Long-Term Stability of Cognitive Behavioral Therapy Effects for Panic Disorder with Agoraphobia: A Two-Year Follow-Up Study

Andrew T. Gloster[1,2], Christina Hauke[1], Michael Höfler[1], Franziska Einsle[1], Thomas Fydrich[3], Alfons Hamm[4], Andreas Sthröhle[5], and Hans-Ulrich Wittchen[1]

[1]Technische Universität Dresden – Institute of Clinical Psychology and Psychotherapy
[2]University of Basel, Switzerland
[3]Humboldt University – Berlin
[4]Ernst-Moritz-Arndt University Greifswald
[5]Charité – Universitätsmedizin Berlin

PAPER ACCEPTED AND PUBLISHED IN BEHAVIOUR RESEARCH AND THERAPY –
FOR CONFIDENTIAL INTERNAL USE ONLY

Gloster, A. T., Hauke, C., Höfler, M., Einsle, F. Fydrich, T., Hamm, A., … A., Wittchen, H. (2013). Long-term stability of cognitive behavioral therapy effects for panic disorder with agoraphobia: A two-year follow-up study. *Behaviour Research and Therapy, 51*, 830-839. Doi: http://dx.doi.org/10.1016/j.brat.2013.09.009

Corresponding Author:
Andrew T. Gloster
University of Basel
Division of Clinical Psychology and Epidemiology
Missionsstrasse 62A
CH-4055 Basel
Switzerland
Tel: ++41-61-267-0275
Email: andrew.gloster@unibas.ch

Abstract

Objective: Cognitive-behavioral therapy (CBT) aims to help patients establish new behaviors that will be maintained and adapted to the demands of new situations. The long-term outcomes are therefore crucial in testing the durability of CBT.

Method: A two-year follow-up assessment was undertaken on a subsample of n = 146 PD/AG patients from a multicenter randomized controlled trial. Treatment consisted of two variations of CBT: exposure in situ in the presence of the therapist (T+) or on their own following therapist preparation (T-).

Results: Both variations of CBT had high response rates and, overall, maintained the level of symptomatology observed at post-treatment with high levels of clinical significance. Effect sizes 24 months following treatment were somewhat lower than at the 6-month follow up. Once patients reached responder status, they generally tended to remain responders at subsequent assessments. Differences were observed for patients that obtained additional treatment during the follow-up period. Expert opinion and subjective appraisal of treatment outcome differed. No robust baseline predictors of 2-year outcome were observed.

Conclusion: Most patients maintain clinically meaningful changes two years following treatment across multiple outcome measures. Approximately 1/3 of patients continued to experience meaningful residual problems.

Keywords: CBT; Panic; Agoraphobia; Long Term Follow-Up; Exposure; Non-responder

Long-Term Response to CBT for Panic Disorder with Agoraphobia: A Two-Year Follow-Up Study

Cognitive Behavioral Treatment (CBT) operates under the premise that patients learn new ways of responding to the internal and situational stimuli that combine to create impairment. Implicit is the assumption that the newly acquired skills and behaviors are maintained over time and are readily adaptable to the demands of new situations. That which patients learn, if properly applied, is maintainable after the treatment has ceased and – at least theoretically – for the rest of their life. The permanency of what is learned in therapy (i.e., skills, behaviors, etc.) is believed to be one of the reasons that the effects of CBT tend to be superior to pharmacological approaches at follow-up assessments (Barlow, Gorman, Shear, & Woods, 2000; Cottraux et al., 1995; de Beurs, van Dyke, Lange, & van Balkom, 1999).

Critical examination of the degree to which patients maintain their gains requires long-term follow-up assessments. Regrettably empirical data on the long lasting (i.e., at least two years) effects of treatment are hard to obtain. As a result, knowledge about the longer term effects of treatment lag behind our understanding of the immediate efficacy of interventions. The studies that do exist provide some support for the long-term efficacy of CBT across numerous disorders (Butler, Chapman, Forman, & Beck, 2006), yet some evidence derived across three anxiety disorders suggests that effects can begin to recede as soon as one year after treatment (Durham, Higgins, Chambers, Swan, & Dow, 2012).

CBT for Panic Disorder with Agoraphobia (PD/AG) represents an especially important test for the long-term effects of CBT given that studies using CBT for this disorder have documented some of the highest efficacy rates in the literature and is considered the gold standard for this disorder. As such, it may be expected that the stability of gains in the treatment

of this disorder should be particularly high. Indeed, evidence from multiple studies suggests that treatment gains are maintained through at least 6 months on average (Craske, Brown, & Barlow, 1991). Fava (1995, 2001) observed high long-term (2 - 14 years) rates, but this group referenced only those who successful responded to the original treatment. A recent meta-analysis found that the controlled effect size (i.e., in comparison to a wait list) at follow up assessments is meaningfully lower than average controlled effect sizes at post-treatment (Sánchez-Meca, Rosa-Alcázar, Marín-Martínez, & Gómez-Conesa, 2010). Similarly, one of the longest follow-up studies followed 189 patients with panic disorder for up to 14 years following several different randomized trials and concluded that the short term effects are unrelated to long term outcomes (Durham et al., 2005). Similar conclusions were reached in a 15 year follow-up assessment following pharmacological treatment of panic disorder (Andersch & Hetta, 2003). To the degree these observations are generalizable, the long-term treatment effect of CBT for PD/AG are called into question.

Even less is known about the characteristics of patients' change beyond effect sizes and response rates. That is, the percentage of patients that maintain their gains, improve or worsen has not been clearly established. Do these groups differ 6 months and 24 months following treatment and is there a subgroup of patients who need more time to respond (during the time between 6 and 24 months following treatment)? Likewise, little is known about the percentage of patients that report residual symptoms and how these residual symptoms influence global functioning. Relatedly, studies show that patients seek additional treatment even after successful treatment (Durham et al., 2012). If replicated, this finding raises two additional issues. First, it is unclear if these patients are those who have relapsed, or if they sought treatment for reasons other than panic symptomatology. Second, it is unclear if opinions of successful treatment outcome concur between independent experts and patient's subjective appraisal. If not, perhaps the

disconnect suggested by Durham's work between successful outcome and additional treatment may be explained by the yet untested possibility that patients differ from experts in what constitutes a successful outcome. The answers to all of these issues offer important clues about the processes that unfold in the time following treatment – the exact period in which our theories assume that patients generalize the newly learned material.

The influence of procedural variations of treatment delivery (e.g., how exposure is administered during treatment) on the long-term response is also unknown. It is unclear if variance due to treatment variations dissipates over time. It is feasible or even probable that the influence of what is learned overshadows any differences in how it was learned. To our knowledge, however, no information exists regarding differences in long-term effects that result from systematic procedural variations in the delivery of CBT. Instead, previous studies have concentrated on how CBT compares to a different treatment package or to an augmented CBT with one or more different components added.

This study is an extension of a previous study (Gloster et al., 2011) and aimed to examine these issues by looking at the durability, pattern, and characteristics of treatment effects from a standardized CBT for PD/AG two years following the end of treatment. All patients examined were treated with a highly efficacious standardized CBT, as measured at post-treatment and the 6-month follow up (Gloster et al., 2011). We are therefore able to investigate the long-term outcomes using a large sample of patients, all of whom received a standardized efficacious treatment. Additionally, all patients in this study were diagnosed with panic disorder and agoraphobia. This is important because the presence of agoraphobia has been found to influence the outcome of panic disorder (Williams & Falbo, 1996) and most previous studies included patients diagnosed with panic disorder with or without agoraphobia. In one exception, patients were treated with either a 14-session, 7-session, or group CBT and followed for two years

(Marchand, Roberge, Primiano, & Germain, 2009). Significant improvements across all treatment modalities and dependent measures were found at post-treatment, one-year and two-year follow-up, with high end-state functioning achieved by 57% of the patients. Given the assumed importance of agoraphobic avoidance, the treatment in this study targeted this factor and concentrated heavily on exposure (both interoceptive and in situ in multiple situations), did not include explicit components of breathing retraining or logical disputation, and increased anxiety in the in situ exposures with interoceptive exercises when patients reported no or insufficient anxiety response (Lang et al., 2012). Conceivably, these characteristics supported inhibitory learning, which may have facilitated adaptive fear responding (Craske et al., 2008) and this study was designed to examine the long-term effects of these factors.

Importantly, this study also examined a procedural variation of CBT that may inform about how long-term treatment gains can be maximized. This approach builds upon other studies that examine the global effects of CBT compared to another treatment modalities or treatment packages. Specifically, the treatment examined in this study utilized two procedural variations of CBT (i.e., exposure in situ in the presence of the therapist vs. planned in the therapy room), but were identical in content (Gloster et al, 2009; Lang et al., 2012).

Additional strengths of this study that built on previous studies were the large sample size that allowed us to examine patterns and issues not amenable to smaller samples and the inclusion of outcome variables above and beyond the frequency of panic attacks. This is important because evidence has cumulated that panic attacks, panic disorder, and agoraphobia are often independent (Craske et al., 2010; Wittchen et al. 2008; Wittchen, Gloster, Beesdo-Baum, Fava, & Craske, 2010), that agoraphobic avoidance is as important if not more so for long-term outcome (Fava et al., 1995), and recent emphasize on how one reacts to the attacks as opposed to the occurrence of the attacks themselves (Eifert & Forsyth, 2005).

Building on an efficacious treatment and using a relatively large sample, the aim of this study was to examine the effects of a standardized CBT for PD/AG two years following the end of treatment. In particular, we examined the durability and pattern of multiple outcomes while examining the differential effect of procedural variations, occurrence of additional treatment, concordance between the patients' subjective evaluation of outcome and expert raters, and prediction of long-term outcome.

Method

Design

All patients were part of the multicenter Mechanisms of Action in CBT (MAC) RCT for Panic Disorder with Agoraphobia (PD/AG) (Gloster et al., 2011). The 24-month follow-up assessment (FU-24) was conducted on a subset of four[1] of the original eight study centers in the MAC Trial. These centers were the largest study cities, thereby allowing maximal utilization of resources. All eligible patients from these four centers (n = 198) were contacted and n = 146 (73.7%) agreed to partake in the FU-24. In this paper, the effects of two treatment groups were also examined: (T+) therapists accompanied patients during in situ exposure or (T-) patients were prepared for in situ exposure but executed the exposure on their own. The participation rate in the FU-24 did not differ between the treatment groups: T+ (81/105, 77.1%) and T- (65/93, 69.9%), chi square (1) = 1.34, $p > .05$. The patients in the study centers selected did not differ from the remaining patients in the other centers on any clinical variable at baseline or during the therapy in terms of treatment integrity (results available on request).

All eligible patients were contacted by phone and invited to the FU-24 follow-up interview. Patients who were unreachable were sent letters. In the case that a participant was unwilling to participate in the full interview, a short version of the interview was offered that captured the most essential information necessary for the main outcomes reported in this study

[1] Berlin Adlershofen, Berlin Charite, Dresden, Greifswald

(i.e., items assessing the number of panic attacks, avoidance, impairment, overall change compared to pre-treatment, whether they obtained additional treatment since the FU-6, and the expert rated CGI). Because of this, the sample size varies across analyses.

Intervention

The standardized CBT manual (Lang et al., 2012) was administered in 12 sessions and implemented over 6 weeks. The treatment consisted of psychoeducation, individualized behavior analysis of the person's presenting problems, interoceptive exposure, standardized exposure in situ (i.e., bus, shopping mall, and forest), individualized exposure in situ, skills to cope with anticipatory anxiety, and relapse prevention. Patients received one of two CBT variations that were identical except for the way in which the exposure in situ sessions were administered. During all exposure sessions, the importance of entering situations without engaging in safety behaviors was stressed and practiced. Further details about the therapy and study design have already been published (Gloster et al., 2009; Gloster et al., 2011; Lang et al., 2012).

Therapists

Therapists were advanced-level clinical psychology graduate students and post-docs experienced in CBT of anxiety disorders. Therapists were only allowed to see patients in the study if they completed an extensive three-day training and passed a test consisting of role-plays of the critical aspects of the manual. Therapists were trained to see patients in both variants of the active treatment. Weekly supervision and videotaping of all sessions was implemented to maintain therapy integrity and identify violations of the protocol.

Assessors

Assessors were advanced graduate students in clinical psychology. The assessors were not explicitly informed about the patient's treatment condition (T+ or T-), but it was impossible to guarantee that the patient did not reveal details about the treatment such that it could be inferred.

Prior to beginning the study, assessors took part in a three-day training, testing, and subsequent certification identical to the efficacy study (Gloster et al., 2011). The training included extensive practice in the proper administration of the instruments in which all interviews were schooled and trained (e.g., role-play, and detailed discussion of common rating scenarios) to the standard of no variability for the CGI and only 3 points on the SIGH-A across all exam cases. Most raters (80%) were certified as study assessors on the first try, the others passed the exam in form of an additional rating after feedback was given on the first. Regular supervision was conducted to maintain consistent strategies across assessors and address questions.

Assessment

The measures used in this study consist of outcome variables, described first, and those assessment measures used in statistical prediction models of outcome, and the instrument used for diagnostics.

Structured Interview Guide for the Hamilton Anxiety Scale (SIGH-A; Shear et al., 2001), is a 14-item interview commonly used to assess the severity of a broad range of anxiety symptoms. The SIGH-A has demonstrated high values of inter-rater and test-retest reliability and is commonly used in outcome studies.

Clinical Global Impressions Scale – Severity Subscale *(CGI; Guy, 1976).* The CGI is a clinician-rated scale that measures the overall severity of a disorder, with scores that range between 1 (*no disorder*) and 7 (*among the most severely ill patients*). The scale queries for information across the facets of panic symptoms, anxiety, anticipatory anxiety, avoidance, and overall functional level before making the global rating. Scores on the CGI are sensitive to change in panic treatment (Barlow et al., 2000; Gloster et al., 2011).

Mobility Inventory (MI; Chambless, Caputo, Jasin, Gracely, & Williams, 1985). Mobility Inventory (MI). The Mobility Inventory is a self-report questionnaire that measures the degree of

agoraphobic avoidance across 27 situations, each of which is rated with respect to being in that situation alone or accompanied by another person. The mean score of the alone subscale (range 1-5) are reported. Scores of the MI are highly reliable and sensitive to change (Chambless et al., 1985, 2011).

Panic Agoraphobia Scale (PAS; Bandelow, 1999). The PAS is a self-report questionnaire that measures the severity of panic attacks, avoidance, anticipatory anxiety, disability, and worries about health. Scores on the PAS have been demonstrated to have good reliability and sensitivity to change (Bandelow, 1999). The PAS total score was not originally conceived as a primary outcome measure, but is presented to facilitate comparisons with other studies.

Acceptance and Action Questionnaire II (AAQ- II; Bond et al., 2011), is a 7-item unidimensional self-report measure for experiential avoidance and psychological flexibility. Items are rated on a 7-point-Likert-scale. High scores reflect more psychological flexibility. The AAQ-II explains unique variance in PD/AG patients and was shown to be sensitive to treatment effects (Gloster et al., 2011). In this study, the AAQ-II was examined as a predictor of treatment outcome.

Anxiety Sensitivity Index (ASI; Reiss, Peterson, Gursky, and McNally, 1986) is a self-report instrument assessing anticipatory fear and sensitivity to anxiety symptoms. 16 items are rated on a 5-point-Likert-scale. Internal validity is good (Cronbach's alpha from .82-.92; Alpers and Pauli, 2002). Studies have found the ASI to mediate treatment outcome in PD (Smits, Powers, Cho, Telch, 2004). In this study, the ASI was examined as a predictor of treatment outcome.

Beck Depression Inventory-2nd Ed. (BDI-II, Beck, Steer, and Brown, 1996; German version by Hautzinger, Keller, and Kühner, 2006), a 21-item self-report questionnaire, measures depression symptoms according to DSM-IV criteria. Participants rate symptom severity on a 4-

point-Likert-Scale with respect to the past two weeks. Internal validity is good (Cronbach's alpha from .89-.93.). In this study, the BDI-II was examined as a predictor of treatment outcome.

Finally, diagnosis was established using the ***Composite International Diagnostic Interview*** (CIDI) with subsequent clinician review. The standardized computer-administered personal CIDI is administered by expert interviewers and systematically assesses all DSM-IV disorders. For additional clarification, a clinician then reviewed the diagnoses. The diagnoses derived by the CIDI have been demonstrated to be reliable and valid (Essau & Wittchen, 1993; Lachner et al, 1988; Reed et al., 1998; Robins et al., 1988; Wittchen 1994; Wittchen & Pfister, 1997.)

Statistical Analysis

Two possible samples were available for these analyses. These were the full sample of patients originally enrolled in the active treatment groups of the study (overall sample; n=301) or the subset of patients approached for the FU-24 follow up from the four focal centers (n=198). Overall, 146 of the 198 eligible patients (73.7%) from these centers provided data for the FU-24 (FU-24 subset; n=146). Of note, the number of patients who participated in the FU-24 exceeded the participation rate of this group at the 6-month follow-up (136/198 = 68.7%). Analyses were run for both the overall sample and FU-24 subset. The sample size varies slightly across analyses due to administration of a short version of the interview for patients unwilling to partake in the full interview and due to random missing values, and n's are noted in the tables.

First, we investigated whether a) those patients from centers who were chosen to participate in the FU-24 differed from those patients at the other study centers not approached (198 eligible vs. 103 from other centers) and b) actual participants differed from those that refused participation in the FU-24 follow up (146 vs. 52 who did not participate). The first comparison (a) did not reveal any significant differences ($p > .05$ in linear regression) at baseline,

post-treatment, or FU-6 values of any of the primary outcome measures. The second comparison (b) revealed slightly lower HAM-A (25.0 vs. 23.8, $F = 4.2$ (1, 299), $p < .05$) and higher CGI values (5.2 vs. 5.4, $F = 4.7$ (1, 299), $p < .05$) than non-participants at baseline, but interestingly not at post-treatment or FU-6.

The goal of the next set of analyses was to determine the treatment efficacy 24 months following treatment. To take into account the full information from all 301 patients at all assessments and to address selective dropouts and missing values across treatment groups and assessment, we fitted multilevel mixed models (Skrondal & Rabe-Hesketh, 2004). These models make much weaker assumptions than conventional complete case and LOCF analysis. In particular, they allow for systematic missingness according to the values at other assessments of the outcome (Wood, Hillsdon & Carpenter, 2005). To address missingness differentially for treatment status (T+ and T-), time (assessments at BL, intermediate, post, FU-6 and FU-24) and their combinations, we specified the models saturated for the combined effects of treatment and time by using dummy variables for the associated main effects and interactions. This adjusts for both the within and between treatment group effects in case of differential missingness (Wood, Hillsdon & Carpenter, 2005). We decided to model time as discrete rather than continuous by fitting discrete time mixture models. This was done because of the non-equidistant design with qualitatively different assessment points offers no one meaningful dimensional metric for time. Moreover, we were interested in the exact changes from one assessment to another and this was possible with the discrete time mixture models. A random intercept parameter was specified, while the other model parameters were specified as fixed and robust standard errors were calculated (Royall, 1986). A between group difference was noted at baseline for the variable CGI. Between-group differences at the other time points for which a significant difference was found (i.e., intermediate, post, and FU-6) were much larger and can only be partially explained by this

baseline difference. Beginning with the intermediate assessment, T+ had lower means than T-. We concluded therefore that the baseline values do not play a meaningful role after intermediate and a negligible role in the main goal of predicting long-term stability as assessed from post-assessment on. For all other outcomes, randomization worked well and the groups did not differ significantly (see Gloster et al., 2011). Effect sizes were determined as differences in means divided by the pooled standard deviation at baseline among all 369 study participants.

For each outcome we calculated all within group treatment effects as well as predicted means (by group and time) from the model coefficients and their 95% confidence interval and p-values. The observed values were comparable in size and in significance and led to identical conclusions. Values from the post-assessment and 6-month follow-up assessment are included in order to facilitate comparison across assessment intervals.

Given differences between those who agreed to participate at FU-24 and those that did not with respect to some baseline scores (but not post-treatment or FU-6), we conducted a sensitivity analysis to assess whether missingness had occurred due to non-considered factors. For this, we repeated the entire analysis using only the 146 patients who completed the FU-24 assessment. Results led to identical conclusions as compared to the full dataset (N = 301) analysis (results available upon request).

Cutoff values used to define response were: SIGH-A: $\geq$ 50% reduction; CGI: categories of "mild", "borderline", or "no" disability; MI-alone subscale: $\leq$ 1.8; Panic-attacks: 0 panic attacks in the past week; and PAS total score: $\geq$ 50% reduction. Derived from previous studies, the cutoffs for Hamilton/ SIGH-A (e.g.,Heldt et al., 2007), CGI (e.g., Barlow, Gorman, Shear, & Woods, 2000), and panic attacks (e.g., Pollack, Mangano, Entsuah, Tzanis, & Simon, 2007) have been widely used in treatment studies. For the MI, the cutoff value represents a midway point between normative and agoraphobic samples: the cutoff value was approximately 1.5 SD's below

the reported means of samples diagnosed with Agoraphobia and 2 SD's above the mean of a normative control group (Chambless et al., 1985). For the PAS, several response criteria have been reported in the literature; the 50% reduction is considered conservative (Bandelow, Baldwin, Dolberg, Andersen, & Stein). All of these definitions have been consistently used within the analyses of the MAC research network (e.g., Gloster et al., 2011; Reif et al., 2013).

Clinical significance of change was determined using the reliable change index (RCI; Jacobson & Truax, 1991). The RCI takes the reliability and standard deviation of the measure into account and requires that change on any given measure must exceed that of the RCI. Cross tabulations and associated Fisher's exact test were calculated to examine the effect of further treatment and the concordance between expert opinion and subjective evaluation of outcome. Finally, baseline predictors of change between baseline and FU-24 were examined with linear regression. For all linear regressions, the robust Huber-White sandwich matrix was used for calculation of 95% confidence interval and p-values (Royall, 1986).

Interactions between the reported need of additional treatment at FU24 and objective need (CGI response as rated by therapists) on outcomes were assessed again with linear regressions (dimensional outcomes) and logistic regressions (response outcomes), respectively.

All analyses were conducted with Stata, version 12.1, and the XTMIXED procedure was used (Stata Corp, 2012). Significance was set at the .05 level. Global null hypotheses (therapy being overall inefficient across all outcomes) where rejected if any of the p-values from individual tests was lower than the test level of 0.5 divided by the number of tests (Bonferroni-correction). In exploratory analyses where groupings of patients resulted in very small cell sizes, p-values below .10 are also reported.

Results

Two-Year Outcome: Effect Size

Collapsing across treatment conditions (T+ and T-), patients reported clear improvement with large within group effect sizes at FU-24 compared to pretreatment levels (global test rejected, $p < .005$; d's range 0.78 to 3.63 across outcomes; see Table 1). Patients in the T+ and T- group reported largely similar within group effect sizes across all outcomes. The only significant statistical difference between the T+ and T- groups at FU-24 occurred in the level of agoraphobic avoidance (MI: difference in $d = 0.37$, $p < .05$), where the T+ group reported less avoidance of situations than the T- group.

Two-Year Outcome: Response Rates

Given that the multi-level data analysis produces only population level estimates, analyses of treatment response were necessarily limited to the $n = 146$ patients who completed the FU-24 assessment. Based on this group and using LOCF, the number of patients (collapsing T+/ T-) that achieved responder status 24 months following treatment was sizable, with differences across outcomes: HAM-A: 52.1%; CGI: 85.6%; number of panic attacks: 63.0%, and MI: 67.4%. The response rates for the T+ and T- conditions did not differ significantly on any outcome ($p > .05$), though the percentage of responders based on the MI was slightly more than 10% higher in the T+ group (73.8%) vs. T- group (59.4%).

Two-Year Outcome: Clinical Significance

We assessed the clinical significance of the observed change using the reliable change index (Jacobson & Truax, 1991) on the two broadest outcome measures in our data: PAS and CGI. Once again, these analyses were limited to those participants who completed the FU-24 assessment. Based on the RCI metric, 75.3% ($n = 73$) of the patients obtained clinically significant change in panic and agoraphobia symptoms as measured by the PAS. This percentage was not significantly different in the T+ (74.5%, $n = 38$) and T- (76.1%, $n = 35$) conditions. With respect to global functioning, 67.6% ($n = 94$) obtained clinically significant change as measured

by the CGI. The percentage of patients obtaining clinically significant levels of change was again about 10% higher in the T+ condition than the T- condition (72.7%, n = 56 vs. 61.3%, n = 38).

Change Between 6- and 24-Month Follow-Ups: Effect Size

Compared to levels measured 6 months following treatment, neither T+ nor T- achieved further improvements 18 months later. Although the overall effect sizes and response rates of both T+ and T- were excellent, on average the level of symptomatology as measured by effect size at FU-24 began to recede towards the post-treatment values and was significant for the outcomes HAM-A, MI, and CGI (all p's < .05). The T+ group reported a significant worsening of symptoms between FU-6 and FU-24 on PAS and CGI (p's < .05), whereas patients of the T- condition reported worsening between FU-6 and FU-24 only on the MI (p < .05).

Stable Gains, Improvement, and Worsening

In order to determine the stability of change following post-treatment, we calculated the percentage of patients who met definitions of positive/ negative status at successive time points of post-treatment, FU-6, and FU-24 among those patients who participated in the FU-24 assessment (see Table 2). The percentage of patients who retained positive status was consistently high across all outcome measures and time points (range 69.4% to 96.4%). In contrast, the percentage of patients who retained negative status between successive assessments varied greatly (12.5% for CGI between FU-6 and FU-24; 73.3% for the MI between FU-6 and FU-24). The percentage of patients that improved (negative status followed by positive status) was also highly variable across outcomes, with the highest levels of improvement observed for the CGI. The percentage of patients who worsened was higher than desired; with values ranging from extremely low (3.6% on the MI from Post to FU-6) to clinically disconcerting values (30.7% on the HAM-A from FU-6 to FU-24).

Proportion and Impact of Seeking Additional Treatment

Because of previous reports citing that many patients seek additional treatment following even successful outcomes (Durham, Higgins, Chambers, Swan & Dow, 2012) the residual symptomatology we observed at FU-6 (Gloster et al., 2011), and reports of increased probability of seeking additional treatment in the presence of high levels of residual symptomatology – especially agoraphobic avoidance (Fava, Zielezny, Savron, & Grandi, 1995), we asked the patients if they had "received any additional treatment" since the end of the trial (i.e., FU-6). In the context of the interview (i.e., the questions before all dealt with PD/AG and the treatment during the study), the question implied but did not specifically ask whether the treatment was for the PD/AG. In total, 42 of 112 respondents (37.5%) endorsed having obtained further treatment during this 18-month period. This rate did not differ between the T+ and T- conditions (23 of 64, 35.9% in T+; 19 of 48, 38.6% in T-, chi2(1) = 0.16, p = .693). First we tested the interactions between objective response (CGI) and the necessity of additional treatment on dimensional and response outcomes at FU24. No evidence was found for any of these interactions (all p's > .05). Second, similar to Durhan et al., we examined the pattern of outcomes grouped by additional treatment. As expected, those considered responders by independent expert raters using the CGI at FU-24 (Table 3, Columns A and B) had better outcomes than those considered non-responders on the CGI (Table 3, Columns C and D). As can also be seen in Table 4, a distinctive pattern arose for those patients considered responders from expert raters and those considered non-responders. For those patients considered responders from the expert raters, those without additional treatment generally faired better than those with additional treatment (Table 4, Columns A and B). The opposite pattern emerged for those patients considered non-responders by the expert raters. For these patients, those that had additional treatment generally had better outcomes than those who did not have additional treatment (Table 3, columns C and D). These patterns were observed for both dimensional and categorical analyses.

Finally, we explored these findings to determine whether those patients who obtained additional treatment differed from those that did not in their severity level at FU-6. We predicted that those with higher levels of severity would have been more likely to seek treatment. However, this could not be substantiated. Although those patients who sought additional treatment had consistently higher values across all assessments than those that did not obtain treatment, these differences were not significant on any of the five outcome measure (all p's > .05, estimated effect sizes <0.4) at FU-6.

Concordance Between Expert Raters and Subjective Opinion on Long-Term Outcome

Next, we asked all patients whether they currently needed additional treatment. Overall, 40.7% of the patients endorsed this question. In order to determine the concordance between subjective appraisal and independent expert assessment, these answers were compared to their response status on the CGI. Two forms of disagreement were possible: subjectively indicating the need for further treatment despite reaching response status as judged by the expert raters or indicating no further need for additional treatment despite being rated as a non-responder by the expert raters. The overall concordance between subjective opinion and expert rater was 68%, with 32% disagreement. As can be seen in Table 4, most disagreement occurred in column B (expert rated as a responder, yet subjective need for additional treatment). It can also be noted, that progression from left to right in the table also reveals generally increasing residual symptomatology.

Predictors of Long-Term Change

In a final step, we attempted to identify patient variables at baseline that might predict change in general functioning (CGI) and panic/agoraphobia symptomatology (PAS) from baseline to FU-24. Towards this end we selected variables previously identified within the literature. Specifically, we tested the predictive value of sex, age, number of panic attacks, panic

symptomatology (PAS), agoraphobic avoidance (MI), clinical global impression (CGI), psychological flexibility (AAQ-II), anxiety sensitivity (ASI), and depressive symptoms (BDI-II). When collapsing across treatment conditions (T+/ T-), no tested predictor emerged as significant for either outcome variable. Next, we tested the interactions between baseline predictor and treatment group on outcome status. No evidence was found for any of these interactions (all p's > .05).

Discussion

This study examined the two-year outcome of manualized CBT for PD/AG in a large multi-center randomized trial. Overall, excellent outcomes were observed compared to the level of symptomatology with which patients presented prior to treatment. The effect sizes two years following treatment were somewhat lower than 6 months following treatment, but generally as good as or better than the levels immediately following treatment (i.e., post-treatment). In contrast, the rates of responders were equivalent to or greater than the rates 6 months following treatment. The change observed was also clinically meaningful, with between 2/3 (CGI) and 3/4 (PAS) of patients reporting change that exceeded critical values on the reliable change index. Thus, on average, the efficacious effects of this treatment observed at post-treatment and 6-month follow-up were maintained two years following the end of treatment.

The observed effect sizes, response rates, and percentage of patients with reliable change were consistent with and sometimes better than previous two-year outcome studies for panic disorder with and without agoraphobia (e.g., Craske, Brown, & Barlow, 1991; Fava et al., 2001; Marchand, Roberge, Primiano, & Germain, 2009). This is notable given some of the characteristics of the treatment: targeting of avoidance and concentration on exposure (both interoceptive and in situ in multiple situations), temporal compression (two appointments per week), and exclusion of the explicit components of breathing retraining or logical disputation.

Thus, the elements included in the treatment appear to have a salient long-term effect for most patients, conceivably by increasing inhibitory learning (Craske et al., 2008).

Despite the overall positive results observed in our data two years following treatment, changes between FU-6 and FU-24 suggested that outcomes ceased to continue to improve and in some outcomes showed marginal worsening. Presently it is unclear if the observed receding suggests a type of regression to the mean, whether the values would continue to deteriorate with more time, or whether the values would stabilize at or around these values. Differences patterns between the T+ and T- during this period are likely a function of the absolute value of outcomes at the FU-6 assessment. In other words, the T+ condition had somewhat better outcomes than T- at FU-6 and therefore had more room to regress to the norm. The notable exception to this is the MI, where the T+ group maintained their level of avoidance whereas the T- statistically worsened. This difference is potentially of importance, as the degree of agoraphobic avoidance has been the most consistently observed difference between the two groups (Gloster et al., 2011) and thus suggests one of the mechanisms by which the two treatment groups differ. Such a position is supported by previous studies that point to the importance of reducing agoraphobic avoidance (Fava, 1995). Until this can be further empirically verified, however, this hypothesis remains speculative.

Having witnessed the marginal reduction of treatment gains – however small – we felt it important to examine and document the stability/ fluctuation of outcome status across assessment time points beyond means and effect sizes. Towards this aim, we examined whether patients met criteria for response for two consecutive time periods: post treatment to FU-6 and FU-6 to FU-24. Examining participants in this way revealed several important patterns. First, stability (i.e., retained positive/ negative status) was much more common than fluctuation (i.e., improved or worsened), with at least 2/3 of the sample in the stable categories for most outcome measures (MI

rates were even higher). Second, among those patients who fluctuated in status, the percentage that improved was always greater than the percentage that worsened. This held true across each time period and outcome measure. Taken together, these analyses suggest that most respond positively, maintain their positive status, and if transition occurs it is usually in a positive direction. However, the percentage that worsened or retained negative status is higher than desired. Indeed, we agree with recent calls to specifically address this group of treatment non-responders and treatment-resistant patients with targeted research (Pollack et al., 2008).

With the aim of further understanding residual symptomatology, we examined how responder status on global functioning (CGI) interacted with whether or not the patients obtained further treatment since the FU-6. Overall, one-third of the patients obtained additional treatment. We failed to find a significant difference between those who obtained additional treatment and those that did not with respect to their levels of symptomatology at FU-6. Whereas we cannot say whether the additional treatment specifically addressed PD/AG, we do assume that these patients did not judge themselves to be free of mental distress or else they would not have sought out additional treatment. About half of the patients were considered responders by expert raters (CGI) and did not need additional treatment. These patients continued to fair better than all other groups. Durham et al (2012) referred to these patients as having obtained a "sustained recovery" group. The rates of sustained recovery in this study were higher than those reported by Durham et al. for patients followed for up to 14 years (approximately 50% vs. 38%) A second group of patients, consisting of about one-third of this sample, needed and obtained additional treatment and this likely contributed to these patients being judged as responders by the expert raters. The rates of patients considered non-responders by the expert raters was much smaller, and in some cases so small that the statistics should be considered with caution. Within this group, slightly less than half received additional treatment, yet remained a non-responder. Durham labeled this

group "treatment resistant". The rates of treatment resistant patients were lower in this study than in patients followed for up to 14 years (approximately 5% vs. 19%, Durham et al., 2012). It remains an open question how the rates observed in this study would continue to look better than the values reported by Durham 12 years into the future (suggesting a differentiation of treatment effects) or recede (indicating a loss of potency over increasing time periods). These findings are inconsistent with Durham et al. who found that those who obtained additional treatment were worse off. Similar to Durham et al., however, we observed meaningful levels of residual symptomatology in all groups except the responders who did not obtain additional treatment. Once again, however, it should be noted that the group of patients categorized as non-responders was small.

We also examined at FU-24 whether outcome as assessed by expert raters and subjective appraisal of outcome were in agreement. Overall, the concordance between expert raters and subjective judgments were good. The mean levels of residual symptomatology increased consistently across the four groups. Of particular interest to us were the disagreements between expert raters and subjective report. Most disagreement came from those patients categorized as responders by the expert raters, yet the patients indicated the need for additional treatment. The other disagreement (non-responders who indicated they didn't need additional treatment) was very small and should be interpreted with caution. If replicable, however, the interesting question arises as to why some patients feel they need more therapy while others do not, despite comparable levels of symptomatology (i.e, groups B and C did not significantly differ from each other). It is further unclear what leads the group of patients to feel they need additional treatment although the expert raters judged otherwise. Remarkably, this group accounted for approximately one-third of the sample.

Finally, we examined predictors of long-term change (BL-FU-24). The predictors we tested established from previous research did not prove to be robust. Baseline severity level of PD/AG symptomatology, the most consistent baseline predictor from previous studies, was not significant in this study. These results surprised us to some degree, especially given that this data set was one of the larger to examine these relationships and thus is presumed to have adequate statistical power. Methodologically, variance may have been restricted to a greater extent in this study than previous studies due to the highly standardized conditions and the fact that the overall response to treatment was very positive. Thus, baseline levels of these variables may not be nearly as relevant in predicting outcome two years later as the processes that occurred during specific phases of treatment (Cammin-Nowak et. al, 2013; Emmerich et al., in prep; Gloster, et al., in press).

This study is limited in several important ways. First, not all patients from the original study sample were approached for the FU-24 follow-up assessment. Although statistical controls failed to find differences and multi-level analyses utilized to model the complete sample, this remains a source of potential bias. Second, of the eligible patients not all were willing to participate in the FU-24 follow-up assessment. Statistical tests found a difference between the patients who agreed and declined participation on some variables at baseline, but not at post-treatment or FU-6. Nevertheless, its effect on unobserved variables important for the treatment can not be excluded. Third, with the goal of maximizing participation rates, a short form of the interview and assessment was available for participants that did not wish to complete a full assessment battery. Although this did indeed increase overall participation rates, it simultaneously led to fewer responses on some questionnaires. Given the difficulty securing the participation of patients so long after treatment – especially when residual problems are present – this cost was deemed acceptable. Finally, as is common in a psychological study of this sort, it

could not be completely ruled out that the patients revealed details of their treatment condition to the assessor. This potential treat to the blinding should be considered when interpreting the results.

Despite these limitations, it can be concluded that on average patients maintained clinically meaningful effects two years following treatment across multiple outcome measures. The effects were somewhat lower than effects observed 6 months following treatment. Our examination of procedural variation revealed that, overall, both treatment variations had positive outcomes. The only difference between the two treatment variations at FU-24 was in the degree of agoraphobic avoidance, which was lower in the T+ group. The greater improvement in agoraphobic avoidance in the T+ condition is consistent with our previous findings (Gloster et al., 2011). Avoidance behavior was one of the main targets of this treatment and residual levels of avoidance have been found to be a risk factor for relapse (Fava, Zielezny, Savron, & Grandi, 1995). As such, we cautiously conclude that this difference, while small, may be of clinical importance.

Enthusiasm is tempered, however, by what appears to be approximately one third of the sample with meaningful residual difficulties. These rates are largely consistent with rates observed in previous studies. We present detailed analysis of these residual issues seldom presented in previous long-term studies, with the goal of refining the discussion in the field. To borrow a term from a similar phenomenon in the treatment of depression, we wish to actively work towards improving the "hidden third" (Schlaepfer et al., 2012). Yes, the overall picture is good. Yet, additional treatment options are likely needed when our current treatments fail to work for these patients.

In conclusion, patients generally obtained meaningful and lasting change through two years following treatment. The most salient difference between these variations was a continued

greater reduction in agoraphobic avoidance, which was targeted as the central aim of new learning within this therapy and is considered by some to be a central if not the central maintain factor (Fava, Zielezny, Savron, & Grandi, 1995; Powers et al., 2004). This bodes well for the permanency presumption of change underlying CBT. Whereas the percentage of patients who met criteria for positive response two years following treatment was good across outcomes, there is still significant room for improvement. Clinical scientists should strive to develop interventions for the subgroup of patients that do not adequately respond and ultimately work towards developing interventions that will achieve even better long-term outcomes.

Acknowledgments:

Funding/Support: This work is part of the German multicenter trial "Mechanisms of Action in CBT (MAC)". The MAC study is funded by the German Federal Ministry of Education and Research (BMBF; project no. 01GV0615) as part of the BMBF Psychotherapy Research Funding Initiative.

Centers: Principal investigators (PI) with respective areas of responsibility in the MAC study are V. Arolt (Münster: Overall MAC Program Coordination), H.U. Wittchen (Dresden: Principal Investigator (PI) for the Randomized Clinical Trial and Manual Development), A. Hamm (Greifswald: PI for Psychophysiology), A.L. Gerlach (Münster: PI for Psychophysiology and Panic subtypes), A. Ströhle (Berlin: PI for Experimental Pharmacology), T. Kircher (Marburg: PI for functional neuroimaging), and J. Deckert (Würzburg: PI for Genetics). Additional site directors in the RTC component of the program are G.W. Alpers (Würzburg), T. Fydrich and L.Fehm (Berlin-Adlershof), and T. Lang (Bremen).

Data Access and Responsibility: All principle investigators take responsibility for the integrity of the respective study data and their components. All authors and co-authors had full access to all study data. Data analysis and manuscript preparation were completed by the authors and co-authors of this article, who take responsibility for its accuracy and content.

Acknowledgements and staff members by site: Greifswald (coordinating site for psychophysiology): Christiane Melzig, Jan Richter, Susan Richter, Matthias von Rad; *Berlin-Charite (coordinating center for experimental pharmacology):* Harald Bruhn, Anja Siegmund, Meline Stoy, Andre Wittmann; *Berlin-Adlershof:* Irene Schulz; *Münster (Overall MAC Program Coordination, Genetics and Functional Neuroimaging):*Andreas Behnken, Katharina Domschke, Adrianna Ewert, Carsten Konrad, Bettina Pfleiderer, Peter Zwanzger *Münster (coordinating site for psychophysiology and subtyping):*, Judith Eidecker, Swantje Koller, Fred Rist, Anna Vossbeck-Elsebusch; *Marburg/ Aachen (coordinating center for functional neuroimaging):*, Barbara Drüke, Sonja Eskens, Thomas Forkmann, Siegfried Gauggel, Susan Gruber, Andreas Jansen, Thilo Kellermann, Isabelle Reinhardt, Nina Vercamer- Fabri; *Dresden (coordinating site for data collection, analysis, and the RCT):* Franziska Einsle, Christine Fröhlich, Andrew T. Gloster, Christina Hauke, Simone Heinze, Michael Höfler, Ulrike Lueken, Peter Neudeck, Stephanie Preiß, Dorte Westphal; *Würzburg Psychiatry Department (coordinating center for genetics):* Andreas Reif; *Würzburg Psychology Department:* Julia Dürner, Hedwig Eisenbarth, Antje B. M. Gerdes, Harald Krebs, Paul Pauli, Silvia Schad, Nina Steinhäuser; *Bremen*: Veronika Bamann, Sylvia Helbig-Lang, Anne Kordt, Pia Ley, Franz Petermann, Eva-Maria Schröder. *Additional support was provided by the coordinating center for clinical studies in Dresden (KKS Dresden)*: Xina Grählert and Marko Käppler.

The RTC project was approved by the Ethics Committee of the Medical Faculty of the Technical University of Dresden (EK 164082006). The neuroimaging components were approved by the Ethics Committee of the Medical Faculty of the Rheinisch-Westfälische Hochschule University Aachen (EK 073/07). The experimental pharmacology study was approved by the Ethics Committee of the state of Berlin (EudraCT: 2006-00-4860-29).

The study was registered with the ISRCTN: ISRCTN80046034.

References

Andersch, S., & Hetta, J. (2003). A 15-year follow-up study of patients with panic disorder. *European psychiatry : The journal of the Association of European Psychiatrists, 18*(8), 401-408.

Bandelow, B. (1999). *Panic and Agoraphobia-Scale (PAS)*. Seattle: Hogrefe & Huber Publishers

Bandelow, B., Baldwin, D. S., Dolberg, O. T., Andersen, H. F., & Stein, D. J. (2006). What is the threshold for symptomatic response and remission for major depressive disorder, panic disorder, social anxiety disorder, and generalized anxiety disorder? *Journal of Clinical Psychiatry, 67,* 1428-1434.

Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2000). Cognitive-behavioral therapy, Imipramine, or their combination for panic disorder: A randomized controlled trial. *Journal of American Medical Association, 283*(19), 2529-2536.

Beck, A.T., Steer, R.A., & Brow,n G.K. (1996). *Manual for the Beck Depression Inventory-II.* San Antonio, TX: Psychological Corporation.

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., . . . Zettle, R. D. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire-II: a revised measure of psychological inflexibility and experiential avoidance. *Behavior therapy, 42*, 676-688. doi: 10.1016/j.beth.2011.03.007

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review, 26*(1), 17-31.

Cammin-Nowak, S., Helbig-Lang, S., Lang, T., Gloster, A.T. Fehm, L., Gerlach, A., Ströhle, A., …Wittchen, H.U. (2013). Specificity of homework compliance effects on treatment

outcome in CBT: Evidence from a controlled trial on panic disorder and agoraphobia. *Journal of Clinical Psychology.*

Chambless, D. L., Caputo, G. C., Jasin, S. E., Gracely, E. J., & Williams, C. (1985). The mobility inventory for agoraphobia *Behaviour Research and Therapy, 23*(1), 35-44.

Chambless, D. L., Sharpless, B. A., Rodriguez, D., McCarthy, K. S., Milrod, B. L., Khalsa, S., & Barber, J. P. (2011). Psychometric properties of the Mobility Inventory for Agoraphobia: Convergent, discriminant, and criterion related validity. *Behavior Therapy, 42,* 689-699.

Cottraux, J., Note, I. D., Cungi, C., Legeron, P., Heim, F., Chneiweiss, L., . . . Bouvard, M. (1995). A Controlled-Study of Cognitive-Behavior Therapy with Buspirone or Placebo in Panic Disorder with Agoraphobia. *British Journal of Psychiatry, 167,* 635-641.

Craske, M., Brown, T., & Barlow, D. Behavioral treatment of panic disorder: A two-year follow-up. *Behavior Therapy, 22,* 289-304.

Craske, M. G., Kircanski, K., Epstein, A., Wittchen, H.-U., Pine, D. S., Lewis-Fernandez, R., Hinton, D., & DSM V Anxiety, OC Spectrum, Posttraumatic and Dissociative Disorder Work Group. (2010). Panic Disorder: A review of DSM-IV Panic Disorder and proposals for DSM-V. *Depression & Anxiety, 27,* 3-112.

Craske, M. G., Kircanski, K., Zelikowsky, M., Mystkowski, J., Chowdhury, N., & Baker, A. (2008). Optimizing inhibitory learning during exposure therapy. *Behaviour Research and Therapy, 46,* 5-27.

de Beurs, E., van Balkom, A. J., Van Dyck, R., & Lange, A. (1999). Long-term outcome of pharmacological and psychological treatment for panic disorder with agoraphobia: a 2-year naturalistic follow-up. *Acta psychiatrica Scandinavica, 99*(1), 59-67.

Durham, R. C., Chambers, J. A., Power, K. G., Sharp, D. M., Macdonald, R. R., Major, K. A., et al. (2005). Long-term outcome of cognitive behaviour therapy clinical trials in central

Scotland. *Health Technology Assessment, 9*(42), 1-174.

Durham, R. C., Higgins, C., Chambers, J. A., Swan, J. S., & Dow, M. G. T. (2012). Long-term outcome of eight clinical trials of CBT for anxiety disorders: Symptom profile of sustained recovery and treatment-resistant groups. *Journal of Affective Disorders, 136*(3), 875-881. doi: Doi 10.1016/J.Jad.2011.09.017

Emmrich, A., Wittchen, H.-U., Lueken, U., Einsle, F., Gloster, A. T., Helbig-Lang, S., Beesdo-Baum, K. (in review). Exposure exercise completion in the treatment of panic disorder with agoraphobia: The role of depressive symptomatology.

Eifert, G. H., & Forsyth, J. P. (2005). *Acceptance & commitment therapy for anxiety disorders: A practitioner's treatment guide to using mindfulness, acceptance, and values-based behavior change strategies.* Oakland, CA: New Harbinger.

Essau, C. A., & Wittchen, H-U. (1993). An overview of the Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research, 3,* 79-85.

Fava, G. A., Zielezny, M., Savron, G., & Grandi, S. (1995). Long-term effects of behavioural treatment for panic disorder with agoraphobia. *British Journal of Psychiatry, 166,* 87-92.

Fava, G. A., Rafanelli, C., Grandi, S., Conti, S., Ruini, C., Mangelli, L., & Belluardo, P. (2001). Long-term outcome of panic disorder with agoraphobia treated by exposure. *Psychological Medicine, 31,* 891-898.

Gloster, A. T., Klotsche, J., Gerlach, A. L., Hamm, A., Ströhle, A., Gauggel, S., Kircher, T., Alpers, G., Deckert, J., & Wittchen, H.-U., (in press). Timing Matters: Mediators of Outcomes in Cognitive Behavioral Therapy for Panic Disorder with Agoraphobia Depend on the Stage of Treatment. *Journal of Consulting and Clinical Psychology.*

Gloster, A. T., Klotsche, J., Chacker, S., Hummel, K., & Hoyer, J. (2011). Assessing psychological flexibility: What does it add above and beyond existing constructs? *Psychological Assessment, 23,* 970-982.

Gloster, A. T., Wittchen, H-U., Einsle, F., Hoefler, M., Lang, T., Helbig-Lang, S….Arolt, V. (2009). Mechanism of action in CBT (MAC): Methods of a multi-center randomized controlled trial in 369 patients with panic disorder and agoraphobia. *European Archives of Psychiatry and Clinical Neuroscience, 259, S155-S166.*

Gloster, A. T., Wittchen, H.-U., Einsle, F., Lang, T., Helbig-Lang, S., Fydrich, T., . . . Arolt, V. (2011). Psychological treatment for panic disorder with agoraphobia: A randomized controlled trial to examine the role of therapist-guided exposure in-situ in CBT. Journal of Consulting and Clinical Psychology, 79, 406–420. doi:10.1037/a0023584

Guy, W. (1976). *ECDEU Assessment Manual for Psychopharmacology*. Rockville, MD: U.S. Department of Health, Education, and Welfare.

Hautzinger, M., Keller, F., & Kühner, C. (2006). *Beck Depression Inventar II (BDI 2)*. Frankfurt: Harcourt Test Service.

Heldt, E., Blaya, C., Kipper, L., Salum, G. A., Otto, M. W., & Manfro, G. G. (2007). Defense mechanisms after brief cognitive-behavior group therapy for panic disorder. *Journal of Nervous and Mental Disease, 195,* 540-543.

Jacobson, N. S., & Truax, P. (1991). Clinical-Significance - a Statistical Approach to Defining Meaningful Change in Psychotherapy-Research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.

Lachner, G., Wittchen, H-U., Perkonigg, A., Holly, A., Schuster, P., Wunderlich, U., Türk, D., Garczynski, E., & Pfister, H. (1988). Structure, content and reliability of the Munich-

Composite International Diagnostic Interview (M-CIDI) substance use sections. *European Addiction Research, 4,* 28-41.

Lang, T., Helbig-Lang, S., Westphal, D., Gloster, A.T., & Wittchen, H.-U. (2012). Expositionsbasierte Therapie der Panikstörung mit Agoraphobie: Ein Behandlungsmanual [Exposure-based Therapy of Panic Disorder with Agoraphobia: A Treatment Manual]. Hogrefe: Göttingen.

Marchand, A., Roberge, P., Primiano, S., & Germain, V. (2009). A randomized, controlled clinical trail of standard, group and brief cognitive-behavioral therapy for panic disorder with agoraphobia: A two-year follow-up. *Journal of Anxiety Disorders, 23,* 1139-1147.

Pollack, M., Mangano, R., Entsuah, R., Tzanis, E., & Simon, N. M. (2007). A randomized controlled trial of venlafaxine ER and paroxetine in the treatment of outpatients with panic disorder. *Psychopharmacology, 194*, 233-242.

Pollack, M. H., Otto, M. W., Roy-Byrne, P. P., Coplan, J. D., Rothbaum, B. O., Simon, N. M., & Gorman, J. M. (2008). Novel treatment approaches for refractory anxiety disorders. *Depression and Anxiety, 25,* 467-476.

Powers, M. B., Smits, J. A. J., & Telch, M. J. (2004). Disentangling the effects of safety-behavior utilization and safety-behavior availability during exposure-based treatment: A placebo-controlled trial. *Journal of Consulting and Clinical Psychology, 72,* 448–454. doi:10.1037/0022-006X.72.3.448

Reed, V., Gander, F., Pfister, H., Steiger, A., Sonntag, H., Trenkwalder, C., . . . Wittchen, H.-U. (1998). To what degree does the Composite International Diagnostic Interview (CIDI) correctly identify *DSM–IV* disorders? Testing validity issues in a clinical sample. *International Journal of Methods in Psychiatric Research, 7,* 142–155. doi:10.1002/mpr.44

Reif, A., Richter, J., Straube, B., Höfler, M., Lueken, U., Gloster, A.T., …Deckert, J. (2013). MAOA and mechanisms of panic disorder revisited: From bench to molecular psychotherapy. *Molecular Psychiatry.* doi: 10.1038/mp.2012.172

Reiss, S., Peterson, R.P., Gursky, D.M., and McNally, R.J. (1986). Anxiety sensitivity, anxiety frequency, and the prediction of fearfulness. *Behavior Research and Therapy, 24,* 1-8.

Robins, J. N., Wing, J., Wittchen, H-U., Helzer, J. E., Babor, T. F., Burke, J., et al. (1988). The Composite International Diagnostic Interview: An epidemiologic instrument suitable for the use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry, 45,* 1069-1077.

Royall, R. (1986). Model robust inference using maximum likelihood estimators. *International Statistical Review. 54,* 221-226.

Sánchez-Meca, J., Rosa-Alcázar, A. I., Marín-Martínez, F., & Gómez-Conesa, A. (2010). Psychological treatment of panic disorder with or without agoraphobia: A meta-analysis. *Clinical Psychology Review, 30,* 37-50.

Schlaepfer, T. E., Agren, H., Monteleone, P., Gasto, C., Pitchot, W., Rouillon, F., Nutt, D. J., & Kasper, S. (2012). The hidden third: Improving outcome in treatment-resistant depression. *Journal of Psychopharmacology, 0,* 1-16.

Skrondal, A., & Rabe-Hesketh, S. (2004). Generalized Latent Variable Modeling. Chapman & Hall.

Shear, M. K., Vander Bilt, J., Rucci, P., Endicott, J., Lydiard, B., Otto, M. W., et al. (2001). Reliability and validity of a Structured Interview Guide for the Hamilton Anxiety Rating Scale (SIGH-A). *Depression and Anxiety, 13*(4), 166-178.

Smits, J. A. J., Powers, M. B., Cho, Y. R., & Telch, M. J. (2004). Mechanism of change in cognitive–behavioral treatment of panic dis- order: Evidence for the fear of fear

meditational hypothesis. *Journal of Consulting and Clinical Psychology, 72,* 646–652. doi:10.1037/0022- 006X.72.4.646

StataCorp: Stata Statistical Software, Release 12.1. College Station, TX: Stata Corporation 2012.

Williams, S. L., & Falbo, J. (1996). Cognitive and performance-based treatments for panic attacks in people with varying degrees of agoraphobic disability. *Behaviour Research and Therapy, 34*(3), 253-264.

Wittchen, H-U. (1994). Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): A critical review. *Journal of Psychiatric Research, 28,* 57-84.

Wittchen, H.-U., Gloster, A. T., Beesdo-Baum, K., Fava, G. A., & Craske, M. G. (2010). Agoraphobia: A review of the diagnostic classificatory position and criteria. *Depression & Anxiety, 27,* 113-133.

Wittchen, H.-U., & Pfister, H. (1997) DIA-X Interview. Instruktionsmanual zur Durchführung von DIA-X-Interviews (Instruction manual for the DIA-X-Interview). Swets & Zeitlinger, Frankfurt.

Wittchen, H.-U., Nocon, A., Beesdo, K., Pine, D. S., Hofler, M., Lieb, R., & Gloster, A. T. (2008). Agoraphobia and panic: Prospective-longitudinal relations suggest a rethinking of diagnostic concepts. *Psychotherapy and Psychosomatics, 77,* 147–157. doi:10.1159/000116608

Wood, A. M., Hillsdon, M., & Carpenter J. (2005). Comparison of imputation and modeling methods in the analysis of a physical activity trial with missing outcomes. *International Journal of Epidemiology, 34,* 89-99.

Table 1. Model-based (mixed effects models) predicted means and effect sizes for main outcomes for T+ (n=163) & T-(n=138)

		Treated		T+		T-		Effect sizes within treated		Effect sizes within T+		Effect sizes within T-		Effect sizes betweeen T+ and T-	
Outcome	**Assessment**	**N**	**Mean**	**N**	**Mean**	**N**	**Mean**	**ES**	**p-value**	**ES**	**p-value**	**ES**	**p-value**	**ES**	**p-value**
HAM-A Total	Pre	301	24.5	163	24.7	138	24.2	*Ref.*		*Ref.*		*Ref.*		*Ref.*	
	Intermediate	301	24.5	/	/	/	/	/	/	/	/	/	/	/	/
	Post	301	12.8	163	12.8	138	12.9	-2.21	0.000	-2.26	0.000	-2.16	0.000	0.11	0.527
	FU 6 months	301	10.4	163	10.2	138	10.6	-2.69	0.000	-2.77	0.000	-2.59	0.000	0.18	0.350
	FU 24 months	301	13.1	163	13.3	138	12.8	-2.17	0.000	-2.17	0.000	-2.18	0.000	-0.01	0.981
	FU 24 - Post	301	0.2	163	0.5	138	-0.1	0.04	0.818	0.09	0.740	-0.02	0.903	-0.12	0.732
	FU 24 - FU6	301	2.7	163	3.1	138	2.2	0.52	0.003	0.60	0.026	0.41	0.053	-0.18	0.591
CGI Total	Pre	301	5.3	163	5.4	138	5.2	*Ref.*		*Ref.*		*Ref.*		*Ref.*	
	Intermediate	301	4.8	163	4.8	138	4.9	-0.64	0.000	-0.81	0.000	-0.43	0.000	0.38	0.013
	Post	301	3.7	163	3.6	138	3.8	-2.32	0.000	-2.59	0.000	-2.00	0.000	0.59	0.003
	FU 6 months	301	2.8	163	2.7	138	2.9	-3.63	0.000	-3.91	0.000	-3.29	0.000	0.61	0.013
	FU 24 months	301	3.1	163	3.1	138	3.1	-3.09	0.000	-3.25	0.000	-2.90	0.000	0.35	0.290
	FU 24 - Post	301	-0.5	163	-0.5	138	-0.6	-0.77	0.000	-0.66	0.005	-0.90	0.000	-0.24	0.468
	FU 24 - FU6	301	0.4	163	0.5	138	0.3	0.54	0.002	0.66	0.006	0.40	0.123	-0.26	0.454
# panic attacks	Pre	301	2.6	163	2.7	138	2.4	*Ref.*		*Ref.*		*Ref.*		*Ref.*	
	Intermediate	301	2.0	163	2.2	138	1.8	-0.24	0.000	-0.22	0.004	-0.26	0.001	-0.04	0.691
	Post	301	1.1	163	1.2	138	1.0	-0.62	0.000	-0.66	0.000	-0.58	0.000	0.07	0.553
	FU 6 months	301	0.5	163	0.4	138	0.5	-0.89	0.000	-0.99	0.000	-0.79	0.000	0.20	0.110
	FU 24 months	301	0.7	163	0.9	138	0.5	-0.78	0.000	-0.77	0.000	-0.78	0.000	0.00	0.987
	FU 24 - Post	301	-0.4	163	-0.3	138	-0.5	-0.15	0.024	-0.12	0.242	-0.19	0.023	-0.07	0.575
	FU 24 - FU6	301	0.3	163	0.5	138	0.0	0.12	0.066	0.21	0.031	0.01	0.906	-0.20	0.108
MI-Unaccompagnied	Pre	301	3.0	163	3.0	138	2.9	*Ref.*		*Ref.*		*Ref.*		*Ref.*	
	Intermediate	301	2.8	163	2.8	138	2.7	-0.27	0.000	-0.29	0.000	-0.24	0.000	0.05	0.503
	Post	301	2.0	163	1.9	138	2.1	-1.20	0.000	-1.38	0.000	-0.98	0.000	0.40	0.000
	FU 6 months	301	1.5	163	1.4	138	1.6	-1.77	0.000	-1.89	0.000	-1.61	0.000	0.28	0.042
	FU 24 months	301	1.6	163	1.5	138	1.7	-1.61	0.000	-1.77	0.000	-1.40	0.000	0.37	0.039
	FU 24 - Post	301	-0.3	163	-0.3	138	-0.4	-0.41	0.000	-0.39	0.002	-0.42	0.000	-0.03	0.840
	FU 24 - FU6	301	0.1	163	0.1	138	0.2	0.16	0.036	0.12	0.291	0.21	0.030	0.09	0.527
PAS Total	Pre	301	27.8	163	28.4	138	27.1	*Ref.*		*Ref.*		*Ref.*		*Ref.*	
	Intermediate	301	23.1	163	23.5	138	22.6	-0.48	0.000	-0.50	0.000	-0.46	0.000	0.04	0.692
	Post	301	14.5	163	14.4	138	14.6	-1.37	0.000	-1.43	0.000	-1.29	0.000	0.14	0.231
	FU 6 months	301	8.9	163	8.4	138	9.6	-1.94	0.000	-2.05	0.000	-1.80	0.000	0.25	0.063
	FU 24 months	301	11.3	163	11.7	138	10.9	-1.69	0.000	-1.71	0.000	-1.67	0.000	0.04	0.839
	FU 24 - Post	301	-3.2	163	-2.7	138	-3.7	-0.32	0.001	-0.28	0.065	-0.38	0.003	-0.10	0.609
	FU 24 - FU6	301	2.4	163	3.3	138	1.3	0.25	0.015	0.34	0.019	0.14	0.328	-0.20	0.311

Note: **/ = Not assessed at this time point; * Difference of within effect sizes**

Table 2: Status Across Successive Measurement Points

	Retained Positive Status		Worsened		Retained Negative Status		Improved	
	Post - FU6	FU6 - FU24	Post - FU6	FU6 - FU24	Post - FU6	FU6 - FU24	Post - FU6	FU6 - FU24
Outcome	% (n)	% (n)	% (n)	% (n)	% (n)	% (n)	% (n)	% (n)
HAM-A	84.5% (49)	69.4% (43)	15.5% (9)	30.7% (19)	49.1% (27)	61.5% (16)	50.9% (28)	38.5% (10)
CGI	89.7% (52)	89.8% (79)	10.3% (6)	10.2% (9)	35.0% (21)	12.5% (3)	65.0% (39)	87.5% (21)
PAS	89.3% (50)	80.7% (50)	10.7% (6)	19.4% (12)	42.1% (24)	36.4% (8)	57.9% (33)	63.6% (14)
MI	96.4% (54)	87.5% (56)	03.6% (2)	12.5% (8)	38.5% (15)	73.3% (11)	61.5% (24)	26.7% (4)

Note. Status defined according to response definitions: HAM-A (50% reduction from BL); CGI ("mild" or less); PAS (50% reduction from BL); MI (≤ 1.8)

Table 3: Proportion Seeking Additional Treatment Following Post-treatment and Impact

	Expert Rated as Responder at FU-24 (CGI)						*Expert Rated as Non-Responder at FU-24 (CGI)*						
Outcome	(A) No Additional Tx			(B) Received Additional Tx			(C) No Additional Tx			(D) Received Additional Tx			Significant Contrasts
<u>Dimensional</u>	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	
HAM-A (mean)	54	10.8	9.0	34	12.6	9.2	8	24.3	11.2	5	22.4	14.7	A,B < C
PAS	52	8.6	8.9	26	14.0	10.7	6	26.2	13.1	5	15.8	5.8	A < B,C,D; B,D* <C
P. Attacks	58	0.5	1.1	32	0.4	0.9	7	4.0	2.4	5	0.8	1.3	A,B,D < C
MI	51	1.4	0.6	29	1.7	0.8	6	2.2	1.1	5	2.2	0.5	A<B*,C*,D; B<D*
<u>Response Rate</u>	Ntot	N	%	Ntot	N	%	Ntot	N	%	Ntot	N	%	
HAM-A	54	34	63.0	34	22	64.7	8	0	0.0	5	2	40.0	A,B > C
PAS	52	41	78.8	26	15	57.7	6	1	16.7	5	4	80.0	A > B*, C; D* > C
P. Attacks	58	42	72.4	32	24	75.0	7	1	14.3	5	3	60.0	A,B > C
MI	51	43	84.3	29	18	62.1	6	3	50.0	5	1	20.0	A > B,C*,D

Note: n's vary across questionnaires due to missing values; significant contrasts were listed if $p < .05$; given the small cell sizes of some comparisons trends (*) were also listed when $p < .10$

Table 4: Concordance Between Subjective Definition and Expert Rating of Treatment Outcome

Outcome	Expert Rated as Responder at FU-24 (CGI) (A) Don't Need Additional Tx			(B) Need Additional Tx			Expert Rated as Non-Responder at FU-24 (CGI) (C) Don't Need Additional Tx			(D) Need Additional Tx			Significant Contrasts
<u>Dimensional</u>	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	
HAM-A (mean)	59	9.6	8.6	31	15.2	8.6	2	12.0	4.3	11	25.7	11.9	A< B; A,B,C < D
PAS	56	8.0	7.9	26	14.9	11.4	3	17.3	7.6	10	24.6	15.2	A< B,C,D; B < D*
P. Attacks	64	0.4	0.8	31	0.7	1.4	3	1.3	1.5	11	3.3	2.8	A,B,C* < D
MI	57	1.5	0.6	27	1.6	0.8	3	2.1	1.0	10	2.2	0.9	A,B < D
<u>Response Rate</u>	Ntot	N	%	Ntot	N	%	Ntot	N	%	Ntot	N	%	
HAM-A	59	41	69.5	31	15	48.4	2	1	50.0	11	1	9.1	A > B*,D; B > D
PAS	56	45	80.4	26	15	57.7	3	2	66.7	10	4	40.0	A > B*,D
P. Attacks	64	49	76.6	31	21	67.7	3	1	33.3	11	3	27.3	A,B > D
MI	57	43	75.4	27	22	81.5	3	1	33.3	10	4	40.0	A*,B > D

Note: n's vary across questionnaires due to missing values; significant contrasts were listed if $p < .05$; given the small cell sizes of some comparisons trends (*) were also listed when $p < .10$