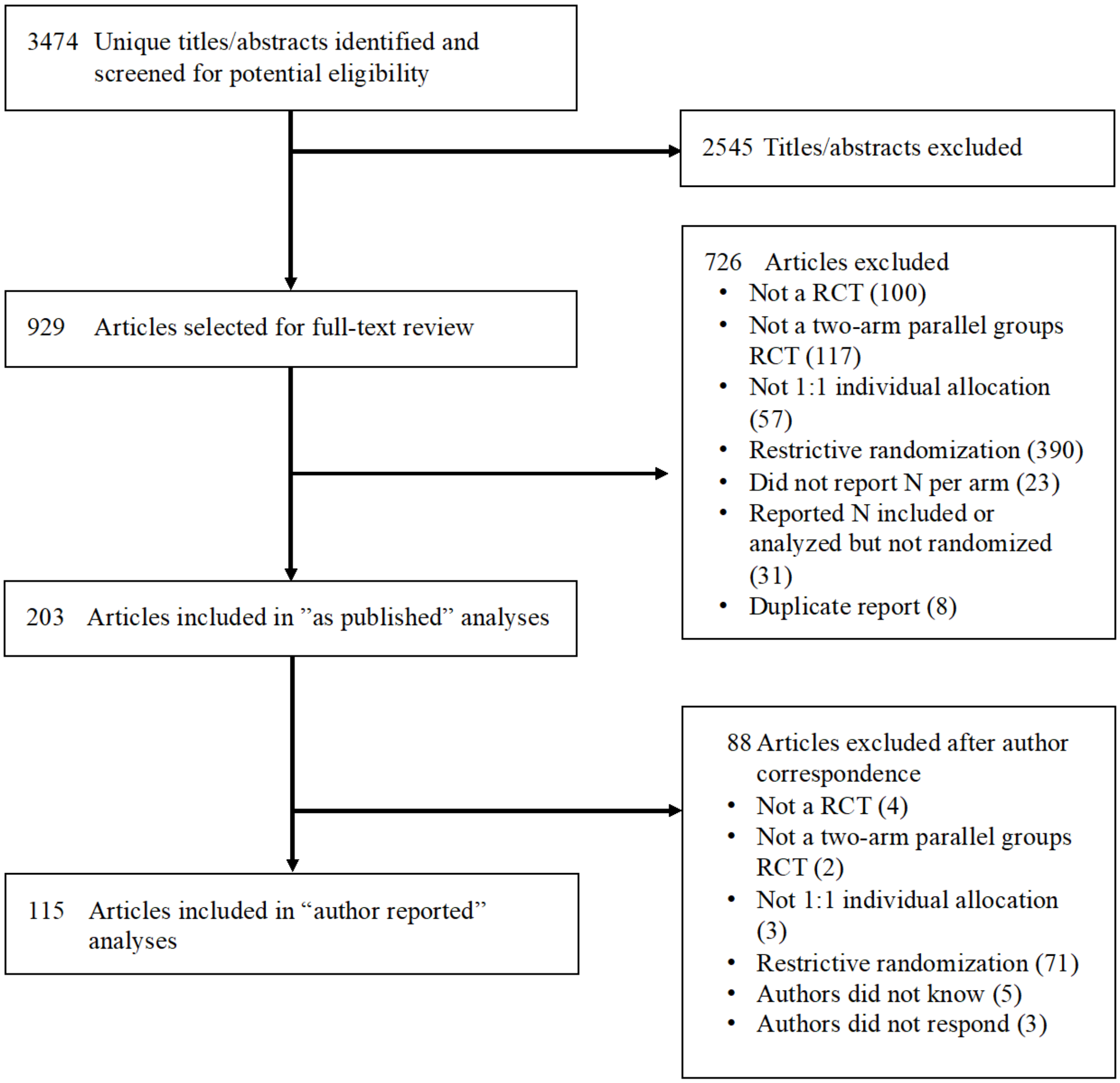
Members of our research team judged that finding 60% or more of trials within the 50% prediction interval would be a meaningful difference. Thus, to determine the number of RCTs to target and the search period for our study, we first calculated the number of included RCTs that would be needed for 80% power to find a statistically significant difference if there were 10% more RCTs than expected within the 50% prediction interval for the difference of participants in the two trial arms. A one-tailed binomial test was selected over a two-tailed binomial test because only a unidirectional effect was being tested. For a one-tailed binomial test with α = 0.05, 158 included RCTs would be needed. Because the consequence of overpowering the study would be additional labour and not risk to human participants and because of the uncertainty involved in predicting the number of eligible RCTs that would be identified in an actual search, we rounded this number up to approximately 200 RCTs.

To estimate the necessary search period to be able to include approximately 200 RCTs, we did a preliminary assessment of the proportion of title and abstract citations generated from the search that would likely result in included RCTs. Thus, during the planning phase of the study in 2017, we examined 100 citations from 2016 and determined that, upon full review, 10 (10%) would likely be eligible, and we generated a 95% confidence interval for this based on Agresti and Coull’s (1998) approximate method for binomial proportions (5.5% to 17.4%). Based on the lower end of the 95% confidence interval (5.5%), we estimated that a search starting 10 years prior to the start of the study (January 2007), which generated 3474 citations, would generate approximately 191 eligible RCTs as a conservative estimate. Thus, we searched from January 2007 to September 2017.

RESULTS

Of the 3,474 unique titles and abstracts identified via the PsycINFO search, 2,545 were excluded after title and abstract review and 726 after full-text review, resulting in 203 publications that met eligibility criteria for the study and were included “as published”. Of these, 22 (11%) explicitly stated in the article that simple randomization had been used; 181 (89%) did not describe the randomization procedure beyond stating that assignment to groups was done randomly or describing a procedure that did not allow us to determine if simple randomization had occurred or if there were restrictions. We emailed all 203 authors to confirm simple randomization or clarify if not described in the published article. We received responses from 200 (99%), and we identified 115 trials with “author verified” simple randomization, of which 18 had described the randomization procedure as “simple” in the published article. We excluded 88 trials from analyses of trials with “author verified” simple randomization because they did not use simple randomization or were otherwise ineligible; of these, 4 trials had reported simple randomization in the published article. See Figure 1.

Figure 1. Flow chart of selection of eligible trials



RCTs Characteristics and Randomization Mechanisms

See Appendix 4 for characteristics of each of the 203 included RCTs and Table 1 for a summary. Of the 203 RCTs, 169 (83%) were conducted in North America (Canada, Puerto Rico, United States). Forty-eight of the included RCTs (24%) were published between 2007 and 2010, and 155 trials (76%) between 2011 and 2017. The RCTs were published in 27 journals with 39 RCTs (19%) in *Health Psychology* and 78 (38%) in the *Journal of Consulting and Clinical Psychology*. Trial characteristics were similar when only the 115 trials with author verified simple randomization were considered.

As shown in Table 1, only 8 of the 203 RCTs (4%) reported using centralized randomization procedures with complete randomization concealment; 77 (38%) reported using a local randomization method, including 55 (27%) that used a computer- or web-based random number generator and 22 (11%) that used a manual procedure (e.g., random number table, coin tossing, playing cards, dice). The remaining 118 RCTs (58%) did not specify the method of randomization that was used (see Appendix 4). This was similar for the 115 trials with author verified simple randomization.

Table 1

Characteristics of Total Included Randomized Controlled Trials (N = 203) and Trials Verified as Using Simple Randomization (N = 115)

Variable	*Total Included Trials: n (%)*	*Trials Verified to Have Used Simple Randomization: n (%)*
First author location		
North America[1]	169 (83%)	96 (83%)
Europe[2]	23 (11%)	12 (10%)
Australia and New Zealand	7 (4%)	3 (3%)
Middle East[3]	2 (1%)	1 (1%)
Asia[4]	2 (1%)	2 (2%)
Top 3 journals with most included RCTs		
Journal of Consulting and Clinical Psychology	78 (38%)	43 (37%)
Health Psychology	39 (19%)	19 (17%)
Psychology of Addictive Behaviors	11 (5%)	----------
Rehabilitation Psychology	----------	7 (6%)
2017 Journal Impact Factor ≥ 3	131 (65%)	73 (63%)
Total N Randomized		
0 to 100	106 (52%)	59 (51%)

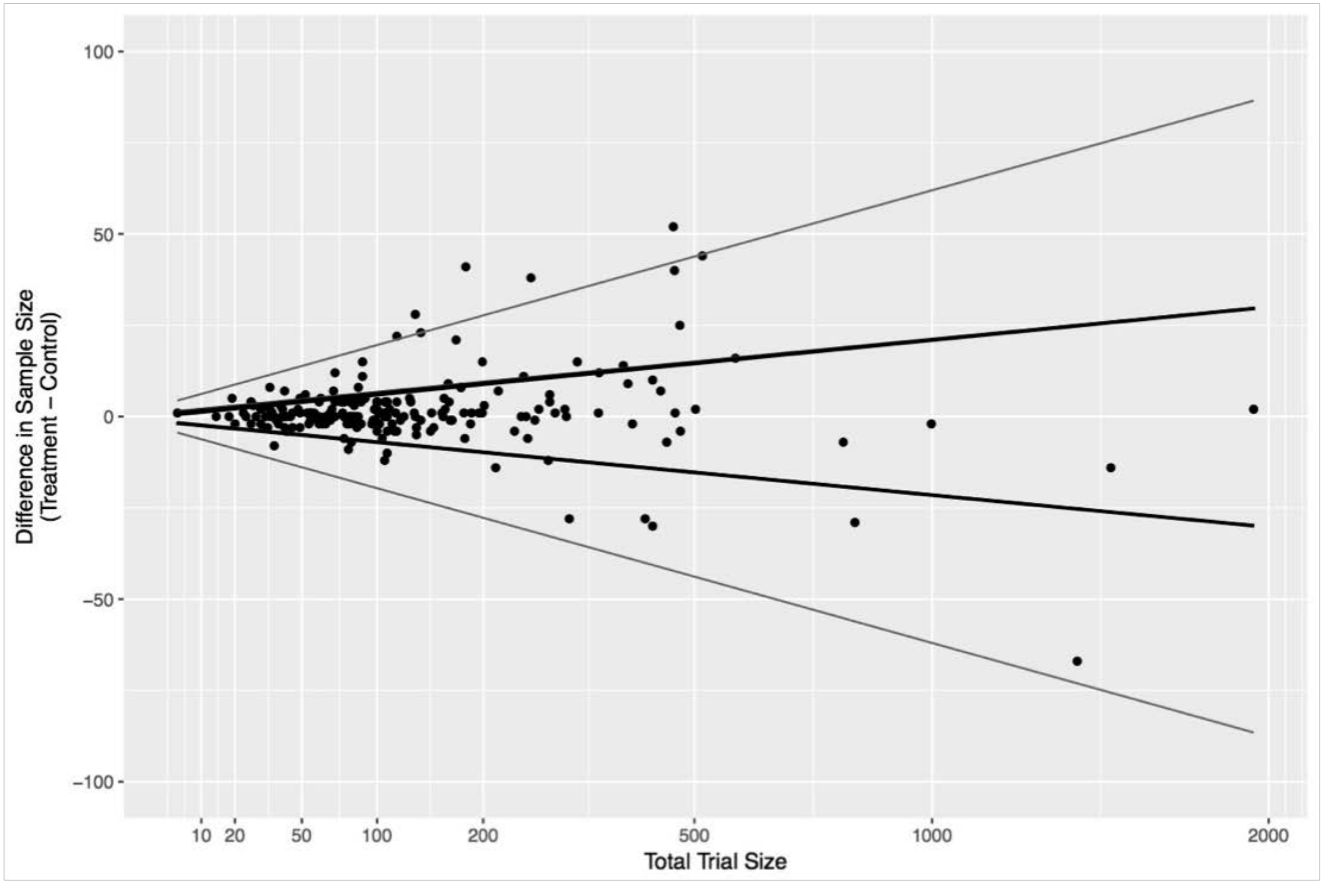
101-200	53 (26%)	33 (29%)
≥ 201	44 (22%)	23 (20%)
Method of Random Sequence Generation:		
Centralized – external to investigators	8 (4%)	5 (4%)
Local – computer or web-based generator	55 (27%)	33 (29%)
Local – manual (e.g. random number table, coin toss, playing cards, dice)	22 (11%)	17 (15%)
Not specified	118 (58%)	60 (52%)

1. Canada, Puerto Rico, and United States of America.
2. Belgium, England, France, Germany, The Netherlands, Norway, Spain, Sweden United Kingdom, and Wales.
3. Iran and Israel
4. Japan and Taiwan.

Differences in Trial Arm Sample Sizes

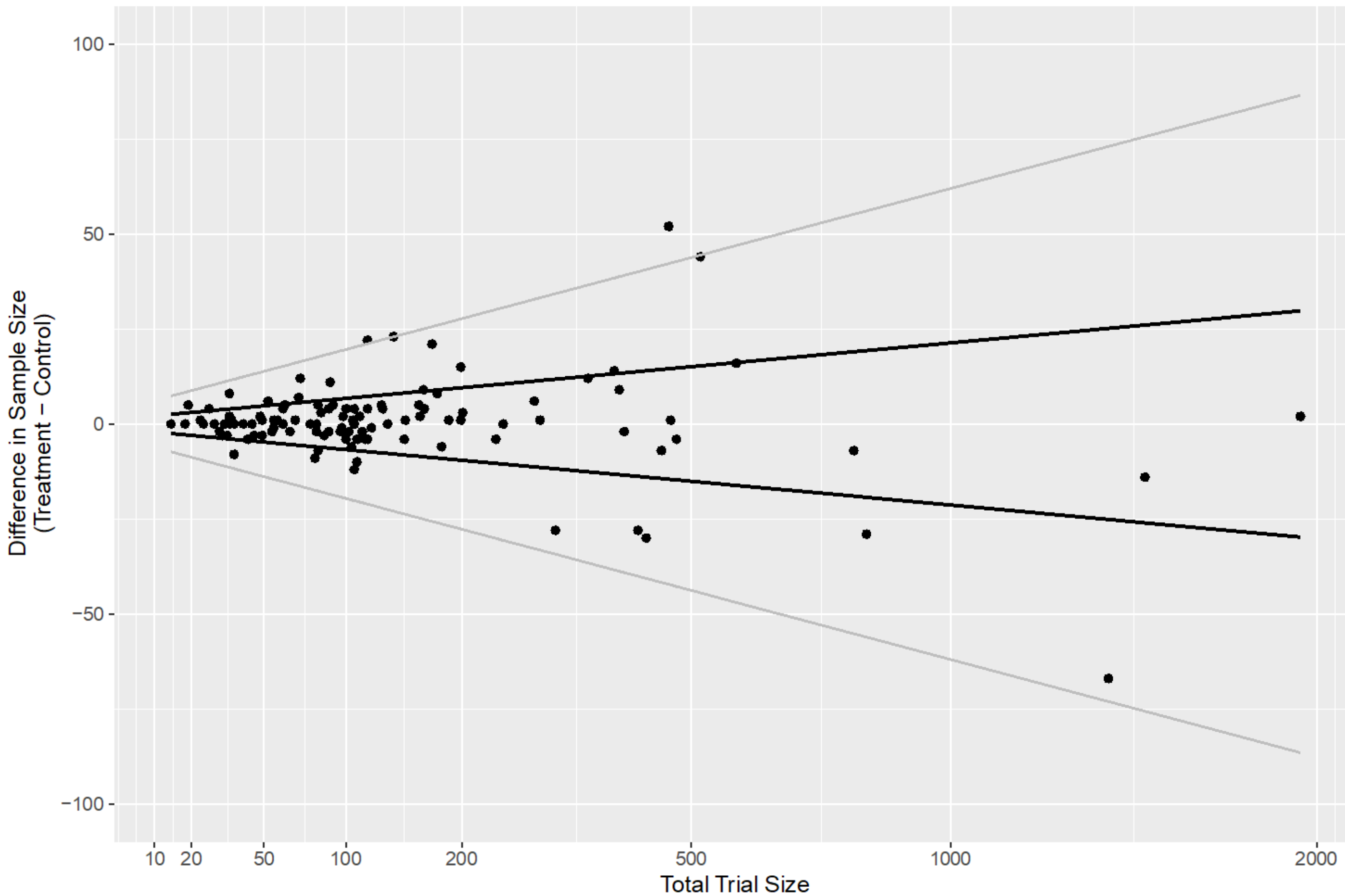
In the 203 included trials evaluated as published, the magnitude of the difference in the sample sizes of the two trial arms was within the 50% prediction interval for 163 trials (80%); for 34 trials (17%) the difference was between the 50% and 95% prediction intervals; only 6 trials (3%) had differences beyond the 95% prediction interval χ^2 (2, N = 203) = 74.96, p < .001 (Figure 2).

Figure 2. Differences in treatment and control arm sample sizes compared to total trial size, on the square root scale, for the N = 203 included trials as published. Dark gray lines contain 50% prediction interval; light gray lines contain 95% prediction interval.



In the main analysis, restricting to the 115 trials for which simple randomization was verified by authors, the number of trials with smaller differences between trial arm sample sizes than would be expected by chance was significantly greater than would be expected (χ^2 (2; N = 115) = 34.54; $p < .001$); 89 of 115 (77%) studies were within the 50% prediction interval; 23 (20%) were between the 50% and 95% prediction intervals; and three (3%) were beyond the 95% prediction interval. See Figure 3.

Figure 3. Differences in treatment and control arm sample sizes compared to total trial size, on the square root scale, for the N = 115 included trials with simple randomization as verified by authors. Dark gray lines contain 50% prediction interval; light gray lines contain 95% prediction interval.



For 62 trials with author-verified simple randomization published in *Health Psychology* or the *Journal of Consulting and Clinical Psychology*, 51 (82%) were within the 50% prediction interval, 10 (16%) were between the 50% and 95% prediction intervals, and only 1 (2%) was beyond the 95% prediction interval $\chi^2(2, N = 78) = 45.61$, $p < .001$ (figure not shown).

Among the 115 trials with author-verified simple randomization, there were 34 published from 2007-2010 and 81 from 2011-2017. For 2007-2010, 27 trials (79%) were within the 50% prediction interval, $\chi^2(2, N = 34) = 12.08$, $p = .002$. For 2011-2017, 62 trials (77%) were within the 50% interval, $\chi^2(2, N = 81) = 27.44$, $p < .001$. There was not any indication that continuous year of publication was associated with likelihood of a difference outside of the 50% interval (r = -0.03, 95% confidence interval -0.21 to 0.15).

Table 2 shows results from the analysis of factors that may be associated with greater likelihood of balanced trial arms. Results suggest that larger published trials may be less likely to have balanced group sample sizes than smaller trials and that higher impact factor may be associated with greater likelihood of reporting balanced sample sizes. There were, however, only 26 trials with differences outside of the 50% prediction interval, and confidence intervals were very wide for all variables in the model.

Table 2

Trial and journal factors associated with the likelihood of balanced trial arm sample sizes (within 50% prediction interval) among trials with author-verified simple randomization (N = 115)

	n within interval / n (%)	**Unadjusted odds ratio (95% CI)**	**Adjusted odds ratio (95% CI)**
N Randomized			
0 to 100 (reference)	48/59 (81%)	----------	----------
101 to 200	26/33 (79%)	0.85 (0.30 to 2.46)	0.73 (0.24 to 2.18)
≥ 201	15/23 (65%)	0.43 (0.15 to 1.27)	0.27 (0.08 to 0.94)
Simple Randomization Reported Adequately in Article			
Unclear in article (reference)	75/97 (77%)	----------	----------
Reported adequately	14/18 (78%)	1.03 (0.31 to 3.44)	1.23 (0.32 to 4.69)
Journal Impact Factor			
< 3	30/42 (71%)	----------	----------
≥ 3	59/73 (81%)	1.69 (0.69 to 4.10)	2.44 (0.88 to 6.75)

DISCUSSION

Among 203 trials published between 2007 and 2017 in APA-affiliated journals and described as two-arm parallel group RCTs without any restrictive randomization noted, the sample sizes in the two trial arms were more balanced than would have been plausible by chance if simple randomization had actually occurred in all of the trials. Among the 203 included trials, the differences in the sample sizes for the two trial arms were within the 50% prediction interval for these differences in 80% of trials. A query of trial authors resulted in 115 trials being classified as having used simple randomization and 88 trials being excluded. Results, however, were similar; 89 of 115 (77%) trials were within the 50% prediction interval.

Multivariate logistic regression analysis found that larger trial size may be associated with a decreased likelihood of reporting smaller relative differences between groups (differences within the 50% prediction interval). Compared to trials that randomized 100 participants or fewer (81% within 50% prediction interval), trials that randomized more than 200 (65% within interval) had an odds ratio of 0.27 (95% CI 0.08 to 0.94) times for being inside the prediction interval. A higher impact factor in the journal where trials were published may be associated with a greater likelihood of balanced sample sizes (odds ratio 2.44, 95% CI 0.88 to 6.75), although this was not statistically significant. These findings should be interpreted with great caution however, and used only as hypotheses; they came from post-hoc analyses with a small number of trials, and all estimates included very wide confidence intervals.

Reporting of randomization procedures in included trials was generally very poor. Of the 203 trials initially included, most only described that the trial had been randomized; only 18 (9%) accurately reported that simple randomization had been conducted or provided an explanation of the randomization procedure that would allow this to be determined definitively.

Furthermore, only 85 publications (42%) specified the mechanism by which random assignments had been generated (e.g., computer-generated numbers, coin tosses).

It is not possible to determine why the differences in the sizes of the trial arms in the 115 RCTs we reviewed with verified simple randomization were smaller than would be plausible if unrestricted simple randomization had occurred. However, our logistic regression results hint at the possibility that smaller sample size and journal impact factor may be associated with greater balance. One plausible explanation for this is that authors, reviewers and journal editors misconceive sample size imbalances as indicative of a flawed randomization process or poor trial quality. This could result in something akin to publication bias, which traditionally is understood to reflect preferences for publishing trials that find that interventions are effective (e.g., Franco, Malhotra, & Simonovits, 2014). In the present scenario, it is possible that small trials with imbalances are less likely to be published and that higher impact journals are more likely to publish trials with balanced group sizes. Greater balance in larger trials could also plausibly be consistent with subversion of randomization processes (Koletsi, Pandis, Polychronopoulou, & Eliades, 2012; Schulz, 1995). Although intentional subversion might seem unlikely, this phenomenon that has previously been described as witnessed by participants in a workshop on clinical trials (Schulz, 1995). Additionally, a study which anonymously surveyed 2000 psychologists found that many admitted to engaging in questionable research practices such as failing to report all dependent measures in publications and stopping data collection after achieving desired results (John, Loewenstein, & Prelec, 2012). If subversion does occur, it would be more likely to happen in small trials where manipulation of the status of only a few participants would balance trial arms; in large trials, where sample sizes may differ substantially by chance, this would require manipulation on a much larger scale.

The factors that we identified may explain some, but not all, of the excessive balance that we identified. Trial arm sample sizes were overly balanced even in large trials and in lower impact journals. Consistent with the poor reporting that we identified, it is possible that some investigators conducting trials do not fully understand randomization methods or how they should be reported. Anecdotally, in our author queries, in several instances, authors responded that they used simple randomization but provided an accompanying explanation that was inconsistent with this. In those cases, we followed up and clarified. It is possible, however, that some number of other investigators may have erroneously responded that they used simple randomization, despite our explanation, without providing us with enough information to identify a possible error. The poor reporting and implausibility of the degree of balance between sample sizes of different trial arms that we found in the present study add to concerns that have been raised previously about the conduct and reporting of trials of psychological interventions. Very few RCTs published in top clinical psychology journals, for instance, are prospectively and adequately registered, which raises the risk of bias from selective outcome reporting or non-publication of trials, depending on the trial results (Azar, et al., 2015; Cybulski, Mayo-Wilson, & Grant, 2016).

APA journals, including *Health Psychology* and the *Journal of Consulting and Clinical Psychology*, which publish far more trials than other APA journals, should implement steps to ensure enforcement of current RCT reporting guidelines, including CONSORT and JARS (Moher et al., 2001; Appelbaum et al., 2018; Schulz et al., 2010). The JARS guidelines, which are intended to be used for psychology studies published in APA journals, include one item regarding reporting of random assignment methodology which requires that the “procedure used to generate the random assignment sequence, including details of any restriction (e.g., blocking, stratification)” be reported (Appelbaum et al., 2018). The CONSORT Statement includes two items specific to trial randomization reporting: “method used to generate the random allocation sequence” and “type of randomization; details of any restriction (such as blocking and block size)”, as well as a recommended flow diagram, in order to easily understand trial group allocation processes (Moher et al., 2001).

Limitations

There are limitations that should be considered when interpreting the results of this study. First, the study search was exclusive to APA-affiliated journals, and relevant trials of psychological treatments published in non-APA affiliated journals were not included. Second, our search strategy relied on free text, and it is possible that we may not have encountered all eligible trials published in APA journals. Third, we are not able to determine if the findings resulted from poor reporting of restrictive randomization practices, inappropriate description of non-randomized trials as randomized, or subversion of randomization results. We conducted a post-hoc analysis to identify possible contributing factors. These raised hypotheses to consider, but the results should be interpreted with great caution given that they were post-hoc and that the study was not adequately powered to support robust analyses of this type. Indeed, it would be not be feasible to conduct such a study in APA journals; we included studies over a 10-year period, and going back further would reduce interpretability of the findings. Fourth, the quality of reporting of trial methods that we described in the present study is not representative of all trials in the journals we examined, because it is possible that trials with different design elements, such as cluster trials or trials that used restrictive randomization, may have been better reported. Fifth, we did not publish a protocol prior to initiation of our study, although methods were developed *a priori*.

Conclusion

The findings of this study suggest that in studies published in APA-affiliated journals, different trial arms in two-arm parallel group RCTs were far too close in size to have plausibly occurred if simple randomization had been properly conducted. Results did not differ among trials where authors verified the use of simple randomization, nor among more recently published trials. Post-hoc analyses suggested that larger trial sample size may be associated with less balance but that journal impact factor may be related to greater balance. These findings could be potentially explained by a publication bias type of phenomenon or by investigator manipulation. Future studies should determine how researchers, peer reviewers, and editors interpret balanced versus imbalanced sample sizes in trials, and efforts should be made to provide education on this issue. It is possible that the overly balanced trial arm sample sizes that we found could have resulted from misconceptions on the value of balance and publication bias, some degree of intentional subversion of the randomization process to attain similar group sizes when true randomization led to imbalances, or poor reporting and misunderstanding of randomization processes, even when an explanation was provided. Regardless of the reason, our findings raise serious concern that some psychological intervention research may either be too poorly reported to understand what has been done in trials or that some trials described as randomized may not truly be fully randomized. Clinicians, policy makers and the public depend on high-quality research to make decisions about the best psychological treatments for different problems; research practices that can introduce bias reduce confidence in the ability to use trial evidence for this purpose. It is important that peer reviewers and journal editors insist on adherence to reporting guidelines. At the same time, more focused education on principles of RCTs in graduate training could help address misconceptions related to randomized processes and group sample sizes, as well as the importance of reporting trial results clearly and transparently.

Authors’ Contribution:

MCA, DBR, and BDT were responsible for the study conception and design; data acquisition, analysis, and interpretation; and drafting and revising the manuscript. MC was responsible for the design and implementation of the search strategy. AWL was responsible for data analysis and interpretation. All authors reviewed and approved the final version of the manuscript.

Declaration of Interest Statement: The authors report no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices Statement: All data associated with this publication are openly accessible in the Appendices associated with this article at [url].

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119-126.

Altman, D. G., & Bland, J. M. (1999). How to randomise. *BMJ*, *319*(7211), 703-704.

Altman, D. G., & Bland, J. M. (2005). Treatment allocation by minimisation. *BMJ, 330*(7495), 843.

Altman, D. G., & Doré, C. J. (1990). Randomisation and baseline comparisons in clinical trials. *The Lancet*, *335*(8682), 149-153.

American Psychologocial Association. (2017). APA Clinical Psychology Journals. Retrieved from http://www.apa.org/pubs/journals/browse.aspx?query=subject:Clinical+Psychology&type=journal&sort=TitleAsc

American Psychological Association Task Force. (2006). APA Presidential Task Force on Evidence Based Practice. *American Psychologist, 61,* 271-285.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology. Why do we need them? What might they be? *American Psychologist*, *63*(9):839-51.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3.

Azar, M., Riehm, K. E., McKay, D., & Thombs, B. D. (2015). Transparency of outcome reporting and trial registration of randomized controlled trials published in the Journal of Consulting and Clinical Psychology. *PLoS One, 10*(11), e0142894.

Azar, M., Riehm, K. E., Saadat, N., Sanchez, T., Chiovitti, M., Qi, L., …Thombs, B. D. Evaluation of journal registration policies and prospective registration of randomized clinicla trials of nonregulated health care interventions. *JAMA Internal Medicine, 179*(5), 624-632.

Berger, V. W., & Bears, J. D. (2003). When can a clinical trial be called 'randomized'? *Vaccine, 21*(5-6), 468-472.

Chan, A.-W., & Altman, D. G. (2005). Epidemiology and reporting of randomised trials published in PubMed journals. *The Lancet, 365*(9465), 1159-1162.

Cybulski, L., Mayo-Wilson, E., & Grant, S. (2016). Improving transparency and reproducibility through registration: The status of intervention trials published in clinical psychology journals. *Journal of Consulting and Clinical Psychology, 84*(9), 753.

DerSimonian, R., Charette, L. J., McPeek, B., & Mosteller, F. (1982). Reporting on methods in clinical trials. *New England Journal of Medicine, 306*(22), 1332-1337.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publiation bias in the social sciences: unlocking the file drawer. *Science, 345(6203)*, 1502-1505.

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M., Granger, C. B. *Fundamentals of Clinical Trials, Fifth Edition* (pp. 123-146). Switzerland: Springer International Publishing.

Grimes, D. A., & Schulz, K. F. (1996). A "randomized" controlled trial without randomization. *American Journal of Obstetrics & Gynecology, 175*(1), 240-241.

Hewitt, C. E., & Torgerson, D. J. (2006). Is restricted randomisation necessary? *BMJ, 332*(7556), 1506-1508.

Jadad, & Enkin. (2007). Randomized controlled trials: the basics. *Randomized Controlled Trials: Questions, Answers, and Musings, Second Edition* (pp. 1-11). Oxford, UK: Blackwell Publishing.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Koletsi, D., Pandis, N., Polychronopoulou, A., & Eliades, T. (2012). What's in a title? An assessment of whether randomized controlled trial in a title means that it is one. *American Journal of Orthodontics and Dentofacial Orthopedics, 141*(6), 679-685.

Lachin J. M. (1988). Statistical properties of randomized clinical trials. *Controlled Clinical Trials*, *9*(4), 289-311.

Matts, J. P., Lachin, J. M. (1988). Properties of permuted-block randomization in clinical trials. *Controlled Clinical Trials, 9*(4), 327-344.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 339, b2535.

Moher, D., Schulz, K. F., Altman, D. G., & The Consort Group. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*, *357*(9263), 1191-1194.

Mosteller, F., Gilbert, J. P., & McPeek, B. (1980). Reporting standards and research strategies for controlled trials: agenda for the editor. *Controlled Clinical Trials, 1*(1), 37-58.

US News and World Report. (2016). Best Graduate Clinical Psychology Programs. Retrieved from https://www.usnews.com/best-graduate-schools/top-health-schools/clinical-psychology-rankings

Schulz, K. F. (1995). Subverting Randomization. *JAMA*, *274*, 1456-1458.

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, *8*(1), 18.

Schulz, K. F., Chalmers, I., Grimes, D. A., & Altman, D. G. (1994). Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA, 272*(2), 125-128.

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Jama, 273*(5), 408-412.

Schulz, K. F., & Grimes, D. A. (2002). Unequal group sizes in randomised trials: guarding against guessing. *The Lancet, 359*(9310), 966-970.

Simon, S. D. (2001). Is the randomized clinical trial the gold standard of research? *Journal of Andrology, 22*(6), 938-943.

Smith, P. J., Moffatt, M. E., Gelskey, S. C., Hudson, S., & Kaita, K. (1997). Are community health interventions evaluated appropriately? A review of six journals. *Journal of Clinical Epidemiology, 50*(2), 137-146.

Thombs, B. D., Levis, A. W., Azar, M., Saadat, N., Riehm, K. E., Sanchez, T. A.,…Kimmelman, J. (2020). Group sample sizes in non-regulated health care intervention trials described as randomized controlled trials were overly similar. *Journal of Clinical Epidemiology*, *120*, 8-16.

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically supported treatment: Recommendations for a new model. *Clinical Psychology: Science and Practice*, 22(4), 317-338.

Wei, L., & Lachin, J. M. (1988). Properties of the urn randomization in clinical trials. *Controlled Clinical Trials, 9*(4), 345-364.