Eur Arch Psychiatry Clin Neurosci (1990) 240:103–108

European Archives of Psychiatry and Clinical Neuroscience

Self- and Observer Assessment in Anxiolytic Drug Trials: A Comparison of Their Validity

W. Maier[1], M. Albus[2], R. Buller[1], D. Nutzinger[4], D. Shera[3], and P. Bech[5]

[1]Department of Psychiatry, University of Mainz, Untere Zahlbacher Strasse 8, W-6500 Mainz, Federal Republic of Germany
[2]Department of Psychiatry, University of Munich, Nußbaumstrasse 7, W-8000 Munich, Federal Republic of Germany
[3]Massachusetts General Hospital, Biostatistics, Fruit Street, Boston, MA 02114, USA
[4]Department of Psychiatry, University of Vienna, Währinger Gürtel 18–20, A-1090 Wien, Austria
[5]Frederiksborg Amts Centralsygehus, 48, Dyrehavevej, DK-3400 Hillerød, Denmark

Received January 4, 1990

Summary. Self-rating scales are considered to be less useful for comparing different treatments in anxiety patients than observer-rating scales. However, the empirical evidence for this assumption is not adequate. A self-rating inventory of 35 items related to anxiety was perfectly parallel with an observer-rating inventory. Both instruments were used in the Cross National Collaborative Panic Study to compare the efficacy of imipramine, alprazolam and placebo in an 8-week drug trial in a sample of 1168 outpatients. The variance of the self-rating assessments was about two times higher. Both scales were equally sensitive to change; however, the measurement of change by means of the self-rating scale was slightly less consistent. The discriminative power of the observer-rating scale between placebo and active treatment was two to three times higher than that of the self-rating scale; consequently the observer-rating procedure provides a more valid instrument when the efficacies of different anxiolytic treatments are compared between different groups of patients.

Key words: Panic disorder – Rating scales – Drug trials – Placebo

Introduction

Observer-rating as well as self-rating scales are widely used in controlled drug trials for the measurement of drug effects. Self-rating scales document the patient's direct and uncommitted view of the complaints, which is the major focus for treatment. Self-rating scales have only a limited application in severe psychiatric conditions and in non-compliant patients; in contrast, however, they have a somewhat unlimited applicability in minor disorders such as the anxiety disorders. They might serve as outcome measures by weighting the complaints of the patient in a manner different from the observer-rating scales. As self-rating and observer-rating scales refer to the measurements of similar or identical psychopathological entities, there is no reason to establish a hierarchical relationship between the two types of scales, at least under conditions of equal applicability.

Beyond these aspects of apparent validity the major discriminating yardstick for measures of changes in drug trials are two validation criteria: the sensitivity to change and the discriminative power between two treatments with different efficacy. It is generally agreed that the self-rating scales have a relatively low discriminative power because of their wider range of variance (Angst et al. 1989). A review article by Glass et al. (1987) reported that the Hamilton Anxiety Scale (Hamilton 1969), which is rated by physicians, is more useful in detecting differences in drug effects between benzodiazepines and placebo in anxiety disorders than the anxiety subscales of the Syndrome Checklist 90 (SCL-90) (Derogatis et al. 1973), which is rated by patients. However, this is not necessarily a disadvantage of self-rating scales because the efficacy of all those benzodiazepines reviewed is not proven convincingly; therefore, the observer-rating scale might over-report positive findings. Since no further reports of the estimate validity of self- and observer-rating scales in anxiety have been published in the literature, it is necessary to compare the discriminative power between placebo and treatment of active compounds with well established efficacy.

Another aspect limits the comparability between observer and self-rating scales: usually the two types of scales administered simultaneously in a drug trial are different with regard to the number, the content, the time frame that has to be taken into account as well as the scoring of the particular items of the scale; additionally self-rating scales are less elaborate and the item description is less specific compared with observer-rating scales. Self-rating scales may therefore measure a different entity than observer-rating scales, even if both refer to the identical psychopathological concept. To clarify this

Offprint requests to: W. Maier

point validation studies with parallel scales for the patients' and the observers' assessment are necessary.

The Cross National Collaborative Panic Study – Phase II – (CNCPS) (Klerman 1988) provides a unique opportunity to compare the validity of measuring change between observer and self-rating anxiety scales. The two major features of this study enabling the comparison are:

1. Placebo is tested against alprazolam and imipramine in an 8-week drug trial; both of these active compounds have been shown to be effective in the treatment of panic disorders in several previous drug trials (Ballenger et al. 1988; Lydiard and Ballenger 1988);

2. An inventory of 35 items used as a physician-rated scale (CRAS) and in a perfectly parallel version used as a self-rating scale (PRAS) was applied during this trial after the items had been validated cross-sectionally in a previous study (Ballenger et al. 1988); both scales were developed by Marks and Sheehan, referring to a previously proposed self-rating scale (Marks and Matthews 1979).

A comparison of different measures of the treatment effect requires criteria of validity for the measurement of change. No convention on those criteria of validity has been established up to now; the following three criteria are proposed as indicators of the clinical validity of the measurements of change and will be used as the criteria of validity:

1. Sensitivity to change by using a global rating of change by a rater as the criterion; a criterion for sensitivity should be rated independently of the measure to be validated. In order to consider self- and observer-rating in a comparable manner global self-rating is used as criterion for the observer-rating scale and global observer-rating is used as the criterion for the self-rating scale.

2. Precision of measurement to be evidenced:

A. By the variance of change scores where lower variance is interpreted as a higher degree of precision if the same quality is assessed

B. By the consistency between measures tapping the same quality rated by the same subject (either physician or patient); the correlation between two self-rating measures or the correlation between two observer-rating scales might be indicators of the consistency of either measure.

3. Discriminative power between two treatments with different efficacy as evidenced by previous drug trials; alprazolam and imipramine are significantly more effective in treating panic disorder than placebo according to previous drug trials (Lydiard and Ballenger 1987; Ballenger et al. 1988); consequently, the comparison of placebo with active (imipramine or alprazolam) treatment measured by the rating scale to be validated may serve as an indicator of the discriminative power of the rating scale applied.

Patients and Methods

The study procedure, diagnostic procedures and diagnoses were described extensively in the paper by Albus et al. (this issue).

Ratings and Reliability

The following rating scales were developed for the CNCPS (Ballenger et al. 1988) and used in this report: a global rating by the physician for the overall improvement relative to baseline and a global rating by the patient for the overall improvement relative to baseline (both global ratings ranged from 0 to 10, with 0 worst, 10 best and 5 indicating no change); an analogue global rating for the overall improvement by the patient; the CRAS (35 items) and the PRAS for anxiety (35 items). The CRAS and the PRAS as well as the training for the assurance of the reliability are described in Albus et al. (1990). In addition, the Hamilton Anxiety Scale (HAS) was rated by the physician and the Syndrome Check List 90 (SCL-90) was rated by the patient.

The CRAS and the PRAS were both developed for drug trials in panic and anxiety disorders; both scales are intended to measure the severity of panic disorder. Both scales are identical by the number of items ($n = 35$), the content, the description and the scoring of the items; each item has five levels ranging from 0 (no) to 4 (severe). The items tap all symptoms defining and related to anxiety including panic attacks, phobias, depression and obsessive-compulsive symptoms. The reliability of the Structured Clinical Interview for DSM-III Diagnoses-Upjohn Version (Spitzer and Williams 1983) – SCID-UP – and of the severity scales was monitored during common training of the investigators of all centres participating before starting the study and by the requirement for all participating centres to tape at least three SCID-UP interviews. The tapes were blindly evaluated. The reliability turned out to be sufficient for diagnoses and global scores of the scales (achieved by summing up the item scores) used in this report (kappa > 0.70).

The total set of ratings of the severity and of the changes of the symptomatology was administered by the investigator (treating physician) to the patient at baseline (before starting medication) and at the end of week 1, 3, 6 and 8.

Sample Size, Characteristics and Completion Rates

A total of 1168 outpatients with panic disorder recruited in 14 participating psychiatric departments (mainly university hospitals) were randomized after meeting the selection criteria. The sample size for the baseline evaluation was $n = 1102$ as $n = 66$ data sets were incomplete. In all, 1090 patients completed at least 3 weeks of the trial and had completed the set of ratings mentioned above (sample size for week 3 analysis); 812 stayed in the trial 8 weeks (completers) and had a complete data set of ratings; the sample size for end-point analysis was $n = 1090$. The completion rate was different for the three treatments: 82.6% for alprazolam, 69.8% for imipramine and 56.3% for placebo. For all patients the set of rating scales to be analysed in this report is complete. The outcome data by treatment groups are reported in Cross National Collaborative Panic Study – Second Phase Investigators (1990).

Statistical Analysis

The CNCPS phase II was an 8-week treatment trial. Therefore one method of analysis is to refer to the outcome measures obtained by the completers of the CNCPS (completer analysis). Equally the data for patients completing week 6 may be analysed using the week 6 data. However, this type of analysis is affected by the drop-outs during treatment. One way of analysing the data is the completer analysis based on all patients completing the trial after 8 weeks. In addition, another method of analysing the data, based on the endpoint analysis and including all patients evaluable (using the final rating before dropping out for endpoint analysis) was used; the sample for this kind of analysis is defined by the completers of at least 3 weeks of treatment. This procedure of multiple analyses allows monitoring for the bias due to drop-outs.

Criteria of Validity for Rating Scales

Three criteria of validity are tested: sensitivity to change, consistency and the discriminative power.

Sensitivity to change is determined by using the global assessment of change by the physician or the patient as the criterion of validity which is assessed independently from the scale to be validated: the physician-rated scale CRAS is correlated with global improvement rated by the patient and vice versa. The correlation between the measurements of change by the physician-rated CRAS and the physician-rated global improvement score indicates the consistency of the physician's ratings of change during treatment; the consistency of the patient's ratings may be indicated in an analogous manner.

The *discriminative power* of rating scales between placebo and active treatment is determined by two parameters:

A. The F value and the corresponding P value of the treatment factor (binary variable indicating placebo versus imipramine, alprazolam) obtained in an analysis of covariance using the outcome as described by the scale under study as the dependent variable; the baseline score of the scale under study is introduced as the covariate; independent variables are (1) a nominal variable indicating the sites; and (2) the treatment factor as a binary variable (placebo versus imipramine or alprazolam);
B. The effect size of the treatment factor described by a rating scale (i.e. the variance described by the outcome measure explained by the treatment factor as described in A divided by the variance of the total model); this quotient is equivalent to the proportion of the total variance of the outcome measure (residuals) described by a particular scale in the analysis of covariance explained by the treatment factor represented by the scale under study. The effect size is equivalent to the eta-square-coefficient as defined by Kerlinger (1964).

These analyses are carried out for week 1, 3, 6, 8 and for endpoint analysis separately in order to investigate if any of the two rating scales reveals a relevant placebo/verum difference earlier than the alternative scale. All statistical analyses identify alprazolam and imipramine because there is no evidence that they show different efficacy in weeks 6, 8 or at endpoint.

Change scores during treatment measured by both scales are calculated to test the sensitivity to change and the discriminative power. Two major approaches for calculating change scores are used in the literature:

A. Differences of the global scores of a scale (obtained by summing up the items) between the last time point of measurement during treatment and baseline;
B. Residuals which are defined as differences between the observed score after a particular period of treatment and the score to be expected by the baseline level determined by a linear regression analysis.

Approach B is preferred to approach A in the biostatistical literature (Lord 1963; Cronbach and Furby 1970).

Pearson correlations between two pairs of variables are tested for equality using Hotelling's test (modified by Williams) as proposed by Dunn and Clark (1971).

Results

Means and Standard Deviations (Table 1)

The mean raw scores for the PRAS are significantly higher than those of the CRAS (tied t-test as well as Wilcoxon-test, $P \geq 0.001$ for baseline, week 8 and endpoint) indicating the tendency of the patients to give more weight to their complaints than physicians.

Table 1. Means and standard deviations for observer and self-rating scales in the CNCPS

	Mean baseline	SD Baseline	Mean week 8	SD Week 8	Mean endpoint	SD Endpoint
Raw scores						
CRAS						
Placebo	42.73	17.78	17.35	13.55	24.42	19.52
Alprazolam	43.30	19.81	14.52	13.85	16.99	16.20
Imipramine	43.66	17.80	13.12	11.77	16.08	13.80
PRAS						
Placebo	76.52	32.80	31.47	28.55	43.66	35.94
Alprazolam	75.94	35.15	27.86	25.74	32.22	29.81
Imipramine	76.95	34.36	25.08	22.11	32.02	29.17

Table 2. Validity criterion: Sensitivity to change; association between change measured by anxiety scales (defined by difference or alternatively residuals) and global direct assessment of change

	Yardstick: physician's rating of change		Yardstick: patient's rating of change	
	Completer analysis	Endpoint analysis	Completer analysis	Endpoint analysis
CRAS				
Differences	0.44	0.57	0.40	0.52
Residuals	0.75	0.79	0.65	0.71
PRAS				
Differences	0.41	0.51	0.40	0.51
Residuals	0.66	0.71	0.66	0.70

The variances of the raw global scores of the self-rating scale PRAS are higher than those of the CRAS under all conditions listed in Table 1 (F-test for variances at baseline, week 8 and end-point, $P < 0.001$).

Sensitivity to Change and Consistency Across Measures of Change (Table 2)

The yardstick for the sensitivity should be assessed independently of the scale to be validated; consequently the correlations PRAS × physicians' global assessment should be compared with the correlations CRAS × patients' global assessment. The CRAS is equal to the PRAS in the sensitivity to change as indicated in Table 2 by absolute values of the correlation coefficients (0.65 and 0.71 for the CRAS residuals compared with 0.66 and 0.71 for the PRAS residuals; 0.40 and 0.52 for the CRAS differences compared with 0.41 and 0.51 for the PRAS differences); the Hotelling test for comparing correlation coefficients ($P = 0.01$) does not indicate significant differences between the CRAS and the PRAS.

The physician's ratings of change are more consistent than the patients' rating of change, although the differ-

Table 3. Validation criterion: prediction of response; discriminative power of anxiety scales for active vs placebo treatment

	F-values for the treatment factor and percentage of variance explained by the treatment factor (effect size)[a]									
	Week 1		Week 3		Week 6		Week 8		Endpoint	
	F-value	Variance explained (%)	*F*-value	Variance explained (%)	*F*-value	Variance explained (%)	*F*-value	Variance explained (%)	*F*-value	Variance explained (%)
Physician's rating CRAS	40.4***	7.2%	40.3***	16.8%	29.3***	25.1%	18.0***	21.2%	58.9***	37.0%
Patient's rating PRAS	21.7***	3.5%	21.4***	5.7%	14.4***	8.4%	10.6**	10.1%	32.7***	15.2%

[a] Obtained by an analysis of covariance with: dependent variable: anxiety scale (end-point/week 8)
independent variables (stepwise procedure): covariate: anxiety scale (baseline)
centre
treatment (placebo vs imipramine/alprazolam)

ences between the correlation coefficients indicating consistency were only moderate: the correlations of the CRAS residuals (rated by physicians) with the global improvement rated by the physician are higher than the correlations of the PRAS residuals (rated by patients) with the global improvement rated by the patient (Table 2): 0.75 for the CRAS compared with 0.66 for the PRAS when using the completer analysis and 0.79 compared with 0.70 when using the end-point analysis. A similar pattern emerges if differences are used instead of residuals as indicators of change: 0.44 for the CRAS compared with 0.40 for the PRAS when using the completer analysis and 0.57 compared with 0.51 when using the endpoint analysis. Hotelling's test for comparing correlation coefficients rejects the hypothesis of equal correlations for completer and for endpoint analysis on a significance level lower than 0.01 using either residuals or differences as measures of change for CRAS/PRAS.

Discriminative Power (Table 3)

After controlling for the site variable, the placebo treatment is significantly inferior ($P = 0.01$) to active treatment at weeks 1, 3, 6, 8 and at endpoint as measured by both scales (analysis of covariance with significant F values – $P \leq 0.01$ – for the treatment factor): the F values obtained by the analysis of covariance for the treatment factor are significant ($P \leq 0.01$) in the completer and endpoint analysis by either of the two scales (CRAS, PRAS). The F values for the observer-rating scale are higher than those for the self-rating scale by a factor of 1.8 or more, indicating that self-rating scales are likely to miss the verum/placebo difference if smaller sample sizes are used.

The effect size of the treatment variable is defined as the proportion of variance of the total model explained by the treatment variable (placebo versus active treatment). Generally, the variance for the total model described by the observer-rating scale is smaller than the variance of the model described by self-rating scales. The percentage of variance described by the observer-rating scale which is explained by the treatment factor is at least twice as high as for self-rating scales at all time points of measurement (Table 3). Using analysis of variance to compare the effect sizes of both scales the proportion of variance explained by the treatment factor in terms of the observer-rating scales is significantly higher than the corresponding proportion in terms of the self-ratings scales ($P = 0.001$). This statement is valid for all time points of measurement.

Discussion

The degree of variation is higher in self compared with observer-rating scales. This argument is not necessarily an argument against self-rating scales, since variation may be informative. However, CNCPS data indicate that the increased variance of the self-rating scale PRAS is associated with no enhancement of information relevant for testing the efficacy of treatments. The discriminative power between placebo and active treatment with evidenced efficacy is substantially lower for the self-rating scale PRAS compared with the observer-rating scale CRAS.

The loss of discriminative power by self-rating scales in anxiety cannot be attributed to a lower sensitivity to change as there were equal correlations of the change scores derived from the CRAS and the PRAS with the independently rated global improvement. However, the reduced discriminative power of the patients' rating scales may at least partially be explained by the lower degree of consistency of patients' ratings; this is indicated by lower correlations between the two change scores obtained by patient-rated global improvement and the residuals of the PRAS.

The statistical power, defined as the inverse of the β-error (i.e. the likelihood of ignoring a true difference between two groups), is dependent of the variance of the outcome measure; as the variance of the PRAS measures is about twice as high as the variance of the CRAS measures, the statistical power in comparing two treatments drops substantially when the patient-rated PRAS instead of the physician-rated CRAS is used.

The sample size of drug trials comparing different treatment groups should be big enough to be able to con-

trol the α- and β-error; both kinds of error should be fixed according to clinical requirements before starting the trial (Pocock 1984). The required number of patients in the treatment groups to be compared is proportional to the variance (if the α- and β-error and the significant difference to be detected are fixed) (Pocock 1984); in order to control for the α- and β-error at least twice the sample size is necessary in a drug trial using the patient-rated PRAS instead of the physician-rated CRAS on the basis of reasonable values for the α-error (0.01), for the β-error (0.1), and for the minimum of clinically relevant differences between the total scores at baseline and at endpoint (d = 10 scores or higher).

The relative amount of variance of the total model explained by the treatment factor (placebo versus active treatment) is higher if an observer-rating scale procedure tapping the total scope of the anxiety syndrome is used instead of a self-rating procedure: between 21% and 37% of the variance of the outcome variable as described by CRAS is explained in the final stages of the drug trial by the treatment factor compared with a range of 8–15% if the outcome variable is described by the PRAS (weeks 6, 8, end-point). Assessment of this magnitude should take into account that the CNCPS is a multicentre trial including different countries and continents; consequently a high variation across different sites is to be expected. However, the site variable contributes substantially less to the variance of the outcome measured by the observer-rating scale CRAS than the treatment factor.

The reduced discriminative power of self-rating scales in anxiolytic drug trials was recently postulated by Glass et al. (1987) by reviewing placebo-controlled drug trials with benzodiazepines and comparing the frequencies of significant differences measured by the HAS and SCL-90 subscale "anxiety". The results of this review are not convincing for the following reasons:

1. The HAS and the SCL-90 subscale tap different scopes of symptoms by a different number of items and use different scoring procedures.
2. The benzodiazepines minor tranquilizers are not necessarily efficacious compared with placebo.
3. The "meta analysis" procedure of Glass et al. is not generally accepted as a sound method (Wilson and Raclunan 1983).

Furthermore, it is not possible on the basis of this review to decide if the self-rating procedure per se or other aspects of the SCL-90 which are not necessarily related to the self-rating procedure are responsible for the disadvantage of the SCL-90 relative to the HAS. Using the CNCPS phase-II data enables to compare the validity of self- and observer-rating anxiety scales by controlling for all differences between self- and observer-ratings beyond the different rating procedures: the disadvantages of the self-rating procedure per se became apparent in this way. Self-rating induces relatively high variances and inconsistencies as well as a reduced discriminative power for detecting different levels of response to different treatments.

Consequently self-rating scales cannot be recommended for the measurement of drug effects in controlled anxiolytic drug trials when observer-rating procedures are available. The scope of application of self-ratings should go beyond the measurement of the efficacy of drugs. This result does not mean that the observer- rating scales available are perfect and that self-rating scales may not be useful: it cannot be excluded that the higher variances of self-rating scales represent the temporal fluctuations of patients' complaints which are not contributing to differences in treatment effects. Thus, based on these results one cannot argue against the utility of self-rating scales in settings not intending to compare groups of patients by different treaments (e.g. in single case studies).

Acknowledgement. This is a report from the Cross National Collaborative Panic Study, Second Phase. This study consists of a Steering Committee: James H. Coleman, Chairman (Kalamazoo, Mich., USA); Gerald L. Klerman, Co-Chairman (New York, NY, USA). Project Directors: Jose Luis Ayuso (Madrid, Spain); Per Bech (Hillerod, Denmark); Otto Benkert (Mainz, FRG); Sydney Brandon (Leicester, England); Giovanni B. Cassano (Pisa, Italy); Jorge A. Costa e Silva (Rio de Janeiro, Brazil); George C. Curtis (Ann Arbor, Mich., USA); Juan R. de la Fuente (Mexico City, Mexico); Jose Guimon (Bilbao, Spain); Hanns Hippius (Munich, FRG); Yves Lecrubier (Paris, France); Carlos A. Leon (Cali, Colombia); Juan J. Lopez-Ibor Jr. (Madrid, Spain); Heinz Katschnig (Vienna, Austria); Juan Massana (Barcelona, Spain); Mogens Mellergaard (Glostrup, Denmark); Jan- Otto Ottoson (Goteborg, Sweden); Ole J. Rafaelsen (deceased); Raben Rosenberg (Copenhagen, Denmark); Martin Roth (Cambridge, England); Javier Sepulveda (Mexico City, Mexico); Leslie Solyom (Vancouver, Canada); Marco Versiani (Rio de Janeiro, Brazil); Jean Wilmotte (Marchienne-au-Pont, Belgium).

Data processing and preparation of the tape were performed by Dennis B. Gillings, Quintiles (Chapel Hill, NC) and by Philip W. Lavori and the staff of the Biostatistical Unit, Massachusetts General Hospital (Boston, Mass.). Statistical analyses reported in these papers were performed mainly by Philip W. Lavori and the staff of the Biostatistical Unit, Massachusetts General Hospital (Boston, Mass.). Dennis B. Gillings and the staff of Quintiles (Chapel Hill, NC) and A. Marx (Mainz, FRG) contributed statistical advice and analyses.

This study has been sponsored and supported by the Upjohn Company, Kalamazoo, Michigan through its Psychopharmacology Research Unit, Division of Medical Affairs, Robert P. Purpura, Director; Carl P. Lewis; Cathy A. White.

References

Albus M, Maier W, Shera D, Bech P (1990) Consistencies and discrepancies in self- and observer-rated anxiety scales: a comparison between the self and the observer rated Marks-Sheehan scale. Eur Arch Psychiatry Clin Neurosci 240:96–102

Angst J, Bech P, Boyer P, Bruinvels J, Engel P, Helmchen H, Hippius H, Lingjaerde O, Racagni G, Saletu B, Sedvall G, Silverstone JT, Stefanis CN, Stoll K, Woggon B (1989) Consensus Conference on the Methodology of Clinical Trials of Antidepressants, Zurich, March 1988: Report of the Consensus Committee. Pharmacopsychiatry 22:3–7

Ballenger JC, Burrows GD, DuPont RL Jr, Lesser IM, Noyes R Jr, Pecknold JC, Rifkin A, Swinson RP (1988) Alprazolam in panic disorder and agoraphobia: results from a multicenter trial. I. Efficacy in short-term treatment. Arch Gen Psychiatry 45:413–422

Cronbach LJ, Furby L (1970) How should we measure "Change" – or should we? Psychol Bull 74:68–80

Cross National Collaborative Panic Study, Second Phase Investigators (1990) Drug treatment of panic disorder: comparative

efficacy of alprazolam, imipramine and placebo. Br J Psychiatry (in press)
Derogatis LR, Lipman RS, Covi L (1973) SCL-90/an outpatient psychiatric rating scale – preliminary report. Psychopharmacol Bull 9:13–28
Dunn OI, Clark V (1971) Comparisons of tests of the equality of dependent correlation coefficents. J Am Stat Assoc 66:904–911
Glass RM, Uhlenhuth EH, Kellner R (1987) The value of self-report assessment in studies of anxiety disorders. J Clin Psychopharmacol 7:215–221
Hamilton M (1969) Diagnoses and rating of anxiety. Br J Psychiatry Spec Publ 3:76–79
Kerlinger FN (1964) Foundations of behavioural research. Holt Rinehart and Winston, New York
Klerman GL (1988) Overview of the Cross-National Collaborative Panic Study. Arch Gen Psychiatry 45:407–412
Lord FM (1963) Elementary models for measuring change. In: Harris (ed) Problems in measuring change. The University of Wisconsin Press, Madison, pp 21–38
Lydiard RB, Ballenger JC (1987) Antidepressants in panic disorder and agoraphobia. J Affect Dis 13:153–168
Marks IM, Mathews AM (1979) Brief standard self-rating for phobic patients. Behav Res Ther 17:263–267
Pocock SJ (1984) Clinical trials. A practical approach. Wiley, Chichester
Spitzer RL, Williams JBW (1983) Structured Clinical Interview for DSM-III – Upjohn Version (SCID-UP, 12/15/83). New York State Psychiatric Institute, New York, Biometrics Research Division
Wilson GT, Raclunan SJ (1983) Metaanalysis of psychotherapy outcome: limitations and liabilities. J Consult Clin Psychol 51: 54–64