ERROR REPORT #1

Report by Matthew B. Jané

Study Title: Mindfulness-Based Stress Reduction for Stress Management in Healthy People: A Review and Meta-Analysis

Journal: The Journal of Alternative and Complementary Medicine

DOI: 10.1089/acm.2008.0495

Citations: 2879

Summary: The meta-analysis by Chiesa and Serretti (2009) contains substantial data extraction errors (i.e., effect size calculation errors) and analysis flaws that artificially inflate the difference between treatment and control pre-post effects. In this report I walk through each of the calculations for all studies included in the meta-analysis that I have access to. Due to the severe data extraction errors and analysis flaws, all the meta-analytic results in Chiesa and Serretti (2009) are incorrect and exaggerate the effect between MBSR and control groups for both stress and spirituality outcomes.

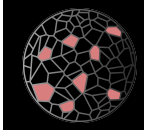


Table of contents

Load necessary R packages

Let's first load in packages for the analysis.

```
library(metafor)
library(ggdist)
library(tidyverse)
library(clubSandwich)
library(MASS)
library(ggdist)
theme_set(theme_ggdist(base_size=15))
```

Tabulating Table 1 of Chiesa and Serretti (2009)

First we will extract the information from table 1 from Chiesa and Serretti (2009). Here is the screenshot of table 1 below:

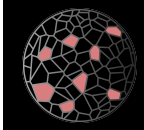
Table 1. Summary of Included Studies

Study	*Meditation/ comparison*	*N*	*Population*	*Duration (weeks)*	*Study design*	*Measure of stress*	*Pre–post within-group effect size*	*Measures of spirituality*	*Pre–post within group effect size*
Astin, 1997[34]	MBSR/waiting list	7/12	University undergraduates	8 weeks; 3-month follow-up	RCT	GSI	n.c.	INSPIRIT	n.c.
Shapiro, 1998[27]	MBSR/waiting list	36/37	Medical and premed students	7	RCT	GSI	0.632/0	INSPIRIT	n.c.
Rosenzweig, 2003[28]	MBSR/waiting list	140/162	Medical and premed students	10	SS-CT	POMS	0.205/−0.339	—	—
Beddoe and Murphy, 2004[35]	MBSR	16	Nursing students	8	UCT	DSP	n.c.	—	—
Cohen-Katz, 2005[36]	MBSR/waiting list	12/13	Nurses	8 weeks; 3-month follow-up	RCT	BSI	n.c.	MAAS	1.959/0.787
Shapiro, 2005[29]	MBSR/waiting list	10/18	Health care professionals	8	RCT	PSS	1.724/−0.303	—	—
Shapiro, 2007[30]	MBSR/weekly meetings	22/32	Therapists in training	10	CT	PSS	1.008/−0.162	MAAS	0.372/−0.396
Jain, 2007[31]	MBSR/relaxation training/waiting list	27/24/30	Medical students, graduate nursing students, undergraduate premed students	4	RCT AC	BSI	1.366/0.911/ 0.272	INSPIRIT-R	0.066/0.074/ −0.027
Klatt, 2008[32]	MBSR/waiting list	22/20	Faculty and staff at a large midwestern university	6	RCT	PSS	2.858/−0.47	MAAS	1.929/0.193
Vieten and Astin, 2008[33]	MBSR/waiting list	13/18	Pregnant women between 12 and 30 weeks gestation	10 weeks; 3 month follow-up	RCT	PSS	0.776/0.041	MAAS	0.253/−0.308

NC, not calculable; MBSR, mindfulness-based stress reduction; UCT, uncontrolled trial; RCT, randomized controlled trial; SS-CT, self selected controlled trial; RCT AC, randomized controlled trial with an active control; CT AC, controlled trial with an active control; GSI, global severity index (of the Hopkins Symptom Checklist 90 Revised); POMS, profile of mood symptoms; DSP, Derogatis stress profile; BSI, brief symptom inventory; PSS, perceived stress scale; INSPIRIT, index of core spiritual experiences; MAAS, mindfulness attention awareness scale; INSPIRIT-r, index of core spirituality–revised.

Within the table I am going to extract the study label `study`, the sample size for treatment (i.e., MBSR) and control (e.g., waitlist) group `n_t` and `n_c`, whether the study is an RCT (`rct=1`) or not (`rct=0`), the pre-post effect sizes for the both treatment (`d_st_tx` and `d_sp_tx`) and control (`d_st_c` and `d_sp_c`) groups for the `st`ress and `sp`iritual outcomes. Below is the final table in R:

```
# data frame of original effects
df <- data.frame(
  study = c("Astin, 1997",
```



```
              "Shapiro, 1998",
              "Rosenzweig, 2003",
              "Beddoe and Murphy, 2004",
              "Cohen-Katz, 2005",
              "Shapiro, 2005",
              "Shapiro, 2007",
              "Jain, 2007",
              "Klatt, 2008",
              "Vieten and Astin, 2008"),
    d_st_tx=c(NA,0.632,0.205,NA,NA,1.724,1.008,1.366,2.858,0.776),
    d_st_c=c(NA,0.000,-0.339,NA,NA,-0.303, -0.162,0.272,-0.470,0.041),
    d_sp_tx= c(NA,NA,NA,NA,1.959,NA,0.372,0.066,1.929,0.253),
    d_sp_c= c(NA,NA,NA,NA,0.787,NA,-0.396,-0.027,0.193,-0.308),
    n_t= c(7,36,140, 16,12, 10,22,27,22,13),
    n_c= c(12,37,162,NA,13,18,32,30,20,18),
    rct=c(1,1,0,0,1,1,0,1,1,1))

  head(df,10)
```

```
                    study d_st_tx d_st_c d_sp_tx d_sp_c n_t n_c rct
1              Astin, 1997      NA     NA      NA     NA   7  12   1
2           Shapiro, 1998   0.632  0.000      NA     NA  36  37   1
3         Rosenzweig, 2003   0.205 -0.339      NA     NA 140 162   0
4  Beddoe and Murphy, 2004      NA     NA      NA     NA  16  NA   0
5         Cohen-Katz, 2005      NA     NA   1.959  0.787  12  13   1
6           Shapiro, 2005   1.724 -0.303      NA     NA  10  18   1
7           Shapiro, 2007   1.008 -0.162   0.372 -0.396  22  32   0
8              Jain, 2007   1.366  0.272   0.066 -0.027  27  30   1
9            Klatt, 2008   2.858 -0.470   1.929  0.193  22  20   1
10  Vieten and Astin, 2008   0.776  0.041   0.253 -0.308  13  18   1
```

Reproducing Data Extraction

In this section I look to reproduce the effect sizes from each of the studies in this table. At the end of this section I will have a summary table for my findings for all the studies. The function below will be used to calculate the pre-post effect sizes (per the equation reported by, Chiesa and Serretti 2009) from the study means and SDs.

```
d <- function(m1, s1, m2, s2, direction = "higher=better"){
  if (direction == "higher=better"){
   return((m2 - m1) / sqrt((s1^2 + s2^2)/2) )
  }

  if (direction == "higher=worse"){
    return(-1*(m2 - m1) / sqrt((s1^2 + s2^2)/2) )
  }
}
```

Reproducing: Astin (1997)

Chiesa and Serretti (2009) reported that the effects for both spiritual and stress outcomes in Astin (1997) were not calculable. Indeed the statistics reported in the paper were insufficient to calculate an exact effect without applying some strong assumptions. Though, it is important to point out that Astin (1997) showed a 64% reduction in stress (SCL-90) scores in the treatment group and a 14% reduction in the control group (see table below).

Table 1. Change in scores on SCL-90-R following treatment (12 experimental subjects, 7 control subjects)

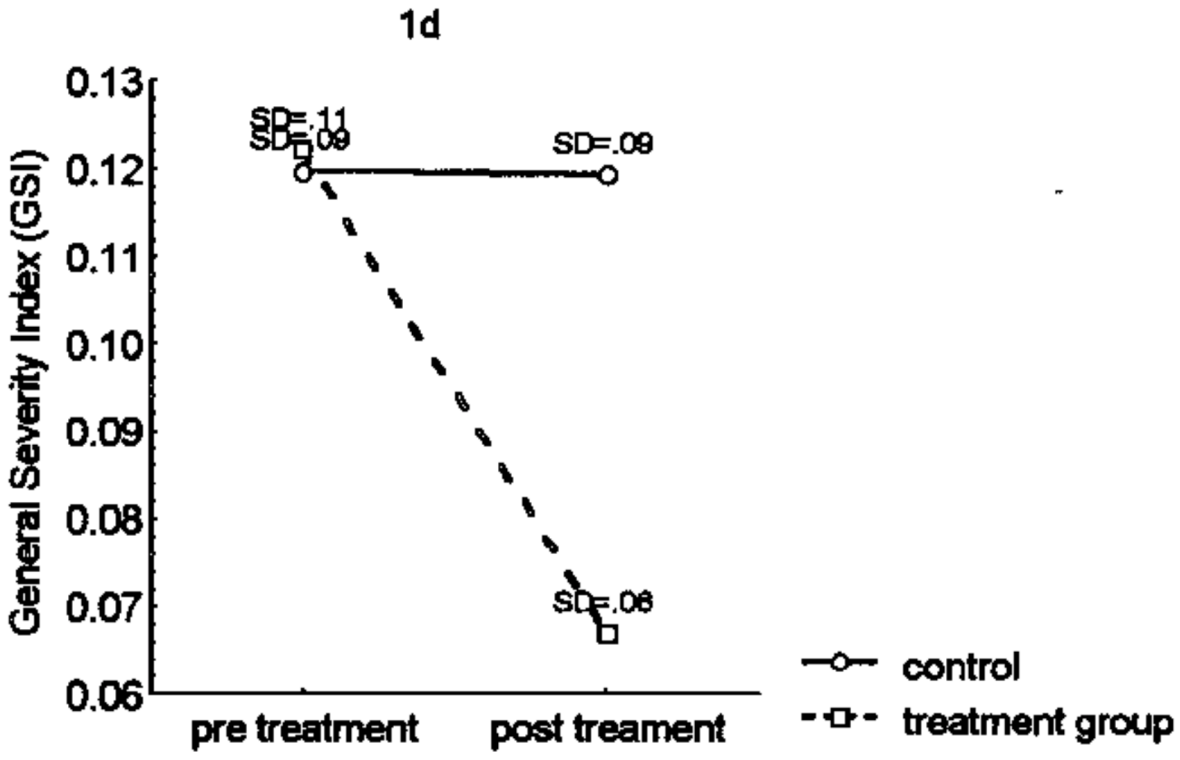
SCL-90-R	Experimental group % reduction	Controls % reduction	F	p<
GSI	64	14	15.87	0.002
Depression	59	7	12.34	0.005
Anxiety	60	10	7.05	0.02
Obsessive-compulsive	59	23	9.55	0.01
Somatization	73	23	16.73	0.005
Interpersonal sensitivity	59	27	7.94	0.05
Psychoticism	76	39	9.27	0.01
Paranoid ideation	73	−1	9.87	0.01
Additional items[1]	73	2	32.20	0.0001
Hostility	32	32	1.70	0.22
Phobic anxiety	58	32	1.13	0.31

[1] Seven items dealing with poor appetite, overeating, sleep disturbances, and feelings of guilt.

In the text, the same beneficial effect of the treatment is seen for the spirituality outcome (INSPIRIT), although again, there is not sufficient information to calculate an effect size.

Reproducing: Shapiro, Schwartz, and Bonner (1998)

The means and standard deviations (SDs) are located in Figure 1 of Shapiro, Schwartz, and Bonner (1998). Particularly panel *d* and *e* show the means and SDs for stress (measured by GSI) and spirituality, respectively. The stress figure is shown below below:



From the table the dots denote the mean and the SDs are labeled above them. For the post-treatment means, we observe a mean of .122 (SD = .11) for the treatment group and .120 (SD=.09) for the control group. For the pre-treatment means, we observe a mean of .066 (SD = .06) for the treatment group and .120 (SD=.09) for the control group. The effect size for stress will be as follows for both groups:

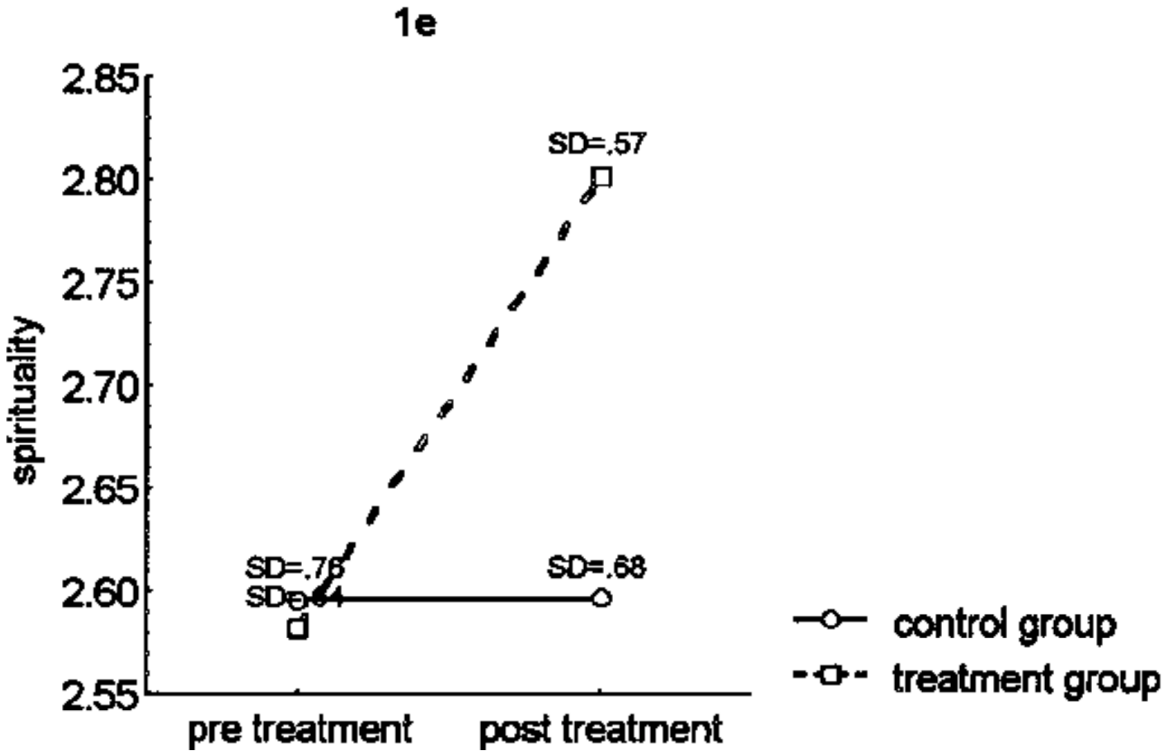
```
# treatment group (stress)
d(m1 = .122, s1 = .11,
  m2 = .066, s2 = .06,
  direction = "higher=worse")
```

```
[1] 0.6320526
```

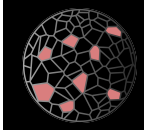
```
# control group (stress)
d(m1 = .120, s1 = .09,
  m2 = .120, s2 = .09,
  direction = "higher=worse")
```

```
[1] 0
```

The effect sizes match what is in Chiesa and Serretti (2009). For the spirituality effects, Chiesa and Serretti (2009) did not code the data and stated that it was “not calculable”, which appears to not be true since panel *e* has the spirituality data:



For the pre-treatment means, we observe a mean of 2.58 (SD = .64) for the treatment group and 2.59 (SD=.76) for the control group. For the post-treatment means, we observe a mean of 2.80 (SD = .57) for the treatment group and 2.59 (SD=.68) for the control group.



```
# treatment group (spirituality)
d(m1 = 2.58, s1 = .64,
  m2 = 2.80, s2 = .57,
  direction = "higher=better")
```

```
[1] 0.3630294
```

```
# control group (spirituality)
d(m1 = 2.59, s1 = .76,
  m2 = 2.59, s2 = .68,
  direction = "higher=better")
```

```
[1] 0
```

This appears to be an error in the data extraction of the meta-analysis as I was able to successfully calculate the effect sizes for both treatment and control group for the spirituality outcome.

Reproducing: Rosenzweig et al. (2003)

Table 1 in Rosenzweig et al. (2003) reports the means and SDs for the Total Mood Disturbance index from the Profile of Mood States (POMS) scale which is used to measure stress:

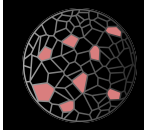
Table 1. *Pretest and Posttest Total and Subscale Scores (Mean ± SD) on the Profile of Mood States (POMS), and Summary Statistical Results*

	MBSR Group[a]				Control Group[b]				
POMS Scales	**Pre**	**Post**	***d*c**	***p*d**	**Pre**	**Post**	***d*c**	***p*d**	**Interaction *p*e**
Tension–Anxiety	14.5±7.2	12.4±7.0	–0.23	0.009	11.3±6.3	13.4±6.9	0.28	0.0008	<0.0001
Vigor–Activity	14.8±5.8	16.3±5.6	0.25	0.006	17.4±5.6	14.2±5.6	–0.47	0.0001	<0.0001
Fatigue–Inertia	10.2±6.3	10.6±6.2	0.06	0.50	8.4±5.3	11.8±6.2	0.49	0.0001	0.0006
Confusion–Bewilderment	10.0±5.6	9.3±4.8	–0.24	0.009	9.1±4.7	9.3±4.8	0.05	0.52	0.02
Depression–Dejection	10.4±10.0	8.8±9.0	–0.15	0.09	8.8±9.0	9.5±8.6	0.07	0.37	0.06
Anger–Hostility	8.5±6.8	7.8±7.3	–0.08	0.38	7.8±7.8	8.9±8.1	0.12	0.13	0.09
Total Mood Disturbance	38.7±33.3	31.8±33.8	–0.18	0.05	28.0±31.2	38.6±32.8	0.30	0.0003	<0.0001

Note: Multivariate $F_{(6,270)} = 7.68$; Wilks' Lambda = 0.85, $p < 0.01$.
[a]$n = 125$. [b]$n = 152$. [c]d is the effect size estimate (standardized mean difference) for pre-post seminar change scores. [d]p values for obtained mean differences using univariate analysis of variance (ANOVA). [e]p values for interaction effects of group by pre-post seminar scores resulted from 2-way ANOVAs for repeated measure design.

Using the Total Mood Disturbance score (i.e., POMS) we can see that for the pre-treatment means, we observe a mean of 38.7 (SD = 33.3) for the treatment group and 28.0 (SD=31.2) for the control group. For the post-treatment means, we observe a mean of 31.8 (SD = 33.8) for the treatment group and 38.6 (SD=32.8) for the control group. In R, let's now calculate the effect.

```
# treatment group (stress)
d(m1 = 38.7, s1 = 33.3,
```



```
  m2 = 31.8, s2 = 33.8,
  direction = "higher=worse")
```

```
[1] 0.2056575
```

```
# control group (stress)
d(m1 = 28.0, s1 = 31.2,
  m2 = 38.6, s2 = 32.8,
  direction = "higher=worse")
```

```
[1] -0.3311465
```

The effect sizes and the sample sizes (treatment group: d=.206, n=125; control group: d=–.331, n=152) are just slightly off from what is reported in Chiesa and Serretti (2009) (treatment group: d=.205, n=140; control group: d=–.331, n=162). Sample sizes appear to be extracted from the abstract, but that is the full sample, not the complete sample. In the table it is clear that these statistics are based on a sample size of 125 and 152 in the MBSR and control group, respectively.

Reproducing: Beddoe and Murphy (2004)

Chiesa and Serretti (2009) reported effects as incalculable. I am unable to gain access to this article so I am also unable to calculate any effects.

Reproducing: Cohen-Katz et al. (2005)

Chiesa and Serretti (2009) reported that the stress outcome was not calculable, however this is not the case. Cohen-Katz et al. (2005) measures stress from the BSI as a dichotomous variable (elevated stress vs not). Instead of continuous scores like we typically get from these outcomes, we can use the dichotomous scores to obtain an odds ratio and convert it to a Cohen's *d*. We can observe the contingency table for pre-post comparisons as implied in the text on pages 31-32 of Cohen-Katz et al. (2005),

```
# treatment group
data.frame(stress = c("elevated","not elevated"),
           pre = c(3,9),
           post = c(1,11))
```

```
        stress pre post
1     elevated   3    1
2 not elevated   9   11
```

```
# control group
data.frame(stress = c("elevated","not elevated"),
           pre = c(7,6),
           post = c(4,9))
```

```
        stress pre post
1     elevated   7    4
2 not elevated   6    9
```

We can then calculate pre-post log odds ratios from the contingency tables and then convert to a Cohen's d with the formula $d = \frac{\log(OR) \times \sqrt{3}}{\pi}$,

```
# treatment group (flip so higher=better)
OR_tx <- (1/11)/(3/9)
d_tx <- (-1) * log(OR_tx)*sqrt(3)/pi
d_tx
```

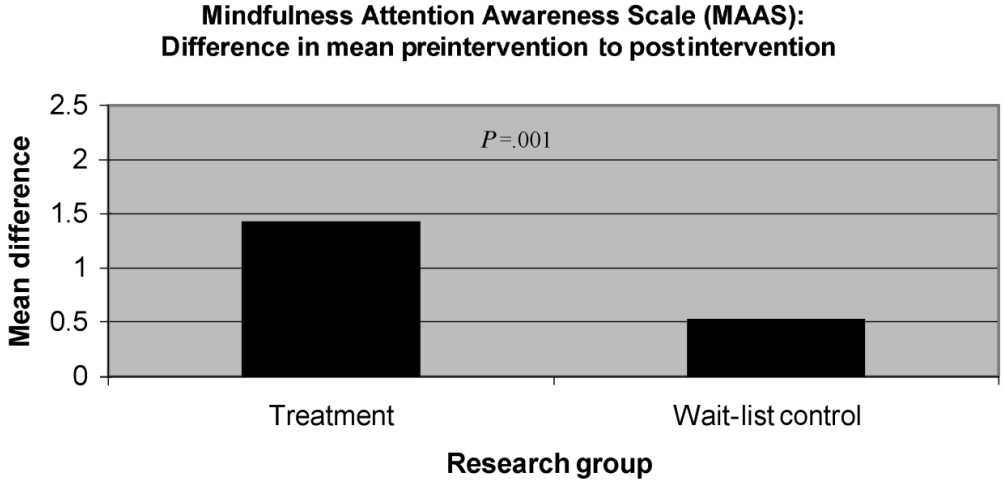
```
[1] 0.7163323
```

```
# control group (flip so higher=better)
OR_c <- (4/9)/(7/6)
d_c <- (-1) * log(OR_c)*sqrt(3)/pi
d_c
```

```
[1] 0.532077
```

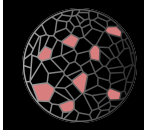
I was therefore able to calculate the effect sizes for treatment and control for the stress outcomes.

The spirituality outcome effects are a bit harder. In Figure 2 of Cohen-Katz et al. (2005) below, they report the pre-post change for treatment and control group for the MAAS along with the p-value (presumably from the mean difference in change scores).



So for the effect sizes, we can start with getting the numerator (i.e., the raw mean change) from figure 2 where the mean change in the treatment group is about 1.4 and mean change in the control group is about 0.5. It is however unclear where the effect sizes of 1.959 and 0.787 come from since the pre and post-intervention standard deviations do not appear to be available. Instead, we will try to capture all possible effect size values based on the available information. The information that is available through the paper text and figures is the following:

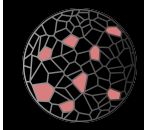
- Treatment group:
  - Pre-Intervention: Mean = 3, SD = unknown (from figure 3)



  - Post-Intervention: Mean = 4.4, SD = unknown (from figure 2 and figure 3)
  - pre-post correlation: r = unknown
  - sample size: n = 12 (from page 29)
- Control Group:
  - Pre-Intervention: Mean = Unknown, SD = unknown (from figure 3)
  - Post-Intervention: Mean = pre-mean + 1.4, SD = unknown (from figure 2)
  - pre-post correlation: r = unknown
  - sample size: n = 13 (from page 29)
- Comparisons of Means:
  - Pre-Pre (treatment vs control): p > .05 (implied from text on page 29)
  - Post-Post (treatment vs control): p = .001 (from text on page 29)
  - Pre-Post (treatment group): p = .004 (from figure 3)

Using the available information, we can attempt to construct a distribution of possible values for the pre-post effect size for the treatment and control group. For all unknown parameter values, we will draw parameter values from a uniform distribution over a plausible range of values. Using this space of parameters (known values are fixed and unknown values are randomly sampled), I will then generate 500,000 possible data sets and calculate the *d* value for treatment and control group for each one. Once the simulated *d* values are obtained, we will only select the ones from datasets that match the p-values reported in the text.

```
# set seed and iteration count
set.seed(1)
iter = 500000

### randomly sampled unknown values ###
# standard deviations
s1_tx <- runif(iter,.5,2.5)
s2_tx <- runif(iter,.5,2.5)
s1_c <- runif(iter,.5,2.5)
s2_c <- runif(iter,.5,2.5)
# pre-post correlations
r_tx <- runif(iter,0,.999)
r_c <- runif(iter,0,.999)
m1_c <- runif(iter,1.5,4.5)

# known values
n_tx <- 12
n_c <- 13
m1_tx <- 3
m2_tx <- 4.4
m2_c <- m1_c + .5
p_tx = .004
p_post = .001
p_pre_thresh = .05 # non-significant threshold

# empty vector of p-values from simulations
p_tx_sim <- c()
p_post_sim <- c()
p_pre_sim <- c()

for(i in 1:iter){
```



```
  # generate pre-post data for treatment group
  df_tx <- mvrnorm(n_tx,
                   mu = c(m1_tx, m2_tx),
                   Sigma = data.frame(pre = c(s1_tx[i]^2, r_tx[i]*s1_tx[i]*s2_tx[i]),
                                      post = c(r_tx[i]*s1_tx[i]*s2_tx[i], s2_tx[i]^2)),
                   empirical = FALSE)

  # generate pre-post data for control group
  df_c <- mvrnorm(n_c,
                   mu = c(m1_c[i], m2_c[i]),
                   Sigma = data.frame(pre = c(s1_c[i]^2, r_c[i]*s1_c[i]*s2_c[i]),
                                      post = c(r_c[i]*s1_c[i]*s2_c[i], s2_c[i]^2)),
                   empirical = FALSE)

  # name pre and post columns
  colnames(df_tx) <- c("pre", "post")
  colnames(df_c) <- c("pre", "post")

  # p-values from simulated data
  p_tx_sim[i] <- t.test(df_tx[,"post"], df_tx[,"pre"], paired = TRUE)$p.value
  p_post_sim[i] <- t.test(df_tx[,"post"], df_c[,"post"], paired = FALSE)$p.value
  p_pre_sim[i] <- t.test(df_tx[,"pre"], df_c[,"pre"], paired = FALSE)$p.value

}

# get locations of p values within rounding error of .04
idx <- which(p_tx_sim > p_tx-.0005 & p_tx_sim < p_tx+.0005 &
             p_post_sim > p_post-.0005 & p_post_sim < p_post+.0005 &
             p_pre_sim > p_pre_thresh)

# calculate pre-post d treatment group
d_tx <- d(m1 = m1_tx, s1 = s1_tx[idx],
          m2 = m2_tx, s2 = s2_tx[idx],
          direction = "higher=better")

# calculate pre-post d in control group
d_c <- d(m1 = m1_c[idx], s1 = s1_c[idx],
          m2 = m2_c[idx], s2 = s2_c[idx],
          direction = "higher=better")

# save simulation
save(d_tx, d_c, file = "cohenkatz2005_simulation.RData")
```

```
load("cohenkatz2005_simulation.RData")

# plot out possible values of d by treatment assignment
ggplot(data = NULL) +
  stat_slabinterval(aes(y = 0, x = d_c),point_interval = "mean_qi",
                    slab_fill = "grey50",slab_alpha = .5) +
  stat_slabinterval(aes(y = 1, x = d_tx),point_interval = "mean_qi",
                    slab_fill = "grey50",slab_alpha = .5) +
```

```
  geom_point(aes(x=0.787,y=0), color = "red3",size = 6, shape = 18) +
  geom_point(aes(x=1.959,y=1), color = "red3",size = 6, shape = 18) +
  geom_vline(xintercept = 0, linetype="dashed", alpha = .5) +
  scale_y_continuous(breaks = 0:1, labels = c("Control","Treatment"),limits = c(-.5,2)) +
  labs(x = "Effect Size (d)", y = "")
```

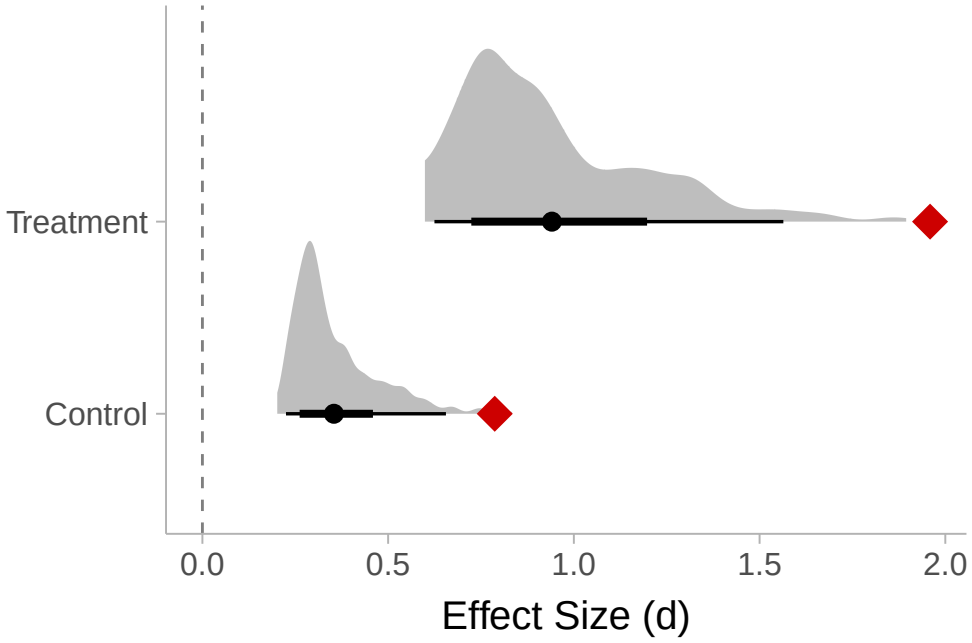


Figure 1: Distribution of possible (simulated) pre-post effect size values (in grey) with the value reported by Chiesa and Serretti (2009) denoted with a red diamond.

The effects reported by Chiesa and Serretti (2009) appear to be very unlikely according to the simulation. I will instead use the means of the two distributions as the approximate effect size for this study: *d*=.354 for the control group and *d*=.941 for the treatment group.

Reproducing: Shapiro et al. (2005)

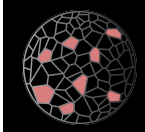
The effect sizes reported by Chiesa and Serretti (2009) for Shapiro et al. (2005) are 1.724 and –.339 for the treatment and control group, respectively. However, in table 1 of Shapiro (2005), the treatment group and control group *both* reduce stress so it does not appear possible that the effect sizes are in different directions. Since Chiesa and Serretti (2009) ensures that beneficial effects of the treatment are positive, both the treatment and control group should see positive effects

Table 1. Means and Statistics for Pre- and Posttreatment

	Mindfulness		Wait-list control		
Primary outcome	Pretreatment	Posttreatment	Pretreatment	Posttreatment	Between-group analyses[a]
Satisfaction With Life	20.80	24.80	23.94	23.83	$F(2, 25) = 3.84$, $p = .06$
Burnout Scale	75.90	68.40	72.94	70.00	$F(2, 25) = 1.69$, $p = .21$
Perceived Stress	28.89	21.22	23.78	22.17	$F(2, 24) = 4.4$, $p = .04$
Brief Symptom Inventory	0.61	0.47	0.56	0.50	$F(2, 22) = 1.44$, $p = .25$
Self-Compassion	16.48	20.15	19.51	20.07	$F(2, 24) = 9.85$, $p = .004$

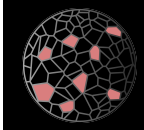
[a]*p* values are based on the results of separate regression analyses and refer to the test of significance for group assignment (treatment or wait list), controlling for the effects of baseline levels of each outcome variable examined.

Unfortunately, the study only reports means and no SDs and so it is unclear how the effects in Chiesa and Serretti (2009) are calculated. We have instead some information from a regression



model that has the form $\text{stress}_{post} = \beta_0 + \beta_1 \text{stress}_{pre} + \beta_2 \text{Tx}$, where Tx denotes the dummy-coded treatment assignment. The p-value in the table is the p-value associated with the β_2 coefficient. Similar to what we did with the previous section (see Reproducing: Cohen-Katz et al. 2005) we can simulate distributions of possible effect sizes based on the information provided in the table.

```
# set seed and iteration count
set.seed(1)
iter = 10000

### randomly sampled unknown values ###
# standard deviations
s1_tx <- runif(iter,2,12)
s2_tx <- runif(iter,2,12)
s1_c <- runif(iter,2,12)
s2_c <- runif(iter,2,12)
# pre-post correlations
r_tx <- runif(iter,0,.999)
r_c <- runif(iter,0,.999)

# known values
n_tx <- 10
n_c <- 18
m1_tx <- 28.89
m2_tx <- 21.22
m1_c <- 23.78
m2_c <- 22.17
p = .04

# empty vector of p-values from simulations
psim <- c()

for(i in 1:iter){

  # generate pre-post data for treatment group
  df_tx <- mvrnorm(n_tx,
                   mu = c(m1_tx, m2_tx),
                   Sigma = data.frame(pre = c(s1_tx[i]^2, r_tx[i]*s1_tx[i]*s2_tx[i]),
                                      post = c(r_tx[i]*s1_tx[i]*s2_tx[i], s2_tx[i]^2)),
                   empirical = TRUE)

  # generate pre-post data for control group
  df_c <- mvrnorm(n_c,
                   mu = c(m1_c, m2_c),
                   Sigma = data.frame(pre = c(s1_c[i]^2, r_c[i]*s1_c[i]*s2_c[i]),
                                      post = c(r_c[i]*s1_c[i]*s2_c[i], s2_c[i]^2)),
                   empirical = TRUE)

  # concatenate treatment and control data
  df_total <- as.data.frame(rbind(df_tx,df_c))

  # name pre and post columns
  colnames(df_total) <- c("pre", "post")
```



```
  # create treatment assignment dummy code vector
  df_total <- cbind(df_total,
                    tx = c(rep(1,n_tx),rep(0,n_c)))

  # estimate regression model
  mdl <- lm(post ~ pre + tx, data = df_total)

  # extract p-value from treatment assignment term from model
  psim[i] <- coefficients(summary(lm(post ~ pre + tx,data = df_total)))[3,"Pr(>|t|)"]

}

# get locations of p values within rounding error of .04
idx <- which(psim > p-.005 & psim < p+.005)

# calculate pre-post d treatment group
d_tx <- d(m1 = m1_tx, s1 = s1_tx[idx],
          m2 = m2_tx, s2 = s2_tx[idx],
          direction = "higher=worse")

# calculate pre-post d in control group
d_c <- d(m1 = m1_c, s1 = s1_c[idx],
          m2 = m2_c, s2 = s2_c[idx],
          direction = "higher=worse")

# save simulation
save(d_tx, d_c, file = "shapiro2005_simulation.RData")
```

```
load("shapiro2005_simulation.RData")

# plot out possible values of d by treatment assignment
ggplot(data = NULL) +
  stat_slabinterval(aes(y = 0, x = d_c),point_interval = "mean_qi",
                    slab_fill = "grey50",slab_alpha = .5) +
  stat_slabinterval(aes(y = 1, x = d_tx),point_interval = "mean_qi",
                    slab_fill = "grey50",slab_alpha = .5) +
  geom_point(aes(x=-.303,y=0), color = "red3",size = 6, shape = 18) +
  geom_point(aes(x=1.724,y=1), color = "red3",size = 6, shape = 18) +
  geom_vline(xintercept = 0, linetype="dashed", alpha = .5) +
  scale_y_continuous(breaks = 0:1, labels = c("Control","Treatment"),limits = c(-.5,2)) +
  labs(x = "Effect Size (d)", y = "")
```

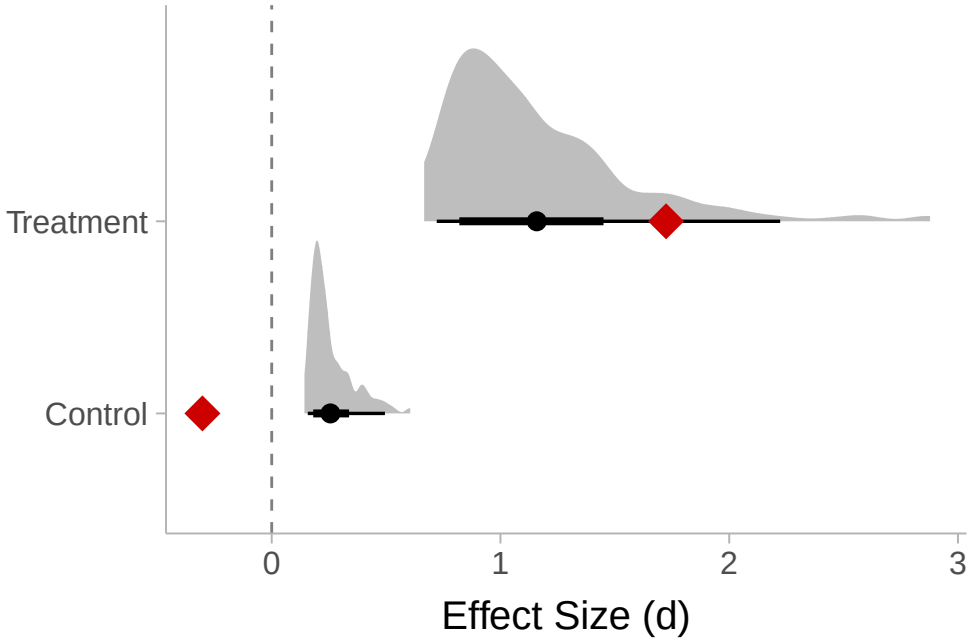


Figure 2: Distribution of possible (simulated) pre-post effect size values (in grey) with the value reported by Chiesa and Serretti (2009) denoted with a red diamond.

As expected, the reported effects by Chiesa and Serretti (2009) are impossible for the control group, but it may be possible for the treatment group. It is quite possible that Chiesa and Serretti (2009) simply coded the wrong direction for the control group, but it is still unclear how the effects are calculated as the paper does provide the necessary information. Due to the inability to reproduce Chiesa and Serretti (2009)'s results, I will use the mean of the simulated effect sizes instead: 1.159 for the treatment group and .257 for the control group.

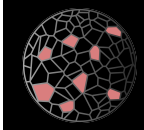
Reproducing: Shapiro, Brown, and Biegel (2007)

Chiesa and Serretti (2009) reported the stress pre-post effect sizes for Shapiro, Brown, and Biegel (2007) as follows 1.008 and –.162 for the treatment and control group, respectively. Chiesa and Serretti (2009) then reported the spirituality effect sizes as .372 and –.396 for treatment and control group, respectively. Table 1 of Shapiro, Brown, and Biegel (2007) has all the necessary information to calculate effect sizes:

Table 1
Mean Scores by Group, Pre-Course (Time 1) and Post-Course (Time 2), and MBSR Intervention Effects

	MBSR		Control		
Variable	Time 1 *M (SD)*	Time 2 *M (SD)*	Time 1 *M (SD)*	Time 2 *M (SD)*	p_{inter}
PANAS positive affect	4.87 (0.75)	5.45 (0.94)	5.14 (0.74)	4.90 (0.95)	.0002
PANAS negative affect	3.09 (0.90)	2.55 (1.01)	3.04 (1.03)	2.99 (0.89)	.04
STAI anxiety, present	3.17 (1.19)	2.18 (1.09)	2.67 (1.11)	2.76 (1.01)	.0005
STAI anxiety, past month	3.43 (0.90)	2.51 (0.77)	3.33 (1.05)	3.44 (1.14)	.0002
PSS perceived stress	24.64 (7.81)	18.36 (5.15)	21.72 (7.14)	22.91 (7.54)	.0001
RRQ rumination	3.42 (0.83)	2.78 (0.63)	3.15 (0.92)	3.11 (0.90)	.0006
SCS self-compassion	18.06 (3.97)	20.92 (3.84)	19.41 (3.75)	19.22 (4.12)	.0001
MAAS mindfulness	3.76 (0.80)	4.01 (0.51)	4.05 (0.64)	3.80 (0.62)	.006

Note. $n = 22$ in MBSR group; $n = 32$ in control group. The p_{inter} column shows the ANOVA Group × Time Interaction Significance Levels. RRQ = Reflection Rumination Questionnaire; PANAS = Positive Affectivity Negative Affectivity Schedule; STAI = State/Trait Anxiety Inventory; PSS = Perceived Stress Scale; SCS = Self-Compassion Scale; MAAS = Mindful Attention Awareness Scale.



From this table we can extract the statistics and calculate the pre-post effect sizes for the stress outcome (Perceived Stress Scale, PSS) and the spirituality outcome (Mindful Attention Awareness Scale, MAAS):

```
## STRESS
data.frame(
  # treatment group
  d_treatment = d(m1 = 24.64, s1 = 7.81,
                  m2 = 18.36, s2 = 5.15,
                  direction = "higher=worse"),
  # control group
  d_control = d(m1 = 21.72, s1 = 7.14,
                m2 = 22.91, s2 = 7.54,
                direction = "higher=worse")
)
```

```
  d_treatment  d_control
1   0.9493459 -0.1620652
```

```
## SPIRITUALITY
data.frame(
  # treatment group
  d_treatment = d(m1 = 3.76, s1 = 0.80,
                  m2 = 4.01, s2 = 0.51,
                  direction = "higher=better"),
  # control group
  d_control = d(m1 = 4.05, s1 = 0.64,
                m2 = 3.80, s2 = 0.62,
                direction = "higher=better")
)
```

```
  d_treatment  d_control
1   0.3726573 -0.3967754
```

The effects from Chiesa and Serretti (2009) were mostly reproducible within rounding error, however, the treatment group in the stress outcome was a bit off (Chiesa reported 1.008, but my calculation shows .949). From what I can tell, this is due to a simple data copying error. When I use SD=7**.15** instead of 7**.81** in the pre-test I recover the 1.008 effect that Chiesa and Serretti (2009) reported (the post-test value also ends in the decimal 5**.15** which makes it appear to be a simple copy and paste mistake).

Reproducing: Jain et al. (2007)

Jain et al. (2007) has three intervention arms: a control, relaxation group, and a meditation group (the treatment). Jain et al. (2007) also had a measure of stress (Brief Symptom Inventory, BSI) and a measure of spirituality (Index of Core Spiritual Experiences). Chiesa and Serretti (2009) reported the effects for the stress outcome as 1.366 for the meditation (treatment) group, .911 for the relaxation group and .272 for the control group. As for the spirituality outcome, Chiesa and Serretti (2009) reported .066 for the meditation (treatment) group, .074 for the relaxation group and -.027 for the control group. Table 1 of Jain et al. (2007) contains the necessary statistics to calculate the effect sizes:

TABLE 1
Unadjusted Pre- and Postintervention Means and Standard Deviations for Control, Meditation, and Relaxation Groups

	Control Preintervention[a]		*Control Postintervention*[a]		*Meditation Preintervention*[b]		*Meditation Postintervention*[b]		*Relaxation Preintervention*[c]		*Relaxation Postintervention*[c]	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Brief Symptom Inventory GSI scores	.59	.43	.46	.52	.64	.40	.22	.17	.74	.52	.35	.31
Daily Emotion Report Rumination scores	3.5	2.4	4.4	3.1	3.9	2.9	2.5	1.9	6.0	3.2	5.0	3.4
Daily Emotion Report Distraction scores	6.7	2.7	7.9	3.1	6.0	3.4	5.2	2.9	8.0	3.4	8.6	3.3
Positive States of Mind Scale scores	16.2	3.5	16.3	3.8	15.0	2.9	17.1	3.0	15.9	3.1	16.8	4.0
Index of Core Spiritual Experiences scores	27.7	7.6	27.5	7.2	28.4	8.0	28.9	7.0	26.8	8.4	27.4	7.8

Note. GSI = Global Severity Index.
[a]$n = 30$. [b]$n = 27$. [c]$n = 24$.

We can then extract the means and SDs and calculate the pre-post effect sizes for each group.

```
## STRESS
data.frame(
  # meditation (treatment) group
  d_treatment = d(m1 = .64, s1 = .40,
                  m2 = .22, s2 = .17,
                  direction = "higher=worse"),
  # relaxation group
  d_relaxation = d(m1 = .74, s1 = .52,
                m2 = .35, s2 = .31,
                direction = "higher=worse"),
  # control group
  d_control = d(m1 = .59, s1 = .43,
                m2 = .46, s2 = .52,
                direction = "higher=worse")
)
```

```
  d_treatment d_relaxation d_control
1    1.366622    0.9110508 0.2724642
```

```
## SPIRITUALITY
data.frame(
  # meditation (treatment) group
  d_treatment = d(m1 = 28.4, s1 = 8,
                  m2 = 28.9, s2 = 7,
                  direction = "higher=better"),
  # relaxation group
  d_relaxation = d(m1 = 26.8, s1 = 8.4,
                m2 = 27.4, s2 = 7.8,
                direction = "higher=better"),
  # control group
  d_control = d(m1 = 27.7, s1 = 7.6,
                m2 = 27.5, s2 = 7.2,
                direction = "higher=better")
)
```

```
  d_treatment d_relaxation   d_control
```

```
1   0.06651901    0.07402332 -0.02701716
```

I was able to exactly reproduce the effect sizes that were calculated by Chiesa and Serretti (2009).

Reproducing: Klatt, Buckworth, and Malarkey (2009)

Chiesa and Serretti (2009) reported the effect sizes for the stress outcome (Perceived Stress Scale, PSS) was 2.858 and –0.47 for the treatment and control group, respectively. They also reported for the spirituality outcome (Mindful Attention Awareness Scale, MAAS) as 1.929 and 0.193 for the treatment and control group, respectively. The statistics needed to calculate the pre-post effect sizes are found in table 2 of Klatt, Buckworth, and Malarkey (2009):

Table 2. Mindfulness, Perceived Stress, and Sleep Scores ($M \pm SE$)

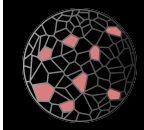
	MBSR-ld ($n = 22$)				Control ($n = 20$)			
Measure	Pre	Post	Diff	p value[a]	Pre	Post	Diff	p value[a]
MAAS	55.14 + 2.90	60.50 + 2.65	–5.36 + 2.02	.0149	63.15 + 3.08	63.70 + 2.57	–0.55 + 1.57	.7294
Perceived Stress Scale	28.09 + 1.19	25.00 + 0.96	3.09 + 0.90	.0025	26.20 + 1.30	25.55 + 1.46	0.65 + 0.86	.4586
PSQI component scores								
Global PSQI	6.73 + 0.69	5.00 + 0.46	1.73 + 0.48	.0018	6.68 + 0.67	5.50 + 0.68	1.18 + 0.39	.0072
Subjective Sleep Quality	1.23 + 0.13	0.91 + 0.15	0.41 + 0.16	.0162	1.22 + 0.14	0.89 + 0.15	0.35 + 0.18	.0692
Sleep Latency	1.09 + 0.20	0.68 + 0.19	0.41 + 0.18	.0355	0.75 + 0.20	0.65 + 0.18	0.10 + 0.14	.4936
Sleep Duration	0.91 + 0.11	0.82 + 0.13	0.09 + 0.09	.3287	1.00 + 0.18	1.00 + 0.19	0.00 + 0.10	1.000
Habitual Sleep Efficiency	0.14 + 0.07	0.18 + 0.11	–0.05 + 0.08	.5758	0.35 + 0.15	0.25 + 0.12	0.10 + 0.07	.1625
Sleep Disturbances	1.41 + 0.11	1.09 + 0.11	0.32 + 0.12	.0157	1.65 + 0.13	1.35 + 0.11	0.30 + 0.15	.0553
Use of Sleeping Medications	0.50 + 0.17	0.27 + 0.10	0.23 + 0.17	.2037	0.50 + 0.22	0.55 + 0.25	–0.05 + 0.18	.7894
Daytime Dysfunction	1.45 + 0.17	1.05 + 0.15	0.41 + 0.16	.0162	1.20 + 0.19	0.85 + 0.15	0.35 + 0.18	.0692

NOTE: MAAS = Mindful Attention Awareness Scale; PSQI = Pittsburgh Sleep Quality Index.
a. Based on paired t tests for the null hypothesis that there is no difference.

According to the caption, the descriptive statistics are reported as Mean$\pm$SE where SE denotes the standard error of the mean. Before we can calculate the effect size, we have to convert the standard errors to standard deviation which we can do simply by multiplying it by the square root of the sample size, $SD = SE \times \sqrt{n}$. It appears that Chiesa and Serretti (2009) skipped that step and instead used the standard errors as standard deviations in the effect size formula, that has the consequence of greatly inflating the effect size as we will see. Let's calculate the effect sizes properly and then try to also reproduce Chiesa and Serretti (2009)'s effect size calculation:

```r
# sample sizes for control (c) and treatment (tx) group
n_c <- 20
n_tx <- 22

## STRESS
data.frame(
  # treatment group
  d_treatment = d(m1 = 28.09, s1 = 1.19*sqrt(n_tx),
                  m2 = 25.00, s2 = 0.96*sqrt(n_tx),
                  direction = "higher=worse"),
  # control group
  d_control = d(m1 = 26.20, s1 = 1.30*sqrt(n_c),
                m2 = 25.55, s2 = 1.46*sqrt(n_c),
                direction = "higher=worse")
)
```

```
  d_treatment d_control
1   0.6093513 0.1051455
```



```
  ## SPIRITUALITY
  data.frame(
    # treatment group
    d_treatment = d(m1 = 55.14, s1 = 2.90*sqrt(n_tx),
                    m2 = 60.50, s2 = 2.65*sqrt(n_tx),
                    direction = "higher=better"),
    # control group
    d_control = d(m1 = 63.15, s1 = 3.08*sqrt(n_c),
                  m2 = 63.70, s2 = 2.57*sqrt(n_c),
                  direction = "higher=better")
  )
```

```
  d_treatment  d_control
1   0.4113868 0.04335779
```

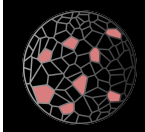
It's clear that the effect sizes are much smaller than what is reported in Chiesa and Serretti (2009). We can try to reproduce Chiesa and Serretti (2009)'s reported effects by using the standard error instead of the standard deviation when calculating the effect sizes:

```
  ## STRESS
  data.frame(
    # treatment group
    d_treatment = d(m1 = 28.09, s1 = 1.19,
                    m2 = 25.00, s2 = 0.96,
                    direction = "higher=worse"),
    # control group
    d_control = d(m1 = 26.20, s1 = 1.30,
                  m2 = 25.55, s2 = 1.46,
                  direction = "higher=worse")
  )
```

```
  d_treatment d_control
1    2.858111  0.470225
```

```
  ## SPIRITUALITY
  data.frame(
    # treatment group
    d_treatment = d(m1 = 55.14, s1 = 2.90,
                    m2 = 60.50, s2 = 2.65,
                    direction = "higher=better"),
    # control group
    d_control = d(m1 = 63.15, s1 = 3.08,
                  m2 = 63.70, s2 = 2.57,
                  direction = "higher=better")
  )
```

```
  d_treatment d_control
1    1.929575 0.1939019
```



I was able to reproduce the values reported by Chiesa and Serretti (2009) when improperly calculating the effect size using the standard errors. However, there is also a mistake in the direction of the effect size for the control group in the stress outcome. Chiesa and Serretti (2009) report negative pre-post effect in the control group even though stress decreased for both groups and therefore both effects should be positive.

Reproducing: Vieten and Astin (2008)

Chiesa and Serretti (2009) reported the stress outcome (Perceived Stress Scale, PSS) pre-post effect sizes for Vieten and Astin (2008) as 0.776 and 0.041 for the treatment and control group, respectively. Chiesa and Serretti (2009) then reported the spirituality outcome (Mindful Attention Awareness Scale, MAAS) effect sizes as 0.253 and –0.308 for treatment and control group, respectively. Table 1 from Vieten and Astin (2008) provides the information necessary to calculate the pre-post effects:

Table 1. Changes within groups from baseline to 10 weeks, and differences in changes between groups controlling for baseline values

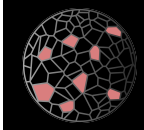
Measure	Value for group												Between-group ANCOVA		Effect size
	Intervention ($n = 13$)						Control ($n = 18$)								
	Baseline		10 wks		Change		Baseline		10 wks		Change				
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	F (2,24)	p	d
Perceived stress	20.1	5.1	15.9	5.7	3.5	5.7	17.1	5.0	16.9	4.6	−0.71	6.2	0.90	0.35	0.39
State anxiety	43.8	12.4	35.4	9.1	6.9	7.6	35.6	10.9	35.6	8.4	−0.35	7.5	4.32	0.04	0.85
Depression	20.4	8.4	16.2	7.3	3.6	5.2	14.2	5.4	17.2	7.4	−4.6	7.3	3.84	0.06	0.80
Negative affect	24.2	5.7	18.2	4.3	5.6	4.5	21.2	5.7	19.9	5.7	0.21	4.5	4.84	0.03	0.90
Positive affect	27.8	7.5	32.4	7.4	2.8	7.9	32.6	6.1	29.5	5.6	−3.0	4.9	3.24	0.08	0.73
Affect regulation	167.1	22.6	152.8	24.0	13.2	16.5	146.3	21.2	143.6	22.2	0.78	13.3	1.51	0.23	0.50
Mindfulness	3.6	0.76	3.8	0.82	0.19	0.45	3.8	0.57	3.6	0.72	−0.23	0.64	2.75	0.11	0.68

This table has some strange anomolies such as the change score means not being equal to the difference in pre and post-test means (e.g., for stress in the intervention group $15.9 - 20.1 \neq 3.5$). Also, the F-statistics with $F(2, 24)$ does not produce the p-values reported in the paper (e.g., for stress if $F(2, 24) = .90$ then $p = 0.4199$). The statistics as they are reported are likely untrustworthy In R, let's calculate the effect sizes from the table:

```
## STRESS
data.frame(
  # treatment group
  d_treatment = d(m1 = 20.1, s1 = 5.1,
                  m2 = 15.8, s2 = 5.7,
                  direction = "higher=worse"),
  # control group
  d_control = d(m1 = 17.1, s1 = 5.0,
                m2 = 16.9, s2 = 4.6,
                direction = "higher=worse")
)
```

```
  d_treatment  d_control
1   0.7950703 0.04163054
```

```
## SPIRITUALITY
data.frame(
  # treatment group
  d_treatment = d(m1 = 3.6, s1 = 0.76,
```



```
                  m2 = 3.8, s2 = 0.82,
                  direction = "higher=better"),
  # control group
  d_control = d(m1 = 3.8, s1 = 0.57,
                m2 = 3.6, s2 = 0.72,
                direction = "higher=better")
)
```

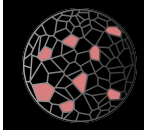
```
  d_treatment  d_control
1   0.2529822 -0.3080023
```

There is only a slight discrepency in the stress outcome where Chiesa and Serretti (2009)'s reported effect was .776 for the treatment group and I calculated .795. Otherwise the effects were reproducible.

Summarizing and visualizing the errors

I highlighted and annotated the corrections in table 1 of Chiesa and Serretti (2009). The edited table below shows three types of errors

- insufficient information/implausible results: effect sizes are not calculable directly from what is reported in the study. Effect sizes were then calculated by simulating plausible data sets from the available information. Then it is determined whether the effects reported by Chiesa and Serretti (2009) are plausible, if not, the mean of the simulated effect sizes will be the effect for that study.
- major error: corrected effect sizes are off by more than $\pm .05$.
- minor error: corrected effect sizes are off by less than $\pm .05$. Also minor errors will include sample size errors (there was only one).



INSUFFICIENT INFORMATION AVAILABLE
Effect sizes estimated via simulations of based on available information

MAJOR ERROR.
Corrected effects are greater than .05 difference from the reported effects

MINOR ERROR
Corrected effects are less than .05 difference from the reported effects (or sample size errors)

TABLE 1. SUMMARY OF INCLUDED STUDIES

Study	*Meditation/ comparison*	*N*	*Population*	*Duration (weeks)*	*Study design*	*Measure of stress*	*Pre–post within-group effect size*	*Measures of spirituality*	*Pre–post within group effect size*
Astin, 1997[34]	MBSR/waiting list	7/12	University undergraduates	8 weeks; 3-month follow-up	RCT	GSI	n.c.	INSPIRIT	n.c.
Shapiro, 1998[27]	MBSR/waiting list	36/37	Medical and premed students	7	RCT	GSI	0.632/0	INSPIRIT	n.c. 0.363/0
Rosenzweig, 2003[28]	MBSR/waiting list	140/162 125/152	Medical and premed students	10	SS-CT	POMS	0.205/−0.339 0.206/-.331	—	—
Beddoe and Murphy, 2004[35]	MBSR	16	Nursing students	8	UCT	DSP	n.c.	—	—
Cohen-Katz, 2005[36]	MBSR/waiting list	12/13	Nurses	8 weeks; 3-month follow-up	RCT	BSI	n.c. 0.716/0.532	MAAS	1.959/0.787 0.941/0.354
Shapiro, 2005[29]	MBSR/waiting list	10/18	Health care professionals	8	RCT	PSS	1.724/−0.303 1.159/0.257	—	—
Shapiro, 2007[30]	MBSR/weekly meetings	22/32	Therapists in training	10	CT	PSS	1.008/−0.162 0.949/-0.162	MAAS	0.372/−0.396 0.373/-0.397
Jain, 2007[31]	MBSR/relaxation training/waiting list	27/24/30	Medical students, graduate nursing students, undergraduate premed students	4	RCT AC	BSI	1.366/0.911/ 0.272	INSPIRIT-R	0.066/0.074/ −0.027
Klatt, 2008[32]	MBSR/waiting list	22/20	Faculty and staff at a large midwestern university	6	RCT	PSS	2.858/−0.47 0.609/0.105	MAAS	1.929/0.193 0.411/0.043
Vieten and Astin, 2008[33]	MBSR/waiting list	13/18	Pregnant women between 12 and 30 weeks gestation	10 weeks; 3 month follow-up	RCT	PSS	0.776/0.041 0.795/0.042	MAAS	0.253/−0.308

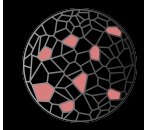
NC, not calculable; MBSR, mindfulness-based stress reduction; UCT, uncontrolled trial; RCT, randomized controlled trial; SS-CT, self selected controlled trial; RCT AC, randomized controlled trial with an active control; CT AC, controlled trial with an active control; GSI, global severity index (of the Hopkins Symptom Checklist 90 Revised); POMS, profile of mood symptoms; DSP, Derogatis stress profile; BSI, brief symptom inventory; PSS, perceived stress scale; INSPIRIT, index of core spiritual experiences; MAAS, mindfulness attention awareness scale; INSPIRIT-r, index of core spirituality–revised.

For 10 studies that had stress outcomes, there were 3 major errors, 2 minor errors, and 1 study with insufficient information/implausible results. For 7 studies that had spirituality outcomes, there were 2 major errors, 1 minor error, and 1 study with insufficient information/implausible results. We can append the data frame constructed at the beginning of the report with the corrected values (denoted with `corr` suffix):

```
df <- df %>%
  mutate(d_st_tx_corr = c(NA,.632, .206,NA,.716,1.159,.949, 1.366,.609,.795),
         d_st_c_corr =  c(NA,   0,-.331,NA,.532, .257,-.162, .272, .105,.042),
         d_sp_tx_corr = c(NA,.363,NA,NA,.941,NA,.373,.066,.411,.253),
         d_sp_c_corr =  c(NA,0,NA,NA,.354,NA,-.396,-.027,.193,-.308),
         n_t_corr = c(7,36,125, 16,12, 10,22,27,22,13),
         n_c_corr = c(12,37,152,NA,13,18,32,30,20,18))
```

The figures below shows the differences between treatment and control group $\Delta = d_{\text{Tx}} - d_{\text{C}}$. Figure 3 shows all the studies with stress outcomes where the circles denote the effect differences Δ and the diamond denotes the sample size weighted mean of Δ.

```
set.seed(1)
library(patchwork)

h_st <- ggplot(data = df) +
  geom_jitter(aes(x = d_st_tx_corr, y = 1, size = n_t),
              height = .2, width = 0, alpha = .4) +
  geom_jitter(aes(x = d_st_tx, y = 0,size = n_t),
              height = .2, width = 0, alpha = .4) +
  geom_point(aes(x = weighted.mean(d_st_tx_corr - d_st_c_corr, n_t_corr+n_c_corr,na.rm=T),
```



```
                y = 1),
            size = 7,color = "black", shape = 18) +
  geom_point(aes(x = weighted.mean(d_st_tx - d_st_c, n_t+n_c,na.rm=T),
                y = 0),
            size = 7,color = "black", shape = 18) +
  theme(aspect.ratio = .4, legend.position = "none") +
  geom_vline(xintercept = 0, linetype = "dashed") +
  geom_vline(xintercept = seq(.25,3,by=.25),
            linetype = "dashed",alpha=.07) +
  scale_y_continuous(breaks = 0:1,
                     labels = c("Reported", "Corrected"),
                     limits = c(-.5,1.5)) +
  scale_x_continuous(breaks = 0:6/2) +
  labs(x = "",
       y = "",
       title = "All studies (Stress)")


hrct_st <- ggplot(data = df %>% filter(rct == 1)) +
  geom_jitter(aes(x = d_st_tx_corr, y = 1, size = n_t),
              height = .2, width = 0, alpha = .4) +
  geom_jitter(aes(x = d_st_tx, y = 0,size = n_t),
              height = .2, width = 0, alpha = .4) +
  geom_point(aes(x = weighted.mean(d_st_tx_corr - d_st_c_corr, n_t_corr+n_c_corr,na.rm=T),
                y = 1),
            size = 7,color = "black", shape = 18) +
  geom_point(aes(x = weighted.mean(d_st_tx - d_st_c, n_t+n_c,na.rm=T),
                y = 0),
            size = 7,color = "black", shape = 18) +
  theme(aspect.ratio = .4, legend.position = "none") +
  geom_vline(xintercept = 0, linetype = "dashed") +
  geom_vline(xintercept = seq(.25,3,by=.25), linetype = "dashed",alpha=.07) +
  scale_y_continuous(breaks = 0:1,
                     labels = c("Reported", "Corrected"),
                     limits = c(-.5,1.5)) +
  scale_x_continuous(breaks = 0:6/2) +
  labs(x = "Treatment-Control Diff in Effects (Δ)",
       y = "",
       title = "RCTs Only (Stress)")

h_st / hrct_st
```

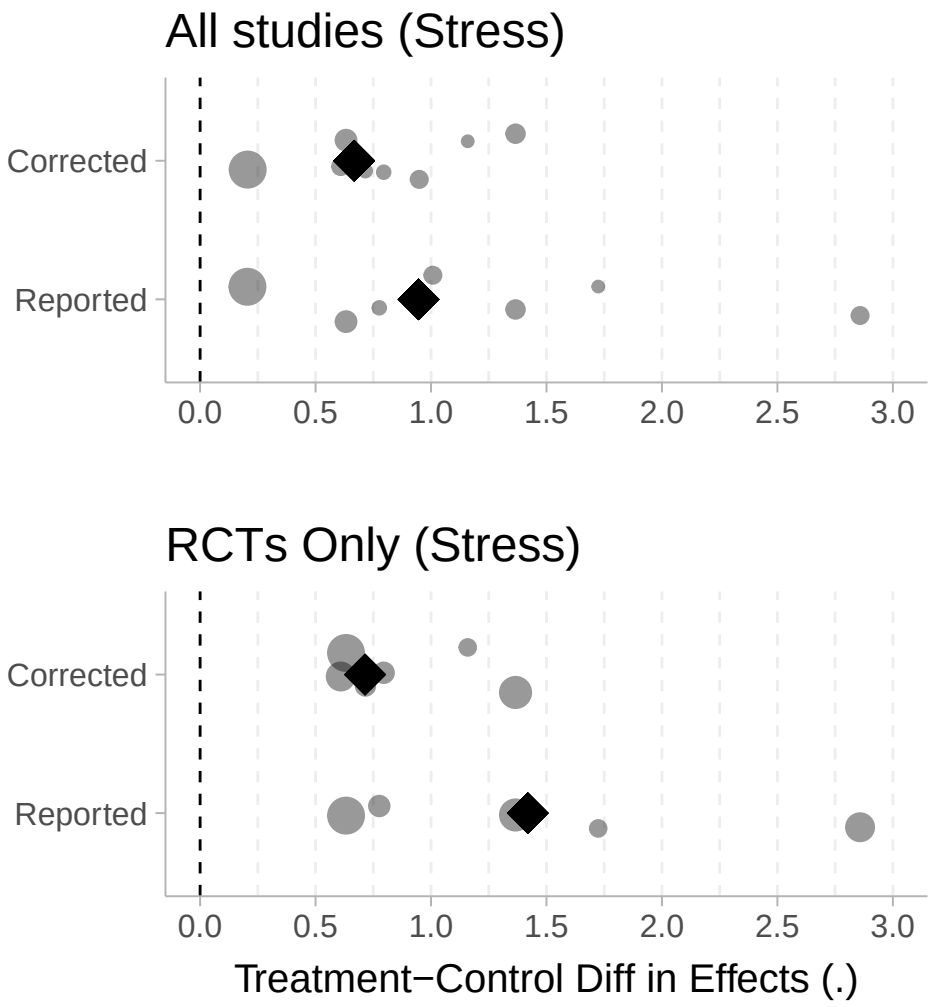


Figure 3: Treatment control contrasts between corrected and reported (by, Chiesa and Serretti 2009) for stress outcomes. We see noticeable changes in the treatment-control difference after correcting the effect sizes for both RCT studies and all studies.

Figure 3 shows the spirituality outcome where the circles denote the study effect differences and the diamond denotes the sample size weighted mean.

```
set.seed(1)

h_sp <- ggplot(data = df) +
  geom_jitter(aes(x = d_sp_tx_corr, y = 1, size = n_t),
              height = .2, width = 0, alpha = .4) +
  geom_jitter(aes(x = d_sp_tx, y = 0,size = n_t),
              height = .2, width = 0, alpha = .4) +
  geom_point(aes(x = weighted.mean(d_sp_tx_corr-d_sp_c_corr, n_t_corr+n_c_corr,na.rm=T),
                 y = 1),
             size = 7,color = "black", shape = 18) +
```

```
  geom_point(aes(x = weighted.mean(d_sp_tx-d_sp_c, n_t+n_c,na.rm=T),
                 y = 0),
             size = 7,color = "black", shape = 18) +
  theme(aspect.ratio = .4, legend.position = "none") +
  geom_vline(xintercept = 0, linetype = "dashed") +
  geom_vline(xintercept = seq(.25,3,by=.25),
             linetype = "dashed",alpha=.07) +
  scale_y_continuous(breaks = 0:1,
                     labels = c("Reported", "Corrected"),
                     limits = c(-.5,1.5)) +
  scale_x_continuous(breaks = 0:6/2) +
  labs(x = "",
       y = "",
       title = "All studies (Spirituality)")


hrct_sp <- ggplot(data = df %>% filter(rct == 1)) +
  geom_jitter(aes(x = d_sp_tx_corr, y = 1, size = n_t),
              height = .2, width = 0, alpha = .4) +
  geom_jitter(aes(x = d_sp_tx, y = 0,size = n_t),
              height = .2, width = 0, alpha = .4) +
  geom_point(aes(x = weighted.mean(d_sp_tx_corr-d_sp_c_corr, n_t_corr+n_c_corr,na.rm=T),
                 y = 1),
             size = 7,color = "black", shape = 18) +
  geom_point(aes(x = weighted.mean(d_sp_tx-d_sp_c, n_t+n_c,na.rm=T),
                 y = 0),
             size = 7,color = "black", shape = 18) +
  theme(aspect.ratio = .4, legend.position = "none") +
  geom_vline(xintercept = 0, linetype = "dashed") +
  geom_vline(xintercept = seq(.25,3,by=.25), linetype = "dashed",alpha=.07) +
  scale_y_continuous(breaks = 0:1,
                     labels = c("Reported", "Corrected"),
                     limits = c(-.5,1.5)) +
  scale_x_continuous(breaks = 0:6/2) +
  labs(x = "Treatment-Control Diff in Effects (Δ)",
       y = "",
       title = "RCTs Only (Spirituality)")

h_sp / hrct_sp
```

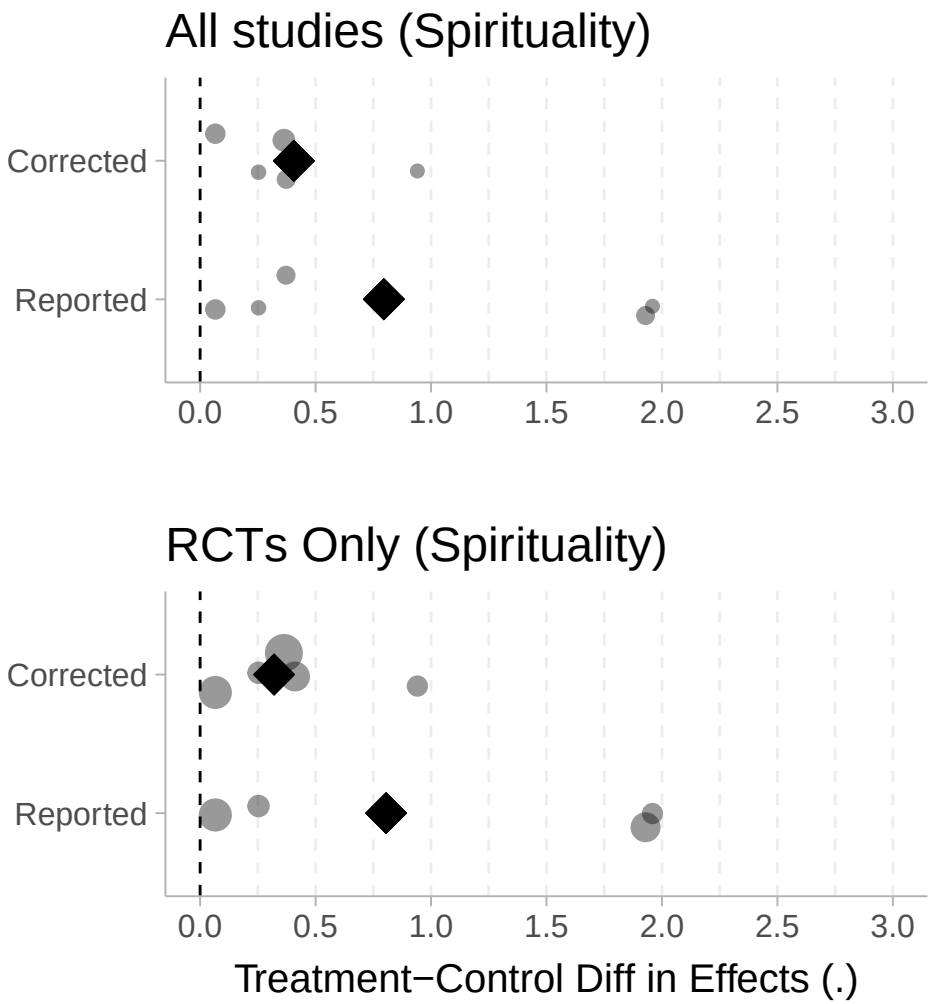
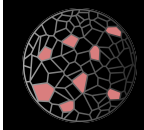


Figure 4: Treatment control contrasts between corrected and reported (by, Chiesa and Serretti 2009) for spirituality outcomes. We see noticeable changes in the treatment-control difference after correcting the effect sizes for both RCT studies and all studies.

As we can see from both Figure 3 and Figure 4, the reported treatment-control difference is substantially larger than the corrected effect sizes for both the stress and spirituality outcomes. This suggests the results reported in Chiesa and Serretti (2009) are artificially large.

Analysis Flaws

Chiesa and Serretti (2009) uses an N-weighted t-test to analyze the contrast between treatment and control pre-post effect sizes. This is an inappropriate statistical method for meta-analysis for two reasons: it does not properly model the heterogeneity in effects and it treats effects as fixed rather than estimated values. The method employed by Chiesa and Serretti (2009) artificially shrinks the standard errors (and thus confidence intervals) of average effects. A standard meta-analysis uses random-effects weights and estimation to calculate the mean



and confidence interval of the *mean* effect size across studies (see the Cochrane handbook https://training.cochrane.org/handbook/current/chapter-10#section-10-10-4).

Instead of comparing the means of pre-post effects independently, they should be modeled jointly since treatment and control pre-post effects coming from the same study tend to be correlated (e.g., for stress, we see a correlation of r=.52).

Conclusion

Due to the severe data extraction errors and analysis flaws, the primary meta-analytic results in table 4 in Chiesa and Serretti (2009) are incorrect and greatly exaggerate the effect between treatment and control groups for both stress and spirituality outcomes.

References

Astin, John A. 1997. "Stress Reduction Through Mindfulness Meditation: Effects on Psychological Symptomatology, Sense of Control, and Spiritual Experiences." *Psychotherapy and Psychosomatics* 66 (2): 97–106.

Beddoe, Amy E, and Susan O Murphy. 2004. "Does Mindfulness Decrease Stress and Foster Empathy Among Nursing Students?" *Journal of Nursing Education* 43 (7): 305–12.

Chiesa, Alberto, and Alessandro Serretti. 2009. "Mindfulness-Based Stress Reduction for Stress Management in Healthy People: A Review and Meta-Analysis." *The Journal of Alternative and Complementary Medicine* 15 (5): 593–600.

Cohen-Katz, Joanne, Susan D Wiley, Terry Capuano, Debra M Baker, and Shauna Shapiro. 2005. "The Effects of Mindfulness-Based Stress Reduction on Nurse Stress and Burnout, Part II: A Quantitative and Qualitative Study." *Holistic Nursing Practice* 19 (1): 26–35.

Jain, Shamini, Shauna L Shapiro, Summer Swanick, Scott C Roesch, Paul J Mills, Iris Bell, and Gary ER Schwartz. 2007. "A Randomized Controlled Trial of Mindfulness Meditation Versus Relaxation Training: Effects on Distress, Positive States of Mind, Rumination, and Distraction." *Annals of Behavioral Medicine* 33: 11–21.

Klatt, Maryanna D, Janet Buckworth, and William B Malarkey. 2009. "Effects of Low-Dose Mindfulness-Based Stress Reduction (MBSR-Ld) on Working Adults." *Health Education & Behavior* 36 (3): 601–14.

Rosenzweig, Steven, Diane K Reibel, Jeffrey M Greeson, George C Brainard, and Mohammadreza Hojat. 2003. "Mindfulness-Based Stress Reduction Lowers Psychological Distress in Medical Students." *Teaching and Learning in Medicine* 15 (2): 88–92.

Shapiro, Shauna L, John A Astin, Scott R Bishop, and Matthew Cordova. 2005. "Mindfulness-Based Stress Reduction for Health Care Professionals: Results from a Randomized Trial." *International Journal of Stress Management* 12 (2): 164.

Shapiro, Shauna L, Kirk Warren Brown, and Gina M Biegel. 2007. "Teaching Self-Care to Caregivers: Effects of Mindfulness-Based Stress Reduction on the Mental Health of Therapists in Training." *Training and Education in Professional Psychology* 1 (2): 105.

Shapiro, Shauna L, Gary E Schwartz, and Ginny Bonner. 1998. "Effects of Mindfulness-Based Stress Reduction on Medical and Premedical Students." *Journal of Behavioral Medicine* 21: 581–99.

Vieten, Cassi, and John Astin. 2008. "Effects of a Mindfulness-Based Intervention During Pregnancy on Prenatal Stress and Mood: Results of a Pilot Study." *Archives of Women's Mental Health* 11: 67–74.