TITLE

The Effect of Cognitive Behavioral Therapy as an Anti-Depressive Treatment is Falling: Reply to Ljòtsson et al. (2016) and Cristea et al. (2016).

RUNNING HEAD

Effect of Cognitive Behavioral Therapy is Falling

AUTHORS

Oddgeir Friborg

Tom J. Johnsen

AFFILIATION

Faculty of Health Sciences, Department of Psychology, UiT The Arctic University of Norway, Tromsø, Norway.

CORRESPONDING AUTHOR

Tom J. Johnsen, Faculty of Health Sciences, Department of Psychology, UiT The Arctic University of Norway, University of Tromsø, N-9037 Tromsø, Norway. E-mail: tjj@psykologtromso.no

ABSTRACT

This paper critically reassess Ljótsson et al.'s (2016) nonlinear reanalysis and review Cristea et al.'s (2016) extension of our original meta-analysis (Johnsen & Friborg, 2015) reporting a decline in the effects of cognitive behavioral therapy (CBT) for treating unipolar depression. Ljótsson fitted a piecewise meta-regression model to the data indicating a halt in the decline from the year 1995 onwards, hence concluding that CBT is not gradually losing its efficacy. We reanalyzed the data for nonlinear time trends and replicated their findings for the 34 studies using Hamilton Rating Scale for Depression as outcome, but not for the 67 studies using Beck's Depression Inventory as outcome. The best nonlinear model was quadratic rather than flat (or linear) from 2001 onwards; which opposes Ljótsson's conclusion of stability in effects. Cristea et al.'s identified additional studies, but their new analyses provided mixed support for a linear decline in CBT effects. They could not dismiss a decline except only in the most stringent analytic condition—namely, when analyzing only 29 randomized controlled trials based on between-group effect sizes solely. Their study includes several questionable methodological choices, so we expand on the discussion of these disparate meta-analytic findings. Of particular concern is the tendency to downplay the fact that when looking at all of the studies together—there is a clear decline in the effects of CBT, which should concern therapy researchers within the field rather than being explained away.

# The Effects of Cognitive Behavioral Therapy as an Anti-Depressive Treatment is Falling: Reply to Ljòtsson et al. (2016) and Cristea et al. (2016).

As the authors of a meta-analysis examining the time trends in the effectiveness of cognitive behavioral therapy (CBT) as an anti-depressive treatment (Johnsen & Friborg, 2015), we are pleased to read critical follow-up papers. The increased focus on historical trends in the effects of psychotherapy is likely to contribute positively to the development and implementation of more efficacious psychotherapies in the future. Some steps towards this end have been taken with the recent publication of two re-reviews of our original findings: the first being a statistical re-analysis (Ljótsson, Hedman, Mattsson, & Andersson, 2016) and the second being a meta-analytic extension (Cristea et al., 2016). We do however note considerable methodological or statistical issues in both papers, which has prompted the current reply.

## Reply to Ljótsson et al. (2016)

Ljótsson et al. (2016) concluded that the decline in CBT effects had stopped falling. They arrived at this conclusion following a meta-regression analysis that had curvature or segmented parameters added to the model. This model ostensibly revealed a leveling off in the decline from 1995 onwards. Hence, they concluded that the CBT treatment effects had not declined during the last 20 years, and that the effects of the current CBT protocols vary around their "true" clinical effect. Given the authors' statement in their final paragraph of the paper, "we did not find any support in their data for their conclusion that the effects of CBT are in decline", one could get the impression that the conclusions by Johnsen

and Friborg (2015) were ill-informed. Inspired by these new analyses, we also reanalyzed the dataset to examine whether we would arrive at similar conclusions.

We first wish to put into perspective the basic message regarding time trends in CBT treatment effects. The predicted decline from 1977 to 2014 in the Beck's Depression Inventory (BDI) effect sizes (ESs), based on Ljótsson's analyses, is a Hedge's $g$ = 1.33 (falling from 2.76 to 1.43). The corresponding decline, according to the linear model, is $g$ = 1.09 (falling from 2.27 to 1.18). The piecewise model predicts a steeper initial fall; differences in the 2014 effects between the two methods is $g$ = .25, which is rather negligible. Comparable reduction in the Hamilton Rating Scale for Depression (HRSD) effects based on the piecewise model is $g$ = 1.69 (falling from 3.29 to 1.60) and $g$ = 1.13 for the linear model (falling from 2.52 to 1.39). The difference in 2014 effects is $g$ = .21, also quite unimportant. What seems to be at stake here is more than just the difference between the two methods in predicting the treatment effects at present. Rather it seems to be the conceptual underpinnings of the linear model, which suggests continual decline. Conversely, the piecewise model provides more solace by indicating a halt in the fall. The current treatment effects are still to be considered good and perhaps even strong; yet, it is puzzling why this happens.

We did not explicitly analyze curvature or segmented time trends in the original paper as did Ljótsson et al. (2016); however, we did not miss this entirely, as is evident in Figure 7 from Johnsen and Friborg (2015). Figure 7 portrays how the

time coefficients change depending on the starting year for study inclusion, even turning positive only when including studies after the year 1995. The choice of 1995 as a breakpoint in Ljótsson's analyses was motivated by this figure, and hence Ljótsson and colleagues did not use an empirical criterion for deciding their breakpoint. Here, we thoroughly examined nonlinear time trends and used a statistical criterion for deciding a regression slope segmentation breakpoint.

*A nonlinear reanalysis*

BDI: We first visually inspected the scattering of the weighted ESs in Figure 1 and noticed that the decline was fairly stable until the year 2001. Moreover, the reported ESs between the years 2001 and 2014 seemed slightly inverse u-curved rather than completely flat, as the piecewise model suggests. In order to examine this possibility, we specified a segmented model consisting of two parts: a linear part describing the whole time period (1977-2014) and a quadratic part describing the time trend following the breakpoint. The fitted weighted least square regression model using random effects model weights from the Johnsen and Friborg (2015) paper was:

$$f(ES) = b_0 + b_1 Y + b_2 Y_{pos} + b_3 Y_{pos}^2$$

Coding of $Y_{pos} = \begin{Bmatrix} Y \; if \; Y \geq 0 \\ 0 \; if \; Y < 0 \end{Bmatrix}$, 0 representing the centered year (breakpoint).

Where $Y$ = year.

The breakpoint was empirically chosen by searching for the publication year that could render both parts of the model to be statistically significant. This only

happened if publication year was centered at the year 2001. Statistical significance for the nonlinear regression parameters is presented in Table 1. The nonlinear model also yielded the highest model fit in terms of the R-square index. Figure 1 illustrates the linear and nonlinear trends visually, including the explanation of the respective amounts of between-study variance.

--- Insert Table 1 and Figure 1 about here ---

In addition to presenting normal standard errors, we also produced bootstrapped error bands based on 5,000 resamplings. Residuals of both models were normally distributed ($Z_{skewness}$ = 1.07 and .86, $Z_{kurtosis}$ = .06 and .11, and both Kolmogorov–Smirnov tests were non-significant). Hence, the bootstrapped confidence intervals overlapped strongly with the model-based intervals.

HRSD: These effects were best described by a piecewise model. The nonlinear model (similar as above), which fit best when centered at the year 2001 ($R^2$ =.298) was not better than the best piecewise model centered at 1998 ($R^2$ = .290). The first part of the segmented nonlinear regression was significant ($b_1$Y = -.080, $p$ = .002); however, this was not the case with the second quadratic part ($b_2Y_{pos}$ = .245, $p$ = .07; $b_3Y^2_{pos}$ = -.011, $p$ = .27) although the coefficients were surprisingly comparable with the nonlinear BDI coefficients.

*Summary and discussion*

These additional analyses indicate that CBT effects, as measured by the BDI, have fallen linearly from 1977 until 2001, and not until 1995 as proposed by Ljótsson

et al. (2016). The fall has been going on for about 24 years, which encompasses half of all studies (33 of 67). From 2001 onward, the treatment effects have not declined further, but stability in the effects cannot be claimed due to the significant segmented quadratic model. This model shows a temporary rise followed by another fall, which may or may not be ongoing. Whatever is true, the major point is that a flattening in the treatment effects of CBT or that the CBT effects now vary around their "true" value, as Ljótsson et al. (2016) conclude, is not well supported by the available data. The segmented nonlinear model with the publication year 2001 as the breakpoint also explained 2.4% more of the variation in the treatment effects than the piecewise model with year 1995 as the breakpoint. We acknowledge Ljótsson and colleagues' effort in addressing nonlinear time trends as it helped gain additional insight into temporal trends. But since they overlooked a significant quadratic curvature in the second part of the segmented model, their conclusions are overstated.

Regarding the HRSD effects, the piecewise model fit the data best; hence, we are left with a mixed picture. There are however good reasons for weighting the BDI outcome data more heavily since the statistical power for detecting nonlinear HRSD trends, with only 34 studies available, is considerably smaller compared to the 67 available BDI studies. The HRSD measure also compares less favorably with the BDI measure in terms of poorer sensitivity to the psychological symptoms of depression related to nonendogenous, atypical depression or personality dysfunctions (Enns, Larsen, & Cox, 2000). HRSD seems, on the other hand, to be more sensitive to somatic symptoms related to endogenous depression. This makes sense since the BDI was specifically designed by the

founder of CBT, Aaron Beck, to identify improvements in attitudinal and cognitive components following therapy (e.g., hopelessness, self-worthlessness, self-dislike, or guilt). We thus consider the BDI to be more valid in evaluating the effects of his therapy than HRSD, which also the large number of clinical trials using the BDI is a testimony of.

What additional points can be made of this reanalysis? First, the present reanalysis do not change the basic message stating that CBT effects have fallen considerably across two and half decades. In fact, the predicted ES for the year 2014 even comes out slightly worse for the segmented nonlinear ($g$ = 1.12), as compared to the linear, model ($g$ = 1.18). Nevertheless, the current ESs are strong, hence CBT is still to be considered as an effective anti-depressive treatment.

Second, it may be wise to include nonlinear time trends in future meta-analyses of therapy studies in order to obtain more accurate information about psychotherapy effects. Since the current reanalysis shows that the nonlinear time trend explains a considerable portion of the between-study treatment variance (almost 30%), future meta-analytic summaries of treatment effects should not dismiss potential time trends.

Third, the psychotherapy research field may profit hugely by establishing a common minimum of variables/measures that is to be included as moderator variables in all future therapy trials. That would not only benefit the individual researcher attempting to analyze reasons for better or worse treatment

outcomes in the study at hand, but also any future meta-analytic attempts at analyzing reasons for time trends in psychotherapy effects more exactly.

Last, since the BDI effects during the last 13 years do not follow a flat trend but rather are in decline again, we believe a weather-climate analogy is an apt comparison: although weather varies across decades, the long-term climate changes (as projected by a linear model) may be regarded as the most reliable indicator.

**Reply to Cristea et al. (2016)**

The study by Cristea et al. (2016) offers a comprehensive extension of the original meta-analysis as they identified a number of additional CBT studies. They also introduce a number of methodological changes that are poorly justified and even incorrect in our opinion. Their analysis offers rather mixed results concerning whether the effects of CBT are in decline or not, which we would like to critically review.

A clear strength of their paper is the identification of 30 additional studies, including 12 randomized controlled trials (RCTs), as compared to the original meta-analysis. This increases the reliability of the time trend coefficient. Unfortunately, they seem to suggest that Johnsen and Friborg (2015) had somehow missed these studies, when in fact they had simply revised the inclusion criteria by including papers published in all languages; in contrast, the original paper included only papers exclusively published in English for the purposes of interpretation. Cristea et al. (2016) also found some inconsistencies in Johnsen and Friborg's (2015) selection of papers—they pointed out that four of the papers should not have been included and questionable calculation of the effect sizes (ESs) for two of the included papers; however, these revisions did not change the original findings.

Cristea et al. (2016) introduce a number of methodological changes to the meta-analysis. For instance, they link the combination of non-RCT and RCT studies to the high degree of heterogeneity noted between studies, which we also believe may be the case. This is why we conducted sub-group analyses for within and

controlled designed studies separately. More troublingly, they argue for excluding non-RCTs from their main analysis because such trials may yield biased findings owing to a plethora of selection biases, which may cause the participating groups to differ at pretest. We fully acknowledge this important objection. However, they also argue that non-RCTs studies may be more correlated with the passage of time than RCT studies may be. They provide no justification for this claim, nor can we conceive of a sensible reason for making it. Why would a potential selection bias (e.g., more motivated patients or more depressed patients) be systematically present solely during the 70s or 80s and not later on? As we regard this possibility as tiny at best, we consider Cristea et al.'s (2016) exclusion of a large array of clinically relevant studies, instead of including them despite the risk of minor time trend biases, to be a major error. Their choice therefore seems to serve a confirmatory purpose.

Second, they object to the use of within-group ESs calculated from pre-post data and to the combination of within- and between group ESs when analysing all studies together. Their argument is that within-group ESs cannot be disentangled from the context in which the study was conducted, which in practise means having a comparison group. While we acknowledge this point, it is important to note that our study (Johnsen & Friborg, 2015) did not rely solely on within-group ESs. As mentioned above, we conducted sub-group analyses, which revealed that the decline extended to the between-group condition. It is important to keep in mind that the vast majority of the calculated within-group ESs was based on randomized clinical trials; however, these studies could not be calculated as between-group ESs as they were compared to other treatment

arms (e.g., medication) or did not include a no-intervention group. Many of these studies thus had a "context" that was not defined by a control group arm. The only method for quantifying the ESs from these relevant treatment arms was to use the within-group formula. A known problem is overestimation of the ES, which may be adjusted for with the correlation between the pre- and post-test measure. The higher this correlation, the lower the within-group ES. In our case, we imputed a large correlation ($r$ = 0.7) for studies not reporting it, thus reducing overestimation risks. But even if these ESs were overestimated it is difficult to conceive of a sensible explanation for why within-group calculated ESs might favor earlier CBT trials compared to later ones, whereas between-group calculated ESs do not, which Cristea et al. (2016) assume. Again, they provide no justification for why this should be the case. Another problem is that Cristea et al. (2016) base their "new" analysis on post-test scores only, whereas we used the standard deviation of the difference score in both the within- and between-group estimations in order to use a comparable denominator. The use of difference scores also corrects between-group ESs for any pre-treatment differences that may occur despite randomization in small sample studies. The meta-analysis of all studies combined was thus more correct in our original approach, whereas Cristea et al. (2016) mix the use of post- and difference scores when analysing all studies together. A final argument for including all available studies is to ensure a substantially larger study pool, which is important for avoiding an underpowered statistical analysis and enabling weaker, yet still clinically important, statistical effects stretching across decades to appear. Studies of clinical effectiveness should, in our opinion, record whether any clinical improvement (or decay) is apparent in both lesser or better defined

contexts. This is well reflected in the long-standing discussion of the use of RCT designs in studies of clinical effectiveness (Persons & Silberschatz, 1998), quote: "*RCT advocates have sacrificed clinical validity in the effort to maximize experimental control*". If ESs do change with time, independently or within a particular context (control group or not), then time trends would still be clinically relevant and thus would need to be addressed. Omitting these studies, as Cristea et al. (2016) do in their "new" analysis, thus represents a larger mistake than including them.

Third, Cristea et al. (2016) consider the use of univariate regression analysis as misleading by increasing the risk of type I error in hypothesis testing. It is important to note that our primary hypothesis exclusively concerned publication year—namely, to what extent an increase in treatment effects was evident across time, as is evident within most other branches of medicine (e.g., publication series of *Advances in Medicine and Biology*). Our approach was not to examine a set of moderators and then select the one(s) that were statistically significant; hence, the univariate regression approach seemed optimal. We did conduct multiple two-way interaction tests between publication year and the moderators (i.e., *time* x *moderator*), hence these tests were prone to type 1 error. But since support of these tests would weaken the temporal (or time) hypothesis, any appropriate statistical adjustments would only make the rejection of the temporal hypothesis less likely. Cristea et al. (2016) conducted a so-called "full model" meta-regression analysis that included all moderators in a multivariate fashion. Moreover, they retained all variables in the model even though 10 of the 11 moderators were statistically non-significant, which reduces the degrees of

freedom substantially, which is quite negative for the statistical power in small samples. A defendable reason for conducting multivariable testing, as Cristea et al. (2016) did, would be if: a) theory or previous empirical evidence substantiate the inclusion of such a large array of predictors, b) omitting a moderator would significantly bias the estimation of the time coefficient, and c) the moderator contributes significantly to the explanation of ES. Since studies of temporal development of psychotherapy effects are a completely new endeavour, neither theory nor relevant empirical evidence exist and support such a-priori multivariable models. Estimation biases may nevertheless occur if an omitted moderator correlates positively with time. This was potentially the case for two moderators (i.e., study quality ratings, and type of BDI measure), but none of these contributed significantly to the explanation of between-study ESs. Inclusion of such non-significant variables (and Cristea et al. included 10 variables) would thus introduce a "spurious" adjustment of the regression model. Had our study context been one that embraces multiple hypothetical explanations for the decline, Cristea et al.'s approach had made sense. Since publication year was our sole hypothesis, their objection is irrelevant to the original statistical analysis.

Fourth, they claimed that time trend analyses should be based on "intention-to-treat" (ITT) rather than "completer" data. In our case, we had no *a priori* reason to consider ITT as any better than an analysis based on completers. Although ITT analyses do retain all patients and thus reduce systematic attrition, they can be biased (Lane, 2008) due to undue assumptions of no change among patients that drop out. Cristea et al. (2016) further argue that the ITT procedure may be less

susceptible to time trend effects than completer data, but again provide no justification or evidence for this point. In contrast, our choice of using data from completers was well-informed because this was the only information available in early CBT trials. Since completer data were uniformly reported and the dropout rate from CBT studies is low in general, we consider analyses based on such data as equally (if not more) correct than analyses based on ITT data.

Finally, Cristea et al. combined treatment effects from trials including several subgroups rather than coding selected subgroups according to an *a priori* criterion. This strategy yielded results supporting weaker time trends, which they argued as superior to basing the calculations on a particular group. Our argument for selecting the most severely depressed patient group was to achieve a uniform comparison group rather than merely collapsing a variety of groups to serve as a comparison. This strategy, if anything, should reduce rather than increase study heterogeneity, which was one of their prime concerns. Since baseline severity does not moderate the outcome of CBT for depression, as reported in a meta-analysis by one of the authors (Driessen, Cuijpers, Hollon, & Dekker, 2010), this is another example of poorly justified selection of studies.

The exclusion of CBT studies regarded as outliers is inherently problematic because such studies may represent less frequent but still true observations in the population. Indeed, we examined the unstandardized residuals for the segmented nonlinear time trend model in our reply to Ljótsson et al. (2016), which showed an almost perfect normal distribution (skewness $Z = 0.86$, kurtosis $Z = 0.11$) with no extreme observations. Hence, removal of outliers is

unjustified, particularly Cristea et al.'s (2016) choice to consider one-third of the within studies in their meta-reanalysis as outliers. They justify their choice by branding it "the winners curse", meaning that there is no way for ESs to go but down. Hence, we should not expect anything other than a decline—even after 40 years of time to improve psychotherapy. This is an extremely pessimistic view on psychotherapy as a field, which is highly speculative and is backed by no evidence, to our knowledge, from research on time trends in psychotherapy.

Cristea et al.'s (2016) choice of splitting the study pool according to whether studies were conducted in the US ($k$ = 25) or the rest of the world ($k$ = 20) is also poorly justified with regard to the time trend hypothesis, although it offers some interesting findings in itself.

We agree with Cristea et al. (2016) that meta-analyses are inherently tricky to design and perform because of the considerable heterogeneity of the studies being aggregated. Another challenging but important aspect relates to communication of the findings in an objective, unbiased, and prudent fashion. In this regard, Cristea et al.'s (2016) study appears to deliberately downplay the fact that many (if not most) of their adjusted analytic conditions support our original findings. In fact, it seems that only the most stringent condition—wherein only 29 RCTs based on between-group ESs were included—reliably contradicted our original findings. However, the largest analytic conditions, based on 45 and up to 75 RCTs, mainly supported the original findings. As readers, we are thus left with an obscure picture; the authors seem to selectively favor results that do not confirm a decline and reject those indicating such a

decline. Indeed, they exclude studies they construe as outliers, and studies including inpatients, and prioritize ITT over completers' analyses. Even after doing all this, the original findings were still evident, which led Cristea et al. to disregard within-group studies altogether to achieve the desired non-significance. Remarkably, to achieve this, they had to reduce the largest analytic condition from 75 to 29 studies! Even at this point, a negative time trend was still present (*beta* = -.01, *p* = .22); however, the low statistical power precludes any strong conclusions. In sum, their paper lacks, in our view, a balanced portrayal of the results, which is of major concern because at least four of the authors are, to our knowledge, adherers to or advocates of CBT. Cristea et al.'s (2016) characterization of their own meta-analysis as the "gold standard" analysis is thus not credible given the current criticism.

Despite Cristea et al.'s (2016) removal of within-group ES calculations to achieve a non-significant decline, the fact remains that, whatever the causes and contextual underpinnings, CBT as a treatment has overall suffered a systematic decline in its ability to treat depressive symptoms. In other words, today, fewer patients recover to the same extent as they did in the past. To brush off this important discovery as a spurious observation reminds us of the idiom, "burying one's head in the sand." To illustrate: if a chemotherapy drug exhibited a significant negative time trend in its ability to treat cancerous tumor cells when considering all studies in a meta-analysis, would any right-minded person still consider this drug as efficacious and safe as originally thought? Or, would it be wise to start addressing the problem and discuss ideas about how to improve this trend? This perspective is unfortunately lacking in both of the recent meta-

analytic re-analyses (Cristea et al., 2016; Ljótsson et al., 2016), which is of concern for future improvements.

REFERENCES


Cristea, I. A., Stefan, S., Karyotaki, E., David, D., Hollon, S. D., & Cuijpers, P. (2016). The effects of cognitive behavioral therapy are not systematically falling: a revision of Johnsen & Friborg (2015). *Psychological Bulletin, accepted for publication*.

Driessen, E., Cuijpers, P., Hollon, S. D., & Dekker, J. J. M. (2010). Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A metaanalysis. *Journal of Consulting and Clinical Psychology, 78*(5), 668–680. doi:10.1037/a0020570

Enns, M. W., Larsen, D. K., & Cox, B. J. (2000). Discrepancies between self and observer ratings of depression: The relationship to demographic, clinical and personality variables. *Journal of Affective Disorders, 60*, 33-41.

Johnsen, T. J., & Friborg, O. (2015). The Effects of Cognitive Behavioral Therapy as an Anti-Depressive Treatment is Falling: A Meta-Analysis. *Psychological Bulletin, 141*(4), 747-768.

Lane, P. (2008). Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. *Pharmaceutical statistics, 7*, 93-106. doi:10.1002/pst.267

Ljótsson, B., Hedman, E., Mattsson, S., & Andersson, E. (2016). The Effects of Cognitive Behavioral Therapy for Depression are not Falling: A Re-Analysis of Johnsen & Friborg (2015). *Psychological Bulletin, accepted for publication.*

Persons, J. B., & Silberschatz, G. (1998). Are results of randomized controlled trials useful to psychotherapist? *Journal of Consulting and Clinical Psychology, 66*(1), 126-135.

Table 1. *Comparison of Two Nonlinear Models for Predicting Change in BDI Effect Sizes Across Time.*

| | | *beta* | *p* | *CI* .95 | Bootstrapped *CI* .95 |
|---|---|---|---|---|---|
| 1 Piecewise | $R^2$=.270 | | | | |
| $b_0$ | | 1.4191 | | | |
| Y | | -0,0744 | < .001 | -.1112 -.0376 | -.1130 -.0410 |
| $Y_{pos} \geq 1995$ | | 0,0741 | .010 | .0183 .1300 | .0190 .1340 |
| 2 Nonlinear | $R^2$=.294[a] | | | | |
| $b_0$ | | 1.120 | | | |
| Y | | -.0653 | < .001 | -.0934 -.0371 | -.0907 -.0408 |
| $Y_{pos} \geq 2001$ | | .2070 | .007 | .0586 .3554 | .0684 .3494 |
| $Y^2_{pos} \geq 2001$ | | -.0109 | .043 | -.0215 -.0004 | -.0210 -.0016 |

*Notes*. $b_0$ = intercept, *beta* = unstandardized coefficient, *p* = p-value, *CI* .95 = 95% confidence interval.

Figure 1. *Time Trends for the Different Meta-Regression Prediction Models*