1. (1) Do not use WLC as the control condition, since criterion I requires a placebo or another treatment.
2. (2) Do not use TAU as the control condition, since the methodological problems described above are so

   extensive.

3. (3) Use an active treatment as comparison, preferably one that has been established as effective for the

   disorder in question.

4. (4) Do a proper power analysis before the start of the study and adjust the cell size for the attrition that may

   occur.

5. (5) Use a representative sample of patients, diagnose them using suitable instruments in the hands of trained

   interviewers, and test the diagnostic reliability.

6. (6) Let an independent researcher or agency use an unobjectionable randomization procedure, and conceal

   the outcome of it from all persons involved in the study.

7. (7) Use reliable and valid outcome measures; both the ones that are specific to the disorder and general ones.
8. (8) Use blind assessors and evaluate their blindness regarding treatment condition of the patients they

   assess.

9. (9) Train the assessors properly and measure inter-rater reliability on the data collected throughout the

   study (not just during training).

10. (10) Use three or more properly trained therapists and randomize patients to therapist to enable an analysis

    of possible therapist effect on the outcome.

11. (11) Include at least a 1-year follow-up in the study and assess any nonprotocol treatments that the patients

    may have obtained during the follow-up period.

12. (12) Audio- or videotape all therapy sessions. Randomly select 20% of these and let independent experts rate

adherence to treatment manual and therapist competence.

13. (13) Insert procedures to control for concomitant treatments that patients in the study may obtain

    simultaneously as the protocol treatment.

14. (14) Describe the attrition, do a drop-out analysis and include all randomized subjects in an intent-to-treat

    analysis.

15. (15) Assess clinical significance of the improvement of the primary measures.

Appendix A. Psychotherapy outcome study methodology rating form[3]

Note: If not enough information is given regarding a specific item a rating of 0 is given.

1. Clarity of sample description
0 Poor. Vague description of sample (e.g. only mentioned whether patients were diagnosed with the disorder).

1. 1 Fair. Fair description of sample (e.g. mentioned inclusion/exclusion criteria, demographics, etc.).
2. 2 Good. Good description of sample (e.g. mentioned inclusion/exclusion criteria, demographics,

and the prevalence of comorbid disorders).

2. Severity/chronicity of the disorder
0 Poor. Severity/chronicity was not reported and/or subsyndromal patients were included in the sample.
1 Fair. All patients met the criteria for the disorder. Sample includes acute (o1 yr) and/or low severity.
2 Good. Sample consisted entirely of chronic (41 yr) patients of at least moderate severity.

3. Representativeness of the sample
0 Poor. Sample is very different from patients seeking treatment for the disorder (e.g. there are excessively strict exclusion criteria).
1 Fair. Sample is somewhat representative of patients seeking treatment for the disorder (e.g. patients were only excluded if they met criteria for other major disorders).
2 Good. Sample is very representative of patients seeking treatment for the disorder (e.g. authors made efforts to ensure representativeness of sample).

4. Reliability of the diagnosis in question
0 Poor. The diagnostic process was not reported, or not assessed with structured interviews by a trained interviewer.
1 Fair. The diagnosis was assessed with structured interview by a trained interviewer.
2 Good. The diagnosis was assessed with structured interview by a trained interviewer and adequate inter-rater reliability was demonstrated (e.g. kappa coefficient).

5. Specificity of outcome measures

1. 0 Poor. Very broad outcome measures, not specific to the disorder (e.g. SCL-90R total score).
2. 1 Fair. Moderately specific outcome measures.
3. 2 Good. Specific outcome measures, such as a measure for each symptom cluster.

6. Reliability and validity of outcome measures
0 Poor. Measures have unknown psychometric properties, or properties that fail to meet current standards of acceptability.

1. 1 Fair. Some, but not all measures have known or adequate psychometric properties.
2. 2 Good. All measures have good psychometric properties. The outcome measures are the best

available for the authors' purpose.

7. Use of blind evaluators
0 Poor. Blind assessor was not used (e.g. assessor was the therapist, assessor was not blind to treatment condition, or the authors do not specify).

1. 1 Fair. Blind assessor was used, but no checks were used to assess the blind.
2. 2 Good. Blind assessor was used in correct fashion. Checks were used to assess whether the

assessor was aware of treatment condition.

8. Assessor training

1. 0 Poor. Assessor training and accuracy are not specified, or are unacceptable.
2. 1 Fair. Minimum criterion for assessor training is specified (e.g. assessor has had specific training

in the use of the outcome measure), but accuracy is not monitored or reported.
2 Good. Minimum criterion of assessor training is specified. Inter-rater reliability was checked, and/or assessment procedures were calibrated during the study to prevent evaluator drift.

9. Assignment to treatment
0 Poor. Biased assignment, e.g. patients selected their own therapy or were assigned in another non-random fashion, or there is only one group.
1 Fair. Random or stratified assignment. There may be some systematic bias but not enough to pose a serious threat to internal validity. There may be therapist by treatment confounds. N may be too small to protect against bias.
2 Good. Random or stratified assignment, and patients are randomly assigned to therapists within condition. When theoretically different treatments are used, each treatment is provided by a large enough number of different therapists. N is large enough to protect against bias.

10. Design

1. 0 Poor. Active treatment vs. WLC, or briefly described TAU.
2. 1 Fair. Active treatment vs. TAU with good description, or placebo condition.
3. 2 Good. Active treatment vs. another previously empirically documented active treatment.

11. Power analysis

1. 0 Poor. No power analysis was made prior to the initiation of the study.
2. 1 Fair. A power analysis based on an estimated effect size was used.
3. 2 Good. A data-informed power analysis was made and the sample size was decided accordingly.

12. Assessment points

1. 0 Poor. Only pre- and post-treatment, or pre- and follow-up.
2. 1 Fair. Pre-, post-, and follow-up o1 year.
3. 2 Good. Pre-, post-, and follow-up X1 year.

13. Manualized, replicable, specific treatment programs
0 Poor. Description of treatment procedure is unclear, and treatment is not based on a publicly available, detailed treatment manual. Patients may be receiving multiple forms of treatment at once in an uncontrolled manner.
1 Fair. Treatment is not designed for the disorder, or description of the treatment is generally clear and based on a publicly available, detailed treatment manual, but there are some ambiguities about the procedure. Patients may have received additional forms of treatment, but this is balanced between groups or otherwise controlled.
2 Good. Treatment is designed for the disorder. A detailed treatment manual is available, and/or treatment is explained in sufficient detail for replication. No ambiguities about the treatment procedure. Patients receive only the treatment in question.

14. Number of therapists

1. 0 Poor. Only one therapist, i.e. complete confounding between therapy and therapist.
2. 1 Fair. At least two therapists, but the effect of therapist on outcome is not analyzed.
3. 2 Good. Three, or more therapists, and the effect of therapist on outcome is analyzed.

15. Therapist training/experience

1. 0 Poor. Very limited clinical experience of the treatment and/or disorder (e.g. students).
2. 1 Fair. Some clinical experience of the treatment and/or disorder.
3. 2 Good. Long clinical experience of the treatment and the disorder (e.g. practicing therapists).

16. Checks for treatment adherence

1. 0 Poor. No checks were made to assure that the intervention was consistent with protocol.
2. 1 Fair. Some checks were made (e.g. assessed a proportion of therapy tapes).
3. 2 Good. Frequent checks were made (e.g. weekly supervision of each session using a detailed rating

form).

17. Checks for therapist competence

1. 0 Poor. No checks were made to assure that the intervention was delivered competently.
2. 1 Fair. Some checks were made (e.g. assessed a proportion of therapy tapes).
3. 2 Good. Frequent checks were made (e.g. weekly supervision of each session using a detailed rating

form).

18. Control of concomitant treatments (e.g. medications)
0 Poor. No attempt to control for concomitant treatments, or no information about concomitant treatments provided. Patients may have been receiving other forms of treatment in addition to the study treatment.
1 Fair. Asked patients to keep medications stable and/or to discontinue other psychological therapies during the treatment.

2 Good. Ensured that patients did not receive any other treatments (medical or psychological) during the study.

19. Handling of attrition
0 Poor. Proportions of attrition are not described, or described but no dropout analysis is performed.
1 Fair. Proportions of attrition are described, and dropout analysis or intent-to-treat analysis is performed.
2 Good. No attrition, or proportions of attrition are described, dropout analysis is performed, and results are presented as intent-to-treat analysis.

20. Statistical analyses and presentation of results

1. 0 Poor. Inadequate statistical methods are used and/or data are not fully presented.
2. 1 Fair. Adequate statistical methods are used but data are not fully presented.
3. 2 Good. Adequate statistical methods are used and data are presented with M and SD.

21. Clinical significance

1. 0 Poor. No presentation of clinical significance was done.
2. 1 Fair. An arbitrary criterion for clinical significance was used and the conditions were compared

regarding percent clinically improved.
2 Good. Jacobson's criteria for clinical significance were used and presented for a selection (or all) of the outcome measures, and conditions were compared regarding percent clinically improved.

22. Equality of therapy hours (for non-WLC designs only)

1. 0 Poor. Conditions differ markedly (X20% difference in therapy hours).
2. 1 Fair. Conditions differ somewhat (10–19% difference in therapy hours).
3. 2 Good. Conditions do not differ (o10% difference in therapy hours).