Running head: Uncertainty associated with normative data

Child Neuropsychology, in press

On the "optimal" size for normative samples in neuropsychology: Capturing the uncertainty when normative data are used to quantify the standing of a neuropsychological test score

John R. Crawford

University of Aberdeen

Paul H. Garthwaite

Department of Statistics

The Open University

Address for correspondence: Professor John R. Crawford, School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 3HN, United Kingdom. E-mail: j.crawford@abdn.ac.uk

Abstract

Bridges and Holler (2007) have provided a useful reminder that normative data are fallible. Unfortunately, however, their paper misleads neuropsychologists as to the nature and extent of the problem. We show that the uncertainty attached to the estimated *z* score and percentile rank of a given raw score is much larger than they report and that it varies as a function of the extremity of the raw score. Methods for quantifying the uncertainty associated with normative data are described and used to illustrate the issues involved. A computer program is provided that, on entry of a normative sample mean, standard deviation and sample size, provides point and interval estimates of percentiles and *z* scores for raw scores referred to these normative data. The methods and program provide neuropsychologists with a means of evaluating the adequacy of existing norms and will be useful for those planning normative studies.

Keywords: norms; confidence limits; sample size; normative comparisons

INTRODUCTION

Bridges and Holler (2007) have recently provided neuropsychologists with a useful reminder that normative data can be highly fallible when it is obtained from modestly sized samples. They attempted to quantify the degree of uncertainty surrounding the standing of a raw score as a function of the size of the normative sample. The aims were to provide neuropsychologists with a principled means of evaluating the adequacy of existing normative data and to provide guidance for those planning a normative study. Unfortunately the methods they adopt are not fit for purpose. However, the aims are laudable and so in the present paper we clarify the issues involved and provide appropriate methods to achieve these aims.

The Bridges and Holler (2007) approach

The Bridges and Holler approach is based on calculating a two-sided 95% confidence interval on the normative population mean using the standard formula

$$\mathrm{CI}_{0.95} = \overline{x} \pm t_{0.025}\left(\frac{s}{\sqrt{n}}\right), \qquad (1)$$

where $\overline{x}$ is the normative sample mean, s is the sample standard deviation, n is the sample size, and $t_{0.025}$ is the critical value for t on $n-1$ df.

To quantify the uncertainty associated with the standing of a given raw score they propose substituting the lower and upper limits on the mean into the formula for a z score in place of the sample mean. That is

$$\text{Lower 95\% confidence limit on 'z'} = \left(x - \overline{x}_U\right)/\,s\,, \qquad (2)$$

and

$$\text{Upper 95\% confidence limit on 'z'} = \left(x - \overline{x}_L\right)/\,s\,, \qquad (3)$$

where *x* is a given raw score, and $\overline{x}_U$ and $\overline{x}_L$ are the upper and lower limits on the mean obtained using formula (1). They suggest that this method gives a 95% confidence interval on the *z* score corresponding to a given raw score (we place inverted commas around *z* in the formula because, as will be shown, it is not justified to denote this quantity as a *z* score). They also convert the lower and upper limits on *z* to percentiles to obtain lower and upper limits on the percentile for the raw score.

To illustrate, take the worked example set out by Bridges and Holler of Johnny, a nine year old boy, who obtained a raw score of 18 on the Hooper Visual Organization Test. The normative sample for nine year old boys consisted of 20 persons and the normative sample mean and standard deviation was 23.23 and 3.16 respectively (Demsky, Carone, Burns, & Sellers, 2000). Thus Johnny's *z* score was -1.66 and the score is estimated to be at the 5th percentile. The lower and upper 95% limits on the normative mean are 21.75 and 24.71. Substituting these into the formula for *z*, that is, applying equations (2) and (3), gives a supposed lower limit on *z* of -2.12 and an upper limit of -1.19 (note that is it the upper limit on the mean that is used to obtain the lower limit on *z*). Using a table of areas under the normal curve the lower limit on *z* corresponds to the 2nd percentile and the upper limit to the 12th percentile.

These results suggest that, when attempting to quantify the standing of a raw score, there is a worryingly high degree of uncertainty associated with the use of normative data based on modestly sized samples. In this example, it is possible that a raw score of 18 could be more than two standard deviations below the normative population mean (i.e., very extreme) or just over one standard deviation below the mean (low, but not unusual). As we will see, however, the uncertainty is often considerably greater than even these figures suggest.

Bridges and Holler present tables and graphs of the width of these confidence limits for published pediatric normative data on four neuropsychological tests – the Boston Naming Test, (BNT`; Kirk, 1992a) the Rey Auditory-Verbal Learning Test (RAVLT`; Bishop, Knights, & Stoddart, 1990), the Rey-Osterreith Complex Figure Test (ROCFT`; Kirk, 1992b), and the Hooper Visual Organization Test (HVOT`; Demsky, Carone, Burns, & Sellers, 2000). They also use their results to develop guidelines for the "optimal" size of normative samples in neuropsychology.

The core problem with the Bridges and Holler approach

As is common, Bridges and Holler assume that the scores of a normative sample are drawn from a normal distribution. They state, "population distributions are assumed to be normal" (p. 537). The problem is that they then proceed as if this ensures that all quantities derived from these scores will also have a normal distribution; this is erroneous. In Appendix 1 we show, using basic algebra, that the equations for Bridges and Holler's confidence limits (equations 2 and 3) reduce to

$$\text{Lower 95\% confidence limit on 'z'} = \left(\frac{x-\bar{x}}{s}\right) - \left(\frac{t_{0.025}}{\sqrt{n}}\right). \qquad (4)$$

and

$$\text{Upper 95\% confidence limit on 'z'} = \left(\frac{x-\bar{x}}{s}\right) + \left(\frac{t_{0.025}}{\sqrt{n}}\right). \qquad (5)$$

The quantities obtained from (4) and (5) have no obvious distribution and, despite denoting them as *z*, Bridges and Holler give no reason for us to believe they are quantiles from a standard normal distribution. Indeed, in one special case, we can determine the nature of the distributional form of these quantities and it is not standard normal. When a score is at the mean of the normative sample the first term in the right hand side of (4) and (5) drops out and we have

$$\text{Lower 95\% confidence limit on 'z'} = -t_{0.025} / \sqrt{n}. \qquad (6)$$

and

$$\text{Upper 95\% confidence limit on 'z'} = t_{0.025} / \sqrt{n}. \qquad (7)$$

It can be seen that, in this special case, the quantities are simply quantiles from a scaled *t* distribution with a variance determined by the normative sample size. A standard *t* distribution has a variance of $v/(v-2)$ if it has *v* degrees of freedom. As *n* increases, $t/\sqrt{n}$ will tend to a normal distribution but its variance will tend to $1/n$. Thus the normal distribution to which it tends is not the *standard* normal distribution and so, even asymptotically, the quantity should not be called a *z* score.

This special case is particularly important because it is used in Tables 1 to 4 and Figures 1 and 2 of Bridges and Holler (2007) to illustrate the uncertainty surrounding scores referred to the normative data sets for the AVLT, BNT, Hooper and Rey-Osterreith figure. Note that, in this special case (equations 6 and 7), the upper and lower limits do not depend upon the sample data at all, apart from the sample size. Thus, when Bridges and Holler point to the “remarkable similarity” (p. 532) between results for the Boston Naming Test and the AVLT, this similarity is in no way remarkable but simply reflects the fact that the sample sizes are similar across the two sets of normative data. Specifically, in each of Tables 1 to 4 of Bridges and Holland, columns 7 and 8 can be obtained simply by calculating $t_{0.025} / \sqrt{n}$ and the last two columns are then obtained using standard normal tables where, as before, $t_{0.025}$ is the critical value for *t* on $n-1$ df.

The effect of error in estimating the normative population standard deviation

Bridges and Holler's attempt to allow for error in estimating the normative population mean in order to capture the uncertainty over the standing of a raw score. The approach they adopt does not achieve their aims but, in any case, there is a further problem. In order to quantify the uncertainties arising from using a normative sample to estimate the population standing of a score, it is necessary to allow for error in estimating the normative population mean *and* standard deviation.

The need to allow for error in estimating the population standard deviation can be illustrated by a simple thought experiment. Suppose that normative data for a neuropsychological test were obtained from a modestly sized sample (say n = 20) of nine year old children and that the normative sample mean and standard deviation were 40 and 10 respectively. Suppose also that interest is in the standing of a raw score of 25. The raw score converts to a z score of -1.5 and so the score is estimated to be exactly one and a half standard deviations below the population mean; using a table of areas under the normal curve, the score is therefore estimated to be at the 6.7th percentile. Further suppose that, by chance (unlikely though it is), the sample mean exactly matched the population mean, i.e., the true mean for nine year olds on this test is 40. With a normative sample of this size there is going to be a high degree of uncertainty surrounding the population standard deviation of the test. Based on the sample n and standard deviation, the 95% confidence interval on the population standard deviation is from 7.6 to 14.6 (the sampling distribution of the variance follows a chi-square distribution on $n-1\ df$ and this allows us to set confidence limits on the population variance or standard deviation).

Suppose that the true population standard deviation for this test was 12, which is well within the bounds of possibility given the confidence interval. Converting the raw score of 25 to a *z*-score based on this population standard deviation yields a score of -1.25. Thus the raw score is much less extreme than the normative sample indicates; 10.6 % percent of the normative population are expected to obtain a lower score (i.e., the score is at the 10.6th percentile). Conversely, suppose that the population standard deviation was only 8 (again this is well within the bounds of possibility). Then the *z* score for the raw score of 25 is -1.875 and in this scenario the score is much more extreme than is indicated using the normative sample data, it is at the 3rd percentile.

Another consequence of error in estimating the population standard deviation from a sample standard deviation is that the uncertainty over the standing of raw scores will increase as a function of the distance of the raw score from the mean. This can also be illustrated without recourse to statistical theory by another thought experiment. Suppose that, as in the previous example, the mean and standard deviation for a normative sample on a neuropsychological test is 40 and 10 respectively and again, for simplicity, let us assume that the sample mean equals the population mean. Take a score close to the mean, say 38, and an extreme score, say 20. Then the *z* scores corresponding to these raw scores are -0.2 and -2.0 respectively. Now suppose that the population standard deviation is 11, i.e., the sample standard deviation is an underestimate of the true standard deviation. Then, using the population standard deviation, the *z* score corresponding to raw scores of 38 and 20 are -0.18 and -1.82. In both cases the use of the sample standard deviation in place of the population standard deviation would lead us to believe that individuals obtaining these raw scores were performing more poorly than is in fact the case. However, it can be seen that the effects are much more marked on the more extreme score. If the population standard deviation is smaller then the sample standard deviation, say 9, then the level of performance will be artificially inflated. Again, however, the effects are more marked on the extreme scores (in this latter scenario the *z* scores calculated using the population standard deviation as the divisor are -0.22 and -2.22).

In summary, the fallibility of normative data leads to much greater uncertainty over extreme scores than over scores close to the mean. It follows that a method, such as that proposed by Bridges and Holler (2007) that does not account for this effect cannot quantify the uncertainty.

Simultaneously allowing for error in estimating the normative mean and standard deviation

The foregoing examples make it clear that, in order to estimate the uncertainty surrounding the standing of a given raw score (i.e., its *z* score or percentile rank), the error in estimating the population standard deviation must be taken into account. However, in these examples the error in estimating the population mean was ignored, which is clearly inappropriate. To solve the problem at hand, we need to simultaneously allow for the error associated with estimating the population mean *and* standard deviation.

Fortunately, statistical methods developed by Crawford and Garthwaite (2002) can be used to tackle this problem. In commenting on these methods Bridges and Holler (2007) state that they "are the equivalent of applying statistical Band-Aids ® to our problem… Calculating confidence intervals around an examinee's score does nothing to address the use of the normative sample mean as a perfect parameter, rather than a sample estimate" (p. 538). This statement is unfortunate because the methods developed in Crawford and Garthwaite (2002) are exactly those required to address the problem at hand. That is, the methods quantify the uncertainty introduced when statistics from a normative sample are used to estimate the standing of a raw score either in terms of its standard deviation units from the population mean (i.e. a *z* score) or in terms of the score's percentile rank.

The methods are based on non-central *t*-distributions and, in their original form, provide 95% confidence limits on the percentage of the population expected to obtain a score lower than a given score. Crawford and Garthwaite (2002) refer to these limits as "confidence limits on the abnormality of a score" but it can be seen that this is just an alternative way of referring to confidence limits on the percentile rank of a score. It would also be useful to express the uncertainty on a *z* score metric as Bridges and Holler attempt to do. Crawford, Lynch and Garthwaite (submitted) have recently addressed this problem and shown that limits on *z* can readily be obtained as an intermediate step in calculating Crawford and Garthwaite's (2002) limits on a score's percentile rank. For the convenience of the reader, Appendix 2 brings together the derivation of the limits provided by Crawford and Garthwaite (2002) with information on the calculations involved and the further observations made by Crawford et al. (submitted).

The uncertainty over the standing of a raw score as a function of sample size and extremity of the score

As a first illustration of the correct method of quantifying the uncertainty over the standing of a raw score we return to the case example of Johnny. Recall that Johnny's raw score of 18 was estimated to be 1.66 standard deviations below the population mean (i.e., a raw score of 18 converts to a *z* score of -1.66). Bridges and and Holler's 95% confidence interval on *z* was from -2.12 to -1.19, with corresponding confidence limits on the percentile of from the 2nd to 12th percentile. Applying the methods set out in Appendix 2, the correct 95% confidence interval on *z* is from -2.33 to -0.96 and the limits on the percentile are from the 1st to 17th percentile. It can be seen that the interval on *z* is considerably wider than that provided by Bridges and Holler's method: the interval spans 1.39 standard deviations as opposed to 0.93 standard deviations. It can also be seen that the correct interval is very wide in absolute terms; with normative samples of this size (n = 20 in the example) there is considerable uncertainty over the standing of a raw score.

Tables 1a to 1c were constructed to compare further the two sets of confidence limits. These tables record the width of limits on *z* and the percentile rank for a raw score as a function of the size of the normative sample (the sample *n* is varied from 5 to 500) and as a function of the extremity of the raw score. Table 1a records the limits for a score at the normative mean, Table 1b for scores one standard deviation below the mean, and Table 1c for scores two standard deviations below the mean. Note that, because Bridges and Holler's (2007) method do not vary with the extremity of the score, the entries for the width of their confidence limits are identical across the three tables. It can be seen from Table 1a that, when the score is at the mean of the normative sample, Bridges and Holler's intervals are wider than the correct intervals obtained using Crawford and Garthwaite's method when the normative sample is modest in size. As the sample size increases their limits converge on the correct limits so that, when the normative sample *n* is 500, they are, for all intents and purposes, indistinguishable from the latter.

For a score one standard deviation below the sample mean (Table 1b), Bridges and Holler's limits are too narrow (the exception being for a very modest normative sample size of 5 where they remain too wide). It can also be seen that their limits differ from the correct limits even for large samples. The discrepancies between Bridges and Holler’s limits and the correct limits become even more pronounced with more extreme scores, as is illustrated in Table 1c which records the limits for a score 1.5 standard deviations below the normative sample mean. In this scenario the limits are too narrow, even for a normative sample of 5.

The formula for confidence intervals proposed by Bridges and Holler is not derived through rigorous mathematics so properties of the method were not immediately transparent. However, the results in Table 1a show that their confidence intervals are too wide when the raw score equals the mean of the normative sample. Also, as Tables 1a to 1c illustrate, the widths of their intervals do not change as the raw score deviates from the normative sample mean. This happens because their method fails to account for the uncertainty in estimating the normative population’s standard deviation from a normative sample. Confidence intervals should become wider as the difference between the raw score and the normative sample mean increases. Tables 1b and 1c show that, as the difference between the raw score and the normative sample mean increases, the correct width of a confidence interval soon exceeds the widths of the intervals given by Bridges and Holler, unless the sample size is very small. In a later section we report simulations where we fix the difference between the raw score and the normative *population* mean, rather than the normative sample mean. We find that, for any sample size, the coverage of the intervals given by Bridges and Holler is too small, on average, for any non-zero difference between the raw score and the normative population mean.

Insert Tables 1a to 1c about here

Point estimates of the percentile rank for a given raw score or z score

The aim of Bridges and Holler's paper was to quantify the uncertainty associated with the use of normative sample data. This is also the primary aim of the present paper and so, up to this point, the focus has been on *interval* estimates of the true standing (i.e. the true percentile rank or *z* score) for a given raw score. However, it is also appropriate to discuss the issue of *point* estimates of the percentile rank for a given raw score. In Bridges and Holler's paper the point estimate for the percentile rank of a raw score is given as the probability for the score following its conversion to a *z* score. Thus, for example, if a normative sample has a mean of 50 and standard deviation of 10, the *z* score for a raw score of 40 is -1.0 and a table of the areas under a normal curve can be used to convert this quantile to a probability. Multiplying the probability by 100 gives a point estimate of the percentile rank for the raw score: in this case the percentile rank is 15.9 (i.e., 15.9% of the population are expected to obtain a score lower than 40).

Although widely used, the foregoing point estimate is *not* the optimal estimate of the percentile rank for a given raw score. When a *z* score obtained using a sample mean and standard deviation is referred to a table of areas under the normal curve to obtain a probability the mean and standard deviation are treated as though they were the population mean and standard deviation; the result is that the percentile rank will more often than not be exaggerated (scores above the mean will be estimated to have a higher percentile rank, scores below the mean will be estimated to have a lower percentile rank). This occurs largely because errors in estimating the standard deviation do not have a symmetric effect on the percentile rank. (It is also the case that the sampling distribution of the variance is not symmetric.) To help understand this, suppose the estimate of the percentile rank is 2.5. Error in the standard deviation may mean that this is an overestimate, but it cannot be overestimated by more than 2.5, as a percentile rank cannot be negative. In contrast, it could be underestimated by as much as 97.5 (A percentile rank must be less than 100.)

To obtain the optimal point estimate for the percentile rank for a given raw score requires use of a (central) *t*-distribution rather than *z*. When a raw score is subtracted from a *sample* mean and divided by a *sample* standard deviation the resultant quantity is a *t*-variate rather than a *z*; it is distributed as *t* on $n-1$ df , where *n* is the size of the sample used to provide the normative mean and standard deviation (Crawford & Howell, 1998).

Based on the foregoing, Crawford and Howell (1998) proposed a significance test, in the form of a modified *t*-test, for comparison of an individual to a normative sample. That is, their method can be used to test whether an individual's score is sufficiently low such that the null hypothesis that it is an observation from the normative distribution can be rejected. However, the one-tailed *p* value for this test is also a point estimate of the abnormality of a patient's score. Thus, if the *p* value was say 0.030 then it is estimated that (*p*100=) 3% of the normative population would obtain a lower score; that is, the score is estimated to be at the 3rd percentile. A brief mathematical proof of this statement can be found in Crawford and Garthwaite (2006). In addition, as the original formula presented by Crawford and Howell (1998) was couched in terms of a significance test, we set out a variant on the formula below that is explicitly framed in terms of obtaining a point estimate of the percentile rank corresponding to a given raw score:

$$\text{Estimated percentile rank for } x = 100 \times \Pr\left(t > \frac{x-\overline{x}}{\sqrt{s_x^2\left(\frac{n+1}{n}\right)}} \right), \qquad (8)$$

where x is the raw score, $\overline{x}$ is the normative sample mean, s_x is the normative sample standard deviation, n the size of the normative sample, and the probability, Pr( ) is the one-tailed probability for t on $n-1$ df .

This method of estimating the percentile rank of a given raw score is technically correct regardless of the size of the normative sample because all normative data in neuropsychology are based on samples rather than on the whole of the relevant population (that is, we work with normative sample statistics rather than population parameters). However, when the normative data are obtained from large samples, such as is the case with Wechsler tests for example, it is acceptable to treat the normative sample statistics as population parameters because the former will give very accurate estimates of the latter. Thus using *z* to estimate the percentile rank of a given raw score will give results that are, for all practical intents and purposes, indistinguishable from the technically correct method based on *t*-distributions (a *t*-distribution on large df approaches *z*).

For example, suppose a normative sample mean and standard deviation was 100 and 15 respectively, that the normative sample consisted of 300 persons and we want to estimate the percentile rank for a score of 80. Using Crawford and Howell’s test to compare this score to the normative sample yields a *t* of -1.331 on 299 df; the one-tailed probability for this *t* is 0.0921 and thus the score is estimated to be at the 9.21th percentile. Expressing the score as a *z* score (-1.333) and referring to a table of the areas under the normal curve the score is estimated to be at the 9.12th percentile, this is only trivially different from the optimal estimate.

The problem is that, as Bridges and Holler have illustrated, the size of samples used to establish norms in neuropsychology are often fairly modest in size. (Note that, because norms are normally stratified by age and determined separately for each age group, it is the sample size used for each age band that is relevant here, not the overall size of the normative sample). For example, the mean, median and modal sample size for the 54 normative samples considered by Holler and Bridges were 27.8, 23, and 20 respectively.

We suggest that with normative samples of this size it is not justifiable to use *z* as a means of obtaining a point estimate of the percentile rank of a score since, as noted, this involves treating the sample statistics as population parameters. Instead Crawford and Howell's method should be employed. To illustrate the difference between the two point estimates we again return to the hypothetical case of Johnny employed by Bridges and Holler. Johnny's raw score on the Hooper Visual Organization test was 18; this raw score converts to a *z* score of -1.655 based on the normative sample mean and standard deviation of 23.23 and 3.16. As noted by Holler and Bridges, it is thus estimated that this score is just below the 5th percentile (4.98th). Using Crawford and Howell's method the score is estimated to be at the 6.14th percentile; it can be seen that the use of *z* to estimate the percentile has overestimated the abnormality of the score.

An empirical comparison of confidence limits on *z* derived from non-central *t*-distributions with Holler & Bridges (2007) method using Monte Carlo simulation

Having discussed the issue of obtaining a point estimate of the percentile rank of a given score we now return to the main focus of the present paper: that of obtaining interval estimates to quantify the uncertainty associated with the use of normative data. In the foregoing sections a series of specific examples and thought experiments were used to illustrate the issues surrounding the quantification of these uncertainties. In the present section Monte Carlo simulations are used to provide a more detailed and systematic examination of the problem.

These simulations are useful for a number of reasons. First, in comparing the two sets of limits, we have been referring to the limits obtained using Crawford and Garthwaite’s method as the “correct” limits. It would be useful to verify the validity of these methods empirically for readers who have little interest in, or background knowledge of, the underlying statistical theory set out in Appendix 2. Second, using Crawford and Garthwaite’s (2002) methods, the estimates of the degree of uncertainty associated with the use of normative data are very substantial unless the normative sample is very large. The degree of uncertainty may strike many neuropsychologists as surprising, perhaps even unbelievable and so, again, an empirical demonstration may be more compelling than the formal derivation.

Third, we have noted that the Bridges and Holler (2007) method yields quantities with no obvious distribution (except for the special case in which the score is at the normative mean). The simulations will quantify the degree to which the limits used by Bridges and Holler fail to capture the true degree of uncertainty. Fourth, it has been noted that the degree of uncertainty associated with use of normative data will vary as a function of the extremity of the score. Crawford and Garthwaite’s method allows for this variability and therefore their confidence limits should capture the true *z* score (hereafter designated as z^*) 95% of the time, regardless of the extremity of the score. In contrast, Bridges and Holler’s method does not account for this variability and the effects of this can be quantified in the simulation: that is, the number of times Bridges and Holler’s limits capture z^* should decrease as the extremity of z^* increases.

In the first of two simulations we repeatedly drew *n* observations from a standard normal distribution (mean = 0; standard deviation =1): the standard normal distribution represents the population and the *n* observations represent the test scores of a normative sample of size *n*. On each trial we also drew an additional observation from the same distribution: this represents an additional test score, z^*; it is this value that the confidence limits calculated using the normative sample statistics attempt to capture. On each trial the sample mean was subtracted from z^* and the result divided by the sample standard deviation. This created a *z based on the sample statistics* on that trial.

The methods set out in Appendix 2 were then applied to calculate a 95% confidence interval on z^*. Bridges and Holler's (2007) method of setting confidence intervals was also applied. That is, on each Monte Carlo trial, the 95% confidence interval on the mean was computed using the normative sample statistics and the endpoints of this interval were entered as the mean in the formula for *z* (together with z^* and the sample standard deviation for that trial), thereby obtaining a (supposed) 95% confidence interval on z^*. The number of trials in which these two sets of confidence intervals captured z^* was recorded. If the methods are valid, then the percentage of trials on which this occurred should be 95%, save for Monte Carlo variation. Eight different normative sample *n*s were used ranging from 5 through 10, 25, 50, 200, and 300 to 500. One hundred thousand trials were run for each of these sample sizes.

The results of the simulation are set out in Table 2. Considering Crawford and Garthwaite's limits first, it can be seen that, in line with the theory set out in Appendix 2, the confidence limits captured z^* (the true *z* score) on 95% of trials, regardless of the size of the normative sample (the small deviations from 95% are of the order expected solely from Monte Carlo variation).

In contrast, it can be seen that using Bridges and Holler's (2007) method, the percentage of trials in which the limits captured z^* was below 95% regardless of the size of the normative sample: their limits captured z^* on only 89 to 90% of trials. That is Bridges and Holler's method consistently underestimated the uncertainty arising from using normative sample data to estimate the true standing of a score (the limits are too narrow). This first simulation quantifies the *overall* performance of the two methods of capturing the uncertainty in estimating the true standing of a score using normative sample data.

In the second simulation we quantified the performance of the two methods as a function of the extremity of a score. Seven population values of *z* (i.e., z^*) were selected, ranging from 0 (representing a score exactly at the mean of the population) through -0.253 (representing a score only slightly below the population mean, i.e. 40% of the control population would obtain a lower score) through z^*s of -0.675 (25%), -1.036 (15%), -1.282 (10%), and -1.960 (2.5%), to a z^* of -2.326 (representing a very low score; only 1% of the control population would obtain a lower score). We examine only negative *z* scores because (a) the results for positive *z* scores would simply be a mirror image of the results for negative scores, and (b) neuropsychologists will typically be more concerned with evaluating scores that potentially indicate cognitive problems. For each of these z^*, one hundred thousand Monte Carlo trials were run in which a normative sample of size *n* was drawn from the population distribution (a standard normal distribution). The eight sample sizes were the same as those used in the first simulations.

Thereafter the procedure was the same as that used in the first simulation; that is, on each trial a *z* was calculated based on the normative sample statistics and the two different methods of setting confidence intervals were applied. As in the first simulation, the number of trials in which these two sets of confidence intervals captured z^* was recorded. If the methods are valid, then the percentage of trials on which this occurred should be 95%, save for Monte Carlo variation.

The results of the second simulation are set out in Table 3. Considering Crawford and Garthwaite's limits first, it can be seen that, in line with the theory set out in Appendix 2, the confidence limits captured z^* (the population *z* score) on 95% of trials, regardless of the size of the normative sample and the extremity of the score (the small deviations from 95% are of the order expected solely from Monte Carlo variation).

In contrast, it can be seen that using Bridges and Holler's (2007) method, the percentage of trials in which the limits captured z^* (the true *z* score) was below 95% in all scenarios studied, except where the score was exactly at the population mean, i.e., when z^* was zero. It can be seen that the performances of the two methods are identical in this latter scenario. This is because both methods give confidence intervals that include 0 (the true value of z^*) if and only if $|\overline{x}|$ exceeds $t_{0.025}s/\sqrt{n}$. When $|\overline{x}|$ is less than $t_{0.025}s/\sqrt{n}$ the confidence interval of Bridges and Holler will be wider than that of Crawford and Garthwaite, while when $|\overline{x}|$ exceeds $t_{0.025}s/\sqrt{n}$ it will be narrower.

It can be seen that the performance of Bridges and Holler limits worsen appreciably as z^* becomes more extreme. For example, with a sample size of 50, the limits captured z^* on only 89% of trials for a score at the 15th percentile (i.e., $z^* = -1.036$) and this figure falls to 80.1% for a score at the 5th percentile ($z^* = -1.645$). This feature is clearly illustrated in Figure 1 which plots the percentage of trials on which the two methods captured z^* as a function of the extremity of the score (in this latter figure the size of the normative sample is held constant at an intermediate value of 50). It will be appreciated that the worsening performance of the Bridges and Holler's limits as scores become more extreme is problematic as neuropsychologists routinely work with test scores that are well below a normative sample mean.

In summary, the simulation results confirm that Crawford and Garthwaite's method can be used to quantify the uncertainty associated with the use of normative samples to establish the population *z* score (or equivalently, the percentile rank in the population) corresponding to a given raw score. The simulation results also show that, in contrast, the work reported in Bridges and Holler has the potential to mislead neuropsychologists as to the degree of uncertainty associated with the use of normative data: the uncertainty will be greater than they acknowledge except when scores are close to the normative population mean.

Insert Tables 2 and 3 about here

Computer program for quantifying the fallibility of normative data

A computer program for PCs (written in the Delphi programming language) is available that implements the methods covered in the present paper. The program, QUAND.exe (**Q**uantifying the **U**ncertainty **A**ttached to **N**ormative **D**ata), has two options. The first allows neuropsychologists to evaluate *existing* normative data. The user enters the size of the normative sample, the normative mean and standard deviation for the test, and the minimum and maximum obtainable raw scores. Working through the range of raw scores (from minimum to maximum and applying the assumption that raw scores are integers), the program converts each score to *z* and calculates the 95% confidence limits on *z*. It also lists the absolute width of the confidence interval (thus, if the lower limit on *z* is –1.30 and the upper limit is +0.70, then the absolute width of the interval is 2.0; e.g., the interval spans two standard deviations). The program also lists two point estimates of the percentile for each raw score. The first of these is obtained by finding the probability corresponding to *z* and multiplying by 100. Thus, if the *z* for a raw score is –1.645, then the score is estimated to be at the 5th percentile. As noted in a previous section, this point estimate is not the optimal estimate (although widely used). The other point estimate of the percentile rank is the optimal estimate and is obtained by application of Crawford and Howell's (1998) method (see formula 5). Finally, 95% confidence limits on the percentile are reported.

The results reported for this option allow neuropsychologists to quantify the uncertainty associated with normative data they either use already or are considering for use. Although this program option quantifies the uncertainty for a *specific* set of normative data, we believe that its use may also serve a more general purpose of countering any tendency to reify normative data. Furthermore, a demonstration of the alarming degree of uncertainty associated with normative based on modest samples may also serve as a tipping point for those who have been considering abandoning a favorite test for a similar test with more adequate norms.

The second option is aimed at those planning to run a normative study. This option allows the user to examine the uncertainty attached to *z* and the percentile rank as a function of sample size. The only input required is the sample size for the proposed normative sample. The program computes confidence limits (and the absolute width of the confidence interval) on *z* and confidence limits on the percentiles for values of *z* ranging from –3.0 through to +3.0. This allows investigators to quantify the likely uncertainty associated with scores that vary in terms of their deviation (in standard deviation units) from a normative sample mean. The program can be run with a range of potential sample *n*s so that the neuropsychologist can select a sample size that strikes the desired balance between precision and practical constraints. If the practical constraints (time, money) are severe, such that the maximum achievable sample *n* is relatively modest, the lack of precision may be judged to be unacceptable. In such circumstances the appropriate decision would either be not to run the study or, more positively, to seek collaborators or further funding so that a larger sample can be obtained.

Although the calculations required to generate the confidence limits are computationally intensive, the output from both of the options described above is available in a matter of seconds. The program output can be viewed on screen, printed, and saved to a file. There is also the option of adding User Notes (e.g., to keep a record of the source of the normative data); these notes are reproduced in the output from the program. A compiled version of the program can be downloaded (as a zip file) from the following website address: www.abdn.ac.uk/~psy086/dept/QUAND.htm.

The optimal size for normative samples in neuropsychology

Based on the width of confidence limits on normative sample means, Bridges and Holler's analysis led them to produce guidelines on the optimal sample sizes for normative studies in neuropsychology. They suggested that the uncertainty associated with the standing of a given raw score would be unacceptably large for norms based on samples of 50 or less and that samples of between 50 and 70 represented the optimal size for normative samples in neuropsychology. They argued that recruiting samples larger than 70 would be subject to diminishing returns, i.e., the increase in precision over the true standing of a given raw score would be modest. In contrast, our view is that the “one size fits all” approach is inappropriate. On the one hand, useful work has been done with far smaller sample sizes than 50 when the raw score of a case has been quite unusual. On the other hand, when frequent use is to be made of the information gained from a normative sample, we believe that every effort should be expended to make the sample just as large as practical constraints allow. With sample sizes above 70 there will still be considerable uncertainty over the standing of a raw score, particularly for scores that are not close to the mean. For example, from Tables 1b and 1c it can be seen that, with a normative sample size of 100, the confidence interval on *z* for a score one standard deviation below a normative sample mean essentially spans half a standard deviation (0.481) and that, for a score one and a half standard deviations below the mean, the interval spans well over half a standard deviation (0.57). With knowledge of this degree of uncertainty, we would imagine that many neuropsychologists would be uncomfortable using such normative data.

Confidence limits quantifying the uncertainty associated with normative samples versus limits based on measurement error

The present paper is solely concerned with quantifying the uncertainty in evaluating the standing of a given score as a function of the use of normative sample data to estimate population parameters. The confidence limits presented should therefore not be confused with confidence limits based on classical test theory that attempt to quantify the effects of measurement error in a neuropsychological instrument on scores; these latter intervals are of course widely used and featured in most test manuals.

The latter limits do not allow for error in estimating the normative population parameters from sample data. It is worth noting that these confidence limits also treat the reliability of the test as a known, fixed quantity when, in reality, it is subject to error like any other statistic. The uncertainty attached to Cronbach's alpha can be considerable when the sample used to calculate it is modest in size, particularly when the test contains relatively few items (Crawford, 2004; Feldt, 1965).

Thus, when using these latter limits, the neuropsychologist is posing the question "ignoring the error in estimating the population mean, standard deviation and reliability of the test, how much uncertainty is there in an individual's score as a function of measurement error in the instrument?" In contrast, when using the limits presented in the present paper, the concern is solely with the score *in hand*. The more concrete question posed is "for a given (i.e., obtained or potentially obtainable) raw score, how much uncertainty is there in its standing as a function of error in using a normative sample to estimate the parameters of a normative population. To avoid potential misunderstanding, it is not the case that these limits assume no measurement error. The limits are concerned with *observed* scores (which are effected by an amalgam of true score variance and measurement error variance) and the method does not need to "know" the relative contributions of these two sources of variance: the intervals automatically become wider by exactly the right amount to compensate for the presence of measurement error and thus the intervals are exactly correct. Therefore the method allows for the presence of measurement error in the normative data and in individual's scores referred to these norms but is solely concerned with the score in hand. It does not address the issue of what score an individual might obtain on another occasion, but simply quantifies (with confidence limits) the proportion of the normative population who would obtain a score as extreme as the individual obtained.

The two types of limits therefore provide useful but different information. The limits quantifying the effects of measurement error on an individual's score are not relevant when evaluating the adequacy of normative data, although the reliability coefficient used in their calculation is clearly highly relevant in evaluating the neuropsychological instrument itself. If an instrument has low reliability, then it should be avoided; therefore the issue of the adequacy of existing normative data for the test should not arise, nor should there be much interest in planning a normative study.

A caveat and some other considerations

In neuropsychology it is still common for norms to be presented as raw means and standard deviations (for example, all of the normative data sets considered by Bridges and Holler were of this form). When the standing of a raw score is estimated by expressing it as a *z* score, or by using Crawford and Howell's method (1998), it is assumed that the normative data are drawn from an underlying normal distribution. If this assumption is violated the resultant point estimates will be biased. For example, if many members of the normative population score at or close to ceiling on a test, the distribution of raw scores will have high negative skew (and will also be leptokurtic, i.e., more peaked than a normal distribution). The effect will be that *z* scores (and Crawford and Howell's *t* values) will underestimate the rarity of scores above the mean and overestimate the rarity of scores below the mean (Crawford & Garthwaite, 2005; Crawford, Garthwaite, Azzalini, Howell, & Laws, 2006).

The interval estimates (i.e., confidence limits) on the standing of a score presented here are also based on the assumption that scores are drawn from a normal distribution (as are Bridges and Holler's limits although, as noted, they make further unjustified assumptions). It follows that, if this assumption is violated, the confidence limits on the standing of raw scores will not capture the true standing of the scores 95% of the time. Thus, although the limits on the standing of a score are wide (particularly when normative samples are modest and the scores under consideration are extreme), these limits can be seen as representing a best case scenario.

The effect of departures from normality on the ability of these limits to capture the true standing of a score will be greater when normative samples are large. When skew is present the intervals will be misaligned (i.e., centered around the wrong quantity) and this will have a more pronounced effect as the calculated limits become narrower with increasing sample size. On the other hand, when normative samples are large, it is more likely that the scores will be expressed on some form of standard metric (i.e., as *T* scores or IQ scores etc). When this is the case it should be safe to assume that the distribution of raw scores will have been examined and a normalizing transformation applied if required.

Having raised the topic of norms expressed on a standard metric it is worth making explicit that the methods presented here are just as applicable to normative data that is in such a form; i.e. they are *not* limited to normative data in the form of raw means and standard deviations. That is, although norms expressed on standard metrics tend to be based on larger samples than norms expressed as raw means and standard deviations, there are still many normative data sets of the former type that are based on relatively modest samples. Moreover, even with normative samples that would be considered large, it can be seen from Tables 1a to 1c there is still appreciable uncertainty over the population standing of a score, particularly for extreme scores.

Finally, it should be stressed that the size of a normative sample is only one of many factors that should be considered when evaluating or using normative data. For example, potential consumers of normative data need to consider how closely matched the normative sample is in terms of demographic variables to the population they are intended to represent. Moreover, one should be confident that the exclusion criteria used were rigorous enough to exclude potential participants who are cognitively impaired. For more detailed discussion of these and other related issues see Cicchetti (1994), Mitrushina, Boone, Razani and D'Elia, (2005), or Strauss, Sherman and Spreen (2006).

References

Bishop, J., Knights, R. M., & Stoddart, C. (1990). Rey Auditory-Verbal Learning Test: Performance of English and French children aged 5 to 16. *Clinical Neuropsychologist, 4,* 133-140.

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative sudies in pediatric neuropsychology. *Child Neuropsychology*.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284-290.

Crawford, J. R. (2004). Psychometric foundations of neuropsychological assessment. In L. H. Goldstein & J. E. McNeil (Eds.), *Clinical neuropsychology: A practical guide to assessment and management for clinicians* (pp. 121-140). Chichester: Wiley.

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. (2005). Testing for suspected impairments and dissociations in single-case studies in neuropsychology: Evaluation of alternatives using Monte Carlo simulations and revised tests for dissociations. *Neuropsychology, 19,* 318-331.

Crawford, J. R., & Garthwaite, P. H. (2006). Methods of testing for a deficit in single case studies: Evaluation of statistical power by Monte Carlo simulation. *Cognitive Neuropsychology, 23,* 877-904.

Crawford, J. R., Garthwaite, P. H., Azzalini, A., Howell, D. C., & Laws, K. R. (2006).

Testing for a deficit in single case studies: Effects of departures from normality. *Neuropsychologia, 44*, 666-676.

Crawford, J. R., & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist, 12*, 482-486.

Crawford, J. R., Lynch, C., & Garthwaite, P. H. (submitted). The single-case method in cognitive neuroscience: A survey and evaluation of contemporary statistical practice and some recommendations for reform (part 1).

Demsky, Y., Carone, D. A., Burns, W. J., & Sellers, A. (2000). Assessment of visual-motor coordination in 6- to 11-year-olds. *Perceptual and Motor Skills, 91*, 311-321.

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

Kirk, U. (1992a). Confrontation naming in normally developing children: Word retrieval or word knowledge? *Clinical Neuropsychologist, 6*, 156-170.

Kirk, U. (1992b). Evidence for early acquisition of visual organization ability: A developmental study. *Clinical Neuropsychologist, 6*, 171-177.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms and commentary* (3rd ed.). New York: Oxford University Press.

Acknowledgements

We are grateful to Dr Sytse Knypstra of the Department of Econometrics, University of Groningen, The Netherlands, for providing an algorithm that finds the non-centrality parameter of a non-central *t*-distribution given a quantile, its associated probability, and the degrees of freedom. This algorithm is incorporated in the computer program (QUAND.exe.) that accompanies this paper. The work was done while one author (PHG) was a visiting academic at the University of New South Wales.

Table 1a. Confidence limits on *z* and on its corresponding percentile rank calculated using the methods of Bridges and Holler (2007) and Crawford and Garthwaite (2002): results for raw scores at the normative sample mean (i.e., *z* = 0, percentile = 50th)

	Confidence limits on *z*						Confidence limits on percentile			
	Lower limit		Upper limit		Width of interval		Lower limit		Upper limit	
	BH	CG	BH	CG	BH	CG	BH	CG	BH	CG
5	-1.242	-0.877	1.242	0.877	2.483	1.753	10.72	19.04	89.28	80.96
10	-0.715	-0.620	0.715	0.620	1.431	1.240	23.72	26.77	76.28	73.23
25	-0.413	-0.392	0.413	0.392	0.826	0.784	33.99	34.75	66.01	65.25
50	-0.284	-0.277	0.284	0.277	0.568	0.554	38.81	39.08	61.19	60.92
100	-0.198	-0.196	0.198	0.196	0.397	0.392	42.14	42.23	57.86	57.77
200	-0.139	-0.139	0.139	0.139	0.279	0.278	44.46	44.49	55.55	55.51
300	-0.114	-0.113	0.114	0.113	0.227	0.226	45.48	45.50	54.52	54.51
500	-0.088	-0.088	0.088	0.088	0.176	0.175	46.50	46.51	53.50	53.49

Table 1b. Confidence limits on *z* and on its corresponding percentile rank calculated using the methods of Bridges and Holler (2007) and Crawford and Garthwaite (2002): results for raw scores one standard deviation below the normative sample mean (i.e., *z* = -1, percentile = 15.8th)

	Confidence limits on *z*						Confidence limits on percentile			
	Lower limit		Upper limit		Width of interval		Lower limit		Upper limit	
	BH	CG	BH	CG	BH	CG	BH	CG	BH	CG
5	-2.242	-2.067	0.242	0.138	2.483	2.204	1.25	1.94	59.55	55.48
10	-1.715	-1.751	-0.285	-0.214	1.431	1.538	4.31	4.00	38.80	41.54
25	-1.413	-1.476	-0.587	-0.511	0.826	0.965	7.89	7.00	27.85	30.48
50	-1.284	-1.337	-0.716	-0.656	0.568	0.681	9.95	9.06	23.71	25.58
100	-1.198	-1.239	-0.802	-0.758	0.397	0.481	11.54	10.77	21.14	22.42
200	-1.139	-1.169	-0.861	-0.829	0.279	0.340	12.73	12.12	19.47	20.35
300	-1.114	-1.138	-0.886	-0.861	0.227	0.277	13.27	12.75	18.77	19.47
500	-1.088	-1.107	-0.912	-0.892	0.176	0.215	13.83	13.41	18.09	18.61

Table 1c. Confidence limits on *z* and on its corresponding percentile rank calculated using the methods of Bridges and Holler (2007) and Crawford and Garthwaite (2002): results for raw scores 1.5 SDs below the normative sample mean (i.e., *z* = -1.5, percentile = 6.7th)

	Confidence limits on *z*						Confidence limits on percentile			
	Lower limit		Upper limit		Width of interval		Lower limit		Upper limit	
	BH	CG	BH	CG	BH	CG	BH	CG	BH	CG
5	-2.48	-2.80	-0.26	-0.14	2.48	2.66	0.31	0.26	39.81	44.55
10	-2.22	-2.40	-0.78	-0.56	1.43	1.84	1.34	0.81	21.63	28.76
25	-1.91	-2.07	-1.09	-0.92	0.83	1.15	2.79	1.93	13.85	17.97
50	-1.78	-1.90	-1.22	-1.09	0.57	0.81	3.72	2.86	11.20	13.78
100	-1.70	-1.78	-1.30	-1.21	0.40	0.57	4.47	3.72	9.65	11.28
200	-1.64	-1.70	-1.36	-1.30	0.28	0.40	5.06	4.44	8.68	9.73
300	-1.61	-1.66	-1.39	-1.33	0.23	0.33	5.33	4.80	8.28	9.11
500	-1.59	-1.63	-1.41	-1.37	0.18	0.26	5.62	5.18	7.90	8.51

Table 2. Results of a Monte Carlo simulation comparing two methods of capturing the uncertainty associated with the use of normative data to estimate z^* (the *z* that would be obtained for a given raw score had we access to the normative population mean and standard deviation); the size of the normative sample is varied from 5 to 500 (if the methods are sound the intervals should capture z^* on 95% of Monte Carlo trials)

Sample *n*	Holler & Bridges (%)	Crawford & Garthwaite (%)
5	90.47	95.14
10	90.06	94.99
25	89.76	95.00
50	89.65	94.95
100	89.44	94.96
200	89.83	95.09
300	89.67	95.12
500	89.73	95.09

Table 3. Results of a Monte Carlo simulation comparing two methods of capturing the uncertainty associated with the use of normative data to estimate z^* (the z that would be obtained for a given raw score had we access to the normative population mean and standard deviation) *as a function of the extremity of* z^*; the size of the normative sample is varied from 10 to 500 (if the methods are sound the intervals should capture z^* on 95% of Monte Carlo trials)

	Bridges & Holler (%)						Crawford & Garthwaite (%)					
z^* (percentile)	10	25	50	100	200	500	10	25	50	100	200	500
0.0 (50th)	95.02	94.96	95.04	95.01	95.02	95.03	95.02	94.96	95.04	95.01	95.02	95.03
-0.253 (40th)	94.64	94.58	94.68	94.84	94.65	94.79	94.98	94.99	95.07	95.15	94.95	95.00
-0.675 (25th)	92.25	92.40	92.34	92.43	92.33	92.38	95.02	95.00	94.97	95.07	95.00	95.09
-1.036 (15th)	89.04	89.08	88.68	88.68	88.59	88.57	94.95	95.03	94.96	94.99	95.01	94.92
-1.282 (10th)	86.71	85.73	85.87	85.29	85.29	85.34	95.03	95.05	95.04	94.90	94.95	95.02
-1.645 (5th)	81.83	80.77	79.96	80.20	80.13	79.93	94.86	94.91	94.80	95.04	95.01	95.06
-1.960 (2.5th)	77.64	75.90	75.06	74.97	74.87	75.07	94.83	95.08	94.94	94.97	95.01	94.98
-2.326 (1st)	72.50	70.21	69.88	69.56	69.23	69.33	95.07	94.90	94.98	95.12	95.05	95.02

Appendix 1

Algebraic manipulation of Bridges and Holler’s equations for confidence intervals on *z*

As set out in equation (3), Bridges and Holler’s method for obtaining the 95% two-sided upper limit on *z* for a raw score referred to normative data is

$$\text{Upper 95\% confidence limit on 'z'} = \left(x - \overline{x}_L\right)/s$$

However, from equation (1) we have

$$\overline{x}_L = \overline{x} - t_{0.025}\left(\frac{s}{\sqrt{n}}\right)$$

and so the right hand side of equation (3) can be expressed as

$$\frac{x - \left(\overline{x} - t_{0.025}\dfrac{s}{\sqrt{n}}\right)}{s},$$

and this can be simplified to

$$\left(\frac{x - \overline{x}}{s}\right) + \left(\frac{t_{0.025}}{\sqrt{n}}\right),$$

which is reproduced in the text as equation (5). The same procedure yields the equation for the lower limit on *z* (equation 4).

Appendix 2

Derivation of confidence intervals on *z* and percentiles

The confidence intervals used in the present paper to quantify the uncertainty associated with the use of normative samples to estimate the *z* and percentile rank corresponding to a given raw score are derived from a non-central *t*-distribution. They are based on theory developed by Crawford and Garthwaite (2002). At some points we change the terminology from that used by Crawford and Garthwaite (2002) to render it consistent with the terminology and intended uses of these limits as set out in the present paper. The non-central *t*-distribution is defined by

$$T_\nu(\delta) = (Z + \delta) / \sqrt{Y/\nu} \,,$$

where *Z* has a normal distribution with a mean of zero and variance 1, and *Y* is independent of *Z* with a chi-square distribution on ν degrees of freedom. δ is referred to as the non-centrality parameter.

For a specified value x_0 (i.e., a given raw test score), where $x \sim \mathrm{N}(\mu, \sigma^2)$, we require $100(1-\alpha)\%$ confidence intervals on z^* and the percentile rank (P^*), based on normative sample data $\overline{x}$ and s^2, where $\overline{x} \sim \mathrm{N}(\mu, \sigma^2 / n)$ and $\nu s^2 / \sigma^2 \sim \chi^2(\nu)$. (In the present case $\nu = n - 1$). Crawford and Garthwaite (2002) originally set out the method for P^* only but, as pointed out by Crawford, Lynch and Garthwaite (submitted), it is very straightforward to also obtain limits on z^* en route to obtaining these former limits. Put

$$z = \frac{(x_0 - \overline{x})}{s}, \tag{9}$$

Let $z^* = (x_0 - \mu)/\sigma$. Then *z* is an estimate of z^*. That is, *z* computed using the *sample* mean and standard deviation, is an estimate of the *z*, here denoted z^*, that we would obtain were the mean and standard deviation of the normative *population* known. Now

$$z\sqrt{n} = \left(\frac{(\mu - \overline{x})\sqrt{n}}{\sigma} + \frac{(x_0 - \mu)\sqrt{n}}{\sigma} \right) / \sqrt{\frac{s^2}{\sigma^2}} \,,$$

so $z\sqrt{n}$ has a non-central *t*-distribution with non-centrality parameter $\delta = z^* \sqrt{n}$ and ν degrees of freedom. The $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ points of this distribution will depend on the value of δ. Let δ_L denote the value of δ for which the $100(1-\alpha/2)\%$ point is $z\sqrt{n}$. Similarly, let δ_L denote the value of δ for which the $100(\alpha/2)\%$ point is $z\sqrt{n}$. Then $\left(\delta_L/\sqrt{n}, \delta_U/\sqrt{n}\right)$ is a $100(1-\alpha)\%$ confidence interval for z^*.

To obtain confidence limits on P^*, the percentile rank of a given raw score, define $h(\alpha/2;\ z;\ \nu+1)$ and $h(1-\alpha/2;\ z;\ \nu+1)$ by

$$h(\alpha/2;\ z;\ \nu+1) = \Pr(Z < \delta_L/\sqrt{N})\cdot 100$$

and

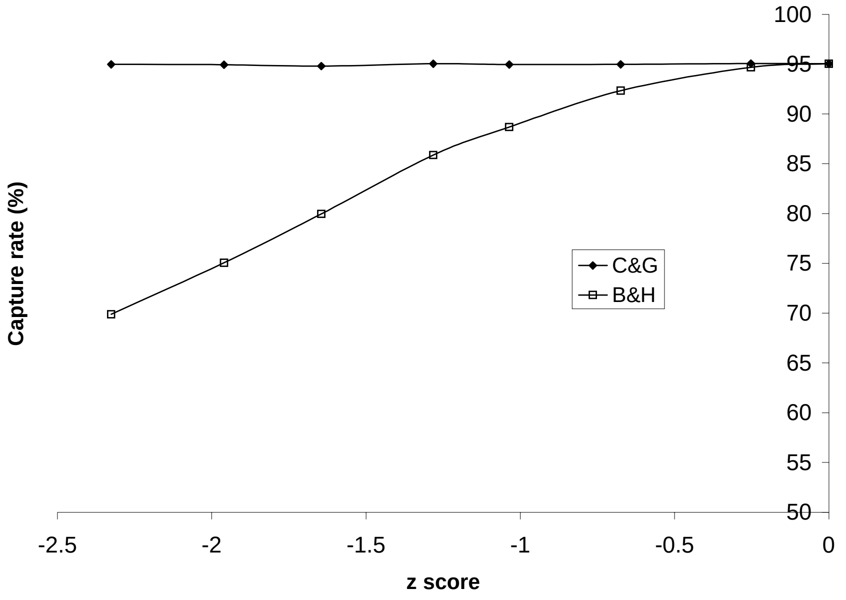
$$h(1-\alpha/2;\ z;\ \nu+1) = \Pr(Z < \delta_U/\sqrt{N})\cdot 100$$

where Z is the standard normal variate [*i.e.* $Z \sim \mathrm{N}(0,1)$]. Then a 100(1-α)% confidence interval for the percentile rank (P^*) is

$(h(\alpha/2;\ z;\ \nu+1),\ h(1-\alpha/2;\ z;\ \nu+1))$.

Figure 1.

Percentage of trials in which the two methods of setting confidence limits captured the population *z* score as a function of the extremity of the population *z* score; results are for a (intermediate) normative sample size of 50; C&G = Crawford & Garthwaite (2002) method, B&H = Bridges & Holler (2007).


100
95
90
85
80
75
70
65
60
55
50
Capture rate (%)
-2.5
-2
-1.5
-1
-0.5
0
z score
C&G
B&H