Self-report measures of depression: Some psychometric considerations

Gregory J. Boyle

This paper examines some aspects of the psychometric adequacy of existing self-report depression measures, so that clinicians might better appreciate their reliability and validity. Issues addressed include the desirability of moderate rather than maximum test-retest and item homogeneity/internal consistency estimates; the distinction between measurement at the surface syndrome level vs. that at the fundamental source state/trait level; the importance of appropriate rather than arbitrary factor analytic procedures; and the need for multivariate measures instead of the single-scale instruments so often employed to measure depression in isolation from interacting emotions such as anxiety or stress. Apart from these more general issues, a number of specific criticisms is considered, along with recommendations for better self-report measures of depression.

Depression has featured throughout history as perhaps the most pervasive of all psychopathology (cf. Boyd et al., 1982). This is partly reflected in the numerous self-report measures which have been devised to quantify human depression. Reviews of the literature pertaining to depression measurement have been undertaken by Becker (1974, pp. 24–27), Levitt & Lubin (1975, pp. 57–73), Mayer (1977), Hughes et al. (1982), as well as by Kazdin & Petti (1982). Among the frequently employed self-report measures are the Minnesota Multiphasic Personality Inventory revised D scale (MMPI-D: Dempsey, 1964), the Beck Depression Inventory (BDI: Beck, 1967, ch. 12; Beck & Beamesderfer, 1974), the Zung Self-Rating Depression Scale (SDS: Zung, 1965, 1971), the Multiple Affect Adjective Check List (MAACL: Zuckerman & Lubin, 1980), the Depression Adjective Check List (DACL: Lubin, 1967), the Institute for Personality and Ability Testing (IPAT) Depression Scale (Krug & Laughlin, 1976), and the Center for Epidemiological Studies Depression Scale (CES-D: Radloff & Locke, 1984).

Several psychometric issues have emerged, however, in the self-report measurement of depression. In clinical work, it is desirable that such instruments be situationally sensitive. Paradoxically, high test-retest correlations may suggest that, while a measure is reliable, it may be insensitive to fluctuations in levels of depression. Additionally, in the past, psychometric opinion was that item homogeneity (internal consistency) of an instrument should be as high as possible. However, modern psychometry attempts to keep item homogeneities at a moderate level, in order that each item adds a new piece of information (breadth) to the measurement. Historically, many self-report measures of depression have been constructed containing items which were little more than paraphrases of each other. Also, some studies have focused on certain attributes of depression (such as fatigue or guilt). Other studies have viewed depression as a psychiatric syndrome. The lack of agreed definition has resulted in many discrepancies of findings reported in numerous studies (Shapiro, 1981). In teasing out the relationships of these underlying attributes to the depression syndrome, appropriate use of factor analysis is an invaluable aid. Frequently, however, factor analytic studies of depression measures have been based on inadequate methodology. In particular, the choice of the number of factors to extract and the method of rotation have been quite arbitrary, resulting in factor solutions which have not attained simple structure.

Accordingly, existing reviews of the reliability and validity of self-report measures of depression, while comprehensive in many respects, are nonetheless defective on several counts. Specifically, they have failed to take into consideration issues of depression
measurement at the surface syndrome level as opposed to the underlying source state/trait level; the need for moderate rather than maximum test–retest and item homogeneity estimates; the importance of appropriate factor analytic procedures; and the desirability for multivariate rather than single-dimension scales. This paper addresses critically these important psychometric considerations in an attempt to sharpen the clinician's awareness of the strengths and limitations of the prevailing plethora of self-report depression measures.

A. Assessment at the syndrome vs. the source state/trait level

Phenomena relating to depression include depressed mood, feelings of helplessness and hopelessness, feelings of loss, feelings of guilt and worthlessness, loss of appetite, sleep disturbance and psychomotor retardation (Radloff, 1977; Roberts, 1980). Some depression phenomena are not necessarily unique to it, however (e.g., guilt—see Shapiro, 1979). The surface syndrome of depression comprises each of these underlying components in varying degrees for different individuals who are clinically depressed. This distinction between measuring depression at the global surface syndrome level vs. the more fundamental source state/trait level has often not been clear in the literature relating to depression measurement. The currently available measures appear to index haphazardly these differing components (source states/trait) depending upon the whims of the particular investigator. The result is that high scores obtained on such an instrument do not necessarily indicate that an individual is clinically depressed. More seriously, some individuals classified as clinically depressed (on the basis of a standardized psychiatric interview such as the Diagnostic Interview Schedule, using the DSM-III, Feighner or Research Diagnostic Criteria—see Robins et al., 1982) may obtain low scores on self-report instruments. While such measures are not intended to serve as diagnostic devices, they are nevertheless used as a convenient means of screening the severity of depression-related phenomena in epidemiological surveys of the general population. They are also used with individuals suffering from clinical depression—see, for example, the numerous studies using the CES-D instrument (e.g., Radloff, 1977; Weissman et al., 1977; Roberts, 1980; Boyd et al., 1982; Roberts & Vernon, 1983; Radloff & Locke, 1984). These discrepancies between scores on a self-report measure of depression and the clinical diagnosis of depression have been accounted for partly in terms of nay saying during the diagnostic interview (where the depressed individual denies depressive feelings or symptoms). Alternatively, nay saying in responding to a self-report questionnaire, or simply difficulty in responding to a questionnaire due to literacy problems, may occur (Boyd et al., 1982).

However, part of the discrepancy appears to lie in the somewhat ad hoc composition of the various self-report measures themselves (Shapiro, 1975, 1981). Some instruments (such as the BDI or SDS) largely quantify symptoms of depression, while others (such as the MAACL, DACL and CES–D) also measure depressed mood. Others (such as the MMPI–D30) assess the extent of depression primarily at the surface syndrome level, rather than in terms of the underlying source states or traits which comprise the syndrome. Psychiatrists and clinical psychologists generally want to have a measure of the severity of the syndrome of depression, rather than of only one or two parts of that syndrome such as depressed mood or psychomotor retardation. Correlations of self-ratings of depressed mood correlate highly with ratings of the severity of the syndrome. However, any clinical or epidemiological survey of the presence or severity of depression is looking for data about the surface syndrome (see, for example, Carroll et al., 1981, for correlations between ratings of depressed mood and intensity of the depression syndrome). It seems that any measure which rates only, say, depressed mood without reference to the other components contributing towards the depression syndrome would be questionable as a clinical or
epidemiological tool. Failure to understand sufficiently the nature of the several attributes or components underlying the depression syndrome has evidently hindered epidemiological and psychiatric studies in the area of depression research. This is illustrated in the disparity between the several measurement instruments mentioned above. In order to measure adequately the surface syndrome of depression, self-report instruments such as the CES-D need to index the full range of components which are the source states or traits underlying the overall syndrome. Measurement of these several components needs to be representative of the relative importance of each in the composition of the depression syndrome. This psychometric requirement is clearly a task for future research as the current instruments do not adequately assess the depression syndrome in terms of the whole range of its underlying components.

The surface syndrome of depression is elucidated by examining the clusters evident in the matrix of intercorrelations of the various attributes of depression. At the intuitive level, this is the kind of operation which the clinician engages in when attempting to diagnose depression. Clearly though, examining the intercorrelations of the various attributes or components of depression does not indicate determinants of depression per se. Whereas psychiatry and clinical psychology depend a good deal upon the classification of depression at the surface syndrome level, psychological research among normal individuals has to accept that syndrome clusters are not so easily identified. Personality research into depression therefore cannot accept the somewhat superficial analysis of the depression syndrome, but rather is compelled on scientific grounds to explore the statistically evident attributes or components of the depression syndrome (e.g. Cattell, 1983, pp. 77–78). Interestingly, Cattell (1979, pp. 73–79, 205–206) has statistically identified seven underlying source components of the surface syndrome of depression (cf. Price et al., 1981). Furthermore, whereas applied clinical psychology and psychiatry examine depression as a more or less pervasive condition bordering on a trait, psychological research (cf. Boyle, 1979a) often investigates phenomena associated with depression as a state. Several of the existing self-report instruments differ in their state–trait properties (See Levitt & Lubin, 1975, pp. 60–63, for a review), and therefore discrepancies in classification of depression phenomena across the different instruments are inevitable on this ground alone. Selection of a particular instrument must therefore depend upon whether investigators are interested primarily in assessing transitory states (such as depressed mood) or, alternatively, pervasive low mood or symptoms associated with clinical depression.

As far as experimental research into personality is concerned, the basic difficulty with measurement at the syndrome level (such as occurs with the MMPI–D) is that the surface syndrome represents a correlational clustering which may have little inherent unity other than in terms of the psychiatric label or classification. Sophisticated multivariate statistical techniques are required to elucidate the fundamental source states and traits which underlie the depression syndrome. While these components of depression may not necessarily be immediately apparent to the ‘eye of the beholder’, they nevertheless have a statistically derived validity which far transcends that of the intuitive observer. From the viewpoint of clinicians, self-report measures of depression are considered to be valid if they define the same individuals as having a disorder as a clinician would diagnose as having such a depressive disorder. For clinicians, the critical question concerning the validity of depression measures is what the sensitivity and specificity of such scales are compared with psychiatric diagnosis (see Boyd et al., 1982). Clearly, better and newer psychometric measures of the depression syndrome will need to be developed, taking into account the existing knowledge about the source components which make up the overall surface syndrome. Only when these separate components are proportionally represented in the measurement devices will there be much chance of reducing the current discrepancies
between the psychiatric diagnosis of depression and the epidemiological classification of depressed individuals.

B. Moderate test-retest and internal consistency estimates

While researchers and clinicians may sometimes be interested in the measurement of depression as a trait, in general the instruments used should be sensitive enough to detect differences in the extent of depression under various circumstances on different occasions. Thus it is desirable for self-report measures of depression to be sensitive to state fluctuations (Thorne, 1974, 1980). This demands that test–retest reliability estimates should not be too high. In the latter event, it is likely that the instrument is not sufficiently sensitive to state fluctuations in depression. For example, Miller & Seligman (1973) reported a test–retest coefficient of 0.74 for a three-month retest interval using the BDI. Since this scale is purported to be essentially a state measure (Levitt & Lubin, 1975, p. 60), this level of stability may suggest a lack of situational sensitivity (cf. Boyle, 1979b, p. 78). Nevertheless, Miller & Seligman failed to report whether the state of their subjects changed between test and retest, and therefore it is not clear whether a value of 0.74 was too high.

If retesting occurs during a depressive episode (which often lasts for several months), high stability (test–retest) estimates would not necessarily indicate that the measure is insensitive. Definitive testing of this question of situational sensitivity of a measure, however, should take place in the developmental phase of a scale's construction, during which the instrument is administered under known levels of depressive induction. Different levels of depressive induction can be produced experimentally in normal subjects, using a variety of methods (Goodwin & Williams, 1982). One such method is the Velten (1968) procedure, which consists of 60 statements designed to produce mild depressive mood. Another method used by Boyle (1984b) and Boyle & Cattell (1984) was to present subjects with a shocking film of road accident victims together with part of an autopsy. However, the less intense Velten procedure was used by Boyle (1984d) in checking on the usefulness of the Differential Emotions Scale (Izard et al., 1974) as a measure of depression or, at least, of depressive mood. Using these procedures, it is possible to elucidate the extent of an instrument's sensitivity to differing levels of depression and/or related components of depression such as depressive mood. Ideally, a self-report measure of depression will have moderate test–retest reliability estimates under different levels of depressive induction, but high retest coefficients should be obtained under identical levels of depression induction for both the initial testing and subsequent retesting. This is an issue which can be explored empirically in normal populations of subjects using simple retest designs (e.g. Boyle, 1983b, 1984c) and control groups.

As for the issue of item homogeneity (internal consistency), Cattell (1973, pp. 357–379, 1978, pp. 289–293, 1982) has addressed this dilemma at length. Boyle (1983a, pp. 389–390) has also discussed this issue, pointing out that a measure should not exhibit excessively high item homogeneity (no greater than about 0.7, as enunciated by Kline, 1979, p. 3), if the instrument is broadly to measure depression. According to Kline:

Lower than this, each part of the test must be measuring something different... A higher correlation than this, on the other hand suggests that the test is too narrow and too specific... if one constructs items that are virtually paraphrases of each other, the result would be high internal consistency and very low validity.

This conclusion remains valid irrespective of the type of questions which the researcher or clinician is interested in asking. Clearly, if a single numerical estimate of a broad range of depression-related phenomena (i.e. the surface syndrome of depression) is required, then a multi-item scale with moderate item homogeneity is appropriate. Even if the investigator is
interested in measuring the relationship between different depression phenomena—such as guilt, sadness, indecisiveness, psychomotor retardation and/or other source components of the surface syndrome—then a wide range of scales broadly tapping each depression component is required. If item homogeneities exceed the 0.7 level, then it is evident that only two or three of the items in a given scale are needed, and that additional items add little if any new information, in which case there is significant item redundancy.

Regrettably, most self-report measures of depression currently in use have unacceptably high levels of item homogeneity/internal consistency. Item homogeneity has been maximized in the belief that test reliability is correspondingly maximized. However, as Allen & Potkay (1983), as well as Lachar & Wirt (1981) have pointed out, either high or low item homogeneity can be related to high or low retest reliability. This observation runs counter to the frequent assertion that item homogeneity should be high for both state and trait measures (cf. Zuckerman, 1983, p. 1084).

The MMPI-D, was reported by Becker (1974, p. 26) as having high item homogeneity (0.88). Dempsey's (1964) attempt to maximize item homogeneity was reasonably successful but, in the process, he inadvertently reduced the breadth of measurement, leaving only a narrow measure of a limited range of depression phenomena. Likewise, Beck et al. (1961) reported a corrected split-half item homogeneity estimate of 0.93 for the BDI. This suggests the scale's narrowness and item redundancy. Evidence on both the retest reliability and item homogeneity of the SDS is lacking (Goodstein, 1975, p. 538; Mayer, 1977, p. 376), along with data on the effects of response sets, and on the validity of the SDS for non-psychiatric settings and with different age groups. Until this information is provided, the SDS cannot be recommended for clinical use in applied settings. However, the MAACL is reported to have high item homogeneity (presumably in the 0.8–0.9 range, cf. Zuckermann, 1983). Similarly, item homogeneities from 0.82 to 0.90 have been reported for the DACL (Mayer, 1977, p. 389). Given item homogeneities up to 0.95 for the IPAT Depression Scale (Krug & Laughlin, 1976), it appears also that this instrument has much item redundancy and is a narrow measure of depression. It seems that Cattell's (e.g. 1973) recommendations on item homogeneity have been ignored by Krug & Laughlin in constructing the IPAT Depression Scale. The fact that LaVoie (1978) applauded its high item homogeneity (p. 859) is an indication of the confusion surrounding this issue in general. Item homogeneity data in no way indicate the reliability of a scale, as shown by Allen & Potkay (1983) and also by Lachar & Wirt (1981). Since Krug & Laughlin provided no retest reliability data for the IPAT Depression Scale, and given its high item homogeneity, the measure cannot be recommended for psychiatric/clinical psychological use at this time. As for the CES-D, high levels of item homogeneity have also been reported, up to about 0.9 (e.g. Radloff, 1977, p. 391; Radloff & Locke, 1984, p. 4). This suggests the considerable item redundancy and narrowness of the CES-D as a measure of depression.

Evidently, none of the self-report measures discussed above is a broad assessment instrument. This has serious implications for current psychological research into depression and for epidemiological studies of depression as a surface syndrome among the general population. Use of these self-report measures with patients diagnosed as clinically depressed cannot be recommended at this time in view of the extreme narrowness of the several instruments discussed.

C. Appropriate factor analytic techniques.

At present, there is a gap between psychiatrists and clinical psychologists concerning the question of validity of self-report measures of depression. In the United States, for example, psychiatric research defines validity as being criterion validity (see Spitzer &
Williams, 1980). Thus self-report measures are considered to be valid if the results of the psychiatric diagnosis are concordant with the questionnaire findings, as Boyd et al. (1982) have indicated. On the other hand, psychologists do not restrict their examination of the validity of a depression measure solely to criterion validity as have psychiatrists in general (cf. Boyle, 1979c). Criterion-keyed instruments such as the MMPI–D (and most other single-dimension depression scales) contain items selected on the basis of responses of a depressed reference group. The adequacy of these items depends on the reliability of the criterion group. Cattell & Kline (1977, p. 54) have argued that selection of items in reference to a criterion-depressed group is tenuous, since psychiatric diagnosis is notoriously unreliable (an observation also made by Weissman et al., 1977, p. 213). Since criterion groups can differ in many ways, a depression measure which is a successful discriminator between depressed and non-depressed groups may be useful in practical terms, but may be theoretically meaningless. Other criterion-keyed depression measures, such as the DACL and BDI, also suffer from these difficulties. Linn (1981) has discussed the psychometric problems inherent in criterion-referenced measurement at some length.

Given that criterion validity is not, by itself, sufficient to establish the true validity of a depression measure, it is necessary to consider seriously other types of validity, such as predictive, concurrent and discriminant validity. Another type of validity which is critically important (at least for psychological research) is construct validity. According to Kline (1979, p. 11):

To demonstrate construct validity a number of hypotheses are set up that would be tenable if the test were valid... showing what is and is not correlated with the test and what groups score high and low. It thus embraces both concurrent and predictive validity... a whole series of results can build up a composite picture, as a mosaic, which overall, demonstrates that a test is valid: each result is seen as fitting into a nomothetic network.

To establish construct validity, it is necessary to not only examine convergent–discriminant correlations, multitrait–multimethod matrices (Campbell & Fiske, 1959) and the effects of experimental variables on test scores, but also to investigate the underlying source states/trait which have a measurement basis in the various self-report depression instruments. This last requirement necessitates the application of appropriate factor analytic techniques in depression research.

The importance of reasonably good factor analytic methodology cannot be overstated. Cattell (1973, pp. 282–287, 1978, 1979, p. 351), Nunnally (1978, pp. 327–436), Kline (1979, pp. 38–41, 1980, pp. 323–324) and Boyle (1984a) have all discussed the attributes of factor analytic procedures, as have many others (e.g. Barrett & Kline, 1982). The work of Cattell is perhaps the most comprehensive in the area of factor analytic methodology. Cattell has suggested eight important conditions for valid factor analytic work. These are:

1. A strategic choice of variables. Thus in factoring the surface syndrome of depression, it is desirable to base the analysis not only on items pertaining to one or two of the source components, but instead to utilize a wide range of items which index all the known source states/trait, which together comprise the syndrome.

2. A minimum sample size of 250 subjects. Generally speaking, smaller samples result in unreliable factor solutions. Some methodologists such as Nunnally (1978, p. 402) have argued for a minimum of 20 subjects per variable. While factor analyses of self-report depression instruments have often utilized large samples (e.g. in the epidemiological surveys with the CES–D), the particular factoring procedures used have frequently left much to be desired. Almost all of the factoring guidelines proposed by Cattell have been ignored in virtually every instance.

3. Determining the correct number of factors by an objective test such as the automated
scree test (Barrett & Kline, 1982), the MAP test (Velicer, 1976), or Revelle & Rocklin's (1979) very simple structure criterion. Use of the old eigenvalues greater than unity (Kaiser-Guttman) criterion is no longer recommended, as the method has been shown to seriously over-extract factors when the number of variables exceeds about 50, and more seriously to under-extract when the number of variables is below about 20 (e.g. Child, 1970; Cattell & Vogelmann, 1977; Horn & Engstrom, 1979).

(4) Fixing communalities by an iterative factoring procedure in accord with the common factor model (Harman, 1976). It is not possible to derive accurate communality estimates without first determining the correct number of factors.

(5) Obtaining a unique oblique rotational solution, taken to maximum simple structure. Boyle (1984c) has discussed this issue at length. Orthogonal rotation (such as varimax) does not achieve simple structure (e.g. Loo, 1979), as it permits only a particular resolution of the numerous outcomes obtainable with oblique rotational methods. In the psychological domain, most variables pertaining to depression are correlated (also true of the physiological variables involved in depression). Use of a topological rotational programme such as Rotoplot (Cattell & Foster, 1963; Cattell, 1978, pp. 141-151) over and above the analytical solution obtained (as say with SPSS) will allow maximum approximation to simple structure to emerge; as demonstrated by Boyle (1983c).

(6) Simple structure should be tested for statistical significance using, for example, the Kameoka & Sine (1978) tables. Almost certainly, there has been much futile discussion in the depression measurement literature concerning non-significant, trivial factors.

(7) As with empirical research in general, it is always desirable to cross-validate one's findings. Similarly, in factor analytic work on depression, it is highly desirable to check on the degree of invariance of the factor pattern solution across different samples.

(8) Finally, for thorough factor analysis, it is also desirable to check on the invariance of the higher-order factor structure, using congruence and salient variable similarity indices (Cattell, 1978).

The inadequacy of arbitrary factor analytic procedures is exemplified in the several inconsistent factorings of the BDI (cf. Weckowicz et al., 1978; Golin & Hartz, 1979, for example). Unfortunately, most factor analytic studies of the BDI and of other self-report measures of depression have been defective on several counts. As with the BDI, the several factor analyses of the SDS (e.g. Rickels et al., 1973; Blumenthal, 1975; Morris et al., 1975) have yielded only inconsistent results. In the strict sense, Blumenthal did not perform a factor analysis, but rather undertook a cluster analysis. Cluster analysis is a superficial grouping of variables (at the syndrome level) based on their intercorrelations. Cattell (1978, pp. 16-17, 45-46) criticized cluster analysis as it fails to elucidate the underlying fundamental source states/traits of the syndrome. Clearly, cluster analysis is fine if all the investigator wishes to know is whether a particular measure of depression is in some way sensitive to aspects of the depression syndrome. However, if a more detailed understanding of which particular aspects of the surface syndrome have a measurement basis in an instrument is required, then cluster analysis is inadequate (see Cattell et al., 1966, ch. 9, for a detailed treatment of the inadequacies of cluster analysis). Moreover, while the Rickels et al. study employed 831 depressed out-patients as their sample, they failed to adhere to any of the other factor analytic guidelines recommended above. They used the dubious principal components plus varimax procedure, and extracted factors on the basis of the outdated eigenvalue greater than unity criterion. Use of principal components is mathematically elegant, but would seem to be psychologically meaningless in view of the common factor model (cf. Harman, 1976). The failure to iterate the factor matrix results in spurious common factor variance being included in the solution (Lee & Comrey, 1979, p. 301). Recent research (Velicer et al., 1982) does suggest that this issue might not be as
critical as first thought. Nevertheless, several factor analyses of self-report depression measures have indicated that, despite typically high levels of item homogeneity, the instruments are basically multidimensional, measuring differing facets of the surface syndrome of depression in a somewhat ad hoc manner. As shown by Hughes et al. (1982), the MMPI-D is multidimensional. Likewise, the CES-D is a composite of various aspects of the depression syndrome, although the components measured differ from those measured in the MMPI-D. Given the high item homogeneities, Radloff (1977) suggested that the CES-D be used as a single measure despite its obvious factorial complexity. Again, the inadequate principal components plus varimax factoring procedure was used in her study. As with the BDI, the surface syndrome of depression is being measured differentially but narrowly by the self-report instruments available. Therefore it is not surprising that scales such as the CES-D have only moderate concordance with the clinical Research Diagnostic Criteria diagnosis of depression (cf. Roberts & Vernon, 1983, p. 45). Evidently, similar scores on such instruments reflect different aspects of the depression syndrome (see Becker, 1974, p. 26).

D. Need for multivariate measures of depression

The use of instruments designed only to index depression is a risky procedure. The difficulty in using such scales is that elevations in other psychological states such as anxiety or stress may go undetected. While scores on a single depression measure may alter due to therapeutic intervention, the greatest effect might involve other unmeasured states. Change in the depression score might even result from its correlation with other states. One can never really be certain that alterations to scores on a single scale of depression are due to alterations in depression itself. Contemporary psychometry of depression phenomena requires a multivariate perspective. Among multivariate measures (albeit of depressed mood) are the Profile of Mood States (POMS: McNair et al., 1971), the Eight State Questionnaire (8SQ: Curran & Cattell, 1976), and the Differential Emotions Scale (DES: Izard et al., 1974). Each of these measures is considered in turn.

The POMS purports to measure six mood states. The first three dimensions are those of the MAACL (anxiety, depression, hostility). The measure comprises 65 adjectives intended to index each of an a priori selected group of moods. No sampling of the overall mood state domain was undertaken and factor analysis was employed only as a justification for choosing the particular dimensions included in the instrument. Item homogeneity estimates ranged from 0.84 to 0.95, which suggests that the POMS comprises narrow scales. Three-week retest correlations ranged from 0.65 to 0.74, which suggests that the scales are not sensitive to transitory fluctuations in depressed mood (Spielberger, 1972, p. 388). Eichman (1978, p. 652) criticized the high face validity of the POMS, and pointed out that the item transparency of the scales makes the measure susceptible to deliberate faking and response sets. Likewise, Weckowicz (1978, p. 651) criticized the lack of sensitivity of the POMS to transient mood alterations. Partial confirmation of the factor structure of the POMS was reported by Jacobson et al. (1978). However, they employed a principal components analysis with varimax rotation. Their factor number was determined by the outmoded eigenvalue greater than unity method. None of the eight recommended factor analytic procedures was employed. It is unlikely, therefore, that their reported factor solution attained maximum simple structure.

Nevertheless, the POMS represents an advance over the MAACL since it attempts to measure additional mood states. It contains a smaller proportion of reversed items than does the MAACL (which has half its items worded in the reverse direction). Boyle (1979b, p. 79) has cast doubts on the validity of reverse-worded items in self-report measures. Such
items were shown to index an entirely different construct, rather than simply the negative pole of a bipolar construct. Blumenthal (1975) has specifically questioned the use of items negating depression. This difficulty has been apparent also in some single measures of depression such as the SDS (Jegede, 1976). Despite these limitations, the POMS potentially has much to offer (at least in the measurement of depressed mood) as compared with the simpler single-scale measures of depression alluded to above.

The 8SQ measures eight important mood states (anxiety, stress, depression, regression, fatigue, guilt, extraversion and arousal). The instrument comprises 12 items per scale which are 'cycled' in order to prevent spurious contiguity effects. Based on several factor analyses along the lines already suggested above, the construct validity of the eight state scales seems satisfactory (Curran & Cattell, 1976, p. 4). However, there is excessively high intercorrelation between the eight scales. Like the MAACL, the range of inter-scale correlations rises up to the mid-0.8 level. Test–retest reliability ranges from 0.88 to 0.96 for immediate retest, and from 0.26 to 0.48 for retest after one week. These test findings accord with state–trait expectations. Thus one would expect high dependability (immediate retest) correlations, but low stability (retest after one week or longer) correlations for a situationally sensitive, but reliable, state measure, as rightly pointed out by Zuckerman (1983). The proviso is, however, that the overall mood on the second measurement occasion differs from that on the first, and substantially. Combination of both Forms A and B of the 8SQ boosts the stability coefficients significantly, which suggests the advisability of using the longer version of the instrument where possible. Against this, however, is the increased measurement time required, which may be unacceptable, not only when testing time is limited but also when emotional states are likely to fluctuate rapidly (i.e. within the time of measurement itself).

Kleinmuntz (1978, p. 783) pointed out that a major weakness of the 8SQ (and indeed of all self-report inventories) is its susceptibility to response sets such as social desirability and even deliberate faking. Kline (1979, p. 170) reviewed the evidence and concluded that the eight states measured in the 8SQ are valid, but that additional mood states need to be added to the instrument. Evidently, the mood state sphere is not adequately measured by the 8SQ—a conclusion also reached empirically by Boyle (1983c).

The 8SQ depression factor is at the second stratum level (Cattell, 1979, p. 205), and is related in a structured manner to seven primary factors of depression (see Cattell, 1979, pp. 73–82). In this light, the single measures of depression such as the MMPI–D, BDI, SDS and CES–D are evidently multidimensional, despite being narrow measures of these dimensions (given their high levels of item homogeneity/item redundancy). These measures appear to be tapping unknown amounts of some or all of the seven primary depression factors elucidated factor analytically in Cattell’s empirical research. The 8SQ Depression scale provides an overall index at the surface syndrome level of these seven primary depression factors. Like the POMS, the 8SQ allows important relationships among different mood states to emerge which otherwise would remain hidden if observation were limited only to a single scale of depression. The major disadvantage of multivariate mood state instruments is, however, that such scales are virtually of no use in investigating the epidemiology of depression. For any given individual, the surface syndrome of depression, which involves a pervasive state–trait condition, may or may not also involve depressed mood as one of its components.

Another multivariate instrument, the DES (see Boyle, 1984d), measures 10 fundamental emotions defined as having neural, neuromuscular-expressive, and experiential components. The DES indexes depressed mood not only in terms of its Sadness scale, but more uniquely in terms of a pattern of fundamental emotions. The 30-item version of the DES has been subjected to several factor analyses, which have suggested that depressed
mood involves the scales of Sadness, Shame, Guilt and Fear (Izard et al., 1974, pp. 18-19). In a subsequent expansion of the DES measure, an additional scale of Hostility (inner-directed) was also shown to be involved in the pattern of fundamental emotions relating to depression. Recent factorings of the DES (e.g. Emde, 1980; Fuenzalida et al., 1981) have demonstrated some support for 6-8 of the 10 postulated mood states (cf. Mosher & White, 1981, p. 62). Unfortunately, these factor analytic results remain somewhat doubtful given the lack of adherence to most of the factor analytic guidelines above, other than, perhaps, the use of an oblique (but analytical only) promax rotation. Regrettably, no attempt to ascertain the extent to which simple structure had been approximated was apparent in any of these factorings of the DES.

Despite the uncertainties surrounding the factor analyses of the DES (what is really required at this point is a confirmatory factor analytic study of the DES structure—see Nunnally, 1978, pp. 347-349, 388-390, 402-403), Izard's claim to have discovered fundamental emotions, which are universally discernible in the facial expression, is appealing. Unfortunately the DES items have high face validity and therefore are necessarily susceptible to the influence of response sets. Also, the instrument has a mere three items per scale, which is not conducive to satisfactory reliability. While the DES may prove useful in classifying depressed patients in terms of a pattern of fundamental emotions, the instrument nevertheless needs an increase in the number of items per scale. However, of the three multivariate measures discussed, the DES (along with the 8SQ) appears to offer the most promise as a measure of the several underlying emotions which seem related to depression. Izard must be commended for his unique conceptualization of depression as a pattern of fundamental emotions.

E. Specific problems with depression measures

Apart from the major areas of weakness discussed in the preceding sections, examination of the existing literature reveals a number of more specific problems with many self-report measures of depression. For example, a study by Meites et al. (1980) demonstrated that the BDI appears to tap an emotionality factor of stability-instability, rather than being primarily a measure of depression. Meites et al. compared the BDI and SDS with the Taylor Manifest Anxiety Scale (TMAS), and with the Eysenck Personality Inventory Neuroticism scale (EPI-N) on a sample of 170 undergraduates. Results indicated that all four instruments were highly intercorrelated, which suggested that all four measures tap the above emotionality factor. According to Meites et al. (p. 431):

The present results call into question the validity of...using a single personality measure...if a single measure of depression were used...the group found to be high on that trait would be likely to score above the mean on a test of anxiety and...on a general neuroticism survey...the results would be confounded.

Meites et al. therefore rightly questioned the validity of the BDI and SDS as measures of depression. Mayer (1977, p. 366) admitted that great variations in correlations (0.2-0.9) between BDI scores and global ratings of depression made both by the psychiatrist and the patient implied that the validity of the instrument was suspect. Kearns et al. (1982) concluded after reviewing the evidence that the BDI should no longer be used for research into depression! Despite these difficulties, the BDI has remained popular (e.g. King & Buchwald, 1982; O'Hara et al., 1982).

Reviews of the validity of the SDS have consistently been negative (e.g. Becker, 1974, p. 25; Goodstein, 1975; Mayer, 1977, p. 370). As Goodstein pointed out (p. 538), the SDS is a psychometrically crude instrument and should not be used to measure depression. The specific limitations of this instrument are too numerous to list here. It is little wonder that
the SDS is insensitive to various levels of depressive severity (Carroll et al., 1973). Claims supporting the reliability and validity of the SDS (e.g. Yanagida & Marsella, 1978) can only be regarded with scepticism. The SDS is an inadequate poor measure of depression, subjected to much incompetent research (e.g. in an examination of the factor structure of the SDS, Steuer et al., 1980, factored the scale on a mere 60 individuals—clearly an insufficient sample size for a valid factor analysis).

The IPAT Depression Scale has also had doubts cast on its validity, since it correlates more highly with five MMPI scales not related to depression than it does with the MMPI-D scale. It also correlates highly (0.8) with the IPAT Anxiety Scale (Mayer, 1977, p. 380; LaVoie, 1978, p. 859). LaVoie pointed out that there are virtually no useful data on its construct validity. Lykken (1978, p. 584) questioned the use of statistics which have not been cross-validated in an attempt to demonstrate the scale’s validity. As Lykken indicated, the criterion validity of the IPAT Depression Scale is unclear. Its ability to differentiate depressed patients from other psychiatric patients, and from normals, using samples other than those involved in the development of the measure remains unknown. As a single scale of depression, this instrument is extremely limited.

Reviews of the MAACL and DACL have also been discouraging (Mayer, 1977, pp. 382–390). Specifically, Mayer concluded that the MAACL lacks adequate discriminant validity, and that the reverse-worded items measure different constructs. The difficulty with reverse-worded items has been noted above (Boyle, 1979b). In this regard, the DACL has only one-third of its items reversed, which is not as problematic as for the MAACL. As with the 8SQ, the scales of the MAACL are too highly intercorrelated. Hence the three scales are not independent. Megargee (1975, p. 490) concluded that the Depression scale of the MAACL probably measures more anxiety than depression. Both the MAACL and DACL are highly subject to response sets (see Kelly, 1975, pp. 488–489; Mayer, 1977, p. 391). Mayer (1977, pp. 389–390) also pointed out that the DACL correlates more highly with scales of the MMPI other than the MMPI-D. Given the large standard deviations in DACL scores, Mayer suggested that the instrument is not adequate for identification of depressed individuals, even though it might detect group differences in depression. McNair (1975, p. 351) also criticized the DACL for its lack of reliability and validity. Levitt & Lubin (1975, p. 70) asserted that mean DACL scores showed a clear progression from depressed patients through non-depressed patients to normal individuals. However, close scrutiny of their data reveals that this effect occurred only for males. Non-depressed female patients obtained a higher mean DACL score than did depressed females! Since there were only 28 male patients, as compared with 100 depressed female patients, and only 56 non-depressed male patients, as compared with 118 non-depressed female patients (on sex differences in depression, see Abramson & Andrews, 1982; King & Buchwald, 1982; Scott et al., 1982), the optimism of Levitt & Lubin for the discriminant validity of the DACL was astonishing. Conclusions drawn from studies using the DACL (e.g. Johnson et al., 1979) must therefore be regarded as dubious.

Conclusion

Taken together, the above considerations cast doubt on the adequacy of much of the psychometric research into depression. Many of the self-report instruments have been shown to have doubtful reliability and validity. Factor analytic findings have been conflicting, suggesting a lack of construct validity and the probable use of ad hoc factoring procedures in many instances. Several recommendations for better measurement of depression, both as a surface syndrome and in terms of its constituent components or attributes, have been proposed. Included also were suggestions regarding the desirability of moderate rather than maximum retest reliability and item homogeneity (internal
consistency) estimates. The use of appropriate criteria for conducting a valid factor analysis was stressed, as well as recourse to multivariate rather than unitary measures of depression. Additionally, a number of more specific criticisms of self-report measures of depression were discussed, with the purpose of alerting clinicians to the strengths and limitations of existing self-report depression scales.

Acknowledgement

The author, who is currently at the Institute of Catholic Education, Australia, is indebted to Dr Wallace Maw, University of Delaware, who suggested that this review be undertaken.

References


Boyle, G. J. (1984b). Effects of viewing a road trauma film on emotional and motivational factors. Accident Analysis and Prevention, 16 (5).


