Methods of assessing personality for epidemiological study

J E J Gallacher

"Personality has different meanings for theologians, philosophers and sociologists, and in psychology it has been used in many ways" wrote G W Allport some 50 years ago. Writing today, he might well have included epidemiologists amongst those challenged to develop their own approach to the problem of personality. Popularly, personality is understood largely in terms of traits or types using categories like Eysenck's "introvert" or "neurotic". In a more technical sense, the term "personality" is a useful rubric for the more complex aspects of psychological function which account for consistent patterns of behaviour characterising individuals. Implied in this definition are concepts like intelligence, emotional disposition, and patterns of social interactions. This distinguishes personality tests such as IQ tests from physiological tests such as galvanic skin response and measures of the social and physical environment such as a life events scale.

Personality assessment began with concern over mental health. The first personality tests were measures of general intellectual performance and were developed to identify children requiring remedial education. Tests of emotional adjustment were then developed for military personnel selection in World War I. These applications illustrate the two main ways personality assessment has been used. In the first, personality is conceived as the dependent variable of a pathological process. In the second, personality is an independent variable predicting behaviour. Early in the history of epidemiology, personality was considered almost exclusively as a dependent variable of a disease process. This confounded the types of legitimate hypothesis largely to those involving mental illness or poor intellectual performance. More recently, however, personality is being increasingly considered as an independent variable in a variety of health related processes, including disease processes, therapeutic processes, and public health programmes. This simple development of thought has facilitated a much greater variety of hypothesis involving personality. For this reason it is becoming increasingly important for the epidemiologist to be familiar with the opportunities afforded by, and the limitations of, personality assessment.

Principles of personality measurement

RATIONALE

The rationale behind assessing personality from an epidemiological point of view is straightforward. The aim is to predict behaviour which is in some way related to disease. Personality assessment is merely a series of methods designed to sample the thoughts, feelings, and actions of the individual in a way that will allow accurate and efficient prediction. Naturally, constant and direct observation is the most accurate method but increasing efficiency by the sampling of behaviour and its antecedents is generally considered to be well worth some loss in accuracy. The range of personality tests reflects the wide variety of ways in which the factors used to predict behaviour are sampled. For example, behaviour itself may be sampled either directly as in a performance test or indirectly as in a self report test. Antecedents of behaviour, such as emotional disposition, may also be of predictive value, but these can only be assessed indirectly as they cannot be observed.

The principles of measurement as applied to personality are no different from those in any other discipline: the measure should be of a clearly defined construct, of known repeatability, of known validity, and administered according to a standard procedure. Two factors, however, have proven to be critical in the shaping of personality assessment. The first is that personality is itself an abstract concept. Although personality tests are designed to predict behaviour, the concepts used to explain the behaviour are abstract. For example, IQ score is based on performance which is directly observable, but the concept of intelligence used to make sense of the score is not. This problem is not unique to personality assessment. A similar problem is faced by physicians in trying to define health and biologists in trying to define life. The abstract nature of personality means that ultimately personality tests measure themselves, as there can be no independent verification of their performance. Nevertheless, not all tautologies are sterile and personality testing has been found to be very useful in spite of this inability to establish its objectivity independently.

The second factor that has shaped personality assessment is that subjects are active participants in the measurement process. Personality tests are done by subjects and not to them. As a result the measurement process itself can contribute towards the test score and may bias it with respect to the hypothesis being tested. For example, scores on IQ tests increase with practice but this does not necessarily signify an increase in IQ!

CLEAR DEFINITION

If these two factors are applied to each of the principles of measurement in turn, the constraints
of personality testing become clear. The need for clear definition of constructs is a difficult problem when the construct is wholly abstract, as there is no observable foundation upon which agreement can be based. This problem has been a major impetus towards the great diversity of theories and models of personality which have been developed.

**REPEATABILITY**

In strict statistical terms the repeatability of personality tests cannot be directly assessed, as truly replicate measurements cannot be made. This is due to the subject participating in and being affected by the measurement process. For example, learning affects the repeatability of performance tests while familiarity with a testing procedure reduces anxiety, affecting the self report of thoughts and feelings. Convenienly, therefore, the consistency of test performance is termed "reliability" and refers to a variety of procedures which attempt to provide an assessment of measurement error. The assumption underlying these procedures is that the precision of an estimate can be gained from a series of repeated measurements.\(^\text{15, 16}\) This argument is applied crudely with regards to personality assessment as, when comparing differences in measurements, no attempt is made to distinguish between biological and measurement variation or between systematic and random variation. Nevertheless, the resulting procedures provide simple and conservative estimates of a test's precision by comparing equivalent measurements of the same construct to identify the level of agreement.

**Test-retest reliability** refers to the repeating of the test on a second occasion. The reliability coefficient in this case is simply the correlation between the two test scores. Of particular importance is the interval between testing, as retest correlations will tend to diminish as the interval lengthens. Practice and learning effects apart, the retest correlation tends to show the generalisability of scores between testing occasions. **Alternate form reliability** refers to the comparison of scores from two alternative forms of the test on the same or equivalent populations. The reliability coefficient is the correlation between the two scores. Although this type of reliability overcomes the problems of test-retest reliability there are few tests which have alternative forms and so in practice it is rarely used.

**Split-half reliability** refers to the comparison of scores between comparable halves of a single test. The reliability coefficient is the correlation between halves of the test and is a measure of the internal consistency of the test. The internal consistency of a test may also be assessed by comparing scores between all items. This is called **inter-item reliability** and may be calculated using a formula derived by Kuder and Richardson.\(^\text{17}\) Internal consistency may also be used to assess the efficiency of measurement but this can be a doubled edged sword: too low a reliability coefficient and more items are needed to improve the consistency of measurement but a very high reliability coefficient suggests some items could be discarded with little loss to accuracy. Therefore reliability coefficients based on the principle of internal consistency are rarely above \(r = 0.8\).

The effect of adding or discarding items is given by the Spearman-Brown formula\(^\text{3}\) and is often quoted in test manuals to show that a reasonable balance has been achieved between consistency of performance and length of test.

**VALIDITY**

Due to the abstract nature of personality the validity of a test can only be established indirectly. This problem is not confined to personality assessment as it also arises in measures of symptoms such as in chest pain questionnaires assessing coronary heart disease.\(^\text{18}\) As with measures of chest pain, the difficulty of establishing the validity of personality tests is considered an acceptable price to pay for the proven worth of a good test. There are three main types of evidence used to assess validity. **Content validity** refers to the contents of the test items. The question asked is "Do they cover the areas needing to be covered?" **Construct validity** refers to the relationships between test scores and those assessing other constructs. The question being asked is "Are the relationships those that would be expected?" **Criterion-related validity** refers to the measure's ability to predict a specific outcome. The question being asked is "Does this test predict, for example, performance or health or job promotion?"\(^\text{19}\)

**STANDARD PROCEDURE**

The importance of standard procedure is emphasised in personality assessment due to the subject's active participation in the measurement process. From the subject's point of view there are two main sets of factors which affect every assessment situation. The first is the test taking habits that the subject brings to the situation.\(^\text{16}\) These personal biases are called "response sets" and include acquiescence (a tendency to accept or reject all items as applying to oneself), evasiveness/extremeness (giving indifferent or extreme responses), social desirability (giving responses that are socially acceptable), and cautiousness (omitting difficult items in a performance test).\(^\text{20}\) The second set of factors is the demand characteristics of the test. Subjects respond to what they think they are required to do. This frequently includes helping the experimenter to achieve the hoped for results, rather than providing a fair test of the hypothesis. An experimenter's own expectation may affect the data, therefore, by producing behaviour which inadvertently affects the subjects and so achieves a self fulfilling prophecy.\(^\text{21}\) Experimenter age, sex, race, and socioeconomic status have all been shown to affect test scores.\(^\text{3}\) Environmental factors are also important—eg, lighting and noise level. Although neither response sets nor demand characteristics can be eliminated their effects can be held constant throughout a study by maintaining standard procedures.

**Methods of assessment**

In spite of the difficulties of personality assessment it has proven capable of predicting behaviour sufficiently accurately to be of scientific use. The wide variety of assessment procedures may be easily understood in terms of four
main types of test comprising performance, self report, structured interview, and projective tests. Each method has its own strengths and weaknesses and applicability to population studies. The main characteristics of the tests cited below are given in the table.

**Performance Tests**
In performance tests, scores are derived from direct observation of the subject performing a task. Performance testing reduces to a minimum issues of observer bias and subject bias as the subject can either perform the task or not. The limitation of performance testing involves generalisability, as few performance tests can be related to more remote aspects of personality. A notable exception to this rule is the Porteus maze test which can be used to assess obsessiosity.22

The lack of generalisability tends to limit performance testing to specific abilities such as numeric and verbal reasoning or dexterity skills. With regard to their use in population studies performance tests tend to be intimidating, as they invariably include items that a subject cannot do, or cannot do well. They can also be long and are usually administered individually. Increasingly performance tests are being computerised, which reduces the need for special training and allows more subjects to be assessed simultaneously. The best known type of performance test is the IQ test, for example the Wechsler adult intelligence scale.23 The debate over racial differences in IQ scores illustrates the difficulties in extrapolating from performance scores to personality.

**Self Report Tests**
Self-report questionnaires provide data which are more generalisable than those from performance tests. The questionnaire format facilitates standardisation of question and response but also limits the flexibility of the instrument. Self-report questionnaires are best used to enable the subject to report thoughts and feelings as well as behaviour in situations which cannot easily be observed directly. The relative ease of standardisation of procedure, the variety of areas which can be covered, and the possibility of group measurement make self-report questionnaires very attractive for use in population studies. Self-report tests which have been used in population studies include the general health questionnaire assessing psychiatric caseness,24 the state-trait anxiety inventory,25 the Beck depression scale,26 and the Jenkins activity survey assessing type A behaviour.27 Although self-report questionnaires minimise bias from the observer, opportunities for bias in the subject remain, with problems of self-report bias and demand characteristics of the measurement situation and the subject’s response sets, such as social desirability and acquiescence discussed earlier. In addition to the self report scales described above, which tend to assess specific aspects of personality, the self report format has also been used in developing more comprehensive assessments of personality. These are what are popularly understood as “personality tests” and include the 16PF (measuring 16 personality factors),28 the Eysenck personality inventory,29 the Myers-Briggs type indicator,30 and the Minnesota multiphasic personality inventory31 among others. Although these tests have been carefully developed they are time consuming and can be intimidating. This limits their applicability to population studies.

**Structured Interviews**
Structured interviews provide a more flexible framework for response than self report tests, as the interviewer can probe the responses to ensure their relevance to the study. In a structured interview questions are asked in order to obtain biographical information and to observe the subject directly. The most widely known structured interview is the one used to assess type A behaviour.32 In this interview the subject is scored by his or her self reports of stress and

<table>
<thead>
<tr>
<th>Type of test</th>
<th>Instrument</th>
<th>Reference No</th>
<th>Construct measured</th>
<th>Number of items</th>
<th>Duration (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance (administered individually)</td>
<td>Porteus maze test</td>
<td>22</td>
<td>Intelligence, Impulsivity</td>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Wechsler adult intelligence scale</td>
<td>23</td>
<td>Obsessivity</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>Self report (can be administered in groups)</td>
<td>Beck depression scale</td>
<td>26</td>
<td>Intelligence, verbal IQ (6 scales)</td>
<td>21</td>
<td>5-10</td>
</tr>
<tr>
<td></td>
<td>Eysenck personality inventory</td>
<td>29</td>
<td>depression, performance IQ (5 scales)</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jenkins activity survey</td>
<td>24</td>
<td>Global personality, introversion/extraversion, neuroticism</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>General health questionnaire</td>
<td>27</td>
<td>Psychiatric caseness, Type A behaviour, direction of aggression</td>
<td>52</td>
<td>20</td>
</tr>
<tr>
<td>Structured interview (administered individually)</td>
<td>Thematic apperception test</td>
<td>36</td>
<td>Structure of personality, social support</td>
<td>20</td>
<td>Variable</td>
</tr>
<tr>
<td></td>
<td>Interview schedule for social interaction</td>
<td>33</td>
<td>availability of attachment, adequacy of attachment, adequacy of social integration</td>
<td>52</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Standardised psychiatric interview</td>
<td>34</td>
<td>Psychiatric caseness, somatic symptoms, cognitive abnormality</td>
<td>Variable</td>
<td>10-60</td>
</tr>
<tr>
<td></td>
<td>Video structured interview</td>
<td>32</td>
<td>Type A behaviour, hostility, time urgency</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>Projective (administered individually)</td>
<td>Rorschach inkblot test</td>
<td>35</td>
<td>Maladaptive personality</td>
<td>10</td>
<td>Variable</td>
</tr>
<tr>
<td></td>
<td>Roseneaig picture frustration study</td>
<td>37</td>
<td>Aggression, direction of aggression, type of aggression</td>
<td>24</td>
<td>Variable</td>
</tr>
<tr>
<td></td>
<td>Thematic apperception test</td>
<td>36</td>
<td>Structure of personality</td>
<td>20</td>
<td>Variable</td>
</tr>
</tbody>
</table>
according to positive gesticulation and voice characteristics. Structured interviews have also been used to assess social networks and psychiatric caseness. From an epidemiological viewpoint the strengths of a structured interview are its flexibility of administration and the fact that it allows direct observations of the subject. The limitations of a structured interview are the requirement for intensive training of the interviewer, the difficulties in maintaining quality control of an essentially subjective scoring procedure, and the need to interview subjects individually. Quality control in terms of consistency between interviewers and consistency over time for the same interviewer can be assessed and standards maintained but it requires careful planning and is a burden on interviewer resources.

PROJECTIVE TESTS
The last category of personality tests are projective tests which require the subject to interpret an ambiguous stimulus, i.e., to project his or her personality onto the stimulus. The response format is generally very flexible as the emphasis is on freedom of expression rather than ease of quantification. Examples of projective tests include the Rorschach inkblot test, the thematic apperception test, and the Rosenzweig picture frustrations study. Projective tests are generally unsuitable for use in population studies, usually require intensive training to be used competently, have poor validity and reliability (except in some specific clinical applications), can be time consuming, and cannot be administered in groups.

Interpreting personality data

TRUSTWORTHINESS
Interpreting personality data involves the two issues of trustworthiness and interpreting causality. Untrustworthy findings can be due to random or systematic error. Sources of error which are peculiar to personality assessment are the response sets and demand characteristics described earlier. Response sets and demand characteristics can have a random element due either to how people are "feeling" on the day of the test or to poor standardisation of measurement procedures. Random error results in the misclassification of individuals due to an increased standard error of measurement and obscures associations. In clinical practice, the potential for misclassification of individuals is so important that specialised training is necessary for the use of personality tests. In epidemiology, which is concerned about classifying groups, any effect of the standard error of measurement can be reduced by increasing sample size, although this will not affect associations being obscured.

Response sets and demand characteristics tend to be consistently expressed by individuals and by experimenters, and so they may introduce a systematic element to measurement error. Systematic error distorts mean values but if the error is of the same magnitude for all groups, the differences between them are not compromised and neither are any associations. In case-controlled, cross-sectional, and cohort studies the assumption of a constant level of systematic measurement error across the sample calls for careful judgement. In intervention studies this assumption may be more confidently made as all spurious differences between groups are taken account of in the randomisation process.

CAUSALITY
Epidemiologists will be more aware than most of the difficulty in inferring causality from complex data. For personality data the critical factor in establishing causality is determining the sequence of events. The propensity for change of personality data means that the sequence of events cannot be assumed when testing a hypothesis. Consequently case-controlled and cross sectional studies are only, generally speaking, of exploratory value. For example, an association of social class and prevalent heart disease may be sufficiently strong to suggest an effect of lifestyle on heart disease. A similar conclusion could not be drawn from an association of stress with prevalent heart disease, since stress score may be affected by health status. The sequence of events is not important when using personality variables as matching variables in the selection of controls or as covariates in statistical analyses. In case-controlled and cross sectional studies, therefore, personality variables are of most value as control variables. For hypothesis testing involving personality, in general the sequence of events should be firmly established in the design of the study. In cohort studies baseline personality measurement may be used to predict incident disease and in intervention studies a psychological intervention may be used to change a personality variable in order to reduce subsequent morbidity or mortality.

Selecting a personality test
Selecting a personality test for use in an epidemiological study is not simply a matter of browsing through test manuals, consulting the local psychology department, or following the literature. Test manuals usually provide inadequate information, as few tests have been developed with an epidemiological application in mind. For example, rarely have tests been developed on representative population samples or with the need to achieve high response rates. Consequently, descriptions of test populations appear to be casual and uninformative. Although most psychologists are eager to be of assistance few are
Methods of assessing personality in epidemiological study

aware of the practical constraints of epidemiological fieldwork in which the time window for collecting personality data is strictly limited. As a result, their desire to see the "best" data collected (from a psychological point of view) can be an enemy of "good" data being collected. Published reports can also be misleading as the confident report that "such and such" a test was used may conceal the fact that only two of the test's four subscales were actually used and these were "modified" according to the "clinical judgement" of the scientists involved. The result of this is that the epidemiologist embarks on the use of a test which is either invalid for the purpose or is impracticable.

To help the epidemiologist select a personality test the model described below shows the main factors to be considered and the sequence of judgements to be made (figure). In the model the population to be studied, the requirements of the personality test under consideration, and the resources available for the study are all independent variables. Their combined effects must be considered on the response rate, quality of data, and personality score in turn. For example, in a study of men aged 60-74 years the use of a long and comprehensive personality test is likely to be tiring and threatening. Many resources in terms of time and manpower are likely to be required, therefore, if such a test is to be used.

In cross sectional and case-controlled studies the personality test used is unlikely to affect response rate. An exception to this is a study of closely knit communities in which the "word gets around". In cohort studies an unfortunate choice of test may, however, adversely affect follow up response rates. In intervention studies a poor choice of test may result in poor compliance and a reduced completion rate.

Although a test may not adversely affect response rates this is no guarantee that it will provide data of consistently good quality. In a postal survey, for example, even if a high response rate is achieved through repeat mailings, many of the responses may be incorrect or incomplete. Usually psychological tests are developed under ideal conditions and so it is assumed that all questionnaires are correct and complete. Therefore there is little or no guide as to how the test will perform in the field.

Finally, the characteristics of the test score may be considered. How will the combination of clinic procedures to be used in the study affect the test score? At what point in the data collection process should personality be assessed? Usually there are no established answers to these questions and no substitute for a feasibility study to assess these likely effects.

Conclusion

In considering personality from first principles through to selecting a test it is clear that personality assessment affords many opportunities for epidemiologists and clinicians alike. It is also clear that including personality assessment in epidemiological studies requires skill and care derived from an appreciation of the limitations of the various assessment procedures. As epidemiologists become increasingly familiar with these issues it is hoped that personality assessment will become increasingly sensitive to the needs of epidemiology.