British Journal of Clinical Psychology (2009), 48, 163–180

The British Psychological Society

www.bpsjournals.co.uk

A convenient method of obtaining percentile norms and accompanying interval estimates for self-report mood scales (DASS, DASS-21, HADS, PANAS, and sAD)

John R. Crawford[1]*, Paul H. Garthwaite[2], Caroline J. Lawrie[1], Julie D. Henry[3], Marie A. MacDonald[1], Jane Sutherland[1] and Priyanka Sinha[1]

[1]School of Psychology, University of Aberdeen, Aberdeen, UK
[2]Department of Statistics, The Open University, Milton Keynes, UK
[3]School of Psychology, University of New South Wales, Sydney, Australia

Background. A series of recent papers have reported normative data from the general adult population for commonly used self-report mood scales.

Aims. To bring together and supplement these data in order to provide a convenient means of obtaining percentile norms for the mood scales.

Method. A computer program was developed that provides point and interval estimates of the percentile rank corresponding to raw scores on the various self-report scales.

Results. The program can be used to obtain point and interval estimates of the percentile rank of an individual's raw scores on the DASS, DASS–21, HADS, PANAS, and sAD mood scales, based on normative sample sizes ranging from 758 to 3822. The interval estimates can be obtained using either classical or Bayesian methods as preferred.

Conclusion. The computer program (which can be downloaded at www.abdn.ac.uk/~psy086/dept/MoodScore.htm) provides a convenient and reliable means of supplementing existing cut-off scores for self-report mood scales.

Self-report scales are widely used in the assessment of anxiety, depression, and related constructs. They can serve as a useful complement to the clinical interview and are generally quick to administer. They also generally have high internal consistency (i.e. they are reliable) and, unlike clinician rating scales, directly assess a patient's psychological state. In a series of recent papers Crawford, Henry and colleagues have obtained normative data from large samples of the general adult population for a

*Correspondence should be addressed to Professor John R. Crawford, School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 2UB, UK (e-mail: j.crawford@abdn.ac.uk).

DOI:10.1348/014466508X377757

number of widely used self-report scales of anxiety, depression, and related constructs (Crawford & Henry, 2003; Crawford & Henry, 2004; Crawford, Henry, Crombie, & Taylor, 2001; Henry & Crawford, 2005; Henry, Crawford, Bedford, Crombie, & Taylor, 2002). The scales for which normative data were made available included: The Depression Anxiety and Stress Scales (DASS; Lovibond & Lovibond, 1995); the short-form version of the DASS (DASS–21; Lovibond & Lovibond, 1995); the Hospital Anxiety and Depression Scale (HADS; Snaith & Zigmond, 1994); the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988); and the Symptoms of Anxiety and Depression Scale (sAD; Bedford & Foulds, 1978).

The normative data from the aforementioned samples were used to generate percentile norms for the various scales, thereby allowing clinicians or researchers to quantify the abnormality or rarity of a patient's scores. These percentile norms provide a useful supplement to previously existing cut-off scores and, by expressing a patient's score as a percentile rank (rather than simply as being above or below a given cut-off), are in keeping with the commonly held view that anxiety and depression should be viewed as dimensional rather than categorical constructs (Crawford, Henry, Crombie, & Taylor, 2001).

The aim of the present study is to bring all these normative data together in one convenient location and to supplement them with normative data obtained from further samples of the general adult population. We decided that, from the point of view of the end-user, the best approach was to develop a computer program to express the raw scores on the various mood scales as percentile ranks. This approach has the advantage that it is quicker and less error prone than referring to multiple, voluminous, sets of conversion tables. With regard to this last point, research shows that clinicians make many more simple clerical errors than we like to imagine when scoring or converting test scores (e.g. Faust, 1998; Sherrets, Gard, & Langner, 1979; Sullivan, 2000).

A second aim of the present study was to provide interval estimates of the percentile ranks corresponding to raw scores on the various scales. When psychologists refer a patient's score to percentile norms, their interest is in the standing (percentile rank) of the patient's score in the normative *population*, rather than its standing in the particular group of participants who happen to make up the normative *sample.* Although, in the present case, the normative samples used to provide the basis of conversion from raw scores to percentile ranks were large, it is still the case that there is uncertainty about these quantities. Thus the percentile rank for a raw score obtained from a normative sample must be viewed as a point estimate of the percentile rank of the score in the population and should be accompanied by an interval estimate. Interval estimates serve the useful general purpose of reminding us that normative data are fallible and serve the specific purpose of quantifying this fallibility (Crawford & Garthwaite, 2002; Gardner & Altman, 1989).

When scores are assumed to be normally distributed (either because the raw scores of the normative sample tend to a normal distribution, or the scores can be transformed to normality, or the scale in question provides standardized scores), then a parametric method developed by Crawford and Garthwaite (2002) can be used to provide an interval estimate of the percentile rank of a test score (see also Crawford & Garthwaite, 2008a). It is possible to calculate this interval because, when test scores are normally distributed, their percentile ranks follow a non-central *t*-distribution. However, this parametric method is not appropriate for the problem at hand because, in normative samples drawn from the general adult population, the distribution of scores on mood scales typically exhibit extreme skew and leptokurtosis; moreover, some self-report scales have a limited range of scores. Alternative methods are therefore required. There are a number of candidate methods available: Because a percentile rank is simply a proportion multiplied by 100, it follows that methods of obtaining interval estimates for proportions can be used to obtain interval estimates of the percentile rank corresponding to each raw score. Details of the classical and Bayesian methods used to obtain interval estimates are set out in the Methods section; the Discussion section considers the interpretations that can be placed on these intervals.

Method

The main sources of data for this study were the samples recruited by Crawford, Henry, and colleagues in the aforementioned studies. These samples, and the additional samples recruited between 2006 and 2008 to augment the normative data, were recruited to be broadly representative of the general adult UK population in terms of the distributions of age, education, and gender (although, in most cases, females were over sampled). Recruitment was through as wide a variety of sources as was practical and included large and small businesses, public service organizations, community centres, and recreational groups. The majority of participants were recruited from urban/ suburban locations although rural/semi-rural dwellers were also represented.

Participants were asked to complete the relevant questionnaire(s) and place them in a sealed envelope. The questionnaires were filled in anonymously; i.e., the participants were asked not to write their name on any part of the questionnaire or envelope. The questionnaires were either collected at a later date by the investigators or returned by the participants by mail (or, more rarely, by hand). The combined refusal/non-return rates ranged from approximately 17% (for the sAD) to 21% (for the PANAS). Ethical approval for the original studies, and for recruitment of the additional samples, was obtained from the Psychology Ethics Committee of the University of Aberdeen.

Summary statistics (sample size, mean age and years of education and so forth) for the normative samples used to generate the percentile norms are presented in Table 1. Brief details of the five self-report mood scales are presented in the next sections.

The Depression Anxiety Stress Scales (DASS)

The DASS (Lovibond & Lovibond, 1995) is a 42 item selfreport measure yielding three scales of 14 items each: Depression, Anxiety, and Stress. Individual items are scored on a four-point scale (0 to 3). The DASS has high reliability, has a factor structure that is consistent with the allocation of the items to subscales, and exhibits high convergent validity with other measures of anxiety and depression (Crawford & Henry, 2003).

The DASS–21

The DASS–21 (Lovibond & Lovibond, 1995) is a short-form of the DASS in which each of the three subscales contains seven (rather than 14) items. The DASS–21 has high reliability, has a factor structure that is consistent with the allocation of the items to subscales, and exhibits high convergent validity with other measures of anxiety and depression (Henry & Crawford, 2005). It has a number of advantages over the full-length version. First, and most obviously, it takes less time to complete (and thus is more acceptable to both patients with limited concentration and busy clinicians). Second, the items retained from the full-length version are generally superior to those omitted and, as

Table 1. Summary statistics for the samples of the general adult population used to generate percentile norms for the various mood scales

Scale	Source of normative data	*N*	(F, M)	Age			Years of Education		
				Mean	*SD*	Range	Mean	*SD*	Range
DASS	Crawford & Henry (2003)	1771	(965, 806)	40.95	15.85	15–91	13.77	3.13	6–26
DASS–21	Henry & Crawford (2004) + 1134 additional cases	2928	(1662, 1266)	40.85	15.98	16–91	13.80	3.07	6–26
HADS	Crawford *et al.* (2001) + 2030 additional cases	3822	(2205, 1617)	41.39	16.50	16–91	14.12	3.15	6–27
PANAS	Crawford & Henry (2004) + 1524 additional cases	2527	(1441, 1086)	42.15	16.52	16–92	14.24	3.20	7–27
sAD	Henry *et al.* (2002)	758	(422, 336)	39.93	15.77	16–91	13.81	3.11	6–24

a result, it has a cleaner factor structure. The downside is that, although the DASS–21 has high reliability, its reliability is a little lower than that of the full-length DASS.

The Hospital Anxiety Depression Scale (HADS)

The HADS (Zigmond & Snaith, 1983) is a 14 item self-report measure which yields two scales (of seven items each): Anxiety and Depression. Items are scored on a four point scale (0 to 3). The HADS has good reliability (Crawford, Henry, Crombie, & Taylor, 2001) and demonstrates high convergent validity with other measures of anxiety and depression (Crawford & Henry, 2003). The depression scale items lay particular emphasis on anhedonia (loss of pleasure); this can be regarded as a positive feature, given that the tripartite theory of anxiety and depression suggests that anhedonia, or lack of positive affect, is important in differentiating depression from anxiety (Clark & Watson, 1991).

The Positive and Negative Affect Schedule (PANAS)

The PANAS (Watson, Clark, & Tellegen, 1988) consists of two 10-item scales designed to provide brief measures of positive and negative affect (PA and NA). Positive and negative affects are construed as broad mood dimensions which are orthogonal to each other. High NA is epitomised by subjective distress and unpleasurable engagement, and low NA by the absence of such feelings. PA represents the extent to which an individual experiences pleasurable engagement with the environment: High PA is characterized by enthusiasm and alertness, whilst low PA is characterized by lethargy and sadness. PA and NA are construed as independent (orthogonal) mood dimensions.

The PANAS test items consists of single mood descriptors (e.g. inspired; excited; nervous; jittery) and respondents are asked to rate the extent to which they have experienced these feelings within a specified time frame, using a five point scale ranging from one ('very slightly or not at all') to five ('very much'). All the data collected in the present study were collected using a time frame of 'during the last week'. The reliabilities of the PANAS scales are high, confirmatory factor analysis has indicated good fit for a two-factor model of PA and NA and the relationships of the scales to measures of anxiety and depression are consistent with Clark and Watson's (Clark & Watson, 1991) tripartite theory (Crawford & Henry, 2004).

The Symptoms of Anxiety and Depression Scale (sAD)

The sAD (Bedford & Foulds, 1978) is another brief selfreport measure of anxiety and depression. It can be used on its own (as in the present study) or as part of a larger inventory, the Delusion-Symptoms States Inventory. Its development was motivated by Foulds and Bedford's hierarchical model of personal illness (Foulds, 1976; Foulds, Bedford, & Csapo, 1975). The anxiety and depression scales each consist of seven items with each item rated on a four point scale. The sAD has been shown to have moderate-to-high reliability, and adequate convergent and discriminant validity as judged by its relationships with other measures of anxiety and depression (Henry, Crawford, Bedford, Crombie, & Taylor, 2002). The results from confirmatory factor analysis of the scale indicates it has a tripartite structure consisting of a general factor of negative affect/general psychological distress, and orthogonal factors of anxiety and depression (Bedford, Henry, & Crawford, 2005; Henry, Crawford, Bedford, Crombie, & Taylor, 2002).

Indices of general psychological distress

In some circumstances it would be useful to have a single measure of general psychological distress for each scale (Crawford, Henry, Crombie, & Taylor, 2001; Spinhoven *et al.*, 1997). Because of this, percentile norms for general distress were derived for each of the mood scales (with the exception of the PANAS) by summing scores from each of their subscales. An exception was made in the case of the PANAS because, as noted, this scale was designed to yield independent (i.e. uncorrelated) scales of positive and negative affectivity. In contrast, the subscales of the other mood scales are moderately to highly correlated (Crawford & Henry, 2003; Crawford, Henry, Crombie, & Taylor, 2001; Henry & Crawford, 2005; Henry, Crawford, Bedford, Crombie, & Taylor, 2002), thereby indicating that deriving a composite measure from them is useful.

Point estimates of percentile ranks

The standard method of obtaining percentile ranks was used (Ley, 1972): That is,

$$\text{Percentile Rank} = \left(\frac{m + 0.5k}{N}\right)100, \tag{1}$$

where m is the number of members of the normative sample obtaining a score lower than the score of interest, k is the number obtaining the score of interest, and N is the overall normative sample size. The percentile ranks thus obtained were then rounded to an integer for scores up to the 99th percentile and to one decimal place for those at or above the 99th percentile.

This latter procedure differs from that used in generating percentile tables for the original normative data. In the original publications the percentiles corresponding to given raw scores were all expressed as integers or, in the case of very high scores, were denoted as $>$99th percentile. On revisiting this for the present paper we realised that it is useful to make a distinction between a patient with a score that is estimated to be exceeded by only one in a thousand members of the general adult population and a patient whose score is estimated to be exceeded by one in a hundred. By adopting the present procedure this form of distinction can be made whilst simultaneously preserving the simplicity of using integer values for the percentile ranks of less extreme scores.

Interval estimates of percentile ranks

As noted, a further difference between the tabled values presented in the aforementioned papers and the present approach is that, in the present case, the aim was to accompany the *point* estimates of the percentile ranks corresponding to raw score with *interval* estimates of these quantities.

Classical (frequentist) interval estimates for percentile ranks

Classical methods of obtaining an interval estimate on a proportion are all based on the fact that the sampling distribution of a proportion follows a binomial distribution. Reviews and quantitative evaluations of the wide range of classical methods can be found in Newcombe (1998) and Brown, Cai and DasGupta (2001). To apply any of these methods would involve designating the number of people in a normative sample scoring below a score of interest as ‘successes’ and the remainder as ‘failures’.

However, for the present problem there is a complication. Although test scores are discrete (i.e. integer-valued), the underlying mood dimensions they index are generally taken to be continuous, real-valued quantities. Thus, a raw score of, say, seven is regarded as a point estimate of a real-valued score which could lie anywhere in the interval 6.5 to 7.4999 (plus an infinite number of additional 9s after the 4th decimal place). Put another way, in principle we could distinguish among individuals obtaining the same raw score were we to introduce tie-breaking items. This assumption of a continuous underlying score motivates the standard definition of a percentile rank (formula 1).

Normative data for mood scales will always contain a sizable number of tied scores; that is, a large number of people in the normative sample will obtain the same raw test score. Indeed, if the normative sample is very large and the data are heavily skewed (as is usually the case), then there could literally be hundreds of such ties for a given raw score. The present problem therefore differs from those dealt with by standard binomial sampling in which there can be no possibility of multiple ties.

One solution would be to simply ignore the problem of ties and obtain an interval based on x number of 'successes' with $x = m + 0.5k$ (where, as noted, m=number scoring below the score of interest and k=number obtaining the score of interest): The number of failures would then be $N - x$. However, regardless of which existing classical method was used to obtain this estimate, it would not incorporate all of the uncertainty involved. That is, the interval should reflect the possibility that, on the underlying continuous score, the number of people scoring lower than the score obtained by the case could be m, $m + 1$, . . ., or $m + k$. Each of these possibilities is given equal probability, as it is reasonable to suppose that differences between the case and the controls with which it is tied are arbitrary.

Crawford, Garthwaite and Slick (2008) have recently developed classical methods that incorporate the additional uncertainty arising from tied scores and this method was applied to the present data. To illustrate, suppose that in a normative sample of 100 people, 89 obtained lower scores than a case and two obtained the same score as the case. Then the point estimate of the percentile rank for the case's score (using formula 1) is 90 and applying the Crawford *et al.* method, the interval estimate is from 82.15 to 95.27. Suppose, however, that 85 obtained lower scores and 10 obtained the same score. The point estimate of the percentile rank is the same as in the foregoing example (90) but the interval estimate is from 79.79 to 97.10; the latter interval is wider because of the increased uncertainty introduced by the larger number of ties (2 versus 10).

The method developed by Crawford *et al.* can be construed as an extension of the standard Clopper-Pearson method (Clopper & Pearson, 1934) of obtaining interval estimates on a proportion to the situation in which there are ties (i.e. some members of the normative sample obtain the same raw score as a case). In practice a mid-p variant of the Clopper-Pearson method is often used in the standard situation where ties are not an issue, because the Clopper-Pearson interval is quite commonly regarded as being too conservative (Brown, Cai, & DasGupta, 2001).

For this reason Crawford *et al.* also offered a mid-p variant of their classical method. For the second numerical example used earlier (in which out of 100 members of a normative sample, 85 scored below a case's score and 10 obtained the same score) the mid-p interval estimate of the percentile rank is 80.30 to 96.80. It can be seen that this interval is narrower, i.e. less conservative than that obtained earlier (79.79 to 97.10).

Bayesian interval estimates for percentile ranks

Bayesian statistics provides an alternative to the classical (frequentist) approach to inference. The essential difference between the classical and Bayesian approaches is that the classical approach treats parameters as *fixed* but unknown whereas, in the Bayesian approach, parameters are treated as random variables and hence have probability distributions.

A Bayesian analysis of the present problem requires specifying a prior distribution (typically using a non-informative prior), and combining this prior with the data (the number of 'successes' and 'failures' as defined earlier) to obtain a posterior distribution (Bolstad, 2007). In the standard approach both the prior and posterior distributions are beta distributions. (A common choice of non-informative prior distribution is the Jeffrey's prior beta (0.5, 0.5)). The interval estimate of a score's percentile rank would be obtained by finding the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution (i.e. the 0.025 and 0.975 quantiles when a two-sided 95% interval is required). These quantiles would then be multiplied by 100 so they are expressed as percentile ranks rather than as proportions.

As was the case for the classical approach, this standard Bayesian approach can be modified to incorporate the additional uncertainty introduced by the presence of multiple ties (Crawford *et al.,* 2008). In the Bayesian case this is achieved by specifying a Jeffrey's prior distribution, with the posterior distribution a mixture of $k + 1$ beta distributions (each with equal weight in the mixture). The parameters of these $k + 1$ beta distributions are set by specifying the number of successes as m, $m + 1$, . . . , and $m + k$.

To illustrate, the Bayesian intervals using the previous examples, the Bayesian interval estimate of the percentile rank for the first example (89 out of 100 scoring below and two at the score) is from 82.99 to 94.74. For the second example (85 below and 10 at the score) the interval estimate is 80.36 to 96.74. These results demonstrate three features of the intervals. First, as was the case for the classical methods, the Bayesian method captures the greater uncertainty in the second example and hence produces a wider interval. Second, it can be seen that the Bayesian and classical intervals show a reasonable degree of convergence (the convergence between the Bayesian and classical approach being particularly close for the classical mid-*p* variant). Indeed the degree of convergence is very close indeed with samples as large as those employed to provide the normative data for the present study. This convergence is reassuring regardless of whether one is classical, Bayesian or eclectic in orientation. Third, it can be seen that, when the sample size providing normative data is modest (as in the present example where $N = 100$), there is considerable uncertainty over the percentile rank of a case's raw score.

One-sided versus two-sided intervals

In practice there will be occasions in which a one-sided interval may be preferred over a two-sided interval. For example, a clinician may be interested in whether a patient's score is less extreme than is indicated by the point estimate but not particularly interested in whether the score is even more extreme (or vice-versa). Both the classical and Bayesian methods developed by Crawford *et al.* (2008) are easily adapted to provide a one-sided limit. However, without prior knowledge of which limit is of interest (the situation here, as the aim is to provide intervals for use by others) it is more convenient to generate $100(1 - (\alpha + (a/2))$ two-sided intervals which then provide $100(1 - \alpha)$

one-sided lower and upper limits. For example, if a 95% lower limit on the percentile rank is required then a 90% two-sided interval is generated: The user then simply disregards the upper limit of the two-sided interval and treats the lower limit as the desired one-sided 95% limit.

Computer program for obtaining point and interval estimates of percentile ranks for raw scores on the various mood scales

The methods for obtaining interval estimates for a percentile rank developed by Crawford *et al.* (2008) are complex and time consuming to calculate. Moreover, for the problem at hand these limits need to be provided for the percentile ranks corresponding to all possible raw scores for each of the mood scales. It therefore makes sense to implement the methods into a computer program so that the limits can be obtained quickly and accurately.

Results

Summary statistics and reliabilities of the mood scales

The summary statistics (mean, median, standard deviation, and range) for the mood scales are presented in Table 2. The equivalent data for the total scores on these scales (used as indices of general psychological distress) are presented in Table 3. These tables also present reliability coefficients (Cronbach's alpha) for the scales; 95% confidence limits on these alphas were computed using Feldt's (1965) formula. For those scales for which no additional normative data were available (the full-length DASS and the sAD), the reliability coefficients are simply those reported in the original studies. For all others they were recomputed using the augmented normative samples.

Table 2. Summary statistics for the mood scales (α=Cronbach's alpha)

Scale		Mean	Median	*SD*	Range	α (95% CLs)
DASS	Depression	3.56	2	5.39	0–40	.95 (.94, .95)
	Anxiety	5.55	3	7.48	0–42	.90 (.89, .90
	Stress	9.27	8	8.04	0–42	.93 (.93, .94)
DASS–21	Depression	3.18	2	4.16	0–21	.90 (.89, .91)
	Anxiety	2.25	1	3.34	0–21	.84 (.83, .85)
	Stress	5.16	4	4.44	0–21	.89 (.88, .90)
HADS	Depression	3.71	3	3.17	0–20	.78 (.77, .79)
	Anxiety	6.42	6	4.00	0–21	.84 (.83, .85)
PANAS	PA	31.72	32	7.38	10–50	.88 (.87, .89)
	NA	17.04	15	6.68	10–49	.88 (.87, .89)
sAD	Depression	1.94	1	2.62	0–16	.85 (.83, .87)
	Anxiety	1.38	0	2.79	0–18	.77 (.75, .80)

Note. NA = Negative Affectivity; PA = Positive Affectivity.

It can be seen that the reliabilities of the mood scales are generally acceptable and in many cases are very high. It can also be seen from the narrowness of the confidence limits on these reliability coefficients that they provide very accurate estimates of the true reliability of the scales (mainly because of the large sample sizes).

Table 3. Summary statistics for total scores (General Psychological Distress; GPD) on the mood scales (α=Cronbach's alpha)

Scale	Mean	Median	*SD*	Range	α (95% CLs)
DASS GPD	18.38	13	18.82	0–121	.97 (.96, .97)
DASS–21 GPD	10.59	7	10.61	0–63	.94 (.94, .94)
HADS GPD	10.13	9	6.39	0–41	.87 (.86, .88)
sAD GPD	3.32	1	5.00	0–33	.88 (.87, .90)

Effects of demographic variables on mood scores

The relationships between the mood scales and demographic variables were examined by computing the Pearson Product Moment correlation between scores on each of the scales and age, years of education, and gender (females coded as 1, males as 2; the latter set of coefficients are termed point biserial correlation coefficients but are computed in the same way as the Pearson coefficient. The correlations with years of education were all modest, ranging from a minimum of -0.01 (for the Anxiety scale of the sAD), to a maximum of 0.11 (for the PA scale of the PANAS).

The correlations with age were also modest, ranging from a minimum of -0.02 (for the Anxiety scale of the full-length DASS) to a maximum of 0.19 (for the NA scale of the PANAS). The possibility of non-linear effects of age was examined using a series of hierarchical multiple regression analyses with the scales as the dependent variables: age was entered as the first predictor, followed by quadratic and then cubic functions of age (i.e. age^2 and age^3). For most scales, the addition of age^2 to the regression model did not lead to a significant change in R^2: A few significant effects were obtained but all were modest in size (the maximum change in R^2 was 0.019 for the Anxiety scale of the sAD). Addition of age^3 led to a significant change in R^2 only for the Depression scale of the HADS but, again, the magnitude of this effect was very modest (R^2 change = 0.006). In summary, age and years of education exerted, at most, a modest influence on scores on the mood scales. Therefore there was little or nothing to be gained by calculating percentile ranks separately for different age groups or different levels of education (and much to be lost in terms of reducing the sample sizes on which the norms would be based).

With regard to the point biserial correlations between the scales and gender, again most of the effects were very modest in size, ranging from a minimum of -0.04,for the Depression scale of the HADS, to a maximum of -0.16, for the Anxiety scale of the HADS (a negative correlation indicating higher scores for females). Crawford *et al.* (2001) reported a modest gender effect on the Anxiety scale of the HADS and provided separate percentile norms for males and females. Although the effects found here with an extended sample were also relatively modest (i.e. gender accounted for only 3.9% of the variance in Anxiety scores) we followed the same procedure: Examination of the percentile ranks for females and males did reveal differences; for example, a score of 12 was at the 87th percentile for females but at the 94th percentile for males. Separate summary statistics for the HADS scales for females and males are presented in Table 4.

Obtaining point and interval estimates of the percentile ranks for raw scores

As previously noted, a computer program for PCs, MoodScore_PRs.exe, was written (using the Delphi programming language) to express a patient's raw scores on the various scales as percentile ranks. The program is free and can be downloaded (either as

Table 4. Separate summary statistics for males and females on the HADS

Scale		Mean	Median	*SD*	Range
Females	Depression	3.81	3	3.30	0–18
	Anxiety	6.96	7	4.19	0–21
	GPD	10.76	10	6.70	0–38
Males	Depression	3.58	3	2.98	0–20
	Anxiety	5.67	5	3.60	0–21
	GPD	9.26	8	5.82	0–41

Note. GPD = General Psychological Distress.

an uncompressed executable or as a zip file) from the first author's web pages at www.abdn.ac.uk/~ psy086/dept/MoodScore.htm (After downloading the program it can be run by clicking on the program in windows explorer, or, if a shortcut to the program has been created on the user's desktop, by clicking on the shortcut icon). The program prompts the user to select the scale they wish to score. The user is then prompted to enter the patient' raw scores for the selected mood scale. When using the HADS the user also needs to identify the patient's gender through the use of radio buttons as the normative data are organized separately for females and males. Finally, there is the option of entering identifying information for the patient (in the form of User's Notes) for future reference.

The output from the program consists of a brief listing (resembling that in Tables 2 and 3) of the summary statistics for the scales; i.e. the mean, median, standard deviation, range, and reliability (users can suppress this listing if they wish). These summary data are followed by User's Notes, if these have been entered, and the point and interval estimates of the percentile ranks for the patient's raw scores. These results can be viewed on screen, saved to a file, and/or printed.

As noted, the provision of a computer program through which clinicians or researchers can obtain point and interval estimates of the percentile ranks for an individual's scores has advantages over the alternative of consulting multiple sets of tables. However, to illustrate the mapping of raw scores to percentile ranks and the accompanying interval estimates, Table 5 provides these data for the Anxiety scale of the DASS–21 (interval estimates having been calculated using the Bayesian method). This scale was chosen mainly because, compared to some of the others, it has a relatively small range of potential raw scores (0 to 21); the GPD scale from the full-length DASS, for example, has a potential range of 0 to 126 (this underlines the advantages of the program over the use of voluminous tables).

Discussion

Despite the widespread use of self-report scales of mood, it is only relatively recently that percentile norms from large samples have been available for many of these scales. We suggest that, by gathering together and supplementing the available normative data, the present study and accompanying program provide a useful and convenient resource for clinical research and practice.

A worked example

The computer program accompanying this paper was designed to be intuitive to use but a brief example may be helpful. Suppose a patient has completed the DASS–21

Table 5. Conversion of raw scores on the DASS–21 Anxiety scale to percentile ranks with accompanying 95% two-sided interval estimate (the interval estimates were obtained using the Bayesian method)

Raw score	Percentile rank (PR)	95% Interval estimate
0	20	1 to 38
1	49	39 to 59
2	65	59 to 71
3	75	71 to 80
4	82	79 to 85
5	86	84 to 88
6	89	87 to 91
7	91	90 to 92
8	93	91 to 94
9	94	93 to 95
10	95	94 to 96
11	96	95 to 97
12	97	96 to 98
13	98	97 to 99
14	98	98 to 99
15	99	98 to 99
16	99.1	98.7 to 99.5
17	99.5	99.1 to 99.7
18	99.7	99.4 to 99.9
19	99.8	99.6 to 99.9
20	99.9	99.7 to 100
21	$>$99.9	99.8 to 100

and obtains raw scores of 4, 14, and 13 on the Depression, Anxiety and Stress scales respectively. A screen capture of the program set up to process these scores is presented as Figure 1. It can be seen that the DASS–21 has been selected using the appropriate radio button, the radio buttons for recording the patient's gender are disabled (because gender effects were minimal for this scale), Bayesian (rather than classical) interval estimates of the percentile ranks have been requested (this being the default option), the interval width has not been changed from its default value of 95%, and the user has entered the patient's raw scores into their respective data fields.

From this setup the results would then be obtained by clicking on the compute button. The results screen (not shown) provides the point estimates of the percentile ranks for the patient's scores with their accompanying interval estimates: The percentile rank of the Depression score is 72 (95% IE = 68 to 75), the PR of the Anxiety score is 98 (95% IE = 98 to 99), and the PR of the Stress score is 93 (95% IE = 91 to 94).

In this example, the Depression score is elevated relative to the average score for the general adult population (i.e. the score is above the 50th percentile) but is not that unusual. That is, a sizable percentage of the general adult population (28%) would be expected to obtain higher scores. In contrast, the Anxiety score is extreme: 98% of the populations are expected to have scores lower than the patient (and thus only 2% would be expected to score higher). Finally, the PR of the patient's total (GPD) score on the DASS (i.e. the sum of raw scores on each of the three DASS scales) is 94 (95% CI = 93 to 95). Note that, in this specific and hypothetical case, the total score is of questionable utility given the discrepancies between the percentile ranks for the different subscales.

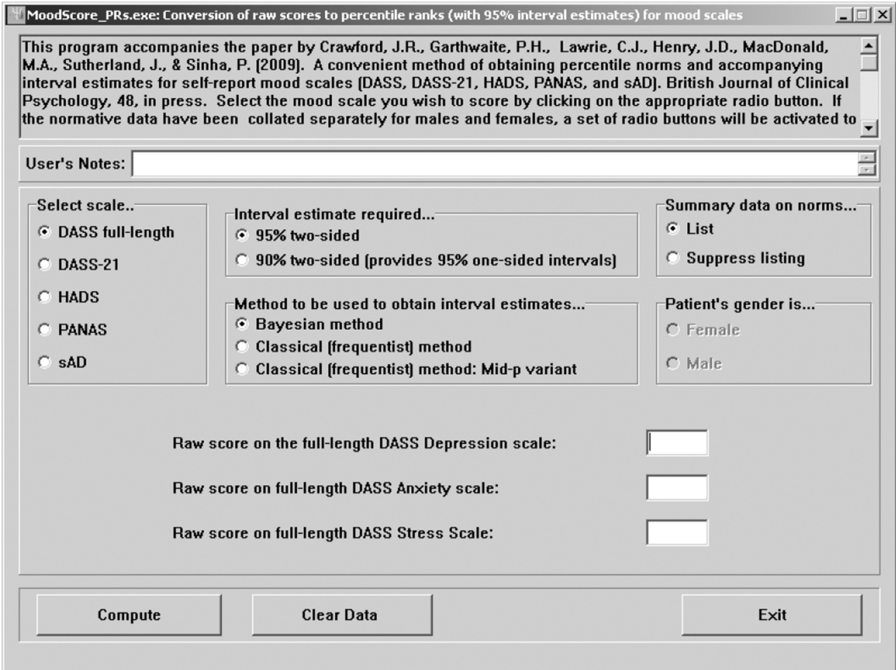

Figure 1. Screen capture of the computer program (MoodScore_PRs.exe) used to express raw scores on various commonly used self-report mood scales as percentile ranks (with accompanying interval estimates); in this example, the program is set up to score the DASS-21 and a 95% (two-sided) Bayesian interval estimate has been selected.

The interval estimates of the percentile ranks in this example are fairly narrow, thereby indicating that the point estimates of percentile ranks of the raw scores obtained using the normative sample provide an accurate estimate of the true percentile ranks of these raw scores in the population. This is generally the case for all of the mood scales included here. However, the normative sample size ($N = 2928$) for the DASS–21 is the largest of all those included. (The overall normative sample size for the HADS is larger, $N = 3822$, but the percentile norms were generated separately for males and females). For the other mood scales the limits will be wider, although still narrow in absolute terms. For example, in the case of the sAD (which has the smallest normative sample of any of the scales; $N = 758$) the 95% limits associated with a percentile rank of 94 on the Anxiety scale are from 92 to 96.

Exceptions to the general rule that the interval estimates of the percentile rank are narrow for these normative data occur for scores that are very low. For example, in the case of the DASS–21 Anxiety Scale (see Table 5) the interval estimate of the percentile rank for a raw score of zero is very wide (from 1 to 38, point estimate = 20). This occurs because very large number of the normative sample obtained a score of zero (that is, there was a very large number of ties) and thus a high degree of uncertainty over an individual's percentile rank. It will be appreciated however, that this is not of much practical concern as there is little need to quantify the uncertainty for low scores.

In interpreting, the percentile ranks obtained from the present normative data it is important to stress that a percentile rank does not have to be very extreme to be a potential cause for concern. For example, Shepherd, Cooper, Brown and Kalton (1966) reported that between 30 and 40% of the general population suffer from anxiety to an extent that would benefit from clinical intervention. Although point prevalence estimates for anxiety and depression in the general population vary from study to study, it is also clear that a sizable percentage exhibit symptoms severe enough to warrant a clinical diagnosis (see Crawford, Henry, Crombie, & Taylor, 2001 for a brief review). In Meltzer, Gill, Petticrew, and Hinds (1995) survey of 10,000 UK households, the (1-week) prevalence of anxiety disorders was 13.9% (this percentage being based on the inclusion of cases diagnosed as mixed anxiety/depression). Prevalence rates for depression tend to be lower but have commonly been reported to be around 3 to 4% (Horwath & Weissman, 1995).

Bayesian versus classical interpretations of the interval estimate on a score's percentile rank

As Antelman (1997, p. 375) notes, the frequentist (classical) conception of a confidence interval is that, 'It is one interval generated by a procedure that will give correct intervals 95% of the time. Whether or not the one (and only) interval you happened to get is correct or not is unknown' Thus, in the present context, the classical interpretation of the interval estimate on the percentile rank for a raw score on (say) the DASS is as follows, 'if we could compute a confidence interval for each of a large number of normative samples collected in the same way as the present DASS normative sample, about 95% of these intervals would contain the true percentile rank of the patient's score'.

The Bayesian interpretation of such an interval is 'there is a 95% probability that the true percentile rank of the patient's score lies within the stated interval'. This statement is not only less convoluted but it also captures what a clinician would wish to conclude from an interval estimate (Crawford & Garthwaite, 2007). Indeed most psychologists who use frequentist confidence limits probably construe these in what are essentially Bayesian terms (Howell, 2002).

For the present problem the frequentist and Bayesian approaches exhibit a high degree of convergence (Crawford *et al.,* 2008). This can readily be verified by the reader by comparing the two sets of interval estimates for scores on the mood scales featured in the present paper. For instance, the intervals for the DASS–21 example reported earlier (in which a patient obtained scores of 4, 14, and 13 on the Depression, Anxiety and Stress scales respectively) were calculated using the Bayesian method. However, the classical intervals are identical. The upshot is that psychologists can place a Bayesian interpretation on the interval estimate of a percentile rank, regardless of whether it was obtained using the Bayesian or classical methods (The classical methods were made available to cater for those who are strongly wedded to the classical approach to inference).

Confidence intervals capturing sampling error versus measurement error

The confidence intervals on the percentile ranks provided here (whether obtained using classical or Bayesian methods) should not be confused with confidence limits derived from classical *test* theory that attempt to capture the effects of measurement error on an individual's score (Crawford & Garthwaite, 2008b).

When the latter intervals are used, the clinician is posing the question 'assuming scores are normally distributed, and assuming no error in estimating the population mean, standard deviation and reliability coefficient of the test, how much uncertainty is there over an individual's score as a function of measurement error in the scale?' (Crawford & Garthwaite, 2008a). In contrast, when using the intervals presented in the present paper, the concern is solely with the score *in hand*. The more concrete question posed is how much uncertainty is there over the standing (i.e. percentile rank) of the score the individual obtained as a function of error in using a normative sample to estimate its standing in the normative population. That is, they do not address the issue of what score an individual might obtain on another occasion, or on a set of alternative, parallel items, but simply provide interval estimates for the percentage of the normative population who would score below the score obtained by the individual.

The Positive Affect (PA) scale of the PANAS

For all the self-report scales reported here, high scores are indicative of disturbance or distress. The one exception to this general feature occurs in the case of the PANAS Positive Affectivity (PA) scale. Although a very high score on this scale is worthy of attention (i.e. manic patients will typically score very highly on PA), the principal clinical concern will be with patients who show very *low* levels of positive affect (i.e. are anhedonic) and thus obtain low percentile ranks. As *high* scores on all the other scales, including the Negative Affectivity scale of the PANAS, are indicators of disturbance, this is potentially confusing. We considered reflecting PA scores so that high scores would indicate low positive affect/presence of anhedonia but decided that this might be even more confusing. Instead, a reminder that attention should be paid to scores with low percentile ranks on the PA scale is included in the output of the computer program.

Provision of comparison standards for group studies

The emphasis in the current paper has been on the use of these normative data with individual patients. However, they have other potential uses. For example, many studies of anxiety and depression in clinical populations (or in particular occupational groups and so forth) do not collect control data from a sample drawn from the general adult population. The summary statistics provided here (Tables 2 and 3) could serve as useful comparison standards against which to compare the means or medians of clinical samples. Alternatively, the median score obtained in a clinical sample could be expressed in terms of its percentile rank in the general population sample using the computer program described earlier (the interval estimates provided should be ignored in such an application as they quantify the uncertainty over the standing of a single score rather than the average score of a second sample). Note that, even if there are marked differences in age or educational level between such clinical samples and the samples of the general adult population reported here, this does not preclude the use of these latter samples for comparison purposes because age, education, and (for the most part) gender do not exert an appreciable effect on scores on the scales.

Future developments

The scales included in this study were selected because they are in widespread use, have good psychometric properties, and are all public domain instruments. However, pursuing a similar exercise for other commonly used self-report mood scales would provide a useful service for clinicians and researchers (although the costs may be prohibitive for tests not in the public domain). We are currently gathering normative data on further self-report scales and will update the program to incorporate these data when completed.

The normative data featured in the present paper have all been gathered from samples of the general adult population. However, it would be useful for clinicians if point and interval estimates for the percentile ranks of mood scores were also available for *clinical* populations. These would provide further context when evaluating an individual's score. Percentile norms could be gathered for clinical populations encountered in general medicine (i.e. in cardiology, oncology, diabetic medicine, etc.) and for neurological populations (traumatic brain injury, stroke etc.) as well as in mental health settings. For example, in the case of the HADS scores of a patient who has suffered a TBI, it would be useful to be able to obtain an estimate of how unusual or otherwise these scores are in the population of patients who have suffered a TBI. We would welcome collaborators in such developments: By pooling data it would be possible to provide clinicians with a single, reliable, and convenient source of normative data for a wide range of scales and populations.

Some researchers may prefer to make normative data available separately, rather than pooling the data as suggested above. Given the importance placed on the use of interval estimates in contemporary biometry, we suggest that any such normative exercise should provide both point and interval estimates for the resultant percentile ranks. It would be particularly important to provide interval estimates if the normative samples were modest in size.

Conclusion

The present normative data and accompanying computer program provides a quick and reliable means of obtaining percentile norms for a range of widely used self-report mood scales. The percentile norms allow clinicians to quantify the rarity or otherwise of the patient's score and are therefore a useful supplement to the traditional cut-off scores available for some (but not all) of these scales. Expressing a patient's score in terms of its percentile rank, rather than simply as below or above a cut-off, is also in keeping with a conception of anxiety and depression as dimensional rather than categorical constructs. Finally, the provision of interval estimates for the percentile rank of a score serves the general purpose of reminding clinicians that all normative data are fallible. It also serves the specific and practical purpose of quantifying the uncertainty over the standing of an individual's score when referred to such data.

References

Antelman, G. (1997). *Elementary Bayesian statistics*. Cheltenam, UK: Elgar.

Bedford, A., & Foulds, G. (1978). *Delusions-symptoms-States inventory State of Anxiety and Depression*. Windsor: NFER-Nelson.

Bedford, A., Henry, J. D., & Crawford, J. R. (2005). Personal disturbance scale: Factor structure confirmed in a large nonclinical sample. *Psychological Reports*, *96*, 107–108.

Bolstad, W. M. (2007). *Introduction to Bayesian statistics* (2nd ed.). Hoboken, NJ: Wiley.

Brown, L. D., Cai, T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*, 101–117.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology*, *100*, 316–336.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404–413.

Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia*, *40*, 1196–1208.

Crawford, J. R., & Garthwaite, P. H. (2007). Comparison of a single case to a control or normative sample in neuropsychology: Development of a Bayesian approach. *Cognitive Neuropsychology*, *24*, 343–372.

Crawford, J. R., & Garthwaite, P. H. (2008a). On the 'optimal' size for normative samples in neuropsychology: Capturing the uncertainty associated with the use of normative data to quantify the standing of a neuropsychological test score. *Child Neuropsychology*, *14*, 99–117.

Crawford, J. R., & Garthwaite, P. H. (2008b). Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *Clinical Neuropsychologist, in press*.

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2008). *On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores.* Manuscript submitted for publication.

Crawford, J. R., & Henry, J. D. (2003). The Depression Anxiety Stress Scales: Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology*, *42*, 111–131.

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, *43*, 245–265.

Crawford, J. R., Henry, J. D., Crombie, C., & Taylor, E. P. (2001). Normative data for the HADS from a large non-clinical sample. *British Journal of Clinical Psychology*, *40*, 429–434.

Faust, D. (1998). Forensic assessment. In comprehensive clinical psychology Volume 4: Assessment. (pp. 563–599). Amsterdam: Elsevier.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability Coefficient Twenty. *Psychometrika*, *30*, 357–370.

Foulds, G. A. (1976). *The hierarchical nature of personal illness*. London: Academic Press.

Foulds, G. A., Bedford, A., & Csapo, K. G. (1975). Class change in the personal illness hierarchy. *British Journal of Psychiatry*, *127*, 316–319.

Gardner, M. J., & Altman, D. G. (1989). *Statistics with confidence- confidence intervals and statistical guidelines*. London: British Medical Journal.

Henry, J. D., & Crawford, J. R. (2005). The 21 Item versions of the Depression Anxiety Stress Scales (DASS–21): Normative data and psychometric evaluation in a large non-clinical sample. *British Journal of Clinical Psychology*, *44*, 227–239.

Henry, J. D., Crawford, J. R., Bedford, A., Crombie, C., & Taylor, E. P. (2002). The Personal Disturbance Scale (sAD): Normative data and latent structure in a large non-clinical sample. *Personality and Individual Differences*, *33*, 1343–1360.

Horwath, E., & Weissman, M. M. (1995). Epidemiology of depression and anxiety disorders. In M. T. Tsuang, M. Tohen, & G. E. P. Zahner (Eds.), *Textbook in psychiatric epidemiology.* New York, NY: Wiley.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Belmont, CA: Duxbury Press.

Ley, P. (1972). *Quantitative aspects of psychological assessment*. London: Duckworth.

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the Depression Anxiety Stress Scales*. Sydney: Psychology Foundation Monograph.

Meltzer, H., Gill, B., Petticrew, M., & Hinds, K. (1995). *The prevalence of psychiatric morbidity among adults living in private households*. London: HMSO.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, *17*, 857–872.

Shepherd, M., Cooper, B., Brown, A. C., & Kalton, C. W. (1966). *Psychiatric illness in general practice*. London: Oxford University Press.

Sherrets, F., Gard, G., & Langner, H. (1979). Frequency of clerical errors on WISC protocols. *Psychology in the Schools*, *16*, 495–496.

Snaith, R. P., & Zigmond, A. S. (1994). *HADS: Hospital Anxiety and Depression Scale*. Windsor: NFER Nelson.

Spinhoven, P., Ormel, J., Sloekers, P. P. A., Kempen, G., Speckens, A. E. M., & VanHemert, A. M. (1997). A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. *Psychological Medicine*, *27*, 363–370.

Sullivan, K. (2000). Examiners' errors on the Wechsler memory scale-revised. *Psychological Reports*, *87*(1), 234–240.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *47*, 1063–1070.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, *67*, 361–370.

Received 27 June 2008; revised version received 29 September 2008