Research on Social Work Practice
1-9

DOI: 10.1177/1049731515592477
rsw.sagepub.com
SAGE

Within-Group Effect-Size Benchmarks for Problem-Solving Therapy for Depression in Adults

Allen Rubin[1] and Miao Yu[1]

Abstract

This article provides benchmark data on within-group effect sizes from published randomized clinical trials that supported the efficacy of problem-solving therapy (PST) for depression among adults. Benchmarks are broken down by type of depression (major or minor), type of outcome measure (interview or self-report scale), whether PST was provided to elderly participants in poor health, and whether an intent-to-treat analysis was conducted. Practitioners can compare these benchmarks to their effect size in providing PST with depressed clients as a basis for deciding whether the way they are adopting or adapting this intervention is satisfactory or needs to be modified or replaced by a different intervention approach. These benchmarks also have potential utility for future implementation research on PST for depression.

Keywords

problem-solving therapy, benchmarking, depression, research supported interventions, implementation science

The progress made in recent decades in the number of randomized clinical trials (RCTs) and meta-analyses providing strong research support for the efficacy of a variety of interventions has been accompanied of late by a growing recognition of the need to study and improve the implementation of these interventions in nonresearch, everyday practice settings. One approach for doing so that is gaining attention in social work (Rubin, 2014) and clinical psychology (Spilka & Dobson, 2015) involves a benchmarking strategy. This article reports a benchmarking study on problem-solving therapy (PST) for depression.

Practitioners working with depressed clients can find various interventions with strong research support regarding their efficacy in treating depression among adults. The website of the Society of Clinical Psychology (http://www.div12.org/psychological-treatments/disorders/depression/), for example, lists the following six psychological treatments as having strong research support for treating depression: behavior therapy/behavioral activation, cognitive therapy, cognitive behavioral analysis system of psychotherapy, interpersonal therapy, self-management/self-control therapy, and PST. However, when practitioners in everyday practice settings implement interventions with strong research support, they cannot assume that they will obtain the same degree of outcome success as was found in the randomized clinical trials (RCTs) that provided the strong research support because the service provision conditions in practice settings tend to be less desirable than the relatively ideal service provision conditions that typify how interventions are provided in RCTs (Embry & Biglan, 2008; Spilka & Dobson, 2015; Weisz, Ugueto, Cheron, & Herren, 2013). For example, practitioners in everyday practice settings—as compared to their counterparts in the RCTs—are likely to have "less ideal training and supervision, larger and more diverse caseloads, more client attendance issues and barriers, and less commitment and adherence to treatment manuals" (Rubin, Parrish, & Washburn, 2014, p. 1). Likewise, agencies are likely to have higher rates of practitioner turnover, fewer funds for service provision, and a larger proportion of ethnic minority clients (Briere & Scott, 2012).

Because of these disparities, some have advocated modifying interventions with strong research support when they are implemented in real-world settings to make them a better fit for the service provision conditions and clientele of those agencies, particularly if the essential and indispensable core elements of the intervention are not modified (Galinsky, Fraser, Day, & Rothman, 2013; Sundell, Ferrer-Wreder, & Fraser, 2013). However, making those modifications means that sufficient intervention fidelity can no longer be assumed, which in turn raises doubt about the effectiveness of the intervention in practice settings. Doubt about the effectiveness also would be warranted—even without any modifications—in light of the

[1] Graduate College of Social Work, University of Houston, Houston, TX, USA

Corresponding Author:
Allen Rubin, University of Houston, 110HA Social Work Building—Room 342, Houston, TX 77204, USA.
Email: arubin2@central.uh.edu

various aforementioned service provision condition disparities between RCTs and everyday practice settings.

Consequently, practitioners in everyday practice settings should not merely assume that the implemented intervention is an appropriate one for their setting only because it has strong research support. Instead, they should monitor pre-to-post client outcomes as one basis for deciding whether the intervention seems to be as good a fit for their setting as they had hoped it would be. Because well-controlled experimental and quasi-experimental designs typically are not feasible in everyday practice settings, most such settings will be limited to assessing the pre-to-post treatment progress made by clients receiving the intervention, only—without a control group. However, if the aggregate progress of intervention recipients is represented by a within-group effect-size statistic akin to Cohen's *d*, that statistic can be compared to the mean within-group effect sizes reported for the experimental groups and the control groups in the RCTs. If the within-group effect size for agency clients approximates the mean within-group effect size of recipients of the intervention in the RCTs, that would supply evidence supporting the notion that the intervention—and the way it is being implemented—is a satisfactory fit for that agency. Conversely, if the agency's within-group effect size is much closer to the RCT control groups' within-group mean effect size, that would indicate the need to further modify the intervention to make it a better fit or perhaps to switch to a different intervention—preferably one with adequate research support, if possible.

It bears emphasizing here that the point of such an assessment would not be to evaluate the efficacy of the chosen intervention, whose efficacy would already have strong research support provided by various RCTs. Thus, the agency's within-group pre-to-post effect size would not purport to establish causality (i.e., by controlling for threats to internal validity); instead, its aim is to provide descriptive data that can be compared to benchmarks from the RCTs that will offer an empirical basis for decisions about whether to continue, modify, or replace the intervention being used when treating depressed clients. Agencies typically make such decisions without conducting controlled experiments or quasi experiments. Comparing their within-group effect size to the mean RCT effect sizes would improve the empirical basis for making such decisions. (Although there is value in an agency assessment that is limited to comparing within-group effect sizes for interventions with strong research support, that is not meant to imply that additional RCT evaluations of such interventions are not needed. Perhaps future studies will contradict the strength of the support in the extant literature. Moreover, interventions that are effective in some contexts might not be effective in other contexts).

With the previously mentioned reasoning in mind, the study reported in this article is the second in a planned series of studies seeking to provide benchmarks enabling agencies to make the foregoing comparisons. Like the previous study, this one examined each individual published report of an RCT that is included in the meta-analyses that provided strong research support for the intervention in question. The previous study focused on interventions with strong research support in the treatment of adult traumatic stress (Rubin et al., 2014). The current study focused on an intervention with strong research support in the treatment of adult depression: PST for depression. The selection of that intervention—as opposed to others with strong research support for treating depression—was based primarily on the lead author's prior work and familiarity regarding that intervention. By calculating and reporting descriptive mean within-group effect size statistics across all studies separately for the RCT experimental group participants and their counterparts in waitlist or treatment as usual control groups, this study seeks to provide benchmarks to which practitioners providing PST for depression in everyday practice settings can compare their own within-group effect sizes. These benchmarks also might be useful in future implementation research studies that seek to discover those service provision characteristics whose within-group effect sizes resemble the benchmarks provided in this study.

PST for Depression

The Society of Clinical Psychology describes PST as an intervention that teaches clients to generate more effective solutions for their problems, particularly regarding pursuing goals and dealing with interpersonal conflict. Practitioners help clients learn and effectively apply problem solving through the following six steps: "1) identifying problems, 2) generating multiple alternative solutions, 3) selecting the best solution from the alternatives, 4) developing a plan, 5) implementing the problem solving tactic, and 6) evaluating the efficacy of problem solving" (Society of Clinical Psychology, p. 1). Cuijpers, van Straten, and Warmerdam (2007) classified three types of PST, including (1) PST that focuses on social problems, (2) PST that emphasizes self-examination to help clients formulate major goals, assess problems in achieving those goals, and engage in efforts to solve controllable problems and accept uncontrollable ones, and (3) PST for primary care settings. The latter type was the case in 6 of the 13 articles included in our sample, in which the participants were elderly and in poor health. Consisting of about 6 treatment sessions, PST in primary care settings tends to be briefer than other forms of PST, which usually range from 8 to 16 sessions.

Methodology

Study Search and Selection

Our first step was to conduct a broad Internet search for meta-analyses and systemic reviews on the efficacy of PST for depression. This included the following databases: Google Scholar, Web of Science, PubMed, and PsycINFO. Four meta-analyses and one systematic review were found. The next step involved a targeted search using the same databases to locate additional RCTs on the efficacy of PST for depression that were published too late to be included in the meta-analyses.

Table 1. Characteristics of Included Studies.

Characteristic	Frequency
Type of problem-solving therapy	
Primary or home care	6
Other	6
Type of depression	
Major	5
Minor	7
Number of treatment sessions	
5 to 7	7
8 to 12	5
Type of control group	
Waitlist	3
Treatment as usual	9
Percentage of women	
4 to 36	2
65 to 79	4
81 to 100	5
Not reported	1
Percentage of ethnic minorities	
Less than 20	4
22 to 42	4
Not reported	4
Intent-to-treat analysis?	
Yes	6
No	6
Percentage attrition if no intent-to-treat analysis	
Less than 10	4
32	1
60	1
Not applicable	6

These searches found 33 RCT studies that were considered for inclusion. However, 20 of these studies were excluded because they did not provide sufficient data to calculate within-group effect-sizes. Five others (of the 20 excluded studies) were excluded when upon a close reading we discovered that the intervention being evaluated either was not really PST at all or was something being called PST but which involved deviations adding components to it or changing it in other ways that made the intervention too different than in the studies that did not change the modality. For example, two of these studies evaluated self-examination therapy and/or bibliotherapy (Bowman, Scogin, & Lyrene, 2010; Bowman, Ward, Bowman, & Scogin, 1996). Another provided PST online, only and with a 42% attrition rate (Ince, Cuijpers, et al., 2013). One made substantial changes to the PST modality and had it provided by lay workers (Chibanda, Mesu, et al., 2011). A fifth excluded study added 8 weekly hour-long (unspecified) psychotherapy sessions to the PST sessions (Hopko, Armento, et al., 2011). PST with. Another two of the 20 studies were excluded because the participants were not depressed or because depression was not measured as an outcome variable. One study was excluded because it was not in the English language.

These exclusions resulted in a preliminary sample of three studies. However, we subsequently decided to exclude one of these 13 studies for the following two reasons: (1) it was the only one of the 13 studies that did not provide the PST in person (providing it instead either via telephone or Skype) and (2) it used a nonblind interview measurement procedure that yielded outlier effect sizes that were at least several times larger than any of the other 12 studies' effect sizes and more than double the next largest effect size. Thus, the remaining 12 studies provided the benchmarks in our findings.

Table 1 displays the number of the 12 included studies broken down by type of PST (whether in primary care), type of depression (major or minor), number of treatment sessions, type of control group (waitlist vs. treatment as usual), percentages of women and minorities, whether an intent-to-treat (ITT) analysis was employed, and the percentage of attrition among PST recipients in studies not conducting an ITT analysis.

Data Collection Process

The following information (presented in the Appendix) was recorded for each study: authors and year of study, whether the intervention was provided as part of primary care or home care (for clients with serious health problems), whether it was combined with treatment as usual or with medication for depression, type of depression (major or minor), number of treatment sessions, whether measurement was blinded, type of control condition (waitlist or treatment as usual), sample size of each group, target population (% Caucasian/White and % Female), whether recipients were elderly and whether they were in poor health, whether an ITT analysis was conducted, percentage attrition in control group if no ITT analysis, and whether reasons for attrition were health or death related. For the benchmark calculations, we recorded the means and standard deviations at pretest and posttest for each outcome measure for treatment recipients and controls. Separate recordings were made for interviewer and self-administered outcome measures of level of depression. The reported interview measure was the Hamilton Depression Rating Scale (HAM-D), which is validated and commonly used in depression outcome research (McDowell, 2006). The most commonly used self-report measure was the Beck Depression Inventory (BDI; Beck, Steer, & Carbin, 1988). We did not record follow-up (after posttest) data because practice settings for whom our benchmarks are provided rarely conduct follow-up assessments of their clients months after the completion of treatment. The two authors of this article independently recorded the data, and the interrater agreement was 98%. The few instances involving discrepancies were resolved via a joint reexamination of the original article and reaching consensus regarding the correct datum to be entered.

Benchmark Calculations

Within-group effect sizes were calculated according to Glass's δ approach (Glass, 1976). In studies of between-group effect sizes, this approach divides the difference between experimental and control group means by the control group standard

Table 2. Aggregate Within-Group Effect-Size Estimates by Type of Depression Measure.

	Self-Report		Interview (HRSD)	
	PST	Control	PST	Control
K	12	11	7	7
g_B	0.925	0.321	2.071	0.738
SE_g	0.034	0.030	0.118	0.082
Minimum g_{Hedges}	0.753	−0.095	0.703	−0.026
Maximum g_{Hedges}	6.178	0.962	5.184	1.860

Note. k = numbers of individual studies; g_B = aggregate effect size; g_{Hedges} = individual study effect size; HRSD = Hamilton-D interview scale; PST = problem-solving therapy.

deviation. For within-group effect sizes, the difference between the pretest and posttest means is divided by the pretest standard deviation. This calculation is done separately to calculate a within-group effect size for the experimental group and a within-group effect size for the control group (Feingold, 2009; Kadel & Kip, 2012; Maier-Riehle & Zwingmann, 2000). So as to account for small sample sizes of some of the RCTs, we adjusted each effect size by calculating Hedge's g, using a formula recommended by the Campbell Collaboration in which the effect size is multiplied by a fraction with three in the numerator and (4 N) −9 in the denominator (Wilson, 2011).

The individual study effect sizes were then averaged across studies by using the following formula (Minami et al., 2008). First, the variance of the individual study effect sizes was estimated as follows: $2(1-r_i)/n_i + \frac{g_i^2}{2n_i}$. Next, the variance and effect size (g_i) were used to estimate the fixed benchmark effect size across studies using the following formula: $g_B = \left[\frac{\Sigma_i g_i}{\delta^2 g(i)}\right]$ over $\left[\Sigma_i \frac{1}{\delta^2 g(i)}\right]$.

In addition, minimum and maximum benchmark effect sizes are reported to show the range of effect sizes because of the potential for skewed distributions in which confidence interval estimates could exceed the minimum or maximum effect size due to the relatively small sample of studies ($N = 12$).

Results

Our first analysis of the data performed separate calculations for BDI self-reports versus other depression self-report scales. Because the aggregate within-group effect sizes were so similar for both of these categories, we decided to combine them to simplify the tables of results, thus facilitating their use by practitioners. We report the interview effect sizes separately in the tables, however, because they were consistently higher than the effect sizes from the self-report scales.

Table 2 displays descriptive statistics on the central tendency and dispersion of the within-group effect sizes for each of the two categories of measures: (1) self-report scales measuring depression and (2) the Hamilton-D interview scale (HRSD). It shows that the aggregate within-group effect size (g_B) is 0.912 for PST recipients when depression is measured by a self-report scale and 2.071 when it is measured in an HRSD interview. The referent aggregate within-group effect sizes for the control groups are much smaller.

Table 3. Aggregate Within-Group Effect-Size Estimates by Type of PST.

		Self-Report		Interview (HRSD)	
		PST	Control	PST	Control
PST	*K*	5	5	3	3
	g_B	1.611	0.233	2.939	1.215
	SE_g	0.129	0.085	0.188	0.124
	Minimum g_{Hedges}	0.753	−0.095	2.793	0.070
	Maximum g_{Hedges}	6.178	0.551	5.184	1.860
PST-Primary Care or Home Care	*K*	6	6	4	4
	g_B	0.869	0.333	1.510	0.372
	SE_g	0.036	0.032	0.151	0.109
	Minimum g_{Hedges}	0.819	0.059	0.703	−0.026
	Maximum g_{Hedges}	1.494	0.962	2.341	1.243

Note. k = numbers of individual studies; g_B = aggregate effect size; g_{Hedges} = individual study effect size; HRSD = Hamilton-D interview scale; PST = problem-solving therapy.

Table 4. Aggregate Within-Group Effect-Size Estimates by Type of Depression.

		Self-Report		Interview (HRSD)	
		PST	Control	PST	Control
Major	*K*	4	4	4	4
	g_B	0.851	0.362	2.752	1.221
	SE_g	0.038	0.033	0.164	0.111
	Minimum g_{Hedges}	0.819	0.344	2.164	0.070
	Maximum g_{Hedges}	6.178	0.962	5.184	1.860
Minor	*K*	7	7	3	3
	g_B	1.312	0.170	1.341	0.153
	SE_g	0.086	0.064	0.170	0.122
	Minimum g_{Hedges}	0.753	−0.095	0.703	−0.026
	Maximum g_{Hedges}	1.834	0.337	2.341	0.547

Note. k = numbers of individual studies; g_B = aggregate effect size; g_{Hedges} = individual study effect size; HRSD = Hamilton-D interview scale; PST = problem-solving therapy.

Table 3 breaks down the same statistics by the type of PST provided, with results showing smaller effect sizes for recipients of PST in primary or home care, who tend to be older and in worse health than other recipients of PST that receive more sessions. (Additional tables are not presented for breakdowns by whether recipients were elderly or in poor health because they would be redundant both conceptually and in their results with Table 3. We also opted not to include a table with a waitlist versus ITT analysis because of the very small number of studies using waitlist controls and the similarity of the results for both categories.)

Table 4 displays the breakdown by the type of depression. It shows a larger interview aggregate effect size for recipients

Table 5. Aggregate Within-Group Effect-Size Estimates by Blinded Versus Not Blinded Assessors.

		Self-Report		Interview (HRSD)	
		PST	Control	PST	Control
Blinded	*K*	7	7	5	5
	g_B	0.907	0.348	1.826	0.339
	SE_g	0.036	0.031	0.162	0.103
	Minimum g_{Hedges}	0.819	0.253	0.703	−0.026
	Maximum g_{Hedges}	6.178	0.962	5.184	1.243
Not Blinded	*K*	4	4	2	2
	g_B	1.109	0.072	2.348	1.437
	SE_g	0.115	0.095	0.172	0.136
	Minimum g_{Hedges}	0.753	−0.095	1.437	0.547
	Maximum g_{Hedges}	1.834	0.337	2.870	1.859

Note. k = numbers of individual studies; g_B = aggregate effect size; g_{Hedges} = individual study effect size; HRSD = Hamilton-D interview scale; PST = problem-solving therapy.

Table 6. Aggregate Within-Group Effect-Size Estimates by Whether ITT Analysis Was Conducted.

		Self-Report		Interview (HRSD)	
		PST	Control	PST	Control
Intent-to-treat (ITT)	*K*	5	5	3	3
	g_B	0.896	0.315	2.062	0.796
	SE_g	0.036	0.031	0.151	0.107
	Minimum g_{Hedges}	0.819	0.059	0.703	−0.026
	Maximum g_{Hedges}	1.826	0.344	2.870	1.860
Not ITT	*K*	6	6	4	4
	g_B	1.304	0.391	2.085	0.657
	Seg	0.128	0.102	0.189	0.127
	Minimum g_{Hedges}	0.753	−0.095	1.437	0.070
	Maximum g_{Hedges}	6.178	0.962	5.184	1.243

Note. *SE* = standard error; k = numbers of individual studies; g_B = aggregate effect size; g_{Hedges} = individual study effect size; HRSD = Hamilton-D interview scale; PST = problem-solving therapy.

with major depression. Table 5 displays the breakdown by whether assessors were blind as to participants' group status. The most noteworthy aspect of that table is the difference in aggregate effect sizes between blinded and not blinded interviews. Table 6 shows similar aggregate effect sizes regardless of whether an ITT analysis was conducted.

The reason the sum of the number of studies (k) in some of these tables exceeds our total of 12 studies is that some studies measured outcome both by self-report and by interview. We aggregated our findings separately for these categories in light of the notion that interviews might be more susceptible to measurement bias and thus yield larger effect sizes. Likewise, we wanted practitioners to be able to compare their effect sizes to a benchmark that fits the way they measure outcome. For example, if they use self-report scales and were to compare their effect size to a benchmark that combined self-report scales and interviews, their benchmarks might be at a comparative disadvantage, and the comparison therefore might underestimate the adequacy of their outcomes.

Practitioners who calculate their setting's own within-group effect size for the type of PST that they adopt or adapt can compare their results to the data in Tables 2 through 6 according to the column (type of depression outcome measure) that applies to their assessment (whether self-report vs. interview) and the row that applies to characteristics of their clients, PST approach, and whether they use an ITT analysis (as opposed to analyzing data on treatment completers, only). In addition to comparing their within-group effect size to the aggregate effect-sizes for PST recipients, they might want to make comparisons to the control group data in the tables. The extent to which their PST recipient effect size more closely approximates the corresponding PST recipients effect size in a table, the greater the grounds for optimism regarding whether their provision of PST appears to be acceptable, and vice versa if it more closely approximates the control group effect size. Before making these comparisons, they should adjust their effect size by the number of their PST recipients, according to the g_{Hedges} formula mentioned previously, in which the unadjusted effect size is multiplied by a fraction with three in the numerator and (4 N) −9 in the denominator.

Discussion

This study calculated and reported descriptive statistics on within-group effect sizes from published studies (all but one being RCTs) that evaluated the efficacy of PST for the treatment of adult depression. The purpose of doing so was to provide benchmarks that practitioners in practice settings can compare to their own within-group PST effect size as a basis for deciding whether the way they are providing the PST is satisfactory or needs to be modified or replaced by a different intervention approach.

Optimism about the adequacy of the practice setting's provision of PST would be supported to the extent that its within-group effect size approximates the aggregate effect size of the PST recipients in this study's tables and the extent to which it exceeds the control group effect sizes in those tables. Conversely, if its effect size more closely approximates the control group aggregate effect sizes in this study's tables that might imply the need to modify its approach to PST or replace PST with a different intervention that might be a better fit for its setting. The importance of calculating the within-group effect size in the practice setting and then comparing it to the appropriate benchmark/benchmarks is based on the recognition that when interventions with strong research support are implemented in everyday practice settings, in all likelihood, they will not be implemented with the same fidelity as they were implemented in the research studies because of differences in training, supervision, staffing, caseload sizes, client attendance issues, and other service provision resources. Moreover, practitioners in everyday practice settings—in keeping with the steps of the evidence-based practice process—might decide that instead of implementing the treatment with precise fidelity it

is more appropriate clinically to modify it to improve its fit with the agency's clientele characteristics and needs (Rubin & Bellamy, 2012).

Although our study employs some meta-analytic techniques, it is not a full-fledged meta-analysis. Its sample does not include unpublished studies because its purpose is to provide benchmarks derived from published studies that provided the strong research support for the efficacy of PST. Thus, the issue of a bias favoring studies whose findings support the efficacy of an intervention is not relevant to the aim of our study. That is, our aim did not address the question of whether PST is, in fact, effective. Instead, recognizing that PST has already been recognized as having strong research support, we merely wanted to provide descriptive benchmark data that will enable practitioners who are providing PST in everyday practice settings to compare their outcomes to those of the studies that have provided the strong research support for PST. For the same reason, we focused on within-group effect sizes instead of between-group effect sizes that are the foci of meta-analyses.

One limitation in our study is that our benchmarks are derived from only 12 studies. Our search identified 333 studies that were considered for potential inclusion. However, for various reasons mentioned earlier, 21 of those studies had to be eliminated. Although 12 is not a large number of studies, we believe that it is—at the time of this writing—the totality of existing studies for which deriving these benchmarks is appropriate. The issue regarding the relatively small number of studies is obviated by the fact that our study did not aim to support the efficacy of PST for depression. Its efficacy was recognized as having strong research support before we embarked on this study, and we simply want to provide practice settings with useful benchmarks that they can compare to the studies that provided that research support.

One caveat regarding comparisons to our benchmarks is whether practice settings triage provision of PST based on elevated pretest depression scores. If they do, a comparison to our benchmarks would be inappropriate because of vulnerability to regression to the mean. Such vulnerability does not apply to RCTs because their use of use random assignment controls for regression to the mean.

When comparing practice setting within-group effect sizes to the benchmarks provided in our study, practitioners should not limit the comparison to our aggregate effect size (g_B), only. Even if a practice setting's adjusted within-group effect size (g_{Hedges}) does not approximate the corresponding aggregate effect size in one of our tables, value might be found in comparing it to the minimum and maximum effect sizes (g_{Hedges}) in the table. For example, suppose the practice setting conducts an ITT analysis that results in an adjusted within-group (g_{Hedges}) effect size of 0.70. That would be notably less than the corresponding aggregate effect size of 0.896 in Table 6. However, it would approximately double the maximum g_{Hedges} effect size of 0.344 of the corresponding control groups. Especially when considering the likelihood that practice settings commonly operate under service provision conditions far less desirable than in RCT studies, the 0.70 effect size in the practice setting might be viewed somewhat favorably and might be deemed grounds for continuing to provide the PST intervention (perhaps with only minor tweaks) even if in their setting they were not achieving outcomes as impressive as those in the research studies.

Some of the control group aggregate effect sizes in our tables approximate what Cohen (1988) has deemed to be of medium strength. This was not surprising, for two reasons. First, the previous study in this series (Authors, 2014) found the same phenomenon in the study of within-group effect sizes for trauma interventions. Second, those authors noted, "it makes sense that the control group effect sizes should be that large, because they represent pre to post change within one group, and not differences between a treatment and control group. Therefore, deeming them to be medium would be to apply an inappropriate standard—one that disregards factors like the impact of contemporaneous events (history) or the passage of time on client improvement. That such factors can in fact explain some of the pre to post improvement is the reason why control groups are used in outcome studies" (p. 9).

Conclusion

This study has provided benchmark data on within-group effect sizes from published studies that provided strong research support for the efficacy of PST in the treatment of adult depression. Researchers conducting future implementation science studies regarding PST for depression might also find our benchmarks useful in that the within-group effect size results of efforts to promote successful implementation of PST in real-world agencies can be compared to our benchmarks. This might enhance the development and eventual testing of hypotheses about what service provision conditions are associated with results that most closely approximate our benchmarks. In light of this potential utility, another implication for future research is to conduct benchmark studies for additional interventions recognized for having strong research support.

Appendix

Table A. Demographics and Effect Sizes for Studies Selected for Inclusion of benchmark Calculations.

	Demographic Characteristics			Clinical Characteristics		Treatment Conditions and Constructs Measured				
Study	*N*	% Female	% Min	Type of Depression	Type of PST	ITT	Group	Measures	Effect Size[a]	*SE*
Arean et al. (1993)	48	77	23	Major	Regular	No	PST	Self-report	1.49	.27
								HRSD	2.79	.42
							Waitlist	Self-report	.40	.23
								HRSD	.07	.22
Arean et al. (2010)	220	ND	ND	Major	Regular	Yes	PST	HRSD	2.87	.22
							TAU	HRSD	1.86	.17
Ciechanowski et al. (2004)	138	79	42	Minor	Home care	yes	PST-HC	Self-report	1.17	.15
							TAU	Self-report	.06	.12
Erdley et al. (2014)	33	36	6	Minor	Regular	No	PST	Self-report	.75	.29
							TAU	Self-report	−.10	.24
Gellis and Bruce (2010)	36	92	6	Minor	Home care	Yes	PST-HC	Self-report	1.45	.34
								HRSD	.70	.26
							TAU	Self-report	.25	.24
								HRSD	.09	.24
Gellis et al. (2008)	62	88	15	Minor	Home care	Yes	PST-HC	Self-report	1.14	.23
								HRSD	2.34	.35
							TAU	Self-report	.25	.18
								HRSD	−.03	.18
Hirai et al. (2012)	23	100	ND	Minor	Regular	No	PST	Self-report	.39	.22
							ND	Self-report	—	—
Kasckow et al. (2014)	45	4	22	Minor	Primary care	No	PST-PC	Self-report	1.02	.25
								HRSD	1.44	.29
							TAU	Self-report	.34	.23
								HRSD	.55	.24
Mynorswallis, Gath, Lloydthomas, and Tomlinson (1995)	60	77	5	Major	Primary care	No	PST-PC	Self-report	1.49	.27
								HRSD	2.16	.33
							TAU	Self-report	.96	.22
								HRSD	1.24	.24
Nezu (1986)	21	81	ND	Minor	Regular	No	PST	Self-report	1.83	.47
							Waitlist	Self-report	−.05	.33
Nezu and Perri (1989)	28	82	ND	Major	Regular	No	PST	Self-report	6.18	1.16
								HRSD	5.18	.98
							Waitlist	Self-report	.55	.30
								HRSD	1.12	.35
Unutzer et al. (2002)	1801	65	23	Major	Primary care	Yes	PST-PC	Self-report	.82	.04
							TAU	Self-report	.34	.03
Vázquez González et al. (2013)	173	100	ND	Minor	Regular	Yes	PST	Self-report	1.83	.17
							TAU	Self-report	.26	.11

Note. ITT = intent-to-treat; *SE* = standard error; PST = problem-solving therapy; HRSD = Hamilton-D interview scale; ND = no data; HC = home care; PC = primary care; TAU = treatment as usual.

[a] g_{Hedges}.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Arean, P. A., Perri, M. G., Nezu, A. M., Schein, R. L., Christopher, F., & Joseph, T. X. (1993). Comparative effectiveness of social problem-solving therapy and reminiscence therapy as treatments for depression in older adults. *Journal of Consulting and Clinical Psychology*, *61*, 1003–1010.

Arean, P. A., Raue, P., Mackin, R. S., Kanellopoulos, D., McCulloch, C., & Alexopoulos, G. S. (2010). Problem-solving therapy and

supportive therapy in older adults with major depression and executive dysfunction. *American Journal of Psychiatry*, *167*, 1391–1398.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*, 77–110.

Bowman, D., Scogin, F., & Lyrene, B. (2010). The efficacy of self-examination therapy and cognitive bibliotherapy in the treatment of mild to moderate depression. *Psychotherapy Research*, *5*, 131–140.

Bowman, V., Ward, L. C., Bowman, D., & Scogin, F. (1996). Self-examination therapy as an adjunct treatment for depressive symptoms in substance abusing patients. *Addictive Behaviors*, *21*, 129–133.

Briere, J. N., & Scott, C. (2012). *Principles of trauma therapy* (2nd ed.). Thousand Oaks, CA: Sage.

Chibanda, D., Mesu, P., Kajawu, L., Cowan, F., Araya, R., & Abas, M. (2011). Problem-solving therapy for depression and common mental disorders in Zimbabwe: Piloting a task-shifting primary mental health care intervention in a population with a high prevalence of people living with HIV. *BMC Public Health*, *11*, 828–837.

Ciechanowski, P., Wagner, E., Schmaling, K., Schwartz, S., Williams, B., Diehr, P., . . . LoGerfo, J. (2004). Community-integrated home-based depression treatment in older adults—A randomized controlled trial [Article]. *Journal of the American Medical Association*, *291*, 1569–1577.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Lawrence Erlbaum Associates.

Cuijpers, P., van Straten, A., & Warmerdam, L. (2007). Problem solving therapies for depression: A meta-analysis. *European Psychiatry*, *22*, 9–15.

Embry, D. D., & Biglan, A. (2008). Evidence-based kernels: Fundamental units of behavioral influence. *Clinical Child and Family Psychology Review*, *11*, 75–113.

Erdley, S. D., Gellis, Z. D., Bogner, H. A., Kass, D. S., Green, J. A., & Perkins, R. M. (2014). Problem-solving therapy to improve depression scores among older hemodialysis patients: A pilot randomized trial [Article]. *Clinical Nephrology*, *82*, 26–33.

Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods*, *14*, 43–53.

Galinsky, M., Fraser, M. W., Day, S. H., & Rothman, J. M. (2013). A primer for the design of practice manuals: Four stages of development. *Research on Social Work Practice*, *23*, 219–228.

Gellis, Z. D., & Bruce, M. L. (2010). Problem-solving therapy for subthreshold depression in home healthcare patients with cardiovascular disease. *American Journal of Geriatric Psychiatry*, *18*, 464–474.

Gellis, Z. D., McGinty, J., Tierney, L., Jordan, C., Burton, J., & Misener, E. (2008). Randomized controlled trial of problem-solving therapy for minor depression in home care. *Research on Social Work Practice*, *18*, 596–606.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3–8.

Hirai, K., Motooka, H., Ito, N., Wada, N., Yoshizaki, A., Shiozaki, M., . . . Akechi, T. (2012). Problem-solving therapy for psychological distress in Japanese early-stage breast cancer patients. *Japanese Journal of Clinical Oncology*, *42*, 1168–1174.

Hopko, D. R., Armento, M. E. A., & 9 others. (2011). Brief behavioral activation and problem-solving therapy for depressed breast cancer patients: Randomized trial. *Journal of Consulting and Clinical Psychology*, *79*, 834–849.

Kadel, R. P., & Kip, K. E. (2012). *A SAS macro to compute effect size (Cohen's d) and its confidence interval from raw survey data*. Retrieved from Analytics.ncsu.edu/2012/SD-06.pdf

Kasckow, J., Klaus, J., Morse, J., Oslin, D., Luther, J., Fox, L., . . . Haas, G. L. (2014). Using problem solving therapy to treat veterans with subsyndromal depression: A pilot study [Article]. *International Journal of Geriatric Psychiatry*, *29*, 1255–1261.

Maier-Riehle, B., & Zwingmann, C. (2000). Effect strength variation in the single group pre-post study design: A critical review [Article in German]. *Rehabilitation (Stuttg)*, *39*, 189–199. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11008276

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires* (3rd ed.). New York, NY: Oxford University Press.

Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C., & Brown, G. S. (2008). Using clinical trials to benchmark effects produced in clinical practice. *Quality and Quantity*, *42*, 513–525.

Mynorswallis, L. M., Gath, D. H., Lloydthomas, A. R., & Tomlinson, D. (1995). Randomized controlled trial comparing problem-solving treatment with amitriptyline and placebo for major depression in primary-care [Article]. *British Medical Journal*, *310*, 441–445.

Nezu, A. M. (1986). Efficacy of a social problem-solving therapy approach for unipolar depression. *Journal of Consulting and Clinical Psychology*, *54*, 196–202.

Nezu, A. M., & Perri, M. G. (1989). Social problem-solving therapy for unipolar depression-an initial dismantling investigation. *Journal of Consulting and Clinical Psychology*, *57*, 408–413.

Rubin, A. (2014). An alternative paradigm for social workers seeking to do intervention research. *Social Work Research*, *38*, 69–71.

Rubin, A., & Bellamy, J. (2012). *Practitioner's guide to using research for evidence-based practice*. Hoboken, NJ: John Wiley & Sons.

Rubin, A., Parrish, D. E., & Washburn, M. (2014). Outcome benchmarks for adaptations of research-supported treatments for adult traumatic stress. *Research on Social Work Practice*, published online before print at rsw.sagepub/content/early/2014/09/10/1049731514547906.refs

Spilka, M. J., & Dobson, K. S. (2015). Promoting the internationalization of evidence-based practice: Benchmarking as a strategy to evaluate culturally transported psychological treatments. *Clinical Psychology Science and Practice*, *22*, 58–75.

Sundell, K., Ferrer-Wreder, L. R., & Fraser, M. W. (2013). Going global: A model for evaluating empirically supported family-based interventions in new contexts. *Evaluation & the Health Professions*, *37*, 203–230.

Unlu Ince, B., Cuijpers, P., & 4 others. (2013). Internet-based, culturally sensitive, problem-solving therapy for Turkish migrants with depression: Randomized control trial. *Journal of Medical Internet Research*, 15. Downloaded from http://www.ncbi.nlm.nih.gov/pubmed/24121307

Unutzer, J., Katon, W., Callahan, C. M., Williams, J. W., Hunkeler, E., Harpole, L., . . . Investigators, I. (2002). Collaborative care management of late-life depression in the primary care setting—A randomized controlled trial [Article]. *Journal of the American Medical Association*, *288*, 2836–2845.

Vázquez González, G. F. L., Otero Otero, P., Torres Iglesias, A., Hermida Garcia, E., Blanco Seoane, V., & Diaz Fernandez, O. (2013). A brief problem-solving indicated-prevention intervention for prevention of depression in nonprofessional caregivers. *Psicothema*, *25*, 87–92.

Weisz, J. R., Ugueto, A. M., Cheron, D. M., & Herren, J. (2013). Evidence-based youth psychotherapy in the mental health ecosystem. *Journal of Clinical Child & Adolescent Psychology*, *42*, 274–286.

Wilson, D. B. (2011). *Calculating effect-sizes. The Campbell collaboration*. Retrieved from http://www.campbellcollaboration.org/artman2/uploads/1/2_D_Wilson__Calculating_ES.pdf