Section of Measurement in Medicine

President J P Shillingford MD

Meeting 20 January 1969

A Growing Edge of Measurement of Feelings [*Abridged*]

Dr R C B Aitken
(*University Department of Psychiatry, Royal Edinburgh Hospital, Edinburgh*)

Measurement of Feelings Using Visual Analogue Scales

Feelings are states of the self, and incorporate moods and sensations. Although a person may appreciate precisely his state on a selected dimension, words may fail to describe the exactness of the subjective experience. The paucity of suitable quantitative terms in common speech limits the amount of information which can be transferred. Continuous phenomena have to be graded in artificial categories. A digital system is imposed on the observer, when the freedom of an analogue system would be welcome.

An understanding of many problems in clinical research presupposes that it is possible to communicate the desired information from patient to clinician in a way amenable to measurement. A working party of the British Association defined measurement as 'the assignment of numerals to things so as to represent facts and conventions about them' (Stevens 1946). For the measurement of feelings, communication based on a simple visual analogue seems appropriate. Lines, with their boundaries clearly defined as the extremes of the feeling, serve well for marking (Hayes & Paterson 1921).

The limitations of an analogue system are no more than those true of words (Fig 1); even speech contains an assumption that the same language is being spoken in order to communicate information, though this assumption may be far from true. The same word can be used with different meanings, and need not imply that people experience the same feeling. The same amount of change may take place, but in only some people will this alter a category term, and then only if the change is from a certain initial value. With verbal scales, it is unknown whether the category sizes or differences between two adjoining ones are the same. Strictly speaking, discrete scores are not additive, and this should preclude certain sensitive statistical procedures. There is a tendency for resolution to be limited since commonly only the central few categories are used. Such scales are inadequate for examination of the exact association to a related concept, such as the physical magnitude of a stimulus. Category scales fail to grasp nuances of feeling.

There is no claim that the use of an analogue scale permits liberal comparisons, as the transfer function is unknown. The same word used by different people need not convey that they experience the same feeling, neither does comparable positioning of marks on lines. A feeling of twice the intensity cannot be inferred by a mark giving twice the score of another. The interpretation of marks on visual analogue scales is governed by the rules under which they are made; as long as positions of marks are referred to, statements appropriate to them can be used. However, com-

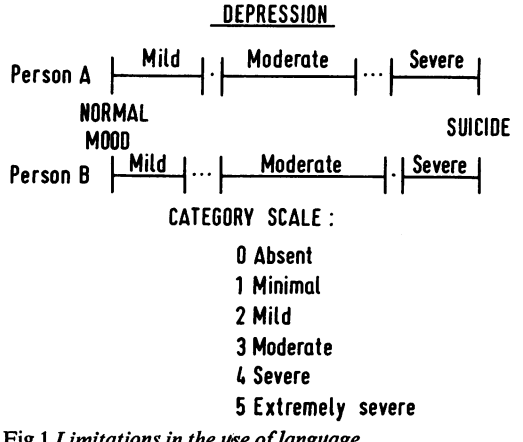

Fig 1 *Limitations in the use of language for the measurement of a feeling*

parisons can be achieved with greater sensitivity than with semantic phrases, particularly between different occasions in the same person.

Analogues should be visual rather than verbal equivalents, otherwise preferred digits – e.g. 0 or 5 – appear too often (Yerkes & Urban 1906). This is a common failing of examination question marking in numbers; it can give the illusion that the resolution is greater than takes place in practice.

Freyd listed in 1923 many advantages of 'graphic rating', and reported simple rules for its use. Let me illustrate the application to the measurement of certain feelings, and in doing so draw attention to aspects of the statistical analysis.

Non-parametric Statistics

Judgments on descriptive material: This requires comparison of created feeling in the observer. It is tedious to rank more than a handful of texts, and it is impossible to do so at one reading if information on more than one item of assessment is requested simultaneously. The use of visual analogue scales can overcome some of these difficulties, and be more sensitive than the direct awarding of scores. Scores from markings on lines can be ranked afterwards, and so be amenable to a variety of statistical procedures (Siegel 1956).

This use can be illustrated by an experiment investigating aspects of aircrew behaviour that might influence the occurrence of flying accidents due to pilot error (Aitken 1962). Structured interviews were conducted, designed to elicit relevant topics, and their content transcribed in typescript. Having read these in random order, judges then marked lines to convey their assessment of the degree of relevance of certain topics in each pilot's case. By calculation of Kendall's coefficient of concordance, reliability was found to be highly significant for each topic. Kruskal-Wallis one-way analysis of variance was used to assess the significance of differences between subgroups of pilots, and Spearman's rank-correlation coefficient to assess association with other scores.

Parametric Statistics

Appropriate transformation to achieve normal distribution: The use of 100 mm lines with measurement to a resolution of 1 mm is convenient. The lines can be placed horizontally on a page, and in my experience are acceptable in gestalt.

Markings dividing lines into two portions give a multinomial distribution of the proportions, which are often skew to one or other end. Nearer to normal for all such distributions can be achieved by the arcsin transformation (Snedecor & Cochran 1967). This can be illustrated by a study where 90 fighter pilots were asked to indicate how apprehensive they might feel in ten defined situations (Aitken 1967). They were presented with 100 mm lines with the extremes defined as maximal relaxation and maximal panic.

Fig 2 gives the frequency distributions from three examples; it illustrates the diversity of skewness for the raw scores, and the improvement toward normality by arcsin transformation. Incidentally, the mean coefficient of variation for the scores in ten different situations was reduced by this transformation from 0·65 to 0·43.

Analysis of variance and convariance: The scores from visual analogue scales are in a form which is ideal for examining the significance of differences in distributions. In a clinical trial of two hypnotic drugs and a placebo, the quality of sleep was assessed with a visual analogue scale (Aitken *et al.* 1965); significant differences between the drugs and the placebo were demonstrated. Similar discrimination was apparent as regards sleep duration estimated by nurses.

Though objective measures of the effects of hypnotics can be made and conclusions drawn as to their relative efficacy, the cardinal discriminator is the patient's subjective experience.

Attitude to conditions in the environment can be determined very simply by using visual analogue scales. Aitken *et al.* (1963) asked subjects how much they preferred a selection of five frequencies of flashing lights; these were pre-

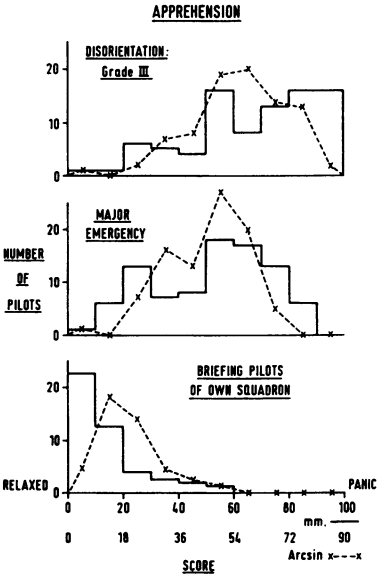

Fig 2 *Histograms of the frequency distributions of scores on apprehension, both raw and arcsin transformed*

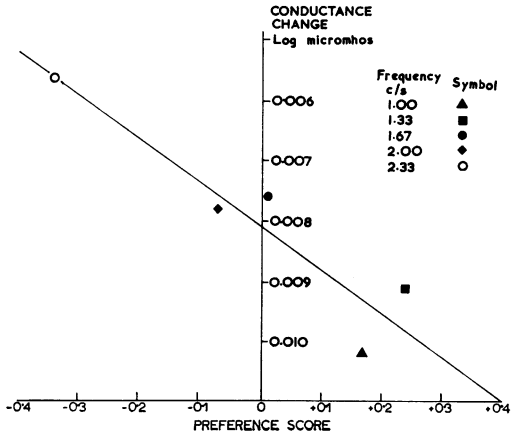

Fig 3 *The relationship between attitude (positive score represents preference) and change in skin conductance when 10 subjects were exposed to five frequencies of flashing light* ($y = -0{\cdot}0072x - 0{\cdot}0081$; $P = 0{\cdot}05$)

sented in pairs, a line being partitioned to convey strength of judgment (Metfessel 1947); the scores were amenable to a sensitive analysis (Schéffé 1952) which gave a clear answer on the most preferred frequency range. Those attitudes could be compared with conclusions obtained from examination of a physiological indicant of arousal (Fig 3).

In another experiment, subjective appreciation of alertness was assessed, and shown to be increased by performance of a task or exposure to irrelevant distraction (Aitken & Gedye 1968). The experiment was conducted to test the hypothesis that time perception was dependent on arousal level. The alertness score was used for the adjustment; but as it accounted for only a negligible part of the variance, no support was provided for the hypothesis.

In each of these three examples – measurement of feelings about sleep, preference and alertness – answers were provided to *practical* questions. Also the variances attributable to other effects such as order of stimulus presentation and their interactions were removed in the analysis. The scores fulfilled the necessary requirements for analysis of variance – homogeneous and independent variance, normal distribution, and sufficient resolution in measurement to provide continuous rather than discrete scores, so allowing the effect of factors to be uniformly additive.

Calculation of regression and correlation: Little can be gained from calculation of correlation between two sets of visual analogue data for the same subjects as it will be unknown how much it reflects only the subject's 'set' to communicate his feelings. The scores from visual analogue scales, particularly if multiple readings are obtained from each subject, are more suitable for examining relationships to other observations or conditions.

In the experiment just mentioned on alertness, the scores were found to be related in rectilinear fashion to change in skin conductance, and in an interesting way. Extrapolation of the regression showed that the centre of the line – normal alertness – corresponded almost exactly to no change in skin conductance; and the end denoting maximal relaxation to the same rate of decrease in skin conductance found in monkeys when they are falling asleep (Richter 1929).

In the experiment to determine attitude to frequency of flashing light, those frequencies most preferred corresponded to those where a greater fall in skin conductance was measured, indicating subjects were less aroused (Fig 3).

In the clinical trial of hypnotics, the product-moment correlation coefficient between subjective rating on quality of sleep and nurses' observation of duration was 0·71 ($P = 0{\cdot}001$). There was also a significant correlation with age ($-0{\cdot}57$: $P = 0{\cdot}025$) – the older the patient the less well he reported he had slept. McGhie & Russell (1962) noted a similar observation.

Good correlation between patient and observer was also noted in a clinical trial of poldine in the treatment of dyspepsia. The active drug undoubtedly reduced gastric acidity (Hunt & Wales 1966), but symptom reduction was no greater than during placebo treatment. The doctors had marked lines independently after clinical interview and their scores correlated highly with those obtained from the patients ($r = 0{\cdot}59$: $P = 0{\cdot}001$).

Let us look in more detail at two experiments in which a resistance to breathing was applied in an external airway, and subjective sensations noted (Aitken 1965).

In the first, the apparatus was rather crude. Five subjects breathed through nine valves with known pressure/flow characteristics with nominal values between 0·25 and 4 in (0·6–10 cm) of water resistance; the apparatus itself had a pressure drop of about 0·25 in (0·6 cm). In the second experiment, the apparatus was more sophisticated. Eight other subjects participated, and only four valves were used. The orders of presentation were selected from factorial designs. The subjects conveyed their assessments on 100 mm lines, the left end defined as detection of the threshold and the right end corresponding to asphyxia (Fig 4).

The results for the two experiments were notably similar. Rectilinear relationships were revealed to the logarithm of the pressure drop interposed, so confirming the application of the Fechner law to this modality. Extrapolation of the regression equation suggested a similar threshold, being 0·26 in (0·6 cm) of water in the

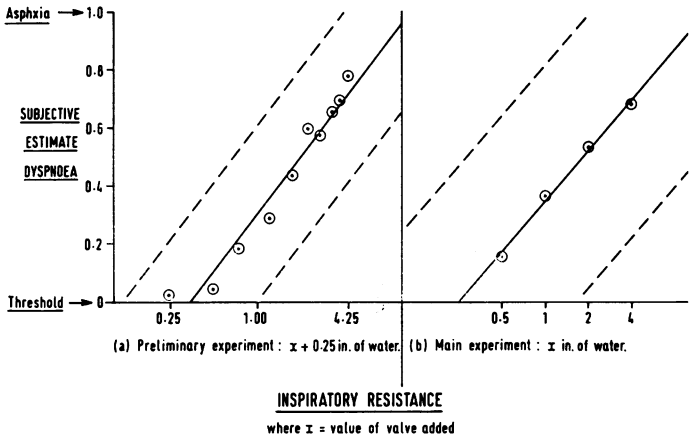

Fig 4 *Relationships between resistance applied to breathing and its subjective appreciation. Each point represents the mean score obtained from* (a) *5 subjects on two occasions,* (b) *8 subjects. The lines represent the regression equations with their 95% confidence limits:* (a) $y=0{\cdot}66\log(x+0{\cdot}25)+0{\cdot}31\ (\pm 0{\cdot}30);\ P=0{\cdot}001$. (b) $y=0{\cdot}58\log x+0{\cdot}34\ (\pm 0{\cdot}47);\ P=0{\cdot}001$

second experiment. This result is virtually identical to that found by Bennett *et al.* (1962) after an experiment using a quite different method of assessment.

Appropriate Resolution in Measurement

When originally describing the advantages of graphic rating, Freyd (1923) wrote: 'The fineness of the scoring method may be altered at will, yielding scores of from 1 to 5 and from 1 to 100.' Critics say that fine resolution is not permitted because it exceeds the discriminatory power of the rater both to appreciate his feeling, and to convey it with this type of perceptual/motor task. Recent researchers on anxiety have used only a resolution of 1 in 10 parts (Lader & Wing 1966, Black 1966, Kellner & Sheffield 1968).

I do not accept this argument. Any excessive resolution can only be assigned to residual error variance, and so cannot bias any conclusion. If it is beyond the appreciation of the rater or beyond his capability to convey it, positioning of the mark cannot occur other than at random, and it is against this that any null hypothesis will be tested.

Resolution of measurement of a 100 mm line to the nearest millimetre is both convenient and appropriate since one part in a hundred is sufficiently sensitive and can be transformed easily from multinomial to near-normal distribution.

People seem to like using this sort of scale and readily understand its requirements: 'It is not a constant source of vexation to the conscientious rater when he finds his judgments falling between the defined points' (Champney 1941). It takes only seconds to obtain a score, and imposes no inconvenience.

Reliability and Validity

Early research quoted by Freyd (1923) indicated high reliability, as shown by close relationship between ratings by different judges. Good reliability has been confirmed with up-to-date methods in the assessment of such a feeling as 'well-being' (Clark & Spear 1964).

I have always observed consistency using this method. For example, the two experiments on assessment of degree of asphyxia gave comparable results. In the first experiment subjects were tested on two occasions; replicates accounted for a negligible part of the observed variance. There were no significant effects due to the order of presentation in either of the two experiments.

Feelings are subjective phenomena, beyond absolute analysis. No final judgment of validity can ever be given – validity can only be relative and examined by empirical comparison. Craik (1943) pointed out that 'use of language itself is based on the principle that any symbolism which works has objective validity'. Validity can be judged by testing predictive, concurrent or content value. Construct validation can only be 'established through a long-continued interplay between observation, reasoning and imagination' (Cronbach 1960). I believe that results obtained on the measurement of feelings using visual analogue scales justify award of this criterion.

Only time will tell whether this technique is at a growing edge of the clinical science of measurement of feelings. The intriguing question is why it has remained so much in the background in its first half century of use.

Acknowledgments: Appreciation is expressed to many colleagues who have encouraged the

author to think about the measurement of feelings, and who have collaborated in enquiries.

REFERENCES
Aitken R C B
(1962) Air Ministry, London: FPRC Memo 192
(1965) MD Thesis, Glasgow
(1967) Proceedings of the International Symposium of Military Psychosomatic Medicine. Societa Italiana di Medicina Psicosomatica, Rome; p 81
Aitken R C B, Ferres H M & Gedye J L
(1963) *Aerospace Med.* 34, 302
Aitken R C B & Gedye J L (1968) *Brit. J. Psychol.* 59, 253
Aitken R C B, Southwell N & Wilmshurst C C
(1965) *Clin. Trials J. (Lond.)* 2, 65
Bennett E D, Jayson M I V, Rubinstein D & Campbell E J M
(1962) *Clin. Sci.* 23, 155
Black A A (1966) *Brit. J. Psychiat.* 112, 557
Champney H (1941) *Child Develop.* 12, 131
Clark P R F & Spear F G (1964) *Bull. Brit. psychol. Soc.* 17, 18A
Craik K J W (1943) The Nature of Explanation. Cambridge; p 27
Cronbach L J (1960) Essentials of Psychological Testing. 2nd ed. New York; p 106
Freyd M (1923) *J. educ. Psychol.* 14, 83
Hayes M H J & Paterson D G (1921) *Psychol. Bull.* 18, 98
Hunt J N & Wales R C (1966) *Brit. med. J.* ii, 13
Kellner R & Sheffield B F (1968) *Brit. J. Psychiat.* 114, 193
Lader M H & Wing L (1966) Physiological Measures, Sedative Drugs and Morbid Anxiety. London; p 99
McGhie A & Russell S M (1962) *J. ment. Sci.* 108, 642
Metfessel M (1947) *J. Psychol.* 24, 229
Richter C P (1929) *Arch. Neurol. Psychiat. (Chic.)* 21, 363
Scheffé H (1952) *J. Amer. statist Ass* 47, 381
Siegel S (1956) Non-parametric Statistics for the Behavioural Sciences. London; pp 184, 202, 229
Snedecor G W & Cochran W G
(1967) Statistical Methods. 6th ed. Ames, Iowa; p 327
Stevens S S (1946) *Science* 103, 677
Yerkes R M & Urban F M (1906) *Harv. psychol. Stud.* 2, 406

Dr A K Zealley and Dr R C B Aitken
(*University Department of Psychiatry, Royal Edinburgh Hospital, Edinburgh*)

Measurement of Mood

Since a person's feelings are, by their very nature, inaccessible to objective scrutiny, it follows that measurement of mood depends to a large extent upon communication by the subject to the observer. When mood lies beyond an arbitrary limit of normality, illness has supervened; diagnosis is then a matter of opinion, based on clinical experience. Having diagnosed an anomaly of mood – be it depression, euphoria or other – it is then desirable to quantify that condition; this will allow observation of intrasubject changes, especially as a result of treatment, and also intersubject comparisons.

Hence there has been interest, particularly in recent years, to develop rating scales of mood; so far, there are two main types – self-rating (or self-description) (Beck *et al.* 1961, Lubin 1965, Shapiro 1961, Zung 1965) and observer-rating (Hamilton 1960, 1967, Medical Research Council 1965). The former avoid professional preconception and prejudice, but require both co-operation and a degree of verbal sophistication on the patient's part. The other type relies entirely on the clinical experience and skill of the staff rater.

In this country, the rating scale introduced by Hamilton (1960) is widely used. It exemplifies attempts to identify 'components' of illness syndromes by factor analysis of clinical ratings (Friedman *et al.* 1963, Kiloh & Garside 1963, Overall 1963): seventeen such 'components' are rated on either three- or five-point scales, preferably by two experienced raters working independently at the same interview. The procedure may take half-an-hour, and is unsuitable for use oftener than once a week.

In a recent report to the Medical Research Council (1965) by its Clinical Psychiatry Committee on the treatment of depressive illness, assessment was carried out using items taken from the Hamilton scale; however, the most important conclusion was based on *overall* ratings of depression by the psychiatrists on a simple five-point scale. It can be assumed with justification that this method of assessment was the one regarded as most appropriate by leading British psychiatric opinion; and accordingly similar ratings were made each week in the studies now to be mentioned, in parallel with patient self-ratings twice daily using a visual analogue scale (Aitken 1969). This scale provides the patient with a language by which to communicate his feelings frequently; its scores are amenable to parametric statistics, allowing precise examination of the significance of any differences.

Clinical Studies of a Visual Analogue Scale in Depressive Illness

In the first instance we examined a broad selection of depressed patients admitted to the Royal Edinburgh Hospital; each patient was asked to mark a horizontal 100 mm line, the ends of which represented normal mood and the extreme of depression respectively. Completion of a fresh recording slip at 12-hourly intervals (morning and evening) throughout their stay in hospital was easily achieved by arranging for the nurse to hand the slip to the patient at 'medicine round' times: it was collected forthwith, as only seconds were required to mark it. No patient failed to grasp the analogue concept of the line.

This technique allowed frequent estimation of the patient's feeling state with minimal inconvenience to himself and to staff, and with likelihood of early detection of change in condition, and subsequent analysis of its significance. In comparison with all the depression-rating techniques described earlier, this method has the advantage of not asking the patient to review his emotional status under numerous – and often