Br. J. clin. Pharmac. (1978), **6**, 15–24

SENSITIVITY AND ACCURACY OF THE VISUAL ANALOGUE SCALE: A PSYCHO-PHYSICAL CLASSROOM EXPERIMENT

C. MAXWELL
Clinical Research Services Limited, 36 Neeld Crescent, London NW4 3RR

1 Twenty-seven intelligent volunteers took part in a classroom experiment on two occasions to assess subjectively, ordinally related volumes of sound (offered in random sequence), using the visual analogue scale.

2 It was simple to use and largely acceptable. Big differences were significant by parametric tests but small ones were sometimes significant only with ordinal ones. Within-subject comparisons were more accurate and more sensitive than those between-subjects. The *t*-test was very robust.

3 Five out of 49 results were erroneously significant; they remained so no matter how the data were handled. It was concluded that this was due to a shift in either perception, cognition or scoring between the two sessions.

4 Arcsine transformations made little difference. The mechanism of these is fully discussed. A conversion to proportional scores resulted in very much improved sensitivity.

5 It is recommended that authors using the visual analogue scale: have valid reason in their own setting for using a transformation; present the distribution, or at least the medians and ranges of their raw scores; also use a simple but different measure so as to demonstrate internal consistency between them.

Introduction

For over 50 years the visual analogue scale and its precursors have been used for the assessment of subjective phenomena (Hayes & Patterson, 1921; Freyd, 1923). Nowadays, this is normally an horizontal line 10 cm long with a statement at one end that the symptom, feeling or attitude could not be stronger and at the other, that it is completely absent (or, in the case of bipolar scales on a single continuum, that the alternative could not be more extreme). After being made to understand what it is meant to signify, subjects make a mark somewhere along its length to indicate their impression of the strength of that feeling. It is customary then to measure in mm the distance of that mark from one end and to use that figure for analysis. Most frequently, the figures are handled by parametric methods; means, s.d., s.e. mean and Student's *t*-tests. More sensitive parametric techniques are also used including analysis of variance (Joyce, Zutshi, Hrubes & Mason, 1975; Hart, Hill, Bye, Wilkinson & Peck, 1976), analysis of covariance (Brodie, McGhee, O'Hara, Valle-Jones & Schiff, 1975) and principal component analysis (Bond & Lader, 1974) and to facilitate these, the raw data are sometimes transformed; $\log_e$ (Bond & Lader, 1974), arcsine (Aitken, 1969; Peck, Bye & Claridge, 1977).

The problems of measuring subjective feelings are highly complex and it is questionable whether the measurements derived from the visual analogue scale are continuous or interval in character notwithstanding the temptation to regard them so as soon as the ruler is applied. Certainly, if the scale is properly understood and properly used the data must at least be ordinal within a single subject; but the results are not treated as within subject date, they are grouped or pooled for between-subject analyses. Do different methods of handling the results (parametric *v* ordinal; within-subject *v* between-subject) produce varying degrees of accuracy and sensitivity?

Furthermore, Aitken (1969) in his studies in anxiety found it necessary to use an arcsine transformation. This transformation is now being applied routinely to visual analogue scale scores in studies in other fields for no better reason than that it has become traditional. It is not always clear whether the arcsine of the square root or the arcsine of the raw score has been used but in either case, might it do any harm if it is used without specific need?

To answer these questions one needs not only subjective assessments but also a 'truth' against which the accuracy and sensitivity can be compared. It was decided therefore to perform an experiment in which volunteers would be exposed to a series of physical stimuli that were ordinally related (though not necessarily parametric) so that they could be assessed

subjectively with the visual analogue scale. The stimuli were different volumes of sound. The lower end of this continuum was firmly anchored (the lowest volume was 'no sound') but the maximum was not demonstrated beforehand so that subjects had to consider not only how loud a particular sound was but also how loud they thought it could possibly be. Varying sound volumes were offered in random sequence on two separate occasions.

Methods

Procedure

So as to allow both within-subject and between-subject comparisons and to partially simulate a clinical trial situation when assessments are made on at least 2 days, the experiment was performed twice (during the first hour of the working day), in the same room, with subjects sitting in the same places; 48 h separated the two series of tests. The purpose and use of the visual analogue scale was described; subjects were told that they would assess the loudness of sounds from a recorder, they should make their mark at the right hand end if they felt that the sound was as loud as they thought the machine could possibly produce, the other end for the opposite and to mark intermediate sounds accordingly. Each subject was given a single assessment sheet with fourteen 10 cm lines on it and space for their names, seating position and date. Subjects could see their earlier scores during each session but they did not have access to the first assessment sheet at the second session. Sound volume was not demonstrated beforehand. Different volumes of sound were offered in a predetermined random sequence that had been adjusted to ensure that each volume would occur at least once in a cycle. A sheet of A4 paper conveniently accommodated fourteen lines and it was felt that spare assessments would assist the randomization and give minimal experience in scoring. For this reason it was predetermined that the earlier of duplicated sound volumes would be discarded and only the latest used for analysis. Subjects raised their hands to show they were ready and the sound was continued until everyone had scored it and showed this by lowering their hands. The sound was then switched completely off for at least 5 s, the machine was reset and the cycle continued.

Sound source

A cassette recorder (Jonan Synchrorecorder 4400SL) was used to record and reproduce a standard tape. The recording consisted of a high pitched interrupted tone (the 'Engaged' signal of the British telephone system) made with the ALC circuit switched in; the tone control was set for maximum treble throughout. The controls on this machine are of the 'slide rule' type and are nominally calibrated 0 to 10. A preliminary experiment was performed with different subjects to ensure that the different settings of the volume control produced ordinally related sounds and that these differences were audibly detectable. This was done by a blind observer declaring that the current volume was louder than, the same as or softer than the last sound as the setting was changed (or not changed) between a pair of adjacent numbers, this being done in random

Table 1 Frequency distribution at each volume setting

First session volume	*0±1*	*2–9*	*10–19*	*20–29*	*30–39*	*40–49*	*50–59*	*60–69*	*70–79*	*80–89*	*90–100*
7					4	3	2	6	6	4	2
6					5	2	3	7	6	2	2
5				3	6	2	8	5	2	1	
4				4	6	2	10	3	0	2	
3		5	13	7	1	1					
2		6	14	5	1	1					
1		8	7	10	2						
0	27										
Second session volume											
7				1	1	4	2	4	6	4	4
6					2	6	0	5	7	3	3
5					4	3	2	8	4	3	2
4					3	5	4	7	6	1	
3			1	8	5	1	6	2	2	1	
2		6	11	7	2						
1		13	10	2	1						
0	26										

order. The machine was not switched off during this procedure and an undoubted discernment of difference was being sought. Each pair of adjacent numbers was so treated and it was definitely established that the volumes were ordinal between 0 and 7 but settings above 7 did not produce confidently discernible differences. At volume 7 in a domestic setting the sound is considered extremely loud and uncomfortable but not distorted.

Subjects

Twenty-seven subjects (19 men, 8 women), aged 25 to 63 years (median 38) of superior intelligence (physicians, and graduates in biological and allied sciences) participated. No effort was made to identify hearing defects. Fully informed consent was obtained (not in writing) by humorous embarrassment so as to act as an illustration and vehicle for discussion in a later seminar on ethics.

Statistics

Student's *t*-tests for paired and unpaired data (comparison of means) were performed regardless of homogeneity of variance. All non-paired comparisons were checked by the more appropriate Fisher-Behrens test; attention is drawn to discrepancies with the *t*-tests only when the 0.05 level is crossed. Within-subject comparisons between the first and second sessions were also checked by ordinal tests; where a McNemar test (which utilizes only the direction of a difference) was significant no further test was done but if either was not significant (NS), the more sensitive Wilcoxon Matched Pairs Signed Rank Test was also performed. Two tailed tests were used throughout. Significance levels are indicated in the tables thus: * 0.05, ** 0.01, *** 0.001 for all but Fisher-Behrens tests. Product moment correlations on means have been performed to show how these, rather than the total data, correlate with the nominal volume settings. Additional statistics are as indicated in the text.

Table 2 Mean raw scores and s.d. at each volume setting

	First session		*Second session*	
Volume	*Mean (mm)*	*s.d.*	*Mean (mm)*	*s.d.*
7	64.6	18.1	68.5	20.4
6	61.7	17.7	66.6	19.5
5	49.4	17.0	63.2	17.9
4	48.2	16.4	58.4	14.8
3	17.5	8.2	42.7	19.3
2	16.1	8.9	15.8	7.9
1	15.9	8.6	11.7	6.6
0	0	0	0	0

Results

Twenty-seven subjects took part in the first experimental session and 26 of them were involved in the second. Table 1 shows for each session, the frequency distribution of scores for each 'volume' (i.e. number on the volume control). It will be noted that everyone scored no-sound as zero. Table 2 shows the mean and s.d. of scores at each volume. It will be noted here that the means are ordinal but in the first series are grouped in three steps (about 62, 49 and 16) whereas they are better spread in the second series with only one major gap between 16 and 43. It happened that in the first series the maximum volume was heard as the third in sequence and with the exception of volume 5, all means in this series relate to sounds heard after the maximum had been offered. In the second series the maximum volume was (by chance) also heard as the third in sequence. The mean scores correlate very well parametrically with the volume settings: $r=0.96$, $P<0.001$, $y=0.20+0.10x$ (regressions are for volume setting on score); for the second series, $r=0.96$, $P<0.001$, $y=0.03+0.08x$.

Table 3 (for the first series) shows the within-series differences between means for pairs of volume settings; Table 4 does the same for the second. In both

Table 3 Comparison of means within the first series

Volume	*6*	*5*	*4*	*3*	*2*	*1*
7	2.0 (NS)†	15.1***	16.4***	47.0***	48.5***	48.6***
6	—	12.3*	13.5**	44.2***	45.6***	45.8***
5	—	—	1.3 (NS)	31.9***	33.4***	33.5***
4	—	—	—	30.7***	32.1***	32.3***
3	—	—	—	—	1.5 (NS)	1.6 (NS)
1	—	—	—	—	—	0.1 (NS)

NS, *P* greater than 0.05.
*, **, ***, significant at 0.05, 0.01 and 0.001 respectively by *t*-test for unpaired data.
† McNemar test significant at 0.05.

cases the zero-scores have been omitted as all these differences are significant. When differences between means exceeded 10 mm they were always significant on *t*-test for unpaired data and often highly so. A couple of differences less than 10 mm were also significant this way. But once in the first series and three times in the second, differences were significant on the McNemar test and not on the *t*-test.

In controlled clinical trials it is frequently necessary to perform between-subject comparisons and for these to be done on more than one occasion. Table 5 compares the mean for each first series volume setting with the mean for each second series volume setting and shows the differences between these two means (first series mean minus second series mean) and the significance level by *t*-test. Thus the first series volume 5 mean minus the second series volume 6 mean was −17 and this was highly significant. The difference between the first 5 *v* second 2 was 34 and also highly significant. But the 5 *v* 5 comparison was significant at −14 (when it should theoretically have been nil) and the 5 *v* 4 difference was −9 (significant) and is in the wrong direction. Thus, in the second series there was a slip or shift upwards compared to the first series (or *vice versa*) and this shift can also be seen in other comparisons; 4 *v* 4, 3 *v* 3, 1 *v* 1. These five results are all significant on *t*-test but wrong; the latter is not significant on the Fisher-Behrens test. The problem of non-significant differences is mentioned in the Discussion but to test the effect of ordinal procedures, the difference between the first 7 and second 4 was also checked by the Mann Whitney U-test corrected for ties which was also non-significant as was the same test without the correction.

To see whether these errors would have been obviated by using within-subject comparisons the analysis was repeated this way. Table 6 shows the results of this approach by both parametric and ordinal tests around the borderline area only (outside results have been omitted to avoid unnecessary clutter, they can only be even more significant and are so). As would be expected from this more appropriate method, extra results become significant: 7 *v* 4, 7 *v* 7, and 6 *v* 4 (this by Wilcoxon Matched Pairs Signed Ranks test only). The erroneous significant results still remain.

The raw data were subjected to two different arcsine transformations to see if either affected the significance levels of between-subject comparisons. The first was the arcsine of the square root (Snedecor & Cochrane, 1967) (arcsine-root) and the other was the arcsine of the raw score (arcsine-raw). All between-subject comparisons were again tested by both *t*-test and Fisher-Behrens test. The significance of differences was almost identical to those in Table 5, the exceptions are detailed here.

In Table 5, the first 5 *v* second 4 comparison is significant by both tests; they remain so with the arcsine-root transformation but neither test is significant with the other transformation. The 2 *v* 1 comparison of raw scores is significant on *t*-test but only borderline on the Fisher-Behrens test ($0=52°48'$, $d=2.058$, $v_1=27$, $v_2=26$). The arcsine-root transformation makes this comparison significant on both tests. The arcsine-raw transformation makes it NS on the Fisher-Behrens but again significant on *t*-test. The 1 *v* 1 comparison of raw scores shows significance on *t*-test but not on Fisher-Behrens test and the arcsine-root does not alter this. The arcsine-raw transformation however inverts this so that it becomes significant on Fisher-Behrens but not on *t*-test. It should be noted that the raw score means are 15.9 and 11.7 respectively (see Discussion).

It was noticeable when marking the record sheets that while the zero position had dutifully been used to record no-sound, the variation in the maximum scores was very great. Some subjects used the right hand end of the line to signify their maximum but others, even when they had been well exposed to it used less than half the line. Indeed, three subjects recorded their maxima at 33, 36 and 39 mm from the zero end. Table 7 shows the distribution of *maximum* scores in each series. In the first, seven subjects made their maximum mark below the halfway point and this reduced to four in the second series.

In view of the wide variation in maxima, it seemed

Table 4 Comparison of means within the second series

Volume	*6*	*5*	*4*	*3*	*2*	*1*
7	1.9 (NS)†	5.4 (NS)†	10.1*	25.8***	52.7***	56.9***
6	—	3.4 (NS)	8.2 (NS)	23.9***	50.8***	54.9***
5	—	—	4.8 (NS)	20.5***	47.4***	51.5***
4	—	—	—	15.7**	42.6***	46.7***
3	—	—	—	—	26.9***	31.0***
2	—	—	—	—	—	4.1*

NS, *P* greater than 0.05.
*, **, ***, significant at 0.05, 0.01 and 0.001 respectively by *t*-test for unpaired data.
† McNemar test significant at 0.05.

Table 5 Differences between first series and second series means

	Volume	7	6	5	4	3	2	1
					Second series			
	7	− 3.94 (NS)	− 2.02 (NS)	1.41 (NS)	6.18 (NS)	21.87***	48.79***	52.91***
	6	− 6.80 (NS)	− 4.88 (NS)	− 1.45 (NS)	3.32 (NS)	19.01***	45.93***	50.05***
First series	5	−19.06***	−17.14***	−13.71**	− 8.94*	6.75 (NS)	33.67***	37.79***
	4	−20.31***	−18.39***	−14.96**	−10.19*	5.50 (NS)	32.42***	36.54***
	3	−50.98***	−49.06***	−45.63***	−40.86***	−25.17***	1.75 (NS)	5.87**
	2	−52.43***	−50.51***	−47.08***	−42.31***	−26.62***	0.30 (NS)	4.42*†
	1	−52.57***	−50.65***	−47.22***	−42.45***	−26.76***	0.16 (NS)	4.28* ‡

NS, P greater than 0.05.
*, **, ***, significant at 0.05, 0.01 and 0.001 respectively by t-test for unpaired data.
† Fisher-Behrens test borderline 0.05.
‡ Fisher-Behrens test NS.
Boxes indicate differences in the wrong direction.

reasonable to convert the raw scores into proportions of the maximum used by each subject (100 × raw score/maximum) and Table 8 shows the means and standard deviations at each volume. Again the means correlate very well parametrically with the volume settings: $r=0.97$, $P<0.001$, $y=0.02 + 0.06x$ for the first; $r=0.97$; $P<0.001$, $y=0.03 + 0.06x$ for the second. The effect of this conversion (in this experiment) is dramatic (Table 9). The five erroneously significant results still remain: no significant differences are lost. But differences between first series 7 and second series 6, 7 *v* 5, and 7 *v* 4 come into significance for the first time. The 6 *v* 5 difference is not only significant but now in the right direction and the 6 *v* 4, 5 *v* 3 and 4 *v* 3 differences are also now significant. Much greater sensitivity has been achieved.

Discussion

These results are at once reassuring and worrying; reassuring because of the generally adequate (though by no means completely satisfactory) sensitivity that this basically crude assessment device seems to produce and worrying because sometimes the statistically significant results were exactly wrong.

When properly used, the unipolar visual analogue scale is marked at one end to signify the complete absence of the feeling and at the other end with an indication of the opposite extreme. Both ends are thought to be anchor points (Joyce *et al.*, 1975) but what happens in between has had to be shown by trial and error. The literature is well endowed with 'trial' (Hayes & Patterson, 1921; Freyd, 1923; Bond & Lader, 1975; Aitken, 1969) but apparently as yet, little or no 'error'. In this study there were five statistically significant results (out of a table of 49) that were indisputably wrong. They were due to an upwards shift in scoring the middle range of sound volumes during the second session relative to the first, or *vice versa*. The possible explanations are not unimportant because of their relevance to clinical research.

Every subject used the zero end of the scale and there can be no doubt about this anchor point in perception, cognition and scoring. But what of the other end, several subjects scored their maximum below the half way point thus virtually shortening the scale; why should this be so? The very first sound was perceived in relation to the subject's concepts of zero sound and his expectation and anticipation of what the maximum sound might be. During the first series they could only estimate the maximum sound by calling on their experience; experience of other such machines. As each score was made, their expectation of there being a still louder sound to come (rightly or wrongly) would effect their use of the maximum position because once that had been used, even louder sounds

Table 6 Within subject mean differences (and s.d.) between the first and second series (first minus second)

		Second series						
	Volume	*7*	*6*	*5*	*4*	*3*	*2*	*1*
First series	*7*	− 4.27 (14.74) NS NS NS	− 2.35 (13.34) NS NS NS	1.08 (12.52) NS NS NS	5.85 (10.52) * NS *	21.54 (16.35) ** *** –		
	6	− 7.23 (14.38) * * –	− 5.31 (13.41) NS NS NS	− 1.88 (12.40) NS NS NS	2.88 (10.03) NS NS *	18.58 (17.08) *** *** –		
	5	−19.23 (18.99) *** *** –	−16.54 (17.21) *** *** –	−14.00 (15.11) *** *** –	− 9.08 (12.40) *** *** –	6.58 (20.81) NS NS NS	33.81 (15.91) *** *** –	
	4			−15.00 (15.04) *** *** –	−10.64 (13.55) *** *** –	5.46 (21.12) NS NS NS	32.38 (16.31) *** *** –	
	3				−41.00 (15.05) *** *** –	−25.31 (18.95) *** *** –	2.00 (8.31) NS NS NS	5.73 (8.29) ** NS ***
	2					−26.42 (20.20) *** *** –	0.50 (8.75) *** NS NS	4.62 (8.57) NS NS **
	1					−25.73 (17.10) *** *** –	−0.04 (9.77) NS NS NS	4.08 (8.62) * * –

Beneath each mean is the statistical significance by (in descending order) paired *t*-test, McNemar test and, if either is not significant at 0.05 (NS), Wilcoxon Matched Pairs Signed Rank test.
*, **, ***, significant at 0.05, 0.01 and 0.001 respectively.
Boxes indicate differences in the wrong direction.

would have to receive the same score. Simultaneously, previous scores would be influencing the location of the newest ones, cognition still being firmly anchored to the zero point. In this experiment, certain physical factors would also be relevant; the distance of a subject from the machine, the competence of his hearing (which was neither tested nor enquired about), influence from neighbours and so on. It must be assumed that the recorder behaved identically on each occasion. In the event, the first series means (even though they correlated extremely well parametrically with the volume control numbers) were grouped into four steps: volume 0; volumes 1, 2 and 3 (around 16 mm); 4 and 5 (around 49 mm); and volumes 6 and 7 (about 63 mm). Within these groupings only the difference between volume 6 and 7 was statistically significant and this by a within-subject test. It is remarkable that the means are still correctly ordinal.

By the end of the first series it would have been reasonable for this group of intelligent people to have concluded that they had now heard the loudest sound and that they would hear none louder in the experiment. At the start of the second session, they would thus be armed with appropriate and relevant experience to assist them in their new session of scoring. They could not see their scores from the first session. Joyce *et al.* (1975) have established in their clinical trial in the pain of arthritis that a sight of earlier scores made no difference to accuracy or consistency. It could well have been that subjects tried to make their second set of scores equate with their earlier ones but they were not requested to do so, they were given the same explanation of the purpose of the scale at the start of both sessions and they did not know at the start that there was going to be a second. It transpired that the maxima were higher this time than previously and there was much better linearity with only one major gap between 16 and 43 mm. There was much better differentiation between closely related sound volumes and it appears that the visual scale is a much better analogue on this occasion. This might be a consequence of experience on cognition; or is this training? Huskisson (1974) believed that training or trained assistants should be available to patients using this device. Certainly these results suggest that a training session would have been helpful and it is a pity, with hindsight, that a third set was not obtained with the results from the first being discarded. But is that the way the scale is used in clinical trials and in clinical research?

For all one knows similar problems might occur in clinical trials though there are important differences. In clinical trials patients tend to start with severe symptoms (high scores) which can usually be expected to ameliorate (become lower scores) be this due to treatment or to non-specific factors. It might be that the patient's memory of his initial suffering acts as a firmer anchor point at the high end than a healthy volunteer's anticipation or expectation of what a maximum might theoretically be. Experience suggests that patients very readily score themselves at the 100 mm mark even if it is obvious to the observer that another patient would be suffering much more when he marked the same place. It is not difficult to believe that a patient can simultaneously comprehend and correctly conceive and use the zero end no matter how much his maximum score is exaggerated. If both these

Table 7 Distribution of maximum scores

	First series	*Second series*
90–100	2	4
80–89	4	5
70–79	8	7
60–69	5	2
50–59	1	4
40–49	4	4
30–39	3	—
20–29	—	—
10–19	—	—
0–9	—	—
Total	27	26

Table 8 Means and s.d. of proportional scores

	First series		*Second series*	
Volume	*Mean (%)*	*s.d.*	*Mean (%)*	*s.d.*
7	97.85	4.36	94.58	11.77
6	93.48	5.51	92.12	9.17
5	74.81	13.47	88.19	8.63
4	72.70	16.75	82.31	11.89
3	27.56	13.05	58.58	18.04
2	24.41	11.63	22.50	10.57
1	23.07	9.81	16.54	8.81
0	0	0	0	0

Table 9 Mean within subject differences of proportional scores between the first and second series (first minus second)

		Second series						
	Volume	*7*	*6*	*5*	*4*	*3*	*2*	*1*
First series	*7*	3.27 (NS)	5.73**	9.66***	15.54***	39.27***	75.35***	81.31***
	6	− 1.10 (NS)	1.36 (NS)	5.29*	11.17***	34.90***	70.98***	76.94***
	5	−19.97***	−17.31***	**−13.38*** **	**− 7.50***	16.23***	52.31***	58.27***
	4	−21.88***	−19.42***	−15.49***	**− 9.61***	14.12***	50.20***	56.16***
	3	−65.02***	−64.56***	−60.63***	−54.75***	**−31.02*** **	5.06 (NS)	11.02***
	2	−70.17***	−67.71***	−63.70***	−57.90***	−34.17***	1.91 (NS)	7.87**
	1	−71.50***	−69.05***	−65.12***	−59.24***	−35.51***	0.57 (NS)	**6.53*** **

NS, P greater than 0.05.
*, **, *** less than 0.05, 0.01 and 0.001 respectively by paired t-test.
Boxes indicate differences in the wrong direction.

conditions are so, the conversion of the raw scores to proportional scores in this study would approximate well to clinical use.

The visual analogue scale is crude and simple yet the scores from it are customarily measured with an accuracy of 1% (and their means taken to decimal places!): very sophisticated mathematical procedures are then engaged. The conversion of any result to numbers (especially when they contain more than one digit) tempts one to believe that they are on a continuous, interval scale. Does it matter if this assumption is wrong? Are non-parametric tests (in this case ordinal) better able to dispel doubt about the results having occurred by chance? These two experimental sessions suggest that it makes little difference most of the time whether parametric or ordinal tests are used. When differences are large they are statistically significant both ways though, as one would expect, within-subject comparisons are more sensitive than between-subject ones. Similarly, when differences are very small, neither way of handling the data will make them significant. There is a borderline area (differences smaller than 10 mm in this experiment) when the choice of test might make a difference to the level of significance and it is suggested that ordinal tests be additionally applied when examining small ones.

A more difficult problem concerns transforming the raw scores. Bond & Lader (1974) employed a $\log_e$ transformation for clearly stated reasons in their study utilizing 16 such scales for measuring as many variables. The distribution of scores on some scales was found to be skewed to the more socially acceptable end of the scales in a population of normals. Certain scales were then reversed to bring the skewness consistently to one end and the transformation applied. Aitken (1969) studying anxiety in fighter pilots in an artificial situation normalized his data by taking the arcsine of the square root because his distributions were sometimes positively skewed, sometimes negatively skewed and sometimes not skewed at all. Huskisson (1974) reporting on pain in arthritis was reassured to find a uniform distribution which another assessment (a simple descriptive pain scale) indicated was correct. In the current experiment, the second series distribution of scores was uniform as it should have been though there was a relative paucity of scores between 90 and 100 mm (see Table 1); the first series had a peak on the left at 10–19 mm.

The arcsine (root) transformation as described by Snedecor & Cochran (1967) and cited by Aitken (1969) helps normalize distributions accumulated at either end of the scale (i.e. both positively and negatively skewed). The raw score in mm is divided by 100, the square root is taken and the arcsine of this is taken. Differences located below 12 mm or above 88 mm are stretched; those between are compressed. Thus statistical significance is dependent not only on

the size of the difference and the test used but also on the location of the raw scores. In this experiment there was very little effect which was because only one sound volume, the second series volume 1, produced a predominance of scores below 12 mm (see Table 1); the mean was 11.7 (s.d. 6.6).

When the distribution of scores is invariably negatively skewed as might be expected with patients scoring maxima on the right, the conversion of raw scores (divided by 100) to the arcsine without taking the square root first (arcsine-raw) will also tend to normalize distributions. This conversion is also used with the visual analogue scale (R.C.B. Aitken, personal communication; C.R.B. Joyce, personal communication). Unlike the arcsine-root transformation this one is not symmetrical, raw score differences located to the right of 77 mm are stretched and those to the left are compressed. Clearly the inappropriate use of either transformation may yield misleading results especially with marginal differences and the amount of confusion caused by this will be dependent on the distribution and location of raw scores. It would assist readers if authors engaging these techniques would also present details of the raw scores and at very least, the ranges and medians. To assist those studying papers with transformed data only, Table 10 shows the critical points and their equivalents in both radians and degrees (decimal) with both transformations.

A manipulation of the original scores that did make a profound improvement to the results was their conversion to proportional scores. With everyone using the zero point, the effect of this is to stretch out the maximum for each subject to 100. Doing this, several results became significant when they were not so beforehand and at least one trend in the wrong direction (NS) was corrected and became significant. This was due to the great increase in some of the measurements, sometimes threefold. Normally this would be very dubious statistically yet the ends justify the means; seven is louder than six and there is no reason why this difference should not be significant. It remains to be seen what effect this conversion would have if the zero end had not been used. In a future experiment it would be interesting to demonstrate the loudest sound first to fix a high anchor point and not to use zero sound thus locating the scores towards the right. This might simulate clinical trial conditions better.

In clinical research the true facts are not normally known and this ignorance is invariably the justification for the study. How should a researcher know if his use of the visual analogue scale on each occasion is reliable? This study has shown that statistical significance is not enough. Several results were not only wrong but also significant. If the true facts had not been known would they have been taken as right *because* they were significant? Very likely. Bond & Lader (1975) had previously validated their particular use of the scale in a well designed experiment appropriate to the use they were going to make of it. Joyce *et al.* (1975) satisfied themselves that the results from the scale were acceptable by comparing them with the results from a four point scale in nominal categories; they suggested that differentiation of a known drug effect from placebo, or better still, the differentiation of two doses of the same drug were also reliable. But the true effects of the drug must be known *a priori*. The present study has shown by experiment that proportional scores yielded the maximum sensitivity though erroneous results remain; it remains to be seen and still to be shown how this might apply to less clinically artificial situations. It is strongly recommended that workers ensure that they have valid reason in their own setting before transforming data and subjecting it to sophisticated analysis. It is also desirable to build into the study a simple check on the measuring device such as a four point scale (Joyce *et al.*, 1975) or a simple global assessment (Maxwell, 1968); inconsistencies between these and the scale results are ignored at one's peril.

Clinical research in general and clinical trials in particular are plagued with the problems of statistical inference. The falsely significant results in this experiment were not errors of statistical inference; they could have become so had the truth not been known. No attention has been paid here to the falsely negative results. Some were Type II errors due to incorrect statistical tests but in the true exploratory study these can always be dismissed as being due to

Table 10 Raw score and transformed equivalents in degrees (decimal) and radians

	Degrees (decimal)		*Radians*	
Raw score (mm)	*Arcsine-root* $y=\sin^{-1}\sqrt{x}$	*Arcsine-raw* $y=\sin^{-1}x$	*Arcsine-root* $y=\sin^{-1}\sqrt{x}$	*Arcsine-raw* $y=\sin^{-1}x$
0	0	0	0	0
12	20.27		0.35	
77		50.35		0.88
88	69.73		1.22	
100	90	90	1.57	1.57

insufficient trial sensitivity be it due to the measuring device, the patient population, inclusion and exclusion criteria and so on. Falsely negative results can be equally or even more dangerous than false positives. Yet all is not hopeless. Overall clinical evaluation is always strengthened by consistency in results; consistency between different methods of measurement and consistency between the results of different studies.

In summary the visual analogue scale is simple to use and large differences can safely be treated by parametric tests; the *t*-test is very robust. Within-subject comparisons are more sensitive and accurate than between-subject ones. Differences smaller than 10 mm may benefit by being examined with the Fisher-Behrens test. Small differences can sometimes be significant with an ordinal test when they are not so with a parametric test. Arcsine transformations can alter levels of significance (especially with small differences) and should only be used when there is clear need or a validation in similar circumstances. Conversion of the raw scores to proportional scores produced much greater sensitivity in this experiment and may be indicated when the whole length of the line has been used by some but not all subjects and one end has been used by everyone. The recognition of a worrying number (5 out of 49) of erroneously significant differences was only possible in this experiment with a knowledge of the true facts. In place of this, it is recommended that an additional clinical measure be used so that consistency can be demonstrated.

Gratitude is expressed to Dr M.H. Lader, Professor Max Hamilton and to the Journal assessors for valuable comment on the manuscript.

References

AITKEN, R.C.B. (1969). Measurement of feelings using Visual Analogue Scales. *Proc. Roy. Soc. Med.*, **62**, 989–993.

BOND, A. & LADER, M.H. (1974). The use of analogue scales in rating subjective feelings. *Br. J. med. Psychol.*, **47**, 211–218.

BOND, A. & LADER, M.H. (1975). Residual effects of flunitrazepam. *Br. J. clin. Pharmac.*, **2**, 143–150.

BRODIE, N.H., McGHIE, R.L., O'HARA, H., VALLE-JONES, J.C. & SCHIFF, A.A. (1975). Anxiety/Depression in elderly patients: a double blind comparative study of fluphenazine/nortriptyline and promazine. *Practitioner*, **215**, 660–668.

FREYD, M.J. (1923). The graphic rating scale. *J. Educational Psychol.*, **14**, 83.

HART, J., HILL, H.M., BYE, C.E., WILKINSON, R.T. & PECK, A.W. (1976). The effects of low doses of amylobarbitone sodium and diazepam on human performance. *Br. J. clin. Pharmac.*, **3**, 289–298.

HAYES, M.H.S. & PATTERSON, D.G. (1921). Experimental development of the graphic rating scale. *Psychology Bulletin*, **18**, 98–99.

HUSKISSON, E.C. (1974). Measurement of pain. *Lancet*, **ii**, 1127–1131.

JOYCE, C.R.B., ZUTSHI, D.W., HRUBES, V. & MASON, R.M. (1975). Comparison of fixed interval and visual analogue scale for rating pain. *Eur. J. clin. Pharmac.*, **8**, 415–420.

MAXWELL, C. (1968). The clinical trials protocol. *Clinical Trials J.*, **5**, 951–956.

PECK, A.W., BYE, C.E. & CLARIDGE, R. (1977). Differences between light and sound sleepers in the residual effects of nitrazepam. *Br. J. clin. Pharmac.*, **4**, 101–108.

SNEDECOR, G.W. & COCHRAN, W.G. (1967). *Statistical Methods*. 6th Ed., p. 327. Iowa: Ames.

(Received August 16, 1977)