Departing from the essential features of a high quality systematic review of psychotherapy : A response to ÖST (2014) and recommendations for improvement

Atkins, Paul W. B., Ciarrochi, Joseph, Gaudiano, Brandon A., Bricker, Jonathan B., James, Donald, Rovner, Graciela, Smout, Matthew, Livheim, Fredrik, Lundgren, Tobias and Hayes, Steven C.

Atkins, P. W. B., Ciarrochi, J., Gaudiano, B. A., Bricker, J. B., James, D., Rovner, G., Smout, M., Livheim, F., Lundgren, T. and Hayes, S. C. (2017). Departing from the essential features of a high quality systematic review of psychotherapy : A response to ÖST (2014) and recommendations for improvement. *Behaviour Research and Therapy*, 97, pp. 259-272. https://doi.org/10.1016/j.brat.2017.05.016

Departing from the Essential Features of a High Quality Systematic Review of Psychotherapy: A Response to Öst (2014) and Recommendations for Improvement

Paul W. B. Atkins
Joseph Ciarrochi
Brandon A. Gaudiano
Jonathan B. Bricker
James Donald
Graciela Rovner
Matthew Smout
Fredrik Livheim
Tobias Lundgren
Steven C. Hayes

Corresponding Author:
Dr Paul W.B. Atkins
Institute for Positive Psychology and Education
Australian Catholic University
Strathfield 2135 NSW
AUSTRALIA
Email: paulw.atkins@acu.edu.au
Tel: +61 2 9701 4741

ABSTRACT

Öst's (2014) systematic review and meta-analysis of Acceptance and Commitment Therapy (ACT) has received wide attention. On the basis of his review, Öst argued that ACT research was not increasing in its quality and that, in contradiction to the views of Division 12 of the American Psychological Association (APA), ACT is "not yet well-established for any disorder" (2014, p. 105). We conducted a careful examination of the methods, approach, and data used in the meta-analysis. Based in part on examinations by the authors of the studies involved, which were then independently checked, 91 factual or interpretive errors were documented, touching upon 80% of the studies reviewed. Comparisons of Öst's quality ratings with independent teams rating the same studies with the same scale suggest that Ost's ratings were unreliable. In all of these areas (factual errors; interpretive errors; quality ratings) mistakes and differences were not random: Ost's data were dominantly more negative toward ACT. The seriousness, range, and distribution of errors, and a wider pattern of misinterpreting the purpose of studies and ignoring positive results, suggest that Öst's review should be set aside in future considerations of the evidence base for ACT. We argue that future published reviews and meta-analyses should rely upon diverse groups of scholars rather than a single individual; that resulting raw data should be made available for inspection and independent analysis; that well-crafted committees rather than individuals should design, apply and interpret quality criteria; that the intent of transdiagnostic studies need to be more seriously considered as the field shifts away from a purely syndromal approach; and that data that demonstrate theoretically consistent mediating processes should be given greater weight in evaluating specific interventions. Finally, in order to examine substantive progress since Öst's review, recent outcome and process evidence was briefly examined.

Keywords: treatment efficacy, randomized clinical trials, acceptance and commitment therapy, cognitive behavior therapy, research methodology, empirically based treatments

Departing from the Essential Features of a High Quality Systematic Review of Psychotherapy: A Response to Öst (2014) and Recommendations for Improvement

The evidence base for the efficacy of Acceptance and Commitment Therapy (ACT) is substantial. ACT is currently listed on the APA Division 12 website as having strong research support for chronic pain and modest research support for depression, mixed anxiety, obsessive compulsive disorder, and psychosis. The website of the Association for Contextual Behavioral Science (https://contextualscience.org/ACT_Randomized_Controlled_Trials) currently lists 171 randomized trials and several dozens more are as of yet uncatalogued because they exist only in non-English versions. Entering even a short list of ACT relevant subject terms in the Web of Science leads to the identification of well over 1,000 articles.

Such a large body of extant research, about 80% which has been produced in the last five years, has led to a series of efforts to summarize and evaluate the ACT and acceptance-based behavior therapy literature and to consider its implications. At least 14 meta-analyses of ACT have appeared since 2009 (see https://contextualscience.org/state_of_the_act_evidence). A recent meta-analysis in the area of anxiety and depression using sequential meta-analytic techniques (Hacker, Stone, & MacBeth, 2016) found that ACT had reached "sufficiency" (i.e. a point at which further research is unlikely to reveal different conclusions) for a large within-treatment effect and a moderate between-group comparison effect in most areas at posttreatment but not superiority over existing evidence-based methods. A-Tjak et al. (2015) and Powers, Zum Vorde Sive Vording, and Emmelkamp (2009) conducted independent meta-analyses and found similar results across a wider range of mental health problems with ACT outperforming control conditions at posttreatment and follow-up for primary outcomes, but with no significant

difference from traditional cognitive behavioral therapy (CBT) more generally. Ruiz (2012), in a meta-analysis focused specifically on comparing ACT to CBT, found that ACT outperformed CBT overall, for depression and for quality of life in the studies analyzed. A recent targeted meta-analysis of studies of substance use disorders (Lee, An, Levin, & Twohig, 2015) found that ACT was statistically superior to active treatment comparisons including CBT, but not when CBT alone was considered. Meta-analyses have also shown that the treatment components of the psychological flexibility model (e.g. acceptance, mindfulness, values) underpinning ACT produce positive and sometimes additive effects (Levin, Hildebrandt, Lillis, & Hayes, 2012).

Against this backdrop, a review by Öst (2014) stands out for its conclusions regarding ACT research. Öst (2014) concluded that the average quality of research in ACT was not improving based on a methodological scale that he developed. In contrast to these conclusions, A-Tjak et al.'s (2015) meta-analysis found that ACT research was improving methodologically according to the same scale. In a recent commentary comparing Öst (2014) with A-Tjak et al.'s (2015) analysis, Hertenstein and Nissen (2015, p. 250) suggest: "It is apparent that the two meta-analyses reach strikingly contrasting conclusions, calling for a critical investigation of the potential reasons for this difference." That is the purpose of the present article.

Gaudiano's (2009a) re-visiting of Öst's (2008) original meta-analysis demonstrates that *average* methodological scores alone do not say much about a research program. The primary question is whether enough high quality studies are available to establish robust scientific conclusions. Methodological ratings thus become most relevant in weighing the additive effects of several studies and their strengths and weaknesses. Such a use of methodological analysis requires very careful attention to the small details. Study-by-study, the ratings need to be relevant, reliable, and examined in detail, rather than in a global or "all-or-none" fashion.

An interest in such details is important in part because Öst (2014) argued broadly that the degree of research evidence for ACT has been systematically over-estimated by the Society of Clinical Psychology (Division 12 of the American Psychological Association) across all disorders it has reviewed. Öst (2014, p. 105) concluded: "ACT is not yet well-established for any disorder." Web of Science shows that the 2014 meta-analysis has already been cited 46 times (the 2008 review has been cited 196 times). Its conclusion stands in juxtaposition to meta-analyses concluding sufficiency has been reached in some key areas (Hacker et al., 2016), the inclusion of ACT on the Substance Abuse and Mental Health Services Administration's National Registry of Evidence-based Practices and Procedures, and the decision by the U.S. Veterans Administration to deploy ACT as an evidence-based method, and to inclusion on the Division 12 evidence-based therapy list itself for multiple specific conditions.

Scholarly criticism is important in science. Indeed, the society of professionals who are primarily responsible for developing ACT, the Association for Contextual Behavioral Science (ACBS; www.contextualscience.org), has several times had Öst speak about his concerns at ACBS conferences, resulting in useful debate and discussion of the issues. Unfortunately, an examination of the Öst (2014) review suggests that there may have been departures from standard practice for systematic reviews as we detail below. These departures from standard practice appear to have contributed to errors across all sections of Öst's review, and to a variety of conclusions that seem to be objectively unjustified in light of the evidence.

In preparing this response, we first asked all lead study authors to comment on their own studies. We then checked the original papers to verify and confirm possible errors in Öst's (2014) analysis. In most cases the author claims were included in this response. The authors reported errors for 48/60 (80%) of the studies. There were 50 errors in Table 1 alone (6.4% of the

total figures reported; see Appendix A) which summarized the methodological specifics of the studies. These were all errors of fact, not interpretation. We have only included errors where the correct facts were reported in the original paper: statements that were shown to be incorrect by additional information that was not in the original manuscript, were counted as being accurately reported. While many of these errors might seem minor if they were just reported in Table 1, the majority of them were against ACT and it seems likely that these errors were also reflected in his meta-analysis and estimates of effect sizes. For example, Öst claimed there was no follow-up data for five studies that in fact did report follow-up data. Presumably, this also meant that incorrect figures were used in the effect size calculations for the meta-analysis (we will explain below why we are using the word "presumably"). The situation appears to be worse for the more interpretive sections of the review such as Tables 11 and 12 where we estimate approximately 12% of the reported figures are incorrect. In this area, we found that all of the errors of interpretation were against ACT.

The present article argues that the pattern and magnitude of errors are serious enough that both the content of Öst's (2014) review and the process used to create it should now be set aside in making decisions regarding the treatment efficacy of ACT and in planning further examinations of this literature. The present paper will also briefly discuss the issues surrounding the development of useful criteria for assessing quality of research across different psychotherapeutic traditions, and will note additional criteria that we believe have been minimized or left out. Finally, we will summarize briefly the current state of the evidence for three disorders that have been most intensively studied.

Providing evidence of error is inherently very detailed work. While we will try to be succinct, in order to evaluate the correctness of our conclusions the reader will need to tolerate

exposure to details that are important primarily when viewed as an overall pattern. Our intention here is to provide sufficient evidence of the problems so that readers can make their own scientific judgment of the 2014 review and so that future recommendations can be made.

Öst's 2014 Review

Öst's (2014) review consisted of four parts: a) selection of studies, b) evaluation of methodological quality of studies, c) a meta-analysis and d) a subjective evaluation of the degree of research evidence for ACT overall and for particular conditions. It is important to be clear on the difference between parts b and d. In part (b), Öst used 22 criteria he developed initially in his 2008 analysis to rate the quality of the studies themselves, whereas in part (d) Öst provided his personal opinions about the APA Division 12 Taskforce criteria for evaluating the quality of evidence for a treatment overall within particular problem areas.

Despite written and face-to-face requests, Öst has not provided us with the actual study by study effect size data used in his meta-analyses. Thus, we have not evaluated his meta-analysis (part c) in this paper. Öst has provided us with his ratings of methodological quality, however, and Tables 11 and 12 of his paper (2014) provide nearly complete data in for his conclusions regarding the strength of research evidence for specific disorders. Thus, our focus will be on the areas where we have the data needed for a careful examination of the paper: parts a, b and d.

Re-examining Öst (2014)

Part a): Selection of Studies

Öst (2014) clearly describes his criteria for inclusion of studies. We have concerns in a few areas. Unlike A-Tjak et al. (2015), Öst (2014) included studies with fewer than 10 participants per cell in the design. Larger studies tend to have smaller effect sizes and higher quality ratings (Barth et al., 2013). An examination of smaller studies can make sense if there is a detailed theoretical attempt to explore innovations, to include research from developing nations or from students, or detect patterns that might be relevant to future research (scientifically, this is primarily why small pilot controlled trials are of interest). If the intent is merely to summarize effects sizes, however, the decision to examine tiny studies necessarily increases variability, and reduces methodological quality.

Second, Öst specifically excluded studies that explored components within ACT, and failed to examine process or mediational evidence in the randomized trials that were included. The effect is that ACT is treated more as a single protocol than as a theoretical model of how to assemble various treatment components linked to the model to address specific problems. Many clinical researchers have argued that examining components and processes of charge are key to advancing evidence-based therapy (Borkovec & Sibrava, 2005; David & Montgomery, 2011; Herbert & Gaudiano, 2005; Kazdin, 2008; Lohr, 2011) and have pointed with concern to weaknesses in these areas in mainstream CBT (Gaudiano, 2005; Kazdin, 2007; Longmore & Worrell, 2007).

Such a focus on mediators and their transdiagnostic application is one of the strengths of the ACT literature, and when specific methodological and strategic approaches are clear and

central in a wing of evidence-based practice and research, they arguably should be considered (S. C. Hayes, 2008). Even the earliest ACT studies attempted to provide data on the treatment's possible mechanisms of action, enabling the first meta-analysis to consider these matters (S. C. Hayes, Luoma, Bond, Masuda, & Lillis, 2006). The same is true of studies targeting treatment elements. Levin et al. (2012) conducted a meta-analysis of 66 laboratory-based component studies targeting specific processes within the ACT model and found broad evidence in support of the psychological flexibility model that undergirds ACT. Systematic reviews should at least mention mediational and component studies that point to the specific processes involved in the intervention, rather than forcing a "protocol for syndromes" model on the field as the very time it is clearly moving away from that very model (S. C. Hayes & Hofmann, in press). For example, A-Tjak et al. (2015) made good use of the process evidence without significantly expanding the size of the review.

Part b): Ratings of Methodological Quality

In his original meta-analysis, Öst (2008) developed a list of criteria against which he believes randomized controlled trials (RCTs) should be evaluated for quality. His rating scale adapted and extended a rating scheme for studies of PTSD presented as a conference poster by Tolin (1999) and included many more criteria than the APA Division 12 Task Force criteria he used to assess the overall degree of research evidence (see part d) below. Öst's 22 criteria for a well-conducted study provide some useful suggestions for ways in which the research community might update the APA Division 12 criteria, but reaching consensus on such things is a communitarian issue that needs to be addressed by scientific organizations, not individual researchers, and the same agreed upon standards need to be applied to all. Presently, the use of Öst's list presents an analytic challenge because: a) some of the operational definitions of given

features are controversial, b) evidence-based therapy is moving toward a more process-based and trans-diagnostic approach which is given short shrift in the list, c) key methodological issues are left off the list, d) over the last eight years this list appears to have been applied to ACT studies and little else, and e) it is unclear whether the scale is reliable and valid.

Fortunately, the A-Tjak et al. (2015) meta-analysis applied the same 22-item scale to the studies they examined. This afforded a unique opportunity to examine the ratings of the two teams. While the two studies used different inclusion criteria, there were 36 studies that were contained in both reviews. Both A-Tjak et al. (2015) and Öst (2014) provided their detailed ratings of methodological criteria to our research team. First, we will consider the magnitude of the ratings provided for the two studies, and then we consider their reliability.

Differences between the two teams. Using paired t-tests category-by-category, the final ratings for the two studies for the list of 36 studies that overlapped were compared. There were significant differences between the two reviews for 8 out of the 22 criteria (36%). In every case, this occurred because Öst (2014) rated ACT studies as poorer methodologically than did A-Tjak et al. (2015) (see Table 1). The overall total was also examined but since Criterion 22 (equality of therapy hours) included a large number of "not applicable" ratings (e.g., if therapy was delivered via bibliotherapy), it was excluded from the total. Considering the overall total of the remaining criteria (1 to 21), Öst's (2014) ratings were 10% lower overall on average than those from A-Tjak et al. ($t(36) = -7.17$, $p < .001$, $\eta^2 = .60$). This appears to explain why Öst (2014) reported that methodological quality had not improved while A-Tjak et al. (2015, p. 34) reported that it had improved from the 2008 analysis.

[TABLE 1 ABOUT HERE]

Why might these differences have occurred? Morina, A-Tjak, and Emmelkamp (2015, p. 252) provide a possible explanation; "Öst was the only rater of the methodological quality of the included studies, whereas our rating was conducted by two of the authors where disagreements were resolved by consensus among four of the authors, a procedure that might decrease potential biases". A-Tjak et al. (2015) used two independent raters for all evaluated studies: Jasper Smits, a highly experienced CBT researcher, and an Associate Editor of the *Journal of Consulting and Clinical Psychology*, who has not to our knowledge previously published work on Acceptance and Commitment Therapy; and Michelle Davis, an advanced graduate student in clinical psychology. Disagreements among their ratings occurred for only 26 of 792 ratings (3%), but when they did the team of investigators who reached consensus included both experienced ACT researchers (e.g., Jacqueline A-Tjak) and well-established researchers who had not done ACT research and had written critical but balanced pieces about ACT in the past (Mark Powers; Paul Emmelkamp).

Öst (2014, p. 106) did check on the reliability of his ratings but reported that "advanced graduate students in clinical psychology received 6 h of training in the use of the scale by the author with various outcomes studies as training examples. Then the students rated a random selection of 20% of the studies." This is a vaguely described and unusual method. It is not clear how many students were involved, whether they all rated the same 20% of the studies, and which outcome studies were used as examples. Furthermore, it is not clear whether the student rating process was part of a graded course and thus subject to demand characteristics (e.g., did the students get grades for agreement). No procedures were reported (e.g., blinding) to ensure the independence of those conducting reliability checks. Because there was no adjustment of ratings

if disagreements were found, all of the ratings reported in the meta-analysis were done by a single scholar.

Reliability of the ratings. We conducted detailed analyses of the reliabilities of raters. For the A-Tjak et al. (2015) review, two raters independently rated all the studies (Smits and Davis). Across all categories the two raters averaged a kappa of .93, which is considered excellent. Twenty-one of the 22 categories had kappas greater than .6, which is substantial or better according to the Landis and Koch (1977) cut-off guidelines. Only one category had moderate agreement that fell below that cut-off.

By contrast, comparing Öst's ratings with the overall ratings published by A-Tjak et al. (2015) or the ratings done by the two individual raters for that study, we found average kappa's of between .35 to .36, which is considered below the cut-off allowing reliable interpretation of data. Only one of the 22 specific categories reached the level of substantial agreement by the Landis and Koch criteria. A chi-square analysis comparing the number of categories reaching substantial or better agreement within the A-Tjak review, as compared to such agreement between Öst and the A-Tjak et al. team, showed that Öst's ratings differed significantly with a very large effect size in the direction of negative ratings against ACT on the part of Öst ($\chi^2(1) = 32.8$, $p < .0001$; $d = 3.43$).

Because of the diversity of the team, the amount of checking, the steps taken to avoid and resolve bias or disagreement, and the high resulting kappas, the A-Tjak et al. (2015) ratings appear to be of proven reliability, scientifically speaking. In contrast, because of the use of a single critic as rater, the small amount of checking, the poor controls over possible bias or lack of independence, the lack of specification of procedures, and the low resulting kappas, Öst ratings do not meet scientific standards for their use.

Part d): Judgments of Quality of Evidence

Here we discuss Öst's judgments regarding quality of evidence for specific disorders. According to the Society of Clinical Psychology, Division 12 of the American Psychological Association, ACT has strong research support for chronic and persistent pain in general, and modest research support for depression, psychotic symptoms, obsessive-compulsive disorder, and mixed anxiety at the time these reviews were last conducted. Öst (2014) disagreed with all of these classifications, arguing that each should be downgraded by one level, although Öst added that ACT also appears to be probably efficacious in tinnitus, and possibly efficacious for drug abuse and stress at work, areas that APA has not yet specifically addressed.

In defending his judgments, Öst (2014, p. 118) argued that "as a BT- and CBT-researcher of more than 40 years I should be allowed to provide my well-founded opinion." While we agree that everybody is entitled to an opinion, we do not agree that broad categorical decisions about whether a treatment is evidence-based according to organizationally established standards should be done by individuals, even experienced ones, or reported in major scientific journals as a substitute for existing impartial review processes established by a disciplinary community.

In judging the degree of empirical support, Öst (2014) made two sets of quality ratings in his article. The first set was based upon his idiosyncratic set of 22 criteria and has already been discussed in the section on ratings of methodological quality. The second set was used to establish the evidence base status of ACT. These ratings made use of the more standard empirically based treatment (EBT) criteria originally developed by the APA Division 12 Task Force (Chambless & Hollon, 1998; Silverman & Hinshaw, 2008). This part of the paper is in some ways the most important in terms of practical implications. Organizations and institutions responsible for funding research and policy regarding mental health are influenced by summary

judgments such as the one Öst provided: “ACT is not yet well-established for any disorder” (2014, p. 105). A Google search for this exact phrase found 205 citations at the time of writing this paper. We have received anecdotal reports of reviewers on research grant applications citing this conclusion by Öst’s review as part of the reason for rejecting funding proposals for future research on ACT. Therefore, it is particularly important that the data used for these ratings be accurate.

[TABLE 2 ABOUT HERE]

Öst (2014) published his detailed ratings in Tables 11 and 12, so it is possible to evaluate them in detail. In this section, we review the studies in detail each of the clinical problem areas listed by Öst, so that readers can determine for themselves whether the Öst review speaks accurately to the state of the evidence in these areas.

Psychiatric Disorders

Depression

For Petersen and Zettle (2009), Öst erroneously compared depression outcomes at discharge between ACT and treatment as usual (TAU) arms and concluded there was no difference. In this study participants were not discharged by the medical staff until “they were deemed to no longer constitute a danger to self or others due to psychiatric and/or substance-related issues” (Petersen & Zettle, 2009, p. 528), and thus depression outcomes were, by design, similar at that time. The main outcome variable was *time-to-discharge* as the authors clearly

stated in their paper. When the correct outcome measure is used, Petersen and Zettle (2009) did indeed show that ACT was more effective than TAU.

Öst incorrectly reported that L. Hayes, Boyd, and Sewell (2011) used the web-based Development and Well-Being Assessment (DAWBA) as the diagnostic method, when their paper reported that they used the clinician rated DAWBA from clinical interviews for recruitment of participants into the study. Öst's criticism of the DAWBA is at odds with psychometric studies showing clinician-rated DAWBA diagnosis generally has good reliability and validity with 92% sensitivity in clinical samples (Goodman, Ford, Richards, Gatward, & Meltzer, 2000) and 90% agreement for depression between clinician only diagnosis and DAWBA only diagnosis (Aebi et al., 2012). Furthermore, while Öst correctly noted that the TAU group received 5 hours less treatment, he failed to note that the mean hours of treatment was not significantly different between ACT and TAU groups (t=1.41), or that TAU also had higher unexplained attrition rates (31.2% in TAU and 13.6% in ACT) whereas most previous studies have shown the opposite pattern of higher attrition in psychotherapy groups (Watanabe, Churchill, Hunot, & Furukawa, 2004). Keeping adolescents in psychiatric outpatient treatment is a critical issue.

Öst incorrectly classified White et al. (2011) as a study investigating the efficacy of ACT on active psychotic symptoms. As the title indicated, that study was a trial of ACT for emotional dysfunction (levels of depression and anxiety) following a psychotic episode, *not* a study of acute psychosis. As the exclusion criteria stated, participants were excluded from the study if they had high levels of current psychotic symptoms. As a result of this misinterpretation of the main point of the study, Öst incorrectly used a measure of psychosis symptoms, the Positive and Negative Syndrome Scale (PANSS), as the primary outcome measure in his meta-analysis and

incorrectly reported in Table 11 no significant difference to the comparison condition. In fact, for depression (the targeted outcome), ACT outperformed TAU for this study.

Given the errors found in Öst's analysis of the aforementioned studies, we believe that the APA Division 12's finding that ACT is (at least) currently probably efficacious for the treatment of depression according to the Silverman and Hinshaw (2008) criteria remains the more reliable and justifiable conclusion. There have now been several additional and largely supportive studies of depression published that might change that categorization in future APA review processes (e.g. Dindo, Marchman, Gindes, & Fiedorowicz, 2015; Lappalainen, Langrial, Oinas-Kukkonen, Tolvanen, & Lappalainen, 2015; Pots et al., 2016; Thekiso et al., 2015).

Psychotic symptoms

For the study conducted by Gaudiano and Herbert (2006), Öst (2014) incorrectly reported no significant differences from TAU. Gaudiano and Herbert (2006, pp. 427-428) reported a significant difference between groups for distress related to hallucinations, which was one of the specified psychotic symptom outcomes analyzed. While there were indeed non-significant effects for some other outcome variables, distress related to hallucinations was a key outcome for the study and is of great relevance to clinicians. Second, Öst (2014) suggested in Table 11 that this study failed criteria 3 and 4 for "inclusion criteria" and "reliable and valid outcome measures" respectively. Both of these ratings are inaccurate. The inclusion and exclusion criteria were clearly specified in detail by Gaudiano and Herbert (2006, p. 419). DSM diagnosis was assessed by the attending psychiatrist. However, the presence of psychotic symptoms was the defined population in this study, and this was verified at baseline by a reliable/valid measure called the Brief Psychiatric Rating Scale (BPRS; Overall & Gorham, 1962). High inter-rater reliability was reported for this measure (ICC = .90). In addition to the BPRS, they also used

other standardized and reliable/valid outcome measures, including the Sheehan Disability Scale (SDS) and Clinical Global Impressions Scale (CGI), and finally they used re-hospitalization status based on insurance records as an objective outcome. These errors in Öst's analysis meant that he omitted results from psychoses outcomes in the calculation of effect sizes in his meta-analysis. In addition, Gaudiano and Herbert found significant differences on the SDS and CGI measures, as well as clinically significant improvements on the BPRS. The article by (Gaudiano, 2009b) reported to Öst that some of these coding errors were in the original 2008 meta-analysis, but inexplicably these same errors re-appeared in Öst's revised 2014 meta-analysis.

Shawyer et al. (2012, p. 112) clearly indicated the inclusion criteria for their study as follows: "having a diagnosis of schizophrenia or related condition based on DSM-IV criteria, aged between 18 and 65 years and having experienced command hallucinations within the previous 6 months that caused distress or dysfunction despite treatment with antipsychotic medication at therapeutic doses." Öst claimed this study did not demonstrate reliable inclusion criteria but it is difficult to imagine what more could have been done other than confirming the diagnosis and presence of the targeted psychotic symptom—command hallucinations—that was the focus of the treatment. Again, after correcting errors in Öst's coding of these studies, we believe that the APA Division 12's conclusion that ACT is currently probably efficacious for the treatment of psychotic symptoms, according to the Silverman and Hinshaw (2008) criteria, is most scientifically justifiable at this time.

Anxiety Disorders

In this category, Öst inappropriately included a study by Wetherell, Afari, Ayers, et al. (2011) that investigated whether ACT could be applied to Generalized Anxiety Disorder in older adults. This study was not an RCT. CBT was not mentioned anywhere in the title nor in the

abstract; and the paper did not conduct any comparisons of the effectiveness of ACT and CBT. For these reasons alone, based on Öst's own criteria for inclusion, this paper should have been excluded from the analysis

The Arch et al. (2012) study provides an interesting illustration of how Öst's exclusive focus upon DSM diagnosis distorts his conclusions. Öst (2014, p. 113) writes: "Finally, there is one study (Arch et al., 2012) on mixed anxiety (panic disorder, GAD, SAD, OCD and specific phobias). The study found no significant difference between ACT and CBT … My evaluation is possibly efficacious which disagrees with Division 12 saying modest research support. Mixed anxiety is not a diagnosis and this study cannot be used as evidence for ACT being efficacious across the five anxiety disorders included in the study." However, at 12-month follow-up, ACT did show significantly lower clinical severity ratings than CBT among completers using blind clinical interviews. Further, the study did not claim to assess ACT and CBT's efficacy in treating individual anxiety disorders. Rather it assessed ACT as a trans-diagnostic treatment *across* the anxiety disorders. In the introduction, Arch et al. (2012, p. 751) wrote: "The current study compares ACT and CBT in a mixed anxiety disorder sample for two reasons. First, ACT (Hayes et al., 1999) originally was developed for the treatment of psychopathology in general rather than a specific disorder in particular. The ACT protocol used in the current study (Eifert & Forsyth, 2005) was designed for application across all of the anxiety disorders, with the content of values-guided behavioral exercises tailored to specific anxiety disorders."

For the study by England et al. (2012), Öst claimed that the study did not use "reliable and valid outcome measures." However, the authors used the structured clinical interview for DSM-IV (SCID) conducted by blinded raters with established interrater reliability to determine diagnostic status as their primary outcome. Öst's overall conclusion for social anxiety disorder

was that ACT is "possibly efficacious" (2014, p.113). Öst agreed that the (Kocovski, Fleming, Hawley, Huta, & Antony, 2013) study was of high quality. When combined with the correct ratings for the (England et al., 2012) study, we disagree with Öst and believe this conclusion should be changed to "probably efficacious" for social anxiety disorder, according to the Silverman and Hinshaw (2008) criteria.

Drug Abuse

Öst reported that the outcome measures in Smout et al. (2010) were not reliable and valid, but the self-report instruments used were accompanied by hair analysis, which is an objective, valid, and reliable outcome measure for methamphetamine use (Han et al., 2015). Also while Öst noted the high attrition for ACT in this study as “astonishing”, he failed to mention that the attrition, which is characteristic of this population, did not differ between the ACT and CBT conditions. Finally, all participants in the study “met DSM-IV criteria for methamphetamine abuse or dependence according to the Mini-International Psychiatric Interview (MINI) substance use module” along with other clear and replicable inclusion criteria.

Öst claimed that Stotts et al. (2012) did not reliably demonstrate inclusion criteria. But they used both the Structured Clinical Interview for the DSM (SCID) to diagnose opioid dependence (which meets Öst’s own criteria) and also an independent psychiatrist’s evaluation. Second, Öst’s method of calculating the effect size was incorrect since the method he applied $[(M_{ACT} - M_{comparison})/SD_{pooled}]$ is appropriate for a continuous outcome whereas this trial used a dichotomous outcome.

Nicotine Dependence

Öst reported the overall outcome of the Bricker, Wyszynski, Comstock, and Heffner (2013) study as non-significant, whereas the study found a significant difference between ACT and Smokefree.gov. Furthermore, contrary to Öst's Table 11, the inclusion criteria were specified clearly in a reliable, valid manner that was fully consistent with web-based smoking cessation intervention trials included in the Cochrane review (Civljak, Stead, Hartmann-Boyce, Sheikh, & Car, 2013). Both Gifford et al. (2004) and Gifford et al. (2011) used comparably reliable and valid inclusion criteria to that of Bricker et al. (2013) and thus we believe they also satisfied criterion 3.

Öst claimed that the Bricker et al. (2013) study did not meet criterion 4 (use of a reliable and valid outcome measure). Self-reported smoking is a standard method for assessing web-based interventions and is fully consistent with web-based smoking cessation intervention trials included in Cochrane reviews (Civljak et al., 2013). False reporting is minimal for low-intensity interventions with no face-to-face contact (e.g. Patrick et al., 1994). Due to cost and low demand characteristics for false reporting, the leading scientific body on tobacco research, the Society for Research on Nicotine & Tobacco, recommends against biochemical confirmation because it has low response rates and is unnecessary in population-based studies with limited face-to-face contact or in studies where the optimal data collection methods are through the mail or telephone (Benowitz, Pomerleau, Pomerleau, & Jacob, 2003). We would not necessarily expect Öst to know all this as it is presumably outside his area of expertise, but this omission again highlights the perils of working alone rather than in a group to make judgments regarding research quality.

Borderline Personality Disorder

Öst (2014, p. 113) argued that ACT should be rated as "experimental" for BPD because "Both studies gave the TAU-treated subjects markedly less therapy hours (see Table 6) and did not fulfill criterion 3." Both of these claims are mistaken with respect to the study by Gratz and Gunderson (2006). First, it was clearly stated in their article that "the average number of hours spent in therapy per week did not differ significantly between groups … (treatment = 3.60, waitlist = 2.95, $t < 1.00$, $p > .10$)" (Gratz & Gunderson, 2006, p. 29). Second, the study extensively specified the list of inclusion and exclusion criteria, including "meeting five or more criteria for BPD and receiving a score of 8 or higher on the Revised Diagnostic Interview for Borderlines" (Gratz & Gunderson, 2006, p. 27). Interestingly, Öst's earlier (2008) study acknowledged that this study met criteria 3, so his analysis here contradicted his earlier assessment of the same study.

Similarly, Morton, Snowdon, Gopold, and Guymer (2012) used the SCID with well-trained research assistants to assess BPD criteria. The diagnosis was also further verified with the referring clinicians, the SCID and a clinical interview with one of the researchers. Öst criticized the study for setting a benchmark of 4/9 criteria for BPD instead of the five required by DSM. On this basis, Öst claimed that "inclusion criteria were not reliably demonstrated." This assessment is entirely inappropriate. The study did not purport to be a study of treatment for BPD as assessed by the DSM, but of treatment for people with four or more criterion symptoms of BPD. The stated focus of this criterion is reliably demonstrating inclusion criteria, not conformity to DSM categories per se. Furthermore, Öst fails to mention that only three of the ACT group and two of the TAU clients met *less* than the full five criteria, and that the average number of BPD criteria met was 6.0 for the ACT group (SD 1.34) and 6.5 for the TAU group

(SD 1.64). While Öst argues the evidence is experimental regarding ACT as a treatment for BPD, we would see it as possibly efficacious on the basis of the Gratz and Gunderson (2006) study.

Somatic Studies and Stress

Pain

Öst (2014) criticized Dahl, Wilson, and Nilsson (2004) on both inclusion criteria and outcome measures, arguing that they had not used a 'structured diagnostic interview,' which is something that does not exist in the area of chronic pain. The inclusion criteria were extremely well specified and clearly replicable. The study was a prevention study focused on people at demonstrably high risk for sick leave utilization. In terms of outcome measures, sick leave was the primary dependent measure and it is difficult to understand how number of sick leave days is not a valid measure of sick leave. Similarly, secondary measures included number of medical visits and a well-validated measure of life satisfaction (Post, van Leeuwen, van Koppenhagen, & de Groot, 2012).

The study by Wicksell, Ahlqvist, Bring, Melin, and Olsson (2008) was also criticized by Öst for inclusion criteria. Their paper clearly specified the inclusion criteria of independently diagnosed whiplash associated disorder and pain for 3 months, together with a host of clear exclusion criteria. The study by Wicksell, Melin, Lekander, and Olsson (2009) was for children experiencing idiopathic pain with a duration greater than three months. The inclusion and exclusion criteria were again well specified and in accord with accepted standards in this area. By definition, idiopathic pain is not associated with a specific psychiatric or medical diagnosis.

In addition, Wicksell et al. (2013) targeted fibromyalgia with patients who fulfilled all of the American College of Rheumatology classification criteria for fibromyalgia (Wolfe et al., 1990) as well as standardized ratings of pain intensity together with a range of clear exclusion criteria. In all three cases, the inclusion/exclusion criteria were clear, and based on widely accepted standards of work in chronic pain. It is scientifically inappropriate to attempt to insert new and idiosyncratic "standards" that are not accepted by the field itself, in the guise of a meta-analysis, which is what occurred here.

For Thorsell et al. (2011), Öst included a question mark regarding the outcome measure. The Numeric Rating Scale that was used is currently one of the most widely used for measuring pain intensity (Ferreira-Valente, Pais-Ribeiro, & Jensen, 2011; Williamson & Hoggart, 2005). Indeed, a similar measure was used by McCracken, Sato, and Taylor (2013), for whom Öst asserted they *had* used a valid outcome measure.

Öst claimed Wetherell, Afari, Rutledge, et al. (2011) did not have reliable inclusion criteria. The study was purposefully designed to include non-malignant chronic pain from many medical sources (e.g. arthritis, fibromyalgia, etc.). Experts in pain research have agreed that the diagnosis and outcomes of chronic non-malignant pain are nonspecific, and rely heavily upon a patient's self-report in the following areas "(1) pain; (2) physical functioning; (3) emotional functioning; (4) participant ratings of improvement and satisfaction with treatment; (5) symptoms and adverse events; and (6) participant disposition" (Turk et al., 2003, pp. 337-338). This was the approach used by Wetherell, Afari, Rutledge, et al. (2011). There is no structured interview to diagnose chronic pain but Wetherell, Afari, Rutledge, et al. (2011) did use SCIDs to characterize comorbid psychiatric diagnoses. Given the high quality of these studies, we would agree with the Division 12 rating that ACT has strong research support (i.e., is well-established)

for “chronic and persistent pain in general” and disagree with Öst’s claim that the evidence is only “probably efficacious.”

Tinnitus

For Westin et al. (2011) Öst criticized the inclusion criteria, but the diagnosis of tinnitus was established using a standardized diagnostic interview under the supervision of an ear-nose-throat physician.

Overweight/Obesity

Öst argued that Lillis, Hayes, Bunting, and Masuda (2009) did not make use of reliable, valid inclusion criteria. There is no SCID diagnosis for being overweight, and the behavioral inclusion criteria were clearly specified in a manner that could easily be replicated. Similarly, clear and easily replicable inclusion and exclusion criteria were listed by Forman, Hoffman, Juarascio, Butryn, and Herbert (2013).

Stress

For Bond and Bunce (2000), Öst excluded the published findings that at post-treatment and follow-up, stress levels (as measured by the General Health Questionnaire) were better for ACT compared with the behavior therapy intervention. This same error occurred in Öst (2008) and again likely contributed to incorrect effect size data in the meta-analysis.

In this section on stress, Öst (2014, p. 110) states “0/7 stress studies diagnosed the participants”, but given that there is no DSM diagnosis available for “stress at work”, this criticism is clearly not justified. As with almost all the other studies that Öst criticized on this criterion, the stress studies listed inclusion and exclusion criteria appropriate for the populations of interest such as being professional staff in defined roles. If such studies are to be included

(along with studies on pain, test anxiety, and other issues which do not have DSM-based diagnoses available), then it is inappropriate to criticize them for lacking something that neither exists nor would be appropriate for the populations of interest.

Summary and Additional Concerns

We have focused this article so far on the factual errors made by Öst. We have not attempted to list the many selective interpretations of data that simply leave out relevant information. In some studies, Öst chose to focus upon the outcome variable that did not change, ignoring clinically crucial outcomes that did improve significantly (Gaudiano & Herbert, 2006; McCracken et al., 2013; Wicksell et al., 2009). In others, he chose to ignore evidence regarding significant reductions in, for example, believability of thoughts indicative of burnout (Bethay, Wilson, Schnetzer, Nassar, & Bordieri, 2013). In still others, he neglected to report critical details. For example, in (Wicksell et al., 2009) Öst reported no significant differences at follow-up, but fails to note that the comparison condition, a state of the art multi-disciplinary approach, continued during the entire follow-up period while the ACT condition did not. Even so, ACT performed significantly better than the control condition on perceived functional ability in relation to pain, pain intensity, and pain related discomfort (intent-to-treat analyses). At post-treatment, before ACT was discontinued, significant differences in favor of the ACT condition were also seen in fear of re/injury or kinesiophobia, pain interference and in quality of life. These kinds of differences almost certainly also impacted the effect size estimates, but as we noted above, we were unable to evaluate that concern because Öst failed to provide us access to the data.

Overall, the severity of the errors and their consistent direction raises significant questions about Öst's (2014) conclusions and the degree to which it was unbiased. Systematic

reviews and meta-analyses rightly carry weight with the public, scientists, funding agencies, and public health institutions. Inaccurate reviews can deprive patients of appropriate care, and distort scientific progress. Thus, it is worth considering how to avoid situations such as those documented here.

The Disciplinary Nature of Methodological Quality Standards

We support the development of standards of desirable methodological quality and efforts to summarize the literature in order to make policy recommendations. In our view, however, this needs to be done as a collaborative activity by the discipline itself. Efforts such as the APA Division 12 EBT list or SAMHSA's NREPP program have well specified and collaboratively agreed upon criteria for evaluating research quality and the extent of empirical support.

Meta-analyses should also rely upon diverse groups of scholars rather than a single individual. No one individual is likely to know enough about all of the many areas in a broad review. It is difficult for a single individual to establish criteria for inclusion that are theoretically neutral, or to apply them accurately and without personal bias. Ironically such a process runs the very risk of unreliability that Öst was so critical about in his 2014 review.

In our view, Öst's approach to rating study quality was unjustifiably saturated with dependence upon the DSM. His criterion 2 refers to "severity/chronicity of the disorder" and his criterion 4 refers to "reliability of the diagnosis in question." By "disorder" and "diagnosis" what is meant and scored by Öst is the use of syndromal diagnosis. In his 2008 review, Öst seemed puzzled by the lack of interest in syndromes among the ACT community: "The descriptive review showed that only half of the ACT studies diagnosed their participants, whereas this was done in all DBT, CBASP, and CBT studies. This is difficult to understand, since there does not

seem to be an ideological resistance to diagnosing among ACT researchers." (p. 312). It should not have been difficult to understand because there is indeed a very long-standing ideological resistance to syndromal diagnosis among behavior analysts and contextual behavioral scientists in favor of a more functional and process-oriented approach (Hayes et al., 1999). In his response to Öst, Gaudiano (2009b, p. 4) noted that "whether or not the sample is defined in terms of the DSM is largely irrelevant to the issues of appropriately describing the sample." Major funders of psychotherapy research such as the National Institute of Mental Health now are also taking a process-focused approach, and no longer encourage definition of samples primarily in terms of the DSM (Insel et al., 2010). In that context, a consensus process by the field itself would be highly unlikely to agree with Öst that high quality definitions of samples require the use of syndromal diagnosis.

We have so far identified 41 incorrect ratings by Öst in his Tables 11 and 12 (see Appendix A). Of these errors, the main area of disagreement concerns what we believe to be an inappropriate and selective interpretation of standards for inclusion criteria. Although it is unclear in the paper what Öst (2014) means by the heading "Inclusion criteria reliably delineated" in Tables 11 and 12, earlier in the paper in reference to Criterion 4 of his 22 personal assessment criteria he states: "In order for ACT-studies to be compared to other therapies regarding the evidence-base it is important that participants are diagnosed, preferably by employing trained interviewers using established interview schedules (or similar instruments) and assessing inter-rater reliability. Looking at the first issue we find that 23 out of 31 (74%) studies of psychiatric disorders, 13/22 (59%) studies of somatic disorders, and 0/7 stress studies diagnosed the participants" (Öst, 2014, p. 110). We have already shown that many of the studies Öst gave a "-" (failed) rating to in Tables 11 and 12, did in fact make use of standardized DSM-

based interviews. If we apply the standard Silverman and Hinshaw (2008, p. 5) definition of "conducted with a population, treated for specified problems, for whom inclusion criteria have been delineated in a reliable, valid manner", at least 26 more studies (43% of the entire sample, Appendix B) would meet this criterion, almost entirely accounting for why Öst's views were so discrepant from the APA Division 12.

It is possible that Öst gave a "-" rating to any study not reporting checks on inter-rater reliability of DSM based interviews, but when we contacted Evan Forman (current Division 12 EBT list Editor) and David Klonsky (former Editor) both stated that they knew of no specific requirement for EBT reviewers to require either that the population be defined by DSM focused diagnostic interviews or that interrater reliability be reported when diagnostic interviews were used. They also noted that not all of the conditions listed on the current EBT website are DSM disorders and that if the researchers used methods that had previously been determined to be reliable and valid methods for defining the sample, this would typically be considered acceptable. Öst (2014) may have not only chosen the narrowest possible definition of good inclusion criteria, but he also applied it to studies (e.g., chronic pain; work stress) that could not possibly satisfy the criterion.

Were a measure of study quality to be created by the ACT scientific community, it would include items on whether process measures were taken and analyzed; whether mediators were assessed; or whether basic science studies were linked to the intervention and its analysis. The comparison of ACT and CBT studies in the Öst (2008) meta-analysis showed these differences clearly: ACT studies generally referred to basic studies and to behavioral principles — that was rare in the CBT studies; the great majority of the ACT studies reported formal mediational results (either in the target article or in later publications linked to the same data set) but none of

the CBT studies did so (Gaudiano, 2009b). Especially as the field shifts away from a purely syndromal approach toward a more trans-diagnostic and process-focused approach (e.g., Insel et al., 2010; Hayes & Hofmann, in press) it seems important for meta-analyses to consider whether a program of research has shown theoretically consistent process evidence.

The field itself needs to decide on such matters, especially as they bear on recommendations by funders and policy makers. It is nearly impossible to avoid bias when a single individual is allowed to define the quality of research, to assess whether research meets those criteria, and to give guidance to governments and agencies about the empirical status of specific applied approaches. Our analysis demonstrates repeated and significant misinterpretations and errors. Reviews of this magnitude are simply too large and complex for a single person to conduct alone no matter how many years of experience they have had. Furthermore, our analysis shows that these errors were systematic. Theoretically diverse multi-person teams, and transparent, collaborative methods of resolving inconsistencies, are necessary for the credibility and accuracy of meta-analytic reports.

Finally, we suggest that journals routinely require those conducting meta-analyses to make their data available in a depository for review by independent scholars as part of the publication process. There are simply too many decisions hidden in the bowels of meta-analytic studies, and major errors can easily go undetected if the data are not made freely available. All rating methods and data need to be 100% reproducible and journals and funding agencies need to make it as easy and affordable as possible for authors to deposit this information in an accessible format.

Recent Evidence

The larger issue underlying the various meta-analyses of ACT is whether it is an evidence-based treatment. We wish to end this article with a brief look at the evidence since Öst's (2014) review. The total amount of good quality research has continued to accumulate. Nine meta-analyses that have appeared since Öst's review, and there are now at least 171 RCT's of ACT (https://contextualscience.org/state_of_the_act_evidence). Increasingly, due to the body of evidence available, meta-analyses are being published in specific areas, as is called for in the new Division 12 standards for evidence-based procedures (Tolin, McKay, Forman, Klonsky, & Thombs, 2015).

These meta-analyses make it clear that in a number of areas, 1) ACT attains better outcomes than wait lists or treatment as usual, 2) ACT is overall at least as good as traditional CBT and other evidence-based methods, and 3) the effects of ACT are at times moderated by different factors than traditional CBT or other evidence-based methods (and vice versa). If these three conclusions are correct it means that ACT now has a place in the range of options to be deployed by evidence-based practitioners.

In addition, nearly 50 mediational analyses are currently available on ACT interventions (https://contextualscience.org/state_of_the_act_evidence) along with an increasing number of studies of treatment moderation. The available evidence suggests that, 4) theoretically coherent ACT processes commonly mediate ACT outcomes and 5) ACT consistent change processes are at times distinct (Niles et al., 2014), even at the level of neurobiological responding (e.g. Burklund, Torre, Lieberman, Taylor, & Craske, 2017). The evidence on change processes (points 4 and 5) is also quite large, so much so that studies that combined several studies in the examination of treatment moderation are beginning to appear (e.g. Niles, Wolitzky-Taylor, Arch,

& Craske, 2017). Thus, the Psychological Flexibility model that underlies ACT seems likely to be of importance to the theoretical development of the field for some time going forward.

Three areas in which these five points can be readily made are in chronic pain, substance use (including smoking), and anxiety disorders. All have been subjected to meta-analyses since Öst's review (Chronic Pain: Veehof, Trompetter, Bohlmeijer, and Schreurs (2016); Substance use: Lee et al. (2015); Anxiety: Bluett, Homan, Morrison, Levin, and Twohig (2014); Hacker et al. (2016)).

Chronic pain. There have been 8 RCTs of ACT for chronic pain since Öst's (2014) review, including two with strong comparison conditions of applied relaxation (Kemani et al., 2015) and pregabalin (Luciano et al., 2014). Both Luciano and colleagues and Trompetter, Bohlmeijer, Veehof, and Schreurs (2015) included wait-list control groups in addition to a comparison treatment to strengthen potential to make efficacy claims. These studies have found that a higher proportion of those who received ACT as compared to other treatments achieved clinically significant reductions in functional disability due to pain (Veehof et al., 2016), and in a more cost-effective way (Kemani, Hesser, Olsson, Lekander, & Wicksell, 2016). Psychological flexibility preceded and mediated reductions in pain disability for ACT recipients but not alternative psychological therapy recipients (Kemani et al., 2016; Trompetter, Bohlmeijer, Fox, & Schreurs, 2015). Pregabalin and group ACT produced equivalent increases in pain acceptance (Luciano et al., 2014). Older adults appear to be more likely to respond to ACT and younger adults to CBT (Wetherell et al., 2016) and ACT may be more effective for those with high psychological wellbeing (Trompetter et al., 2016).

Substance use. Since Öst's review two new trials have appeared, both in the area of smoking (Bricker, Bush, Zbikowski, Mercer, & Heffner, 2014; Bricker, Mull, et al., 2014). Both

found superior outcomes for ACT. For example, In a study of 121 smokers, Bricker, Bush et al., 2014 found that ACT was superior to CBT overall, but the differential odds ratio in favor of ACT was over three times higher than the study overall among participants scoring low on acceptance of cravings at baseline (n = 57), suggesting moderation by ACT relevant processes. Across all areas of substance abuse (Lee et al., 2015), found an effect size of g=.45 (p = .003) favoring ACT at post-treatment in comparison to active treatments,

Anxiety disorders. Since Öst (2014), there have been 4 RCTs of ACT for DSM-defined anxiety disorders, 3 for OCD or illness anxiety disorder and 7 for anxious symptoms among participants recruited for an alternative primary diagnosis or problem. Eight of these have employed active comparison conditions. Studies by Craske et al. (2014) and Hancock et al. (2016) used the strongest designs with both CBT and waitlist comparisons, demonstrating both active conditions effectively reduced anxiety symptoms, with neither more effective. Faster improvements in psychological flexibility predicted reduced social anxiety symptoms in ACT but not CBT (Niles et al., 2014). Higher baseline activity in anterior cingulate regions in response to social threat cues was associated with reduced social anxiety in CBT, whereas hyperactivity in the posterior insular predicted reduced social anxiety within ACT (Burklund et al., 2017). Meta-analyses (Bluett et al., 2014; Hacker et al., 2016) have shown large differences between ACT and wait-list control groups but while effect sizes favor ACT there were no overall differences between ACT and other evidence-based active treatments. Research has increasingly identified moderators of differential response to ACT as compared to CBT, however, suggesting that this global equivalence is misleading if the goal of evidence-based care is personalized treatment. A recent multi-study multi-component analysis of the moderation of treatment drop out, for example Niles et al. (2017, p. 20) found that CBT “appears to be more acceptable to

individuals who even before treatment begins, already perceive that they can control or are motivated to maintain control of their anxiety" while ACT is more effective for who "who do not perceive having control over internal anxiety states" among other related factors (p. 21). In some areas, such as behavioral avoidance, the data are confusing (Davies, Niles, Pittig, Arch, & Craske, 2015; Mesri et al., 2017) but the growing body of moderation work suggests that traditional CBT and ACT are evidence-based approaches that benefit characteristic populations in a differential way, suggesting that both are worthy of inclusion in the armamentarium of evidence-based practitioners.

Conclusion

The Öst (2014) review departed from essential features of a high quality systematic review of psychotherapy. Its most fundamental empirical errors are the use of an idiosyncratic and unvalidated rating scheme that appears not to have been reliably applied. The review contains numerous factual and interpretive errors in the reporting of trials included in the review. In all areas we could review, quality ratings, facts, and interpretations, errors were dominantly biased against ACT trials. Given these serious flaws, in our opinion the Öst (2014) review cannot be relied upon by the field, and should no longer be used or referred to in the evaluation of the ACT research program.

We recommend that future reviews and meta-analyses utilize rating methods that are broadly accepted by the mainstream scientific community and that reporting of included trials be fact checked with the corresponding authors. The use of diverse teams of investigators containing both advocates and critics would further prevent bias, error, or ignorance from inadvertently entering stated conclusions. Full data sets should be immediately available and the

purpose of research programs should be considered fairly. Following these basic procedures will ensure the entire field of behavioral intervention science is conducted with rigor, transparency, and high integrity.

The time for meta-analyses that ask gross outcome questions about ACT in an across the board way is passing into history. In part that is because the literature is too large and the need for reviews in specific areas is much greater and in part it is because gross outcome questions are just not very important scientifically once a treatment method is reasonably well established. Especially in the context of decreasing reliance upon syndromal classification in research, and the recent turn toward process-based therapy and personalized treatment, it is becoming obvious that there are inherent limitations to estimating pooled effect sizes across diverse settings, methodologies, components of intervention, delivery methods, problems conditions, and treatment goals. The era of meta-analyses focused on an overall "horse race" question such as "is ACT better than CBT?" is over. The growing body of moderation and mediation evidence suggests that global questions of that kind are both scientifically and clinically naive. Such questions are not adequate to assess the impact of evidence-based components linked to evidence-based processes.

It is also obvious that CBT itself is changing: for example, it is now common to see CBT protocols adopting acceptance, mindfulness, and values-based methods. ACT is changing, too: for example, behavioral methods that were always part of its treatment model but were artificially put aside for political reasons (e.g., to avoid the claim that ACT outcomes are just due to known behavioral methods) are now more commonly included in ACT protocols. Outcomes in a process-based era need to be advanced by philosophically and theoretically coherent research programs that draw upon data about basic processes in multiple domains (behavioral, cognitive,

biological, social), component analyses, moderation, mediation, and frequently assessed person-specific progress over time. ACT, the psychological flexibility model, relational frame theory, and contextual behavioral science have conceptual and methodological contributions to make to evidence-based care in such an era (S. C. Hayes, 2008), as just the data collected since Öst's review makes clear. Our field needs to learn to focus on the more scientifically and clinically interesting questions, and to adopt high impact research strategies that have a chance to answer them.

References

A-Tjak, J. G. L., Davis, M. L., Morina, N., Powers, M. B., Smits, J. A. J., & Emmelkamp, P. M. G. (2015). A Meta-Analysis of the Efficacy of Acceptance and Commitment Therapy for Clinically Relevant Mental and Physical Health Problems. *Psychotherapy and Psychosomatics, 84*(1), 30-36.

Aebi, M., Kuhn, C., Metzke, C. W., Stringaris, A., Goodman, R., & Steinhausen, H. C. (2012). The use of the development and well-being assessment (DAWBA) in clinical practice: a randomized trial. *Eur Child Adolesc Psychiatry, 21*(10), 559-567. doi:10.1007/s00787-012-0293-6

Arch, J. J., Eifert, G. H., Davies, C., Vilardaga, J. C. P., Rose, R. D., & Craske, M. G. (2012). Randomized clinical trial of cognitive behavioral therapy (CBT) versus acceptance and commitment therapy (ACT) for mixed anxiety disorders. *Journal of Consulting and Clinical Psychology, 80*(5), 750-765. doi:10.1037/a0028310

Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., . . . Cuijpers, P. (2013). Comparative Efficacy of Seven Psychotherapeutic Interventions for Patients with Depression: A Network Meta-Analysis. *PLoS Med, 10*(5), e1001454. doi:10.1371/journal.pmed.1001454

Benowitz, N. L., Pomerleau, O. F., Pomerleau, C. S., & Jacob, P. (2003). Nicotine metabolite ratio as a predictor of cigarette consumption. *Nicotine & Tobacco Research, 5*(5), 621-624. doi:10.1080/1462220031000158717

Bethay, J. S., Wilson, K. G., Schnetzer, L. W., Nassar, S. L., & Bordieri, M. J. (2013). A Controlled Pilot Evaluation of Acceptance and Commitment Training for Intellectual Disability Staff. *Mindfulness, 4*(2), 113-121. doi:Doi 10.1007/S12671-012-0103-8

Bluett, E. J., Homan, K. J., Morrison, K. L., Levin, M. E., & Twohig, M. P. (2014). Acceptance and commitment therapy for anxiety and OCD spectrum disorders: An empirical review. *Journal of Anxiety Disorders, 28*(6), 612-624. doi:10.1016/j.janxdis.2014.06.008

Bond, F. W., & Bunce, D. (2000). Mediators of change in emotion-focused and problem-focused worksite stress management interventions. *J Occup Health Psychol, 5*(1), 156-163.

Borkovec, T. D., & Sibrava, N. J. (2005). Problems with the use of placebo conditions in psychotherapy research, suggested alternatives, and some strategies for the pursuit of the placebo phenomenon. *Journal of Clinical Psychology, 61*(7), 805-818. doi:10.1002/jclp.20127

Bricker, J. B., Bush, T., Zbikowski, S. M., Mercer, L. D., & Heffner, J. L. (2014). Randomized Trial of Telephone-Delivered Acceptance and Commitment Therapy Versus Cognitive Behavioral Therapy for Smoking Cessation: A Pilot Study. *Nicotine & Tobacco Research, 16*(11), 1446-1454. doi:10.1093/ntr/ntu102

Bricker, J. B., Mull, K., Kientz, J. A., Vilardaga, R. M., Mercer, L. D., Akioka, K., & Heffner, J. L. (2014). Randomized, Controlled Pilot Trial of a Smartphone App for Smoking Cessation Using Acceptance and Commitment Therapy. *Drug and Alcohol Dependence, 143*, 87-94. doi:10.1016/j.drugalcdep.2014.07.006

Bricker, J. B., Wyszynski, C., Comstock, B., & Heffner, J. L. (2013). Pilot randomized controlled trial of web-based acceptance and commitment therapy for smoking cessation. *Nicotine & Tobacco Research, 15*(10), 1756-1764. doi:10.1093/ntr/ntt056

Burklund, L. J., Torre, J. B., Lieberman, M. D., Taylor, S. E., & Craske, M. G. (2017). Neural responses to social threat and predictors of cognitive behavioral therapy and acceptance

and commitment therapy in social anxiety disorder. *Psychiatry Research: Neuroimaging, 261*, 52-64. doi:http://dx.doi.org/10.1016/j.pscychresns.2016.12.012
Chambless, D. L., & Hollon, S. D. (1998). Defining Empirically Supported Therapies. *Journal of Consulting & Clinical Psychology, 66*(1), 7-18.
Civljak, M., Stead, L. F., Hartmann-Boyce, J., Sheikh, A., & Car, J. (2013). Internet-based interventions for smoking cessation. *Cochrane Database of Systematic Reviews, 7*.
Craske, M. G., Niles, A. N., Burklund, L. J., Wolitzky-Taylor, K. B., Vilardaga, J. C. P., Arch, J. J., . . . Lieberman, M. D. (2014). Randomized controlled trial of cognitive behavioral therapy and acceptance and commitment therapy for social phobia: Outcomes and moderators. *Journal of Consulting and Clinical Psychology, 82*(6), 1034-1048. doi:10.1037/a0037212
Dahl, J., Wilson, K. G., & Nilsson, A. (2004). Acceptance and Commitment Therapy and the Treatment of Persons at Risk for Long-Term Disability Resulting From Stress and Pain Symptoms: A Preliminary Randomized Trial. *Behavior Therapy, 35*(4), 785-801.
David, D., & Montgomery, G. H. (2011). The Scientific Status of Psychotherapies: A New Evaluative Framework for Evidence-Based Psychosocial Interventions. *Clinical Psychology: Science and Practice, 18*(2), 89-99. doi:10.1111/j.1468-2850.2011.01239.x
Davies, C. D., Niles, A. N., Pittig, A., Arch, J. J., & Craske, M. G. (2015). Physiological and behavioral indices of emotion dysregulation as predictors of outcome from cognitive behavioral therapy and acceptance and commitment therapy for anxiety. *Journal of Behavior Therapy and Experimental Psychiatry, 46*, 35-43. doi:10.1016/j.jbtep.2014.08.002
Dindo, L., Marchman, J., Gindes, H., & Fiedorowicz, J. G. (2015). A brief behavioral intervention targeting mental health risk factors for vascular disease: A pilot study. *Psychotherapy and Psychosomatics, 84*(3), 183-185. doi:10.1159/000371495
England, E. L., Herbert, J. D., Forman, E. M., Rabin, S. J., Juarascio, A., & Goldstein, S. P. (2012). Acceptance-based exposure therapy for public speaking anxiety. *Journal of Contextual Behavioral Science, 1*(1–2), 66-72. doi:http://dx.doi.org/10.1016/j.jcbs.2012.07.001
Ferreira-Valente, M. A., Pais-Ribeiro, J. L., & Jensen, M. P. (2011). Validity of four pain intensity rating scales. *PAIN, 152*(10), 2399-2404. doi:10.1016/j.pain.2011.07.005
Forman, E. M., Hoffman, K. L., Juarascio, A. S., Butryn, M. L., & Herbert, J. D. (2013). Comparison of acceptance-based and standard cognitive-based coping strategies for craving sweets in overweight and obese women. *Eating Behaviors, 14*(1), 64-68. doi:http://dx.doi.org/10.1016/j.eatbeh.2012.10.016
Gaudiano, B. A. (2005). Cognitive behavior therapies for psychotic disorders: Current empirical status and future directions. *Clinical Psychology: Science and Practice, 12*(1), 33-50.
Gaudiano, B. A. (2009a). Öst's (2008) methodological comparison of clinical trials of acceptance and commitment therapy versus cognitive behavior therapy: Matching Apples with Oranges? *Behav Res Ther, 47*(12), 1066-1070.
Gaudiano, B. A. (2009b). Reinventing the wheel versus avoiding past mistakes when evaluating psychotherapy outcome research: Rejoinder to Öst (2009). Retrieved from http://www.psychotherapybrownbag.com/psychotherapy_brown_bag_a/2010/01/the-empirical-status-of-acceptance-and-commitment-therapy-act-a-conversation-amongst-two-prominent-p.html

Gaudiano, B. A., & Herbert, J. D. (2006). Acute treatment of inpatients with psychotic symptoms using Acceptance and Commitment Therapy: Pilot results. *Behav Res Ther, 44*(3), 415-437. doi:http://dx.doi.org/10.1016/j.brat.2005.02.007

Gifford, E. V., Kohlenberg, B. S., Hayes, S. C., Antonuccio, D. O., Piasecki, M. M., Rasmussen-Hall, M. L., & Palm, K. M. (2004). *Acceptance-Based Treatment for Smoking Cessation*: Behavior Therapy Vol 35(4) Fal 2004, 689-705.

Gifford, E. V., Kohlenberg, B. S., Hayes, S. C., Pierson, H. M., Piasecki, M. P., Antonuccio, D. O., & Palm, K. M. (2011). Does Acceptance and Relationship Focused Behavior Therapy Contribute to Bupropion Outcomes? A Randomized Controlled Trial of Functional Analytic Psychotherapy and Acceptance and Commitment Therapy for Smoking Cessation. *Behavior Therapy, 42*(4), 700-715. doi:http://dx.doi.org/10.1016/j.beth.2011.03.002

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The Development and Well-Being Assessment: Description and Initial Validation of an Integrated Assessment of Child and Adolescent Psychopathology. *Journal of Child Psychology and Psychiatry, 41*(5), 645-655. doi:10.1111/j.1469-7610.2000.tb02345.x

Gratz, K. L., & Gunderson, J. G. (2006). Preliminary Data on an Acceptance-Based Emotion Regulation Group Intervention for Deliberate Self-Harm Among Women With Borderline Personality Disorder. *Behavior Therapy, 37*(1), 25-35. doi:http://dx.doi.org/10.1016/j.beth.2005.03.002

Hacker, T., Stone, P., & MacBeth, A. (2016). Acceptance and commitment therapy - Do we know enough? Cumulative and sequential meta-analyses of randomized controlled trials. *Journal of Affective Disorders, 190*, 551-565. doi:10.1016/j.jad.2015.10.053

Han, E., Lee, S., In, S., Park, M., Park, Y., Cho, S., . . . Lee, H. (2015). Relationship between methamphetamine use history and segmental hair analysis findings of MA users. *Forensic Science International, 254*, 59-67. doi:10.1016/j.forsciint.2015.06.029

Hancock, K. M., Swain, J., Hainsworth, C. J., Dixon, A. L., Koo, S., & Munro, K. (2016). Acceptance and Commitment Therapy versus Cognitive Behavior Therapy for Children With Anxiety: Outcomes of a Randomized Controlled Trial. *Journal of Clinical Child & Adolescent Psychology*, 1-16. doi:10.1080/15374416.2015.1110822

Hayes, L., Boyd, C. P., & Sewell, J. (2011). Acceptance and commitment therapy for the treatment of adolescent depression: A pilot study in a psychiatric outpatient setting. *Mindfulness, 2*(2), 86-94.

Hayes, S. C. (2008). Climbing Our Hills: A Beginning Conversation on the Comparison of Acceptance and Commitment Therapy and Traditional Cognitive Behavioral Therapy. *Clinical Psychology: Science and Practice, 15*(4), 286-295. doi:10.1111/j.1468-2850.2008.00139.x

Hayes, S. C., & Hofmann, S. (Eds.). (in press). *Process-based CBT: Core clinical competencies in evidence-based treatment*. Oakland, CA: New Harbinger Publications.

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: Model, processes and outcomes. *Behav Res Ther, 44*(1), 1-25.

Herbert, J. D., & Gaudiano, B. A. (2005). Moving from empirically supported treatment lists to practice guidelines in psychotherapy: the role of the placebo concept. *Journal of Clinical Psychology, 61*(7), 893-908. doi:10.1002/jclp.20133

Hertenstein, E., & Nissen, C. (2015). Comment on 'A Meta-Analysis of the Efficacy of Acceptance and Commitment Therapy for Clinically Relevant Mental and Physical

Health Problems'. *Psychotherapy and Psychosomatics, 84*(4), 250-251. doi:10.1159/000374124
Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry, 167*(7), 748-751. doi:10.1176/appi.ajp.2010.09091379
Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology, 3*, 1-27. doi:10.1146/annurev.clinpsy.3.022806.091432
Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63*(3), 146-159. doi:10.1037/0003-066X.63.3.146
Kemani, M. K., Hesser, H., Olsson, G. L., Lekander, M., & Wicksell, R. K. (2016). Processes of change in Acceptance and Commitment Therapy and Applied Relaxation for long-standing pain. *European Journal of Pain, 20*(4), 521-531. doi:10.1002/ejp.754
Kemani, M. K., Olsson, G. L., Lekander, M., Hesser, H., Andersson, E., & Wicksell, R. K. (2015). Efficacy and Cost-effectiveness of Acceptance and Commitment Therapy and Applied Relaxation for Longstanding Pain: A Randomized Controlled Trial. *Clin J Pain, 31*(11), 1004-1016. doi:10.1097/ajp.0000000000000203
Kocovski, N. L., Fleming, J. E., Hawley, L. L., Huta, V., & Antony, M. M. (2013). Mindfulness and acceptance-based group therapy versus traditional cognitive behavioral group therapy for social anxiety disorder: a randomized controlled trial. *Behav Res Ther, 51*(12), 889-898. doi:10.1016/j.brat.2013.10.007
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.
Lappalainen, P., Langrial, S., Oinas-Kukkonen, H., Tolvanen, A., & Lappalainen, R. (2015). Web-based acceptance and commitment therapy for depressive symptoms with minimal support: a randomized controlled trial. *Behav Modif, 39*(6), 805-834. doi:10.1177/0145445515598142
Lee, E. B., An, W., Levin, M. E., & Twohig, M. P. (2015). An initial meta-analysis of Acceptance and Commitment Therapy for treating substance use disorders. *Drug and Alcohol Dependence, 155*, 1-7. doi:http://dx.doi.org/10.1016/j.drugalcdep.2015.08.004
Levin, M. E., Hildebrandt, M. J., Lillis, J., & Hayes, S. C. (2012). The Impact of Treatment Components Suggested by the Psychological Flexibility Model: A Meta-Analysis of Laboratory-Based Component Studies. *Behavior Therapy*. doi:10.1016/j.beth.2012.05.003
Lillis, J., Hayes, S. C., Bunting, K., & Masuda, A. (2009). Teaching acceptance and mindfulness to improve the lives of the obese: A preliminary test of a theoretical model. *Annals of Behavioral Medicine, 37*(1), 58-69.
Lohr, J. M. (2011). What Is (and What Is Not) the Meaning of Evidence-Based Psychosocial Intervention? *Clinical Psychology: Science and Practice, 18*(2), 100-104. doi:10.1111/j.1468-2850.2011.01240.x
Longmore, R. J., & Worrell, M. (2007). Do we need to challenge thoughts in cognitive behavior therapy? *Clinical Psychology Review, 27*(2), 173-187. doi:10.1016/j.cpr.2006.08.001
Luciano, J. V., Guallar, J. A., Aguado, J., Lopez-Del-Hoyo, Y., Olivan, B., Magallon, R., . . . Garcia-Campayo, J. (2014). Effectiveness of group acceptance and commitment therapy

for fibromyalgia: a 6-month randomized controlled trial (EFFIGACT study). *PAIN, 155*(4), 693-702. doi:10.1016/j.pain.2013.12.029
McCracken, L. M., Sato, A., & Taylor, G. J. (2013). A trial of a brief group-based form of acceptance and commitment therapy (ACT) for chronic pain in general practice: pilot outcome and process results. *Journal of Pain, 14*(11), 1398-1406. doi:10.1016/j.jpain.2013.06.011
Mesri, B., Niles, A. N., Pittig, A., LeBeau, R. T., Haik, E., & Craske, M. G. (2017). Public speaking avoidance as a treatment moderator for social anxiety disorder. *Journal of Behavior Therapy and Experimental Psychiatry, 55*, 66-72. doi:http://dx.doi.org/10.1016/j.jbtep.2016.11.010
Morina, N., A-Tjak, J. G. L., & Emmelkamp, P. M. G. (2015). Reducing Biases in Meta-Analyses: Reply to Hertenstein and Nissen. *Psychotherapy and Psychosomatics, 84*(4), 252-252.
Morton, J., Snowdon, S., Gopold, M., & Guymer, E. (2012). Acceptance and Commitment Therapy Group Treatment for Symptoms of Borderline Personality Disorder: A Public Sector Pilot Study. *Cognitive and Behavioral Practice, 19*(4), 527-544. doi:10.1016/j.cbpra.2012.03.005
Niles, A. N., Burklund, L. J., Arch, J. J., Lieberman, M. D., Saxbe, D., & Craske, M. G. (2014). Cognitive Mediators of Treatment for Social Anxiety Disorder: Comparing Acceptance and Commitment Therapy and Cognitive-Behavioral Therapy. *Behavior Therapy, 45*(5), 664-677. doi:http://dx.doi.org/10.1016/j.beth.2014.04.006
Niles, A. N., Wolitzky-Taylor, K. B., Arch, J. J., & Craske, M. G. (2017). Applying a novel statistical method to advance the personalized treatment of anxiety disorders: A composite moderator of comparative drop-out from CBT and ACT. *Behav Res Ther, 91*, 13-23. doi:http://dx.doi.org/10.1016/j.brat.2017.01.001
Öst, L.-G. (2008). Efficacy of the third wave of behavioral therapies: A systematic review and meta-analysis. *Behav Res Ther, 46*(3), 296-321.
Öst, L.-G. (2014). The efficacy of Acceptance and Commitment Therapy: An updated systematic review and meta-analysis. *Behav Res Ther, 61*(0), 105-121. doi:http://dx.doi.org/10.1016/j.brat.2014.07.018
Overall, J. E., & Gorham, D. R. (1962). The brief psychiatric rating scale. *Psychol Rep, 10*(3), 799-812.
Patrick, D. L., Cheadle, A., Thompson, D. C., Diehr, P., Koepsell, T., & Kinne, S. (1994). The validity of self-reported smoking: a review and meta-analysis. *American Journal of Public Health, 84*(7), 1086-1093.
Petersen, C. L., & Zettle, R. D. (2009). TREATING INPATIENTS WITH COMORBID DEPRESSION AND ALCOHOL USE DISORDERS: A COMPARISON OF ACCEPTANCE AND COMMITMENT THERAPY VERSUS TREATMENT AS USUAL. *The Psychological Record, 59*(4), 521-536.
Post, M. W., van Leeuwen, C. M., van Koppenhagen, C. F., & de Groot, S. (2012). Validity of the Life Satisfaction questions, the Life Satisfaction Questionnaire, and the Satisfaction With Life Scale in persons with spinal cord injury. *Arch Phys Med Rehabil, 93*(10), 1832-1837. doi:10.1016/j.apmr.2012.03.025
Pots, W. T., Fledderus, M., Meulenbeek, P. A., ten Klooster, P. M., Schreurs, K. M., & Bohlmeijer, E. T. (2016). Acceptance and commitment therapy as a web-based

intervention for depressive symptoms: randomised controlled trial. *Br J Psychiatry, 208*(1), 69-77. doi:10.1192/bjp.bp.114.146068

Powers, M. B., Zum Vorde Sive Vording, M. B., & Emmelkamp, P. M. (2009). Acceptance and Commitment Therapy: A Meta-Analytic Review. *Psychotherapy and Psychosomatics, 78*(2), 73-80.

Ruiz, F. J. (2012). Acceptance and Commitment Therapy versus Traditional Cognitive Behavioral Therapy: A Systematic Review and Meta-analysis of Current Empirical Evidence. *International journal of psychology and psychological therapy, 12*(3), 333-357.

Shawyer, F., Farhall, J., Mackinnon, A., Trauer, T., Sims, E., Ratcliff, K., . . . Copolov, D. (2012). A randomised controlled trial of acceptance-based cognitive behavioural therapy for command hallucinations in psychotic disorders. *Behav Res Ther, 50*(2), 110-121. doi:10.1016/j.brat.2011.11.007

Silverman, W. K., & Hinshaw, S. P. (2008). The Second Special Issue on Evidence-Based Psychosocial Treatments for Children and Adolescents: A 10-Year Update. *Journal of Clinical Child & Adolescent Psychology, 37*(1), 1-7. doi:10.1080/15374410701817725

Smout, M. F., Longo, M., Harrison, S., Minniti, R., Wickes, W., & White, J. M. (2010). Psychosocial treatment for methamphetamine use disorders: a preliminary randomized controlled trial of cognitive behavior therapy and Acceptance and Commitment Therapy. *Subst Abus, 31*(2), 98-107. doi:10.1080/08897071003641578

Stotts, A. L., Green, C., Masuda, A., Grabowski, J., Wilson, K., Northrup, T. F., . . . Schmitz, J. M. (2012). A stage I pilot study of acceptance and commitment therapy for methadone detoxification. *Drug Alcohol Depend, 125*(3), 215-222. doi:10.1016/j.drugalcdep.2012.02.015

Thekiso, T. B., Murphy, P., Milnes, J., Lambe, K., Curtin, A., & Farren, C. K. (2015). Acceptance and Commitment Therapy in the Treatment of Alcohol Use Disorder and Comorbid Affective Disorder: A Pilot Matched Control Trial. *Behavior Therapy, 46*(6), 717-728. doi:http://dx.doi.org/10.1016/j.beth.2015.05.005

Thorsell, J., Finnes, A., Dahl, J., Lundgren, T., Gybrant, M., Gordh, T., & Buhrman, M. (2011). A comparative study of 2 manual-based self-help interventions, acceptance and commitment therapy and applied relaxation, for persons with chronic pain. *Clin J Pain, 27*(8), 716-723. doi:10.1097/AJP.0b013e318219a933

Tolin, D. F. (1999). *A revised meta-analysis of psychosocial treatments for PTSD*. Paper presented at the Poster presented at AABT, Toronto.

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically Supported Treatment: Recommendations for a New Model. *Clinical Psychology: Science and Practice, 22*(4), 317-338. doi:10.1111/cpsp.12122

Trompetter, H. R., Bohlmeijer, E. T., Fox, J.-P., & Schreurs, K. M. G. (2015). Psychological flexibility and catastrophizing as associated change mechanisms during online Acceptance & Commitment Therapy for chronic pain. *Behav Res Ther, 74*, 50-59. doi:http://dx.doi.org/10.1016/j.brat.2015.09.001

Trompetter, H. R., Bohlmeijer, E. T., Veehof, M. M., & Schreurs, K. M. (2015). Internet-based guided self-help intervention for chronic pain based on Acceptance and Commitment Therapy: a randomized controlled trial. *J Behav Med, 38*(1), 66-80. doi:10.1007/s10865-014-9579-0

Turk, D. C., Dworkin, R. H., Allen, R. R., Bellamy, N., Brandenburg, N., Carr, D. B., . . . Witter, J. (2003). Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. *PAIN, 106*(3), 337-345.

Veehof, M. M., Trompetter, H. R., Bohlmeijer, E. T., & Schreurs, K. M. G. (2016). Acceptance- and mindfulness-based interventions for the treatment of chronic pain: a meta-analytic review. *Cognitive Behaviour Therapy, 45*(1), 5-31. doi:10.1080/16506073.2015.1098724

Watanabe, N., Churchill, R., Hunot, V., & Furukawa, T. A. (2004). Psychotherapy for depression in children and adolescents. *Cochrane Database of Systematic Reviews*(4). doi:10.1002/14651858.CD005334

Westin, V. Z., Schulin, M., Hesser, H., Karlsson, M., Noe, R. Z., Olofsson, U., . . . Andersson, G. (2011). Acceptance and commitment therapy versus tinnitus retraining therapy in the treatment of tinnitus: a randomised controlled trial. *Behavior Research and Therapy, 49*(11), 737-747. doi:10.1016/j.brat.2011.08.001

Wetherell, J. L., Afari, N., Ayers, C. R., Stoddard, J. A., Ruberg, J., Sorrell, J. T., . . . Patterson, T. L. (2011). Acceptance and Commitment Therapy for Generalized Anxiety Disorder in Older Adults: A Preliminary Report. *Behavior Therapy, 42*(1), 127-134. doi:http://dx.doi.org/10.1016/j.beth.2010.07.002

Wetherell, J. L., Afari, N., Rutledge, T., Sorrell, J. T., Stoddard, J. A., Petkus, A. J., . . . Hampton Atkinson, J. (2011). A randomized, controlled trial of acceptance and commitment therapy and cognitive-behavioral therapy for chronic pain. *PAIN, 152*(9), 2098-2107. doi:http://dx.doi.org/10.1016/j.pain.2011.05.016

Wetherell, J. L., Petkus, A. J., Alonso-Fernandez, M., Bower, E. S., Steiner, A. R. W., & Afari, N. (2016). Age moderates response to acceptance and commitment therapy vs. cognitive behavioral therapy for chronic pain. *Int J Geriatr Psychiatry, 31*(3), 302-308. doi:10.1002/gps.4330

White, R., Gumley, A., McTaggart, J., Rattrie, L., McConville, D., Cleare, S., & Mitchell, G. (2011). A feasibility study of Acceptance and Commitment Therapy for emotional dysfunction following psychosis. *Behav Res Ther, 49*(12), 901-907. doi:http://dx.doi.org/10.1016/j.brat.2011.09.003

Wicksell, R. K., Ahlqvist, J., Bring, A., Melin, L., & Olsson, G. L. (2008). Can exposure and acceptance strategies improve functioning and life satisfaction in people with chronic pain and whiplash-associated disorders (WAD)? A randomized controlled trial. *Cognitive Behaviour Therapy, 37*(3), 169-182. doi:10.1080/16506070802078970

Wicksell, R. K., Kemani, M., Jensen, K., Kosek, E., Kadetoff, D., Sorjonen, K., . . . Olsson, G. L. (2013). Acceptance and commitment therapy for fibromyalgia: a randomized controlled trial. *European Journal of Pain, 17*(4), 599-611. doi:10.1002/j.1532-2149.2012.00224.x

Wicksell, R. K., Melin, L., Lekander, M., & Olsson, G. L. (2009). Evaluating the effectiveness of exposure and acceptance strategies to improve functioning and quality of life in longstanding pediatric pain--a randomized controlled trial. *PAIN, 141*(3), 248-257. doi:10.1016/j.pain.2008.11.006

Williamson, A., & Hoggart, B. (2005). Pain: a review of three commonly used pain rating scales. *J Clin Nurs, 14*(7), 798-804. doi:10.1111/j.1365-2702.2005.01121.x

Wolfe, F., Smythe, H. A., Yunus, M. B., Bennett, R. M., Bombardier, C., Goldenberg, D. L., . . . Clark, P. (1990). The American College of Rheumatology 1990 criteria for the classification of fibromyalgia. *Arthritis & Rheumatology, 33*(2), 160-172.

APPENDICES

Appendix A: Öst Table 1 corrected. Where corrections have been made, the revised figures are on the right in square brackets. Studies without errors reported in this paper are not shown.

Disorder	Study	Comparison	N total	N / cell	Att. Total %	Att. ACT %	N compl.	Compl. / cell	% women	Mean age	# of therap ists	# of weeks	# of sess.	# of hours	F-up months
Depression	Zettle and Hayes (1986)	CBT	18	9[6]	NI	NI	18	9[6]	100	NI	1	12	12	12.0	3[2]
Depression	Zettle and Rains (1989)	CBT	37	12.3	16.2	15.4[NI]	31	10.3	100	41.3	1	12	12	10.8[16]	2
Depression	Hayes et al. (2011)	TAU	38	19	21.1	13.6	30	15	71	14.9	3[8]	NI	NI	20.8	3
Depression	Petersen and Zettle (2009)	TAU	28	14	14	20	24	12	50	37.8	1	4[3]	5	3.1	0
Psychotic symptoms	Gaudiano and Herbert (2006)	TAU	40	20	5.0	5.3	38	19	36	40.0	1	3[1.5]	3	3.0	4
Psychotic symptoms	Shawyer et al. (2012)	Other	73[43]	14.3	9.3	4.8	39	19.5	44	39.8	5	15	15	12.5	6
Math anxiety	Zettle (2003)	CBT	33	16.5	27.3	14.3	24	12	81[83]	30.5	1	6	6	6.0	0[2]
Test anxiety	Brown et al. (2011)	CBT	16	8	31.3	12.5	11	5.5	69	20.2	1[2+]	1	1	2.0	0
GAD	Wetherell, Afari, Ayers, et al. (2011)	CBT	21	10.5	23.8	36.4[0]	16	8	48	70.8	6	12	12	12.0	6
GAD	Haves-Skelton et al. (2013)	CBT	81	40.5	22.2	25.0	25[63]	31.5	65	32.9	11	16	16	18.0	6
OCD	Twohig et al. (2010)	CBT	79	39.5	17.7[16.5]	14.6[12.2]	5[6]	32.5	61	37.0	6	8	8	8.0	3
Mixed group	Arch et al. (2012)	CBT	128	64	33.6	35.1	85	42.5	52	38.0	39	12	12	12.0[14.0]	12
Drug abuse	Luoma et al. (2012)	TAU	133	66.5	24.1[39.1]	29.4[42.6]	79[81]	50.5[81]	46	33.6	2[3]	4[1]	3	6.0	4
Nicotine dependence	Bricker et al. (2013)	Other	222	111	46.4	45.9	119	59.5	38	45.1	0	12	0	NA	0[3]
Trichotillomania	Woods et al. (2006)	WLC	28	14	10.7	14.3	26[25]	13	89	35.0	1	12	10	12.0[10.0]	3
Borderline PD	Gratz and Gunderson (2006)	TAU	24	12	8.3	0[8.3]	22	11	100	33.2	1	14	14	21.0	0
Borderline PD	Morton et al. (2012)	TAU	41	20.5	31.7[22.0]	33.3[14.3]	28[32]	14[16]	93	34.8	3	12	12	24.0	3
Pain	Johnston et al. (2010)	WLC	24	12	41.6	50.0	14	7	63	43.0	1	6	6[self-help]	3.0[self-help]	0
Pain	Wetherell, Afari, Rutledge, et al. (2011)	CBT	114	57	25.4	24.6[12.2]	85	42.5	51	54.9	3	8	8	12.0	6
Pain	Buhrman et al. (2013)	WLC [Placebo]	76	38	19.7	23.7	61	30.5	59	40.1 [49.1]	3	7	2	0.5	6
Tinnitus	Westin et al. (2011)	Other/ WLC	64	21.3	6.3	4.8	60	20	47	50.9	8	10	10	10.0	6[18]
Tinnitus	Hesser et al. (2012)	CBT	99	33	10.1[6]	8.6[11.4]	89	29.7	43	48.5	7	8	8	1.2	12

Overweight/Obesity	Tapper et al. (2009)	WLC	62	31	13[19]	25.8[16]	51	26.5	100	41.0	1	3	3	6.0	3
Overweight/Obesity	Forman, Butryn, et al. (2013)	CBT	48	24	0	0	48	24	100[NI]	32.5	NI	1	1	2.0	0
Diabetes	Gregg et al. (2007)	Other	81	40.5	18.5	16.3	66	33	47	50.9	1	1	1	7.0	0[3]
Stress	Flaxman and Bond (2010a)[2010b]	SIT/WLC	311	155.5	59.2	64.4	127	63.5	72	41.0	1	14	3	9.0	0
Stress	Flaxman and Bond (2010b[2010a])	WLC	107	35.7	38.3	48.6	66	22	NI	39.0	1	2	2	6.0	0[3]
Stress	Lloyd et al. (2013)	WLC	100	50	26.5	29.5	64[100]	32[50]	83	47.0	1	10	3	9.0	6

Appendix B: Öst Tables 11 and 12 revised in line with this review. Studies without errors reported in this paper are not shown. Where corrections have been made, the original figure is shown first, followed by the correction in brackets

Table 11 from Öst (2014)

Study	Comparison condition	WLC	Placebo	TAU	Established treatment	Equivalence analysis	Treatment manuals	Inclusion criteria reliably delineated	Reliable and valid outcome measures	Appropriate data analysis
Depression										
Petersen and Zettle (2009)	TAU			=[>]		0	+	–	+	+
Hayes et al. (2011)	TAU			>			+	–[+]	+	–
Gaudiano and Herbert (2006)	TAU			=[>]		0	+	–[+]	?[+]	+
White et al. (2011) (incorrectly classified as psychosis)	TAU			=[>]		0	+	–	+	+
Shawyer et al. (2012)	Befriending		=			0	+	–[+]	+	+
Anxiety disorders										
Wetherell, Afari, Ayers, et al. (2011)	CBT				=		=[+]	–[+]	+	–[?]
England et al. (2012)	Habitutation (Exposure)				=	0	+	+	–[+]	–
Drug Abuse										
Smout et al. (2010)	CBT				=	0	+	–[+]	–[+]	+
Luoma et al. (2012)	TAU			>			+	–	+	+
Stotts et al. (2012)	Drug couns.				=	0	+	–[+]	+	+
Gifford et al. (2004)	NRT				=	0	+	–[+]	+	+

Gifford et al. (2011)	Bupropion		>		+	—[+]	+	+
Bricker et al. (2013) *Borderline PD*	Smokefree		=[>]	0	+	—[+]	—[+]	?
Gratz and Gunderson (2006)	TAU	>			+	—[+]	+	+
Morton et al. (2012)	TAU	>			+	—[+]	+	+

Table 12 from Öst (2014)

Study	Comparison condition	WLC	Placebo	TAU	Established treatment	Equivalence analysis	Treatment manuals	Inclusion criteria reliably delineated	Reliable and valid outcome measures	Appropriate data analysis
Pain										
Dahl et al. (2004)	TAU			>			+	–[+]	–[+]	+
Wicksell et al. (2008)	WLC	>				+	+	–[+]	+	+
Wicksell et al. (2009)	TAU			=		0	+	–[+]	+	+
Thorsell et al. (2011)	AR				(=)	0	+	–	?[+]	+
Wetherell, Afari, Rutledge, et al. (2011)	CBT			=		0	+	–[+]	+	+
Wicksell et al. (2013)	WLC	>					+	–[+]	+	+
McCracken et al. (2013)	TAU			=		0	+	–[+]	+	+
Westin et al. (2011)	TRT/WLC	>			>		+	–[+]	+	+
Overweight/Obesity										
Lillis et al. (2009)	WLC	>					+	–[+]	+	+
Forman, Hoffman, et al. (2013)	BT				=	0	+	–[+]	+	+
Various Disorders										
Rost et al. (2012)	TAU			>			–[+]	+	+	+
Stress at work										
Bond and Bunce (2000)	IPP/WLC	>	>[not placebo]		[>]		+	–[NA]	+	+

Flaxman and Bond (2010a)[2010b]	SIT/WLC	>	=	0	+	–[NA]	+	+
Flaxman and Bond (2010b)[2010a]	WLC	>			+	–[NA]	+	+
Brinkborg et al. (2011)	WLC	>			+	–[NA]	+	+
Bethay et al. (2013)	ABA		(=)	0	+	–[NA]	+	+
Lloyd et al. (2013)	WLC	>			+	–[NA]	+	+
Lappalainen et al. (2013)	WLC	=		0	–	–[NA]	+	+

TABLES

Table 1: *Comparison of A-Tjak et al. (2015) and Öst's (2014) ratings for each scoring category*

Criteria	Öst (2014)	A-Tjak et al. (2015)	*t*-value
1. Clarity of sample description	1.39	1.72	-2.96*
2. Severity/chronicity of the disorder	1.39	1.42	-.21
3. Representativeness of the sample	1.28	1.69	-4.14***
4. Reliability of the diagnosis in question	.44	.67	-1.67
5. Specificity of outcome measures	1.92	1.89	.44
6. Reliability and validity of outcome measures	1.72	1.92	-2.91*
7. Use of blind evaluators	.39	.58	-1.87
8. Assessor training	.33	.39	-.53
9. Assignment to treatment	1.00	1.08	-1.78
10. Design	1.06	1.03	.27
11. Power analysis	.39	.28	1.28
12. Assessment points	.92	1.06	-1.41
13. Manualized, replicable, specific treatment programs	1.44	1.47	-.21
14. Number of therapists	.67	.88	-2.50*
15. Therapist training/experience	.64	1.17	-3.91***
16. Checks for treatment adherence	.31	.75	-4.09***
17. Checks for therapist competence	.19	.78	-5.39***
18. Control of concomitant treatments	.28	.50	-1.60
19. Handling of attrition	.86	1.39	-3.37***
20. Statistical analyses and presentation of results	1.78	1.89	-1.44
21. Clinical significance	.61	.64	-.27
22. Equality of therapy hours (n=22)	1.55	1.46	.57
Total C1-C21	19.00	23.17	-7.17

Note. N= 36 except for criterion 22 which included 'not applicable ratings for 14 studies. * $p > .05$, *** $p > .001$.

Table 2: S*ilverman and Hinshaw (2008) Criteria for Classifying Evidence-Based Psychosocial Treatments*

Criteria 1: Well-Established Treatments
1.1 There must be at least two good group-design experiments, conducted in at least two independent research settings and by independent investigatory teams, demonstrating efficacy by showing the treatment to be: *a) statistically significantly superior to pill or psychological placebo or to another treatment* *OR* *b) equivalent (or not significantly different) to an already established treatment in experiments with statistical power being sufficient to detect moderate differences* *AND* *1.2 treatment manuals or logical equivalent were used for the treatment* *1.3 conducted with a population, treated for specified problems, for whom inclusion criteria have been delineated in a reliable, valid manner* *1.4 reliable and valid outcome assessment measures, at minimum tapping the problems targeted for change were used, and* *1.5 appropriate data analyses*
Criteria 2: Probably Efficacious Treatments
2.1 There must be at least two good experiments showing the treatment is superior (statistically significantly so) to a wait-list control group *OR* *2.2 One or more good experiments meeting the Well-Established Treatment Criteria with the one exception of having been conducted in at least two independent research settings and by independent investigatory teams*
Criterion 3: Possibly Efficacious Treatments
At least one "good" study showing the treatment to be efficacious in the absence of conflicting evidence
Criterion 4: Experimental Treatments *Treatment not yet tested in trials meeting task force criteria for methodology*