Br. J. clin. Pharmac. (1976), Supplement, 58–60

COMPARATIVE VALUE OF RATING SCALES

MAX HAMILTON
Department of Psychiatry, University of Leeds

There are four types of scales in psychiatry: for assessment of the patient's condition, for diagnosis, for prognosis and for the selection of treatment. This paper is concerned with only the first type and, even then, is restricted to scales for depressive illness.

There are very many such scales now available and Table 1 lists some of them, classified according to their use by an observer or by the patient. This list is by no means exhaustive. It may seem strange that so much effort should have been put into the construction of so many scales, but there are at least two good reasons why this should be so. The first is that we have by no means reached perfection; we surely cannot believe that in the course of a decade, a scale would have been constructed which is incapable of further improvement. The second is that scales are required for different purposes and therefore there can be no such thing as a 'best' scale for all purposes, in all circumstances and for use by all kinds of raters.

There are many ways of classifying scales, but the most important is that shown in Table 1 which distinguishes between scales to be used by an observer (especially a skilled observer) and those used for self-assessment by the patient. On theoretical grounds, the former have most of the advantages, but the latter are often more useful in practice and can be shown to be about as effective in appropriate circumstances.

Table 1 Some scales for assessment of depression

Observer

Hamilton rating scale for depression (1960 and 1967) Cronholm-Ottosson scale (1960 and 1973) Beck depression inventory (1961) Wechsler depression scale (1963) van Praag standardized interview (1965) Rickels OP scale (1967) Zung depression status inventory (1972) Bojanovsky scale (Czech) (1966) FKD scale (Czech) (1966) CPRG scale (Japan) (1966) AMP scales (German) (1965)

Self-assessment

MMPI (D) (1946) Kanter & Standen (1963) Zung self-rating depression scale (1965) Lubin adjective check list (1965) MHQ (D) (1966) Popoff depression inventory (1969) Hamburg scale (German) (1970) von Zerssen scale (German) (1970) Wakefield self-assessment depression scale (1971) Pilowsky questionnaire (1972) Maudsley personality inventory (1956)

A skilled observer, by reason of his training and experience, has standards against which he can evaluate the intensity of any one symptom, whereas a patient has no such standards. 'Severe', 'moderate' and 'mild' have no meaning to him, except in relation to his own experience. A skilled interviewer can penetrate the mask which the patient sometimes holds up, either deliberately or unintentionally, to hide the full picture of his illness. Patients can have many reasons for minimizing their symptoms or emphasizing them.

Futhermore, a rater can observe and assess certain manifestations of illness which the patient would find impossible or extremely difficult to do. For example, a patient cannot assess loss of insight, by definition, and he would find it extremely difficult to assess such symptoms as mild retardation or agitation, or to evaluate hypochondriasis and delusions. An observer can use a scale to cope with all grades of severity of illness, from the mildest to the most severe, whereas a patient can be too ill to be able to complete a form.

Finally, a patient may be unable to fill in a questionnaire if he is insufficiently literate, if his vocabulary is deficient, if he is lacking in concentration and so on.

All these are good reasons why the first type of assessment is better than the second; but there are practical difficulties. The most important is the length of time taken by observer ratings. It is true that a psychiatrist does, or should, spend sufficient time with a patient to enable him to fill in a rating scale, but this becomes very difficult when the scales have to be used frequently. Another important reason is that a sufficient number of trained raters may not be available. Self-assessment scales are particularly useful for surveys and screening procedures.

There are three properties of scales which are of fundamental importance; these are validity, sensitivity and reliability. There are many different ways of defining these but for present purposes it is sufficient to say that validity signifies that different scores on the scale accurately reflect different grades of severity of the phenomena measured. Sensitivity is really a form of validity but differs in that it means that changes in the phenomena are accurately reflected in changes in the score. This is particularly important in therapeutic trials. Reliability signifies that the scale can give rise to measurements which have the minimum interference from random error. There are several different forms of reliability but the most important one is inter-rater reliability. This means that two or more raters can use the scale and obtain closely related results.

Validity

One of the best ways of determining validity is to use the scale on two groups of subjects who are known to differ in a particular characteristic and to demonstrate that the scale clearly shows this difference in the scores obtained for the groups. The Wakefield self-assessment depression scale (SADS) was tested on a group of 'normal' subjects and a group of depressives and it was shown that there was very little overlap in the scores obtained for these two groups (Snaith *et al.*, 1971), only 3% of patients and 7.5% of normals being misclassified by a cut-off score. The same test was carried out with the Hamburg inventory (Kerekjarto & Lienert, 1970), and this gave a biserial correlation coefficient of 0.72. Schwab *et al.* (1967) used the Hamilton rating scale (HRS) on medical patients and depressives and found a bimodal distribution of scores. Downing & Rickels (1972) assessed a group of depressive patients and non-depressive psychiatric patients with the Zung SDS and the Popoff depression inventory and found no significant difference in the scores between these two groups. Sensitivity appears to be high for depressive rating scales in general and there are few references to it in the literature. However, a few authors have reported that the Zung SDS is somewhat lacking in sensitivity.

In order to identify the groups when making such comparisons, it is first necessary to make some clinical decision about them—that is, that they are suffering from a depressive illness or are 'normal'. A simple extension of this use of clinical judgement is to take a group of depressive patients and to judge them as suffering from different degrees of severity of illness. They are then rated on a scale and the scores correlated with the levels of 'global' severity. Using this method with the HRS, Zealley & Aitken (1969) obtained a correlation 0.90 for patients admitted to hospital but this decreased to 0.55 when they were discharged. They found that the visual analogue scale (VAS) gave a correlation of 0.78 for patients on admission which decreased to 0.13 when they were discharged. Bech *et al.* (1975) obtained a correlation of 0.84 for the HRS and 0.77 for the Beck scale. Metcalfe & Goldman (1965) reported that the Beck scale gave a correlation of 0.62 when used in this way. Downing & Rickels (1972) obtained much lower figures for the self-assessment inventories they examined. For the Popoff scale they obtained a validity coefficient of 0.36 for GP patients and 0.28 for psychiatric patients; for the Zung SDS they obtained 0.45 and 0.22 respectively. It is not surprising that the highest figures should be obtained with what are, in effect, two different methods of observer ratings.

Reliability

There seems to have been little interest in examining the reliability of global judgements, but Bech *et al.* (1975) reported a reliability of 0.88. A fair number of papers have reported on the inter-rater reliability of the HRS and the findings range from 0.88 to a surprising 0.98. The Bojanovsky scale gives reliability of 0.92 (Bojanovsky & Chloupkova, 1966). The Wechsler scale has a reliability of 0.88 when the raters interview simultaneously and 0.78 when the ratings are done a week apart (Wechsler *et al.*, 1963).

Concurrent validity

Another way of assessing validity is to compare the scores on different scales; this is known as concurrent validity. It could be said that this is arguing in a circle, but there are occasions on which it would be useful to know which scales could be regarded as equivalent. The HRS is the scale which has been most commonly used for making such comparisons. Schwab *et al.* (1967) found that it correlated 0.74 with the Beck scale, which is almost the same as the 0.72 found by Bech *et al.* (1974). These values are distinctly higher than the 0.51 reported by Tan (1969). Brown & Zung (1972) obtained a correlation of 0.79 with the Zung SDS; this equals the figure obtained by Zealley & Aitken (1969) for the VAS.

Such a high correlation suggests that there is no point in using the HRS, but it must be remembered that this scale gives detailed information on specific symptoms, which is not available with the VAS. However, these authors reported that the correlation between the HRS and the VAS sank to 0.06 for patients when they were discharged. The concurrent validity of self-assessment scales with the HRS is low except for the Wakefield SADS which is 0.89. For example, Tan (1969) obtained a correlation of only 0.25 with the MMPI depression scale. Although not strictly to the point, it is of some interest that Garside *et al.* (1970) obtained a correlation of 0.49 with the MMPI neuroticism scale and a non-significant correlation with the extraversion scale. Nevertheless, 0.49 is higher than 0.25.

The Beck depression inventory was reported by Lubin (1965) to have a correlation with the Lubin adjective check list of 0.40 to 0.66 for various groups. The latter scale also has a correlation with the MMPI depression scale of 0.31 to 0.53 for the same groups. Tan (1969) reported a correlation between the Beck scale and MMPI depression scale of 0.53. The Zung SDS has a correlation with the Popoff scale of 0.71 for GP patients and 0.65 with the psychiatric patients

(Downing & Rickels, 1972). Zung (1972) reported that it had a correlation of 0.87 with his depression status inventory. It would appear that self-assessment scales tend to have a lower concurrent validity among themselves, with some notable exceptions, than do observer-rating scales.

Requirement of a scale

It is obvious that a scale should have high validity, reliability and sensitivity, and it is clear from the above descriptions that a large proportion of the scales available do indeed fulfil these requirements. The choice between them will therefore be determined by other factors. In the case of observer scales, one of the most important is that it should be easy to use. It should not be too short because that gives a relatively low reliability, but it also should be not too long, for then it interferes with the interviewing of the patient. If the interviewer goes through the items in order, he converts the psychiatric interview into an interrogation which is extremely unsatisfactory. If he conducts the interview in the normal manner then he has to keep interrupting while he turns over from one page to another to make an appropriate entry. If he waits until the end of the interview then he finds that he has forgotten what he should report.

The items of the scale should be judged in relation to the particular use. They should cover the range of variables and the grades should be such that there is an adequate spread of scores for all grades of severity. Furthermore, these grades should be such that there is no bunching of scores at one end or the other of the scale, since this causes a lack of discrimination between individuals. The grades must be easily distinguishable. It is easy to assess a number of grades for such symptoms as depression, guilt, and suicidal tendencies, but very difficult to do so for such symptoms as appetite and insomnia. The grades should be relevant to the purpose of the scale; this is the disadvantage of all-purpose scales.

Furthermore, the items should be found with a sufficient frequency in the group being studied otherwise they are, in effect, irrelevant. The items should be distinguishable and defined in such a way that they are not automatically rated together, because this gives rise to spurious high correlations. Finally, the items should be sensitive in the relevant circumstances.

References

BECH, P., GRAM, L.F., DEIN, E., JACOBSEN, O., VITGER, J. & BOLWIG, T.G. (1975). Quantitative rating of depressive states. Correlation between clinical assessment, self-rating scale (Beck) and objective rating scale (Hamilton). *Psychol. Med.* (in the press).

BOJANOSKY, J. & CHLOUPKOVA, K. (1966). Bewertungsskala der Depressionzustande. *Psychiat. Neurol., Basel,* **151**, 54-61.

BROWN, G.L. & ZUNG, W.W.K. (1972). Depression scales: self or physician rating? A validation of certain clinically observable phenomena. *Compr. Psychiat.*, **13**, 361-367.

DOWNING, R.W. & RICKELS, K. (1972). Some properties of the Popoff index. *Clin. Med.*, **79**, 11-18.

GARSIDE, R.F., KAY, D.W.K., ROY, J.R. & BEAMISH, P. (1970). M.P.I. scores and depression. *Br. J. Psychiat.*, **116**, 429-432.

KEREKJARTO, M. von & LIENERT, G.A. (1970). Depressionsakalin als Forschungsmittel in der Psychopathologie. *Pharmakopsychiatrie Neuro-Psychopharmakologie,* **3**, 1-21.

LUBIN, B. (1965). Adjective check lists for measurement of depression. *Arch. gen. Psychiat.*, **12**, 57-62.

METCALFE, M. & GOLDMAN, E. (1965). Validation of an inventory for measuring depression. *Br. J. Psychiat.*, **111**, 240-242.

SCHWAB, J.J., BIALOW, M.R. & HOLGER, C.G. (1967). A comparison of 2 rating scales for depression. *J. clin. Psychol.*, **23**, 94-96.

SNAITH, R.P., AHMED, S.N., MEHTA, S. & HAMILTON, M. (1971). The assessment of the severity of primary depressive illness. *Psychol. Med.*, **1**, 143-149.

TAN, B.K. & UYTERLINDE, A.J. (1969). Een voorlopig onderzoek naar de praktische bruikberheid van drie vertaalde depressieschalen. *Bull. Coord. Comm. Biochem. Onderzooch,* **3**, 49-57.

WECHSLER, H., GROSSER, G.H. & BUSFIELD, B.L. Jr. (1963). The depression rating scale. *Arch. gen. Psychiat.*, **9**, 334-343.

ZEALLEY, A.K. & AITKEN, R.C.B. (1969). Measurement of mood. *Proc. roy. Soc. Med.*, **62**, 993-996.

ZUNG, W.W.K. (1972). The depression status inventory: an adjunct to the self-rating depression scale. *J. clin. Psychol.*, **28**, 539-543.