Brit. J. Psychiat. (1983), **143**, 58–63

The Melancholia Scale and the Newcastle Scales

Item-combinations and Inter-observer Reliability

P. BECH, A. GJERRIS, J. ANDERSEN, S. BØJHOLM, P. KRAMP, T. G. BOLWIG, M. KASTRUP, L. CLEMMESEN and O. J. RAFAELSEN

Summary: The reliability of the total scores on three rating scales (Melancholia Scale and the two Newcastle Scales) and the algorithms leading to the Feighner, Research Diagnostic Criteria, and the DSM-III subtypes of depression have been compared. The degree of inter-observer agreements for the various item-combinations was significantly higher than would be expected by chance. The average agreement for each assessment system ranged from 80 to 93 per cent. This 7 to 20 per cent lack of total agreement probably reflects the limitation of clinical assessments including the influence of halo effects.

In recent years our group has evaluated the Hamilton Scales for depression and anxiety (Bech *et al,* 1979; Gjerris *et al,* 1982) and has developed Scales for Mania and Melancholia (Bech *et al,* 1979; Bech and Rafaelsen, 1980). In the present study we have examined the inter-observer reliability of the Newcastle Scales for the diagnosis of depression (Carney *et al,* 1965; Gurney, 1971). Although these scales are being increasingly used there have been few attempts to measure their reliability and, where this has been done, the number of patients investigated was fairly small (Carney and Sheffield, 1972; Kragh-Sørensen *et al,* 1973).

As the main reason for using diagnostic scales is to improve the reliability of the clinical diagnosis (Bech, 1981), we decided to examine the reliability of the Newcastle Scales in relation to our Melancholia Scale. Moreover, the item-combinations of these scales were compared with the algorithms used in other assessment systems in this field, namely the Feighner criteria (Feighner *et al,* 1972), the Research Diagnostic Criteria (Spitzer *et al,* 1978) and the Diagnostic and Statistical Manual (DSM-III, 1980).

Materials and Method

Rating scales

Quantitative scales for depression: The Bech-Rafaelsen Melancholia Scale (MES) was used to quantify the severity of depressive states (Bech and Rafaelsen, 1980; Bech, 1981).

The scale was developed on the basis of the Hamilton Depression Scale (Hamilton, 1967) with each of the eleven items operationally defined on a five-point scale (Table I). The criterion for item-combination of the MES is the total score; hence, the MES score ranges from 0–44. Satisfactory inter-observer reliability has been demonstrated, with Spearman coefficients from 0.79 to 0.93 for the various raters (Rafaelsen *et al,* 1980).

Qualitative scales for depression

The Newcastle Scale-I (N-I, Carney *et al,* 1965) and the Newcastle Scale-II (N-II, Gurney, 1971) were used for assessing the diagnosis of depression. The *N-I* was slightly modified by Carney and Sheffield (1972) as the item 'distinct quality' was redefined. We have, moreover, developed the item scale steps from a two-point scale (0 = absent and 1 = present) to three points, as 'present' was subdivided into 1 = slight or doubtful and 2 = marked. The number of items is ten (Table I) of which eight items are positively and two items negatively weighted. The number of items of *N-II* (Bech *et al,* 1980) is ten (Table I) of which three items are positively and seven are negatively weighted.

Research criteria systems

Quantitative aspects: The Feighner criteria (Feighner *et al,* 1972), and the DSM-III criteria (DSM-III, 1980)for major versus minor depressive episodes are based on non-statistical algorithms. In the Appendix we have shown how we have translated the research criteria from the item domain covered by the three scales (MES, N-I, N-II). The DSM-III definition of major depression equates to the Research Diagnostic Criteria (Spitzer *et al,* 1978) of probable major depression.

Qualitative aspects: The item-combinations of the Newcastle Scales are based on multivariate statistical models (Bech, 1981) while the RDC concept of Endogenous Major Depressive Disorders or the DSM-III concept of Melancholia are based on non-statistical algorithms. In the Appendix the research criteria systems have been translated from the item domain covered by the MES, N-I, and N-II.

Administration of the scales

Two groups of raters participated: (a) four psychiatrists who had been the rating team in our previous studies (the experienced group) and (b) five psychiatrists who formed 'the less experienced group of raters'. On the basis of a joint interview with the patients each rater completed the scales independently.

The rating procedure took place between 8.15 and 9.00 a.m. The number of raters varied from patient to patient, because of unavoidable absence of one or more raters on certain test days. For group (a) the number of raters was four on 6 occasions, three on 19 occasions, and two on 10 occasions. For group (b) the number was five on 5 occasions, four on 17 occasions, three on 12 occasions, and two on 1 occasion. Thus the mean number of raters per patient was 6.5.

Patients

The patients were admitted to the Department of Psychiatry, Rigshospitalet, Copenhagen, from January 1981 to January 1982. They were inpatients and investigated while still in a depressive state. Patients who showed evidence of organic brain disease or schizophrenia were excluded. Each patient was only examined once and the results are based on 35 patients (24 females and 11 males), aged between 22 and 87 years (median 52 years).

Statistical analysis

The inter-observer reliability of the various assessment systems has been expressed by (1) correlation coefficients and (2) percentage agreement.

(1) Correlation coefficients have been used for comparing the reliability in terms of score distribution of the assessment systems taking the concordance derived from chance agreement into account. The Spearman correlation coefficient (Siegel, 1956) has been used as a non-parametric measure when comparing each rater to the average score of the remaining group of raters. The intraclass correlation coefficient (Bartko and Carpenter, 1976) has been used as a parametric measure for the index of reliability of all raters. The intraclass coefficient does not require the same number of raters per patient and we have used the unbiased expression ICC(U). When measuring the ICC(U) for research criteria systems the number of items rated as present has been used within each system.

(2) Percentage agreement has been used for analysing the individual scores, i.e., agreement among

TABLE I

The scoring sheets for the three rating scales

Melancholia Scale (MES)		Newcastle Scale (1965)	Weighted score	Newcastle Scale (1971)	Weighted score
Retardation (motor)	0–4	No personality deviation	+1	Sudden onset	−6
Retardation (verbal)	0–4	No psychological stressors	+2	Duration of actual episode	−6
Retardation (intellectual)	0–4	Quality of depression	+1	Psychological stressors	+12
Anxiety (psychic)	0–4	Weight loss	+2	Phobias	+8
Suicidal impulses	0–4	Previous depressive episodes	+1	Persistence of clinical picture	−2
Lowered mood	0–4	Agitation/retardation	+2	Reactivity	+14
Self-depreciation	0–4	Anxiety	−1	Mornings worst	−16
Retardation (emotional)	0–4	Nihilistic delusions	+2	Early awakening	−10
Sleep disturbances	0–4	Accusations of others	−1	Motor inhibition	−9
Tiredness and pains	0–4	Feelings of guilt	+1	Delusions	−7
Work and interests	0–4				
Total score	0–44	Total score	−2 to +12	Total score	−56 to +34
Cut off scores:		Cut off scores:		Cut off scores:	
No depression	0–5	Endogenous depression	+6 or more	Endogenous depression	−20 or less
Mild depression	6–14	No endogenous depression	+5 or less	Doubtful endogenous depression	−12 to −19
Moderate depression	15–25			No endogenous depression	−11 or more
Severe depression	26–44				

https://doi.org/10.1192/bjp.143.1.58 Published online by Cambridge University Press

raters for each patient calculated as the ratio of number of rater-agreements on the considered categories to the total number of possible agreements. The agreement has then been expressed both by the average agreement among the raters for all patients and by the proportion of patients allocated to the diagnostic categories for the following levels: all nine raters agree, a maximum of one rater disagrees, and at least half of the raters agree.

The Mann-Whitney test (Siegel, 1956) has been used when analysing independent two-sample cases. The median has been used to express the central tendency, and the 25–75 percentiles to express the dispersion. The level of statistical significance is considered to be $P < 0.05$, two-tailed.

Results

As the results of the less experienced group of raters did not differ from those of the experienced group the results for all nine raters have been combined. As indicated in Table II the intraclass coefficients differed significantly from $r = 0.0$ for all assessment systems. No raters differed significantly from the group of the remaining raters when expressed by the Spearman coefficients of which the range of the nine raters for the various assessment systems is shown in Table II.

The average inter-observer agreement in percentage terms for the three rating scales (MES, N-I, N-II) was 86, 90, and 91, respectively. For the research criteria (Feighner definite depression, DSM-III major depression, RDC endogenous depression, DSM-III melancholia) the average agreement in percentage terms was 84, 93, 85, and 80, respectively. The proportion of patients allocated to the cut-off scores of the various assessment systems for the three levels of agreement is shown in Table III. As can be seen the number of patients within each assessment system varied up to a factor of two according to whether all raters agreed or at least half of the raters agreed.

In Table IV we have compared the rating scale scores with the research criteria. The subgrouping of the patients according to the research criteria has been made when not more than one rater disagreed that the patients fulfilled the criteria. The rating scale scores for

TABLE II

Inter-observer reliability for all nine raters expressed by correlation coefficients

Assessment systems	Intraclass coefficients	P	Spearman coefficients (range)	P
Rating scales				
Melancholia Scale	0.82	<0.001	0.82–0.92	<0.001
Newcastle-I	0.81	<0.001	0.73–0.92	<0.001
Newcastle-II	0.77	<0.001	0.55–0.92	<0.01
Research criteria				
Feighner (definite depression)	0.61	<0.001	0.62–0.84	<0.001
DSM-III (major depression)	0.64	<0.001	0.65–0.87	<0.001
RDC (endogenous depression)	0.65	<0.001	0.60–0.83	<0.01
DSM-III (melancholia)	0.57	<0.001	0.45–0.74	<0.01

TABLE III

The proportion of patients allocated to the considered cut-off scores of the assessment systems for three levels of percentage agreements

	Number of patients		
Assessment systems	All nine raters agree	Not more than one rater disagrees	At least half of the raters agree
Rating scales			
MES ≥ 15	16	19	27
N-I $\geq +6$	5	8	12
N-II ≤ -20	3	4	6
Research criteria			
Feighner: Definite depression	20	22	32
DSM-III: Major depression	26	31	34
RDC: Endogenous depression	15	21	30
DSM-III: Melancholia	8	9	18

https://doi.org/10.1192/bjp.143.1.58 Published online by Cambridge University Press

TABLE IV

Research criteria for depression in rating scale values

Research criteria	Number of patients	Rating scales: Melancholia scale Median (25–75 percentiles)	Newcastle I Median (25–75 percentiles)	Newcastle II Median (25–75 percentiles)
Feighner				
Definite	22	20 (16–25)	5.0 (3.5–6.5)	−14 (−22–0)
Non-definite	13	13 (11–16)	3.0 (1.5–4.8)	−3 (−9–+10)
P		<0.001	<0.05	<0.05
DSM-III				
Major	31	18 (15–23)	4.5 (3.0–6.0)	−8 (−19–+2)
Minor	4	12 (8–16)	4.0 (1.5–7.0)	−5 (−18–+19)
P		<0.05	NS	NS
RDC				
Endogenous	21	20 (15–25)	5.5 (4.0–7.0)	−14 (−23–−4)
Non-endogenous	14	15 (13–18)	3.0 (1.5–4.5)	+2 (−5–+14)
P		<0.05	<0.01	<0.001
DSM-III				
Melancholia	9	25 (21–27)	6.8 (5.5–7.5)	−21 (−29–−12)
Non-melancholia	26	16 (13–19)	3.5 (2.0–5.5)	−3 (−12–+8)
P		<0.001	<0.01	<0.001

each of the nine raters were used for each patient to calculate the median and the 25–75 percentiles. However, when testing the difference between the scores by use of the Mann-Whitney test, the average scores of the nine raters for each patient were utilized. The results (Table IV) showed that the MES scores differed significantly for both the axis of major versus minor depression (Feighner and DSM-III) and the axis of endogenous versus non-endogenous depression (RDC and DSM-III), supporting a MES cut-off score of 15. The N-I and N-II scores differed significantly on the axis of endogenous versus non-endogenous depression (RDC and DSM-III). However, the cut-off scores of N-I emerged as +5 rather than +6, and of N-II as −12 rather than −20.

Finally the correlations of the three rating scales with one another were determined. The results showed a Spearman coefficient of 0.75 ($N = 35$, $P < 0.01$) when the two Newcastle Scales were intercorrelated. When MES was correlated with N-I and N-II, coefficients of 0.41 and of 0.35 were found, both having P between 5 per cent and 1 per cent.

Discussion

The item domain investigated in this study is the one covered by the Melancholia Scale (MES) and the two Newcastle Scales (N-I and N-II). Therefore our results based on the algorithms applied to the research criteria systems for depression (Feighner, RDC, and DSM-III) are limited to the rating scale items shown in the Appendix.

In this study, where the raters have assessed a series of patients simultaneously, the concordance between the raters has been used as a measure of inter-observer reliability. On the basis of intraclass coefficients we demonstrated that there was no difference between the experienced and less experienced group of raters, and that the degree of inter-observer agreements for the various assessment systems was significantly higher than would be expected by chance. Expressed as percentage agreements the assessment systems ranged from 80 to 93. However, a closer analysis showed (Table III) that a maximum of 100 per cent agreement was found in many cases but a 'chance' agreement of 50 per cent emerged in a small number of cases, resulting in the average percentage between 80 and 93. The 7 to 20 per cent lack of complete agreement seemed to be due to a small group of patients who gave 'ambiguous' signals during the interview. This may reflect the 'human' factor of clinical assessment scales, for example the influence of halo effects (Guilford, 1954).

https://doi.org/10.1192/bjp.143.1.58 Published online by Cambridge University Press

Appendix

The algorithms for endogenous depression or melancholia

RDC Endogenous major depressive disorders ≥6 A or B items (at least one A)		DSM-III (1980) Melancholia (= endogenous depression) Both A items plus ≥3 B items	
No.	Items	No.	Items
A (1)	Distinct quality (N-I, 3)	A (1)	Loss of pleasure (MES, 6)
A (2)	Lack of reactivity (N-II, 6)	A (2)	Lack of reactivity (N-II, 6)
A (3)	Morning worst (N-II, 7)	B (1)	Distinct quality (N-I, 3)
A (4)	Pervasive loss of pleasure (MES, 6)	B (2)	Morning worst (N-II, 7)
B (1)	Feelings of self-reproach (MES, 7; N-I, 10)	B (3)	Early morning awakening (N-II, 8)
B (2)	Early morning awakening or middle insomnia (MES 9, N-II, 8)	B (4)	Agitation/retardation (MES, 1, 2; N-I, 6, N-II, 9)
B (3)	Agitation/retardation (MES, 1, 2; N-I, 6; N-II, 9)	B (5)	Weight loss (N-I, 4)
B (4)	Poor appetite (N-I, 4 = 1)	B (6)	Guilt (MES, 7; N-I, 10)
B (5)	Weight loss (N-I, 4 = 2)		
B (6)	Loss of interests (MES, 11)		

The algorithms for definite or major depression

Feighner criteria: Definite depression item A plus ≥ 5 B items		RDC or DSM-III Major depression item A plus ≥ 4 B items	
No.	Item	No.	Item
A	Dysphoric mood (MES, 6)	A	Depressed mood (MES, 6)
B (1)	Poor appetite or weight loss (N-I, 4)	B (1)	Poor appetite or weight loss (N-I,4)
B (2)	Sleep difficulty (MES, 9; N-II, 8)	B (2)	Insomnia (MES, 9; N-II, 8)
B (3)	Loss of energy; fatigue (MES, 10)	B (3)	Agitation/retardation (MES, 1, 2; N-I, 6; N-II, 9)
B (4)	Agitation/retardation (MES, 1, 2; N-I, 6; N-II, 9)	B (4)	Loss of interests (MES, 11)
B (5)	Loss of interests (MES, 11)	B (5)	Loss of energy; fatigue (MES, 10)
B (6)	Feelings of self-reproach (MES, 7; N-I, 10)	B (6)	Feelings of self-reproach (MES, 7; N-I, 10)
B (7)	Diminished ability to think (MES, 3)	B (7)	Diminished ability to think (MES, 3)
B (8)	Suicidal impulses (MES, 5)	B (8)	Suicidal impulses (MES, 5)

There are few published studies in which the inter-observer coefficients of the Newcastle Studies, the Feighner, RDC or DSM-III systems for depression have been examined (Kragh-Sørensen *et al,* 1973; Helzer *et al,* 1977; Spitzer and Williams, 1980). The inter-observer coefficients in these studies are of the same order as those found in our study, i.e. significantly higher than chance agreement. In contrast, it has been found that diagnostic classification systems for depression (e.g. the Eighth Revision of International Classification of Disease, WHO, 1974) with no item definition or standardized item-combinations have an inter-observer agreement which is no better than chance (Beck, 1967; Kendell, 1975; Spitzer and Fleiss, 1974).

The assessment systems concerning the quantitative aspect of depression (e.g., major versus minor) were found to have a significant concordance, i.e. a MES score of 15 or more corresponded to a definite depressive state (Feighner) or to a major depression (DSM-III). The assessment systems concerning the qualitative aspect of depression (endogenous versus non-endogenous depression) were also found to have a significant concordance.

Acknowledgement

The authors wish to thank Ove Aaskoven, H. Lundbeck & Co. Copenhagen, for assistance with data analysis.

References

American Psychiatric Association (1980) *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III). Washington DC.

BARTKO, J. J. & CARPENTER, W. T. (1976) On the methods and theory of reliability. *Journal of Nervous and Mental Disease*, **163**, 307–17.

BECH, P. (1981) Rating scales for affective disorders : Their validity and consistency. *Acta Psychiatrica Scandinavica*, **64**, (Suppl 295), 1–101.

—— BOLWIG, T. G., KRAMP, P. &. RAFAELSEN, O. J. (1979) The Bech-Rafaelsen mania scale and the Hamilton depression scale. *Acta Psychiatrica Scandinavica*, **59**, 420–30.

—— & RAFAELSEN, O. J. (1980) The use of rating scales exemplified by a comparison of the Hamilton and the Bech-Rafaelsen melancholia scale. *Acta Psychiatrica Scandinavica*, **62**, (Suppl 285), 128–31.

—— GRAM, L. F., REISBY, N. & RAFAELSEN, O. J. (1980) The WHO depression scale: Relationship to the Newcastle scales. *Acta Psychiatrica Scandinavica*, **62**, 140–53.

BECK, A. T. (1967) *Depression. Clinical, experimental, and theoretical aspects*. Philadelphia: University of Pennsylvania Press.

CARNEY, M. W. P., ROTH, M. & GARSIDE, R. F. (1965) The diagnosis of depressive syndromes and prediction of ECT response. *British Journal of Psychiatry*, **111**, 659–74.

—— & SHEFFIELD, P. B. (1972) Depression and the Newcastle scales: Their relationship to Hamilton's scale. *British Journal of Psychiatry*, **121**, 35–40.

FEIGHNER, J. P., ROBINS, E., GUZE, S. B., WOODRUFF, JR. R. A., WINOKUR, G. & MUNOZ, R. (1972) Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry*, **26**, 57–63.

GJERRIS, A., BECH, P., BØJHOLM, S., BOLWIG, T. G., KRAMP, P., ANDERSEN, J. & RAFAELSEN, O. J. (1982) The Hamilton anxiety scale: Evaluation of homogeneity and inter-observer reliability in patients with depressive disorders. *Journal of Affective Disorders*, (in press).

GUILFORD, J. P. (1954) *Psychometric Methods*. New York: McGraw Hill.

GURNEY, C. (1971) Diagnostic scales for affective disorders. *Proceedings of the Fifth World Conference of Psychiatry*. Mexico City, p. 330.

HAMILTON, M. (1967) Development of a rating scale for primary depressive illness. *British Journal of Clinical Psychology*, **6**, 278–96.

HELZER, J. E., CLAYTON, P. J., PAMBAKIAN, R., REICH, T., WOODRUFF, R. A. & REVELEY, M. A. (1977) Reliability of psychiatric diagnosis. *Archives of General Psychiatry*, **34**, 136–41.

KENDELL, R. E. (1975) The role of diagnosis in psychiatry. Oxford: Blackwell.

KRAGH-SØRENSEN, P., HANSEN, C. E. & ÅSBERG, M. (1973) Plasma levels of nortriptyline in the treatment of endogenous depression. *Acta Psychiatrica Scandinavica*, **49**, 444–56.

RAFAELSEN, O. J., BECH, P., BOLWIG, T. G., KRAMP, P. &. GJERRIS, A. (1980) The Bech-Rafaelsen combined rating scale for mania and melancholia. In *Psychopathology of Depression* (eds. K. Achte, V. Aalberg and J. Lönnqvist). *Psychiatrica Fennica*, Suppl, 327–31.

SIEGEL, S. (1956) *Non-parametric Statistics*. New York: McGraw Hill.

SPITZER, R. L. & FLEISS, J. L. (1974) A re-analysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry*, **125**, 341–7.

—— ENDICOTT, J. E. & ROBINS, E. (1978) Research Diagnostic Criteria. *Archives of General Psychiatry*, **35**, 773–82.

WILLIAMS, J. B. W. & SPITZER, R. L. (1981) The reliability of the diagnostic criteria of DSM-III. In *What is a Case?* (eds. J. K. Wing, P. Bebbington and L. N. Robins). London: Grant McIntyre, p. 107–14.

WORLD HEALTH ORGANIZATION (1974) Glossary of mental disorders and guide to their classification, for use in conjunction with the International Classification of Diseases, 8th revision. Geneva.

*Per Bech, M.D., *Consultant Psychiatrist*

Annette Gjerris, M.D., *Senior Registrar*

John Andersen, M.D., *Senior Registrar*

Søren Bøjholm, M.D., *Senior Registrar*

Peter Kramp, M.D., *Senior Registrar*

Tom G. Bolwig, M.D., *Professor of Psychiatry*

Marianne Kastrup, M.D., *Senior Registrar*

Lars Clemmesen, M.D., *Research Associate*

Ole J. Rafaelsen, M.D., *Professor of Psychiatry*

Psychochemistry Institute and Psychiatric Department, Rigshospitalet, University of Copenhagen, DK 2100 Copenhagen, Denmark

*Reprint requests.

(Received 13 October 1982; revised 19 January 1983)