The Diagnostic Melancholia Scale (DMS): dimensions of endogenous and reactive depression with relationship to the Newcastle Scales

P. Bech, P. Allerup, L.F. Gram, P. Kragh-Sørensen, O.J. Rafaelsen, N. Reisby, P. Vestergaard and the Danish University Antidepressant Group (DUAG) *

Frederiksborg General Hospital, DK-3400 Hillerød, Denmark
(Received 5 August 1987)
(Accepted 13 October 1987)

Summary

The two diagnostic Newcastle Scales for depression have been evaluated in a drug trial with antidepressants. By use of latent structure analysis (Rasch models) it was found that two dimensions are necessary for describing the diagnosis of depression, one for endogenous features and one for reactive features. Of the depressed patients 50% had a pure endogenous depression, 14% had a pure reactive depression, 32% had mixed endogenous and reactive depression, and 4% had uncertain diagnosis. In the pure endogenous depression group 77% had a monotonically non-decreasing improvement curve during treatment whereas in the other diagnostic categories around 50% had such an improvement.

Key words: Newcastle Depression Scale; Depression dimensions

Introduction

Among the diagnostic scales for depression we have previously analysed the two Newcastle Scales *(Carney et al., 1965; Gurney, 1971; Bech et al., 1983; Roth et al., 1983). In our first study (Bech et al., 1980) we found no significant correlation between the Hamilton Depression Scale (Hamilton, 1967) and the Newcastle Scales when assessing patients prior to antidepressant therapy, thereby stressing the fact that the Hamilton Depression Scale (HDS) and the Newcastle Scales rate two different aspects of depression, severity and diagnosis, respectively. Of the two Newcastle Scales the 1965 version containing 10 items (Carney et al., 1965) is the most frequently used scale in clinical research. The 1971 version (Gurney, 1971)
contains 10 items for the diagnosis of endogenous versus reactive depression, which were, however, originally applied secondary to a subdivision of patients into anxious and depressed (e.g., Kragh-Sørensen et al., 1973; Roth et al., 1983; Mullaney, 1985). We developed more explicit item specifications for the Newcastle Scales than the original instructions (Bech et al., 1983), but we discarded the initial subgrouping where patients with an anxiety score are excluded from a further diagnosis of depression.

The Newcastle Scales have been criticised by Eysenck (1970) for containing two subscales which, by differential weights (positive and negative) to different items, are transferred to only one bipolar dimension of endogenous versus reactive depression. This criticism has, in our opinion, not been conclusively analysed. The background for the criticism is outlined in Fig. 1, where the Newcastle continuum is indicated as a one-dimensional subspace of the two-dimensional classification with the one axis containing items indicative of endogenous depression and the other axis containing items indicative of reactive depression. For this continuum to properly reflect the different positions of the two-dimensional system, the two axes need to show simple mathematical relationships. In fact, it is not otherwise possible to distinguish patients with high scores on both axes from patients with low scores on the two axes — patients d and c in Fig. 1 who received identical Newcastle scores. However, the results found by Carney et al. (1965) seem to indicate that patient d in Fig. 1 does not exist, i.e., the distribution of scores along the Newcastle continuum is bimodal.

The Newcastle Scales, furthermore, contain items for measuring delusions, which narrow their applicability to inpatients, and most of the depressed patients are now treated in general practice. In our opinion endogenous depression is not synonymous with psychotic depression. With this background we have reconsidered the two Newcastle Scales. We selected ten items from the scales of which five items are indicative of endogenous depression, and the other five items are indicative of reactive depression. In an earlier study (Bech et al., 1984a), we found that these items maintained the same level of interobserver reliability when we used equal weighting of items as when we used the differential weights originally found by Carney et al. (1965) or Gurney (1971).

In this study we have, prior to the analysis of relations between the two axes, undertaken two separate item analyses by means of Rasch models in order to examine whether endogenous and reactive scores are sufficient statistics for the two dimensions.

**Methods**

**Patients**

The sample consists of 95 patients who entered a double-blind, 35-day between-group comparison of clomipramine and citalopram. The methodology and results have been reported elsewhere (DUAG, 1986). In this multicentre study 102 hospitalised patients suffering from a major depression (a HDS score of 18 or more, or a score of 9 or more on our HDS subscale, Bech et al., 1975) participated. In the present study we included those patients in whom both Newcastle Scales were completed during the placebo wash-out period, prior to the 35-day active treatment, in total 95 patients. Of the 95 patients 64 were
female and 31 were male. Their ages ranged between 24 and 67 years, median 51 years.

Raters

In total 15 psychiatrists took part in the trial. These raters have been described elsewhere (Bech et al., 1986) with regard to such variables as years of experience in psychiatry, experience with and attitude to rating scales, sex, and place of work. The interobserver reliability expressed by intraclass coefficients of all 15 psychiatrists of the Newcastle Scales on the basis of joint interviews with ten patients was 0.57 for the 1965 version and 0.34 for the 1971 version ($P < 0.01$, DUAG, 1986). However, the number of raters per patient in this reliability study ranged from two to 13. Furthermore, three of the raters together assessed 41 of the 95 patients in the trial. The interobserver reliability of these three raters was 0.69 for the 1965 version of the Newcastle Scales, 0.50 for the 1971 version, 0.62 for the five endogenous items of the Diagnostic Melancholia Scale (DMS), and 0.51 for the five reactive items ($P < 0.01$).

The Diagnostic Melancholia Scale (DMS)

The two axes consist of 10 items in total. All items are scored on a 3-point response scale (0 = not present; 1 = doubtful or very slight; and 2 = mild to severe). The endogenous part contains the following items: quality of depression, feeling worse in the morning, early waking, weight loss, and persistence of clinical picture. The reactive part contains the remaining five items: psychological stressors, reactivity of symptoms, character deviations, somatic anxiety, and duration of current episode. Each axis ranges from 0 to 10, as the total score is measured as item (0–2) multiplied by the number of items (5). It should be emphasised that none of the 10 DMS items are included in our HDS subscale, which measures the severity of depressive states. For the specific instructions of the DMS, see Appendix.

Statistical analysis

The Rasch model (Rasch, 1960, 1966; Fischer, 1974) for multivariate item response calibration is a general theory for the relation between manifest (clinical) item responses and latent (theoretical) dimensions. The link between manifest item responses and the latent dimension is defined through a requirement of statistical sufficiency, i.e., the requirement that the item responses can be additively combined to a total score. If these total scores are sufficient statistics, the next step in the analysis is to see whether there exists a simple mathematical relationship between the scores in the two-dimensional diagnostic system, i.e., to examine whether the two dimensions are parallel in the sense that ‘less’ endogenous means ‘more’ reactive.

We have previously described the mathematical aspects of the general Rasch model when analysing data from rating scales of depression (Bech et al., 1981; Allerup, 1986), and it has been shown how to obtain simple one-dimensional comparisons from the general multivariate Rasch model. Testing procedures are initiated as likelihood ratio tests for the general model (Andersson, 1973) in which case general homogeneity among scale items is investigated. Subsequent tests of the one-dimensional Rasch model are based on subdivisions of items (homogeneity) and patients (transferability) either by means of internal test criteria (score level) or by external patient characteristics such as sex and age. The likelihood ratio tests are all performed at a level of significance of $P = 0.05$.

The DMS items are defined in the Appendix, and it should be noted that the item responses are ordered, i.e., an item response of 2 means that a patient has also obtained a response of 1 on this item. The one-dimensional Rasch model assesses quantitative scorings of these qualitative response rankings, and it is part of the statistical analysis to test whether these scores differ significantly from the equidistant calibration: 0, 1, and 2.

Results

Both the five items constituting the endogenous axis of depression and the five items defining reactive depression passed the tests for the general Rasch model and the subsequent test of one-dimensionality, as all $\chi^2$ values were insignificant at the 5% level of statistical significance. It was estimated that the scoring system for the item categories ‘not present’, ‘doubtful or very mild’, and ‘mild to severe’ was 0.0, 1.2, and 2.0 for the endogenous dimension, and 0.0, 1.5, and 2.0 for
the reactive dimension. However, we could not reject the hypothesis ($P < 0.05$) of equidistant calibration: 0.0, 1.0, and 2.0, and the further analyses were based on this calibration system.

The relative significance of the five items within the endogenous dimension could be evaluated by Rasch analysis in relation to the calibrated item responses: 0, 1, and 2. We found that most patients scored on the item of 'early waking', next came 'quality of depression', 'weight loss', 'persistence of clinical picture', and the most exclusive item was 'feeling worse in the morning'.

Likewise, the five items within the reactive dimension could be ranked by Rasch analysis. We found that most patients scored on the item of 'duration of current episode', next came 'psychological stressors', 'reactivity of symptoms', 'character deviations', and the most exclusive item was 'somatic anxiety'.

It is a powerful property of the Rasch analysis that beyond these rankings, quantitative weights can be obtained. In fact, we found that around 85% of the expected contribution to the endogenous scores was due to the item 'feeling worse in the morning'. For the reactive dimension, around 80% of the expected contribution was due to the items 'character deviations' and 'somatic anxiety'.

The distribution of the Newcastle 1965 and Newcastle 1971 scores of the 95 patients were both clearly unimodal. For Newcastle 1965 we found a mean score of 6.5 (standard deviation: 2.2), the median was 7.0 (range 0.5–11). For Newcastle 1971 we found a mean score of $-25.3$ (standard deviation of 15.7), the median was $-27.3$ (range 15.5–51.5).

In Fig. 2 we show the distribution of scores of the 95 patients in accordance with Fig. 1. Using cut-off scores of 5 on each of the two dimensions it appears that 31 patients were type d patients as defined by Fig. 1. Only four patients were classified as type c patients. It is clear (Fig. 2) that no simple mathematical relationship between the two dimensions emerged. This proves that the two dimensions cannot without loss of information be reduced to a single underlying continuum, as presumed by the Newcastle Scales.

In Fig. 3 we show the distribution (percentage) of patients within the four diagnostic categories derived from our two-dimensional diagram. As can be seen, type a patients (endogenous depression) accounted for 49.5%, type b patients (reactive depression) for 13.7%, type c patients (uncertain diagnosis) for 4.2%, and type d patients (mixed endogenous and reactive depression) for 32.6%. This classification is based on pretreatment ratings (week 0). We have investigated this pretreatment classification in relation to the posttreatment measurements of severity of depression (weeks 1–6) by use of our HDS subscale transformed to the three degrees of severity, 'no depression', 'minor depression', and 'major depression', considered relevant in trials with antidepressants (Bech et al., 1984b). It can be seen from Fig. 3 that the original distribution over the diagnostic categories was found to be very stable over the weeks of treatment, namely around 50% endogenous depression, around 14% reactive depression, around 32% mixed endogenous and reactive depression, and around 4% uncertain depression.

In contrast, the pretreatment Newcastle 1971 classification of 73% endogenous depression and 27% reactive depression was found to vary considerably from week to week and within the three degrees of depressive states (Fig. 4). The same
results were obtained with Newcastle 1965, indicating an association between diagnosis and severity.

Finally, we analysed the predictive validity of our two-dimensional scales (DMS) with regard to response to treatment. We have previously argued (Bech et al., 1984b) that improvement curves for patients with endogenous depression during treatment with antidepressants should have a monotonic shape, indicating a steady time-dependent improvement towards recovery. In contrast, patients with reactive depression should have less clearly time-dependent improvement curves. We have, therefore, analysed the percentage of patients with steady time-dependent improvement curves (i.e., patients who week after week during the trial have no worsening) in each of the three diagnostic categories of endogenous depression (type a), reactive depression (type b), and mixed endogenous and reactive depression (type d). The number of patients in the uncertain group (type c) was too small for statistical analysis.

Using our HDS subscale transformed to three degrees of severity (no depression = complete re-
response, minor depression = partial response, and major depression = no response), the results showed that 76.6% had monotonically non-decreasing improvement curves in the group of endogenous depression, whereas 58.1% and 53.8%, respectively, had monotonically non-decreasing improvement curves in the mixed endogenous–reactive group and the reactive group of patients. The difference between these proportions was statistically significant ($P < 0.05$, Fisher's exact test). However, when end-point analysis by means of the posttreatment outcome scores (HDS subscale) was made, no statistically significant differences were found between the proportions of complete and partial responders over the three diagnostic groups of patients.

When the original Newcastle Scales were used to classify patients no statistically significant differences between the proportions of complete and partial responders in the groups of endogenous and reactive depression were obtained, either as regards non-decreasing improvement curves, or at the end-point analyses.

**Discussion**

This study has shown not only that the items constituting the axes of endogenous and reactive depression fulfill the Rasch model, thereby stipulating dimensions in which the total scores are sufficient statistics. It has also shown that there is no simple mathematical connection between these two dimensions; they are not correlated. Both dimensions are, therefore, needed when classifying depressed patients diagnostically.

In our previous study with the Newcastle 1965 Scale (Bech et al., 1980) we found a unimodal distribution of the scores which we have now replicated in this study. Recently, Carney et al. (1986) replicated their original bimodal distribution, which was also found by Mullaney (1985), but most studies with this scale found a unimodal distribution (e.g., Post, 1972; Rao and Coppen, 1979; Zimmerman et al., 1987). In the present study we found that our 1971 version of the Newcastle Scale also had a unimodal distribution of the diagnostic scores.

The interobserver reliability of the Newcastle Scales expressed by intra-class coefficients in our previous studies ranged from 0.44 to 0.81 (1965 version) and from 0.68 to 0.71 (1971 version) (Bech et al., 1983, 1984a). The interobserver reliability of the DMS has previously been found to be 0.71 for the endogenous scale and 0.58 for the reactive scale (Bech and Allerup, 1986). In the present study the intra-class coefficients were of the same magnitude for all the scales apart from the 1971 version which in the present study had a coefficient 0.34 for all raters. However, for the three raters who assessed nearly 50% of the patients the coefficient was 0.50. All the intra-class coefficients were measured conservatively as unbiased coefficients in accordance with Bartko and Carpenter (1976), and all the coefficients were statistically significant ($P < 0.01$). Furthermore, the intra-class coefficient becomes higher the more variation there is between the scores of the patients. In the present study the variation between the patients was low. In general, the interobserver reliability of the Newcastle Scales is rarely reported in the literature (e.g., Mullaney, 1985). It is of interest that we found around 30% of the patients having a high score on both dimensions, indicating a mixed combination of endogenous and reactive features.

The ten DMS items were selected on the basis of the theory (Bech, 1981) that the construct of
Endogenous and reactive depression should be separated from the construct of severity of depression. In a review of the universe of diagnostic items for depression (Bech and Allerup, 1986) it is shown that the works of Lewis (1934) and Kendell (1968) lack such items as quality of depression, feeling worse in the morning, early waking, and persistence of clinical picture (i.e., around 80% of the items of endogenous depression in DMS). Neither Lewis nor Kendell at the time of their investigations accepted the concept of endogenous depression. In contrast, Paykel (1971) included most of the items of endogenous depression found in DMS in his studies on the classification of depressed patients. In his subsequent studies (Paykel, 1972, 1977) he found that patients with endogenous depression had the best outcome on amitriptyline compared with other subgroups of depression. Both Lewis, Kendell, and Paykel included all five reactive items found in the DMS.

The validity of the Newcastle Scales was considered promising in some studies (e.g., Bech, 1981) but rather negative in other studies (e.g., Katschnig et al., 1986). In this study we found that the construct validity of the DMS was superior to the Newcastle Scales, as no interaction between DMS and severity of depression emerged, in contrast to the Newcastle Scales (Fig. 3). It is difficult to evaluate the predictive validity of diagnostic rating scales in controlled clinical trials because patients who are able to complete the fixed treatment procedure in the obsessive multicentre protocols already by their common feature of adhering to the protocol neutralise their pretreatment differences in the clinical profile of depression (Bech et al. 1984b). However, in the present multicentre study we were able to demonstrate a superiority of DMS compared to the Newcastle Scales concerning an aspect of predictive validity. Hence, a majority of patients with DMS endogenous depression had non-decreasing improvement curves, whereas this was only the case in half of the patients with reactive depression. The mixed group of endogenous and reactive depression seemed closer in this respect to the reactive than to the endogenous group. However, an end-point analysis of outcome to treatment showed no statistically significant differences between these three diagnostic groups of patients. As stated by Wing (1978) it is an attractive theory to regard mixed endogenous and reactive depression as a biological disorder of lower vulnerability than pure endogenous depression, because the biological disturbance in the mixed group only shows clinical manifestations when provoked by psychosocial stressors. The DMS opens possibilities for studies in this field, in contrast to the Newcastle Scales where the concept of psychoprovoked endogenous depression is lacking.

Further studies will clarify the predictive validity of DMS. The conceptual or construct validity of DMS was found acceptable in this study. Its applicability in general practice where most depressed patients are now treated is under investigation.

Appendix

SCORING INSTRUCTIONS FOR THE DIAGNOSTIC MELANCHOLIA SCALE (DMS)

Item 1 Quality of depression (0–2)

This item includes the patient’s experience of the current depressive episode as qualitatively distinct from normal despondency when under adversity or distress, e.g., the death of a loved one. The patient should, therefore, be asked for qualitative feelings different from the range of his or her ordinary affective responses to adversity. It is a difficult item to assess, especially, of course, if the patient denies ever having had severe adversities. It is of importance to ensure whether prior to the current depressive episode, i.e., in the habitual state, the patient has experienced the same kind of symptoms as now or whether the current symptoms are more of a ‘foreign body’, a distinct quality of depression.

If the patient cannot identify himself/herself with the current depressive syndrome, which therefore is conceived as qualitatively distinct from feelings of grief, the score is 2. If the interviewer or the patient are in doubt whether this item is present, the score is 1. If the current depressive episode has been described as ‘ordinary’ tristesse as experienced in adverse situations like death within the family or circle of friends, the score is 0.
**Item 2 Early awakening (0–2)**

Early awakening implies that the patient wakes up at least one hour earlier than usual. The assessment of this item should include the general ('average') sleep pattern during the current depressive episode, and not only the last days. N.B. It is immaterial whether the patient has been using sedative/hypnotic medication or not!

When the interviewer is convinced that during the actual episode as a persistent feature the patient has woken up at least one hour too early, the score is 2. When in doubt, or when the patient has only had this symptom during the last few days, the score is 1. If early awakening has not been present, the score is 0.

**Item 3 Weight loss (0–2)**

When the patient indicates a weight loss of 3 kg or more related to the current depressive episode, or 0.5 kg or more per week during the last 3 weeks, the score is 2. At less pronounced weight loss, the score is 1, and with no weight loss related to the current episode, the score is 0.

**Item 4 Diurnal variation, feeling worse in the morning (0–2)**

Diurnal variation implies that during the current depressive episode the patient has generally ('on average') been most depressed in the morning hours, and that the severity diminishes during the day. The criterion for diurnal variation is not fulfilled if the patient only indicates having a short-lasting amelioration just before going to bed.

When the interviewer is convinced that during the current episode the patient has had diurnal rhythmicity in the severity of his or her symptoms, the score is 2 (cross-examine to take habitual diurnal variation and possible reactivity into account!). When it is doubtful whether a truly autonomous diurnal variation is present, the score is 1. When there is no diurnal rhythmicity, the score is 0.

**Item 5 Persistence of clinical picture (0–2)**

This item implies that the clinical picture, the depressive syndrome, in general has been constant during the current episode, apart from smaller day-to-day variations and/or diurnal variation.

If there has been no significant change (fluctuations between 'good' or 'bad' days or weeks), the score is 2. If it is more doubtful whether persistence has been present during the current episode, the score is 1. If there have been clear fluctuations, the score is 0.

**Item 6 Psychological stressors (0–2)**

Psychological stressors imply any situation or event which is considered by the interviewer to have been a significant contributor to the development of the current depressive episode. The stressors must have appeared within the last 6 months prior to this episode, and may or may not still be present and maintaining the depressive syndrome. The stressors may be worries concerning one's health, worries concerning the health of near relatives or friends, a death of a loved one, interpersonal conflicts in the family or at work, and financial problems. The same stressor will, of course, be a very different experience for the different patients, and the patient's subjective experiences and feelings must be taken into consideration. However, the interviewer's evaluation is decisive.

If psychological stressors have been or are still present, the score is 2. If the interviewer is in doubt, the score is 1. If psychological stressors have not been present, the score is 0.

**Item 7 Reactivity (0–2)**

Reactivity means that the severity of the depressive symptoms waxes and wanes in relation to circumstances, e.g., the patient feels less depressed when something pleasant or positive appears or takes place. The patient thus retains the capacity to react positively when something positive takes place and/or to feel less depressed for a while in good company.

When the interviewer is convinced that reactivity is present, the score is 2. If reactivity is only very transient, the score is 1. If there is lack of reactivity to usually pleasurable stimuli, that is the patient does not feel much better, even temporarily, when something good happens, the score is 0.

**Item 8 Somatic anxiety (0–2)**

Somatic anxiety should be assessed independently of the coexistence of psychic anxiety.
Somatic anxiety includes all physiological concomitants of anxiety: motor tension and/or autonomic hyperactivity (especially palpitations, nausea or vomiting, sweating, and dizziness). It is often difficult to distinguish between somatic anxiety and psychomotor agitation, but in this connection it is immaterial whether agitation is included in the score. It is also immaterial to distinguish between attacks of somatic anxiety and generalized anxiety. It is, however, decisive to assess whether during the current depressive episode the patient has experienced somatic anxiety, and the last week should be stressed most.

If during the last week and/or during the interview the patient has been clearly anxious (experienced motor tension, palpitations, nausea, sweating etc.), the score is 2. If in doubt, the score is 1. If the patient has not experienced somatic anxiety, the score is 0.

Item 9 Duration of current episode (0–2)
Duration of current episode is noted from the time when the patient first experienced a clear change from normal life or mood to the time of investigation. If the illness is phasic, the current episode must have been preceded by a clear illness-free interval of at least 3 months.

If the current episode has lasted 1 year or more, the score is 2. If the episode has lasted between 6 and 12 months, the score is 1. If the episode has lasted less than 6 months, the score is 0.

Item 10 Character neurosis (0–2)
Character neurotic features might have emerged before the current episode from the patient’s choice of spouse or life partner (a peaceful or considerate partner rather than a dominating or self-assertive partner) because character neurotics avoid persons who provoke them. During the current episode the neurotic features might have manifested themselves in the way the patient is presenting his or her complaints, namely by striving for an emotional secondary gain. At the interview this striving can be observed by the patient’s cooperation on an attention-demanding dimension.

If the patient has shown clear signs of character neurosis, the score is 2. If it is more uncertain for the interviewer that the patient has a character neurosis, the score is 1. If the patient has no sign of character neurosis, the score is 0.

References


