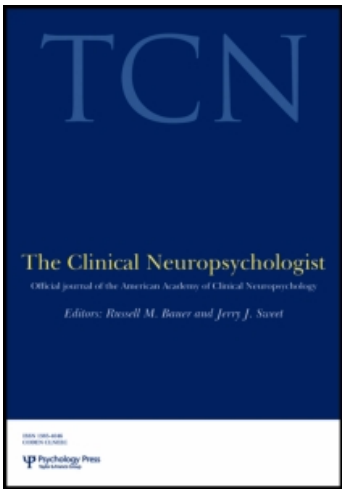

The Clinical Neuropsychologist
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713721659

Percentiles Please: The Case for Expressing Neuropsychological Test Scores and Accompanying Confidence Limits as Percentile Ranks

John R. Crawford [a]; Paul H. Garthwaite [b]
[a] School of Psychology, University of Aberdeen, Milton Keynes, UK [b] The Open University, Milton Keynes, UK

First Published:February2009

To cite this Article Crawford, John R. and Garthwaite, Paul H.(2009)'Percentiles Please: The Case for Expressing Neuropsychological Test Scores and Accompanying Confidence Limits as Percentile Ranks',The Clinical Neuropsychologist,23:2,193 — 204
To link to this Article: DOI: 10.1080/13854040801968450
URL: http://dx.doi.org/10.1080/13854040801968450

The Clinical Neuropsychologist, 23: 193–204, 2009
http://www.psypress.com/tcn
ISSN: 1385-4046 print/1744-4144 online
DOI: 10.1080/13854040801968450

Psychology Press
Taylor & Francis Group

CE PERCENTILES PLEASE: THE CASE FOR EXPRESSING NEUROPSYCHOLOGICAL TEST SCORES AND ACCOMPANYING CONFIDENCE LIMITS AS PERCENTILE RANKS

John R. Crawford[1] and Paul H. Garthwaite[2]
[1]*School of Psychology, University of Aberdeen*
and [2]*The Open University, Milton Keynes, UK*

Many commentators on neuropsychological assessment stress the disadvantages of expressing test scores in the form of percentile ranks. As a result, there is a danger of losing sight of the fundamentals: percentile ranks express scores in a form that is of greater relevance to the neuropsychologist than any alternative metric because they tell us directly how common or uncommon such scores are in the normative population. We advocate that, in addition to expressing scores on a standard metric, neuropsychologists should also routinely record the percentile rank of all test scores so that the latter are available when attempting to reach a formulation. In addition, it is argued that the current practice of expressing confidence limits on test scores on a standard score metric should be supplemented with confidence limits expressed as percentile ranks, because the latter provide a more direct and tangible indication of the uncertainty surrounding an observed score. Computer programs accompany this paper and can be used to obtain percentile rank confidence limits for Index scores (and FSIQs) on the WAIS-III or WISC-IV (these can be downloaded from the following web page: http://www.abdn.ac.uk/~psy086/dept/PRCLME.htm).

Keywords: Percentiles; Percentile ranks; Confidence limits; Neuropsychological assessment; Psychometrics.

INTRODUCTION

The literature on neuropsychological assessment offers little encouragement to express test scores in the form of their percentile ranks. The position of some commentators is unequivocal: for example, Golden, Espe-Pfeifer, and Wachsler-Felder (2002) state that "percentiles should not be used for the purpose of neuropsychological interpretation" (p. 7). Although less extreme, other treatments of this issue typically recommend that test scores be expressed on a standard metric (i.e., z scores or derived scores such as IQs or T scores) and emphasize the disadvantages of percentiles (Crawford, 2004; Lezak, Howieson,

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

Accepted for publication: February 2, 2008. First published online: May 2, 2008.

Loring, Hannay, & Fischer, 2004; Mitrushina, Boone, Razani, & D'Elia, 2005; Strauss, Sherman, & Spreen, 2006).

These disadvantages stem from the fact that, assuming a normal distribution for raw scores, the distance between raw scores is not maintained following conversion to percentile ranks. As a result, many arithmetical and statistical operations on test scores cannot be performed when scores are expressed as percentile ranks. For example, when interpreting scores, neuropsychologists often find it useful to obtain an individual's averaged performance across different measures of a construct. It would be inappropriate to average the percentile ranks of the individual's scores on these measures to obtain of their average standing on the construct (such averaging should be conducted using standard scores). It is also commonly pointed out that examination of the simple difference between the percentile ranks of an individual's scores on two tasks is not helpful in arriving at a formulation because its clinical implications will depend on the absolute value of the two percentile ranks used to create it (Golden et al., 2002; Mitrushina et al., 2005; Strauss et al., 2006).

Most commentators do acknowledge that percentiles, because of their simplicity, may have a useful role in communicating the results of a neuropsychological assessment to other professionals, clients, or carers (Lezak et al., 2004). The general impression thus created is that, given their psychometric expertise, neuropsychologists should not overly concern themselves with percentiles when interpreting test results, but should reserve their use for communication with the psychometrically unsophisticated. Some commentators are vehemently opposed even to this latter role. Bowman (2002) describes percentiles as treacherous, and suggests that their use will do more harm than good because non-psychologists will commonly misunderstand them.

In marked contrast to the positions of Golden et al. (2002) and Bowman (2002), the present paper argues that percentiles should play a central role in the interpretation of neuropsychological test scores. In short we suggest that, because of the various negative commentaries on percentiles, there is a danger of losing sight of the fundamentals: Percentile ranks express test scores in a form that is of greater relevance to the neuropsychologist than *any* alternative metric because they tell us directly how common or uncommon such scores are in the normative population.

With the exception of criterion-referenced tests (which are rarely used in neuropsychology), a raw test score has little, if any, intrinsic meaning. Assuming a normal distribution of raw scores, conversion of the raw score to a standard metric (e.g., z score, or derived score, such as an IQ or T score) moves things along, but only because this standard metric allows us to estimate how common or uncommon the score is in the general population. A score that is one standard deviation below the mean is imbued with meaning solely because we know that only around 16% of the healthy normative population are expected to obtain a lower score; that is, we know it is at the 16th percentile.

Thus we suggest that, when a neuropsychologist is attempting to interpret an individual's test scores to arrive at a formulation, they should simultaneously have sight of the profile of test scores expressed not only on a standard metric but also as percentile ranks. It might be argued that the recording of the scores' percentile ranks is not necessary in the case of experienced neuropsychologists

either because (a) they have developed a direct, internalized, "look up" table that maps derived scores (e.g., IQs or Index scores) onto percentile ranks, or because (b) they can rapidly and with minimal effort convert a derived score to z and thereby obtain the corresponding percentile rank. We do not believe that either of these possibilities applies to many neuropsychologists.

With regard to (a) for many neuropsychologists such a look up table is liable to be sparsely populated, being limited to various landmark scores (i.e., IQs that correspond to one standard deviation or one and a half standard deviations below or above the mean, IQ scores that correspond to the 10th or 5th percentile etc). With regard to (b) this possibility will hold even less frequently. There is no reason to believe that a look-up table mapping z to percentile ranks will be any more populated than that mapping derived scores to percentile ranks. Moreover, the conversion of a derived score to a z is not without effort, particularly in the case of IQ or Index scores (because of the use of a standard deviation of 15, rather than the more convenient value of 10 in the case of T scores). Take the example of an IQ or Index score of 82: the score is 18 points below the mean and dividing by 15 yields a z score of -1.20 which in turn corresponds to the 12th percentile. Our guess is that few neuropsychologists would have immediate access to this percentile rank from either the Index score or its corresponding z score.

To be clear: we do not suggest that neuropsychologists should refrain from trying to develop an internalized detailed mapping of standard scores to percentile ranks. Indeed, such a mapping would be very beneficial. However, one of the simplest ways to develop such a mapping is through the procedure suggested in this paper: that is, by simultaneous and repeated exposure to standard scores paired with their corresponding percentile ranks.

PERCENTILES: PERFIDIOUS OR VIRTUOUS?

Bowman (2002) adopts a particularly extreme position on the role of percentiles in neuropsychological assessment. She ran a study in which students who had completed a course in psychological assessment were asked to estimate the percentile rank corresponding to various IQ scores and to estimate the IQs corresponding to various percentile ranks. The participants made very large errors when estimating IQs corresponding to the percentile ranks: when presented with percentile ranks below the mean (i.e., for percentile ranks < 50) the estimated IQs were much lower than the correct values. For example, when presented with a percentile rank of 19 (corresponding to an IQ of 87), 86% of participants estimated the IQ to be below the true value. Moreover, in many cases the estimated IQs were substantially lower than the correct value. The opposite pattern was observed for percentile ranks greater than 50: the estimated IQs were higher than the actual IQs.

We regard the results of this study as both striking and important. However, we consider their implications are very different from those arrived at by Bowman. She argues that the results demonstrate the "perfidy of percentiles" (p. 295) and concludes that percentiles should be avoided, even for the purpose of communicating results to non-psychologists. In contrast we see the results as underlining our suggestion that psychologists should, for their *own* interpretative

purposes, simultaneously express scores both on a standard metric and as percentile ranks.

In arguing against the use of percentiles, Bowman (2002) focuses on the errors made when the participants attempted to convert from percentiles to IQs. Crucially, however, very large errors were also made when the participants attempted to convert IQs to percentiles. The percentile rank of IQs below the mean was grossly overestimated and the percentile rank of IQs above the mean tended to be underestimated. For example, given an IQ of 80 (corresponding to the 9th percentile), 84% of participants overestimated the percentile rank. Thus, despite a formal (albeit limited) education in psychometrics, participants had a very distorted notion of how unusual an IQ score was in the normative population. This latter set of results is much more worrying than those given prominence by Bowman, particularly given that most neuropsychologists express scores on standard metrics rather than as percentiles.

It should be stressed that the participants in this study were undergraduate psychology students, not trained neuropsychologists. However, the size of these effects were very large, suggesting that the biases revealed are liable to be persistent and require deliberate and diligent effort if they are to be overcome. Thus there is the danger that some neuropsychologists will also exhibit such biases, albeit in an attenuated form (it would clearly be worth examining this issue empirically by repeating the study with clinical neuropsychologists as participants).

The conclusions drawn from Bowman's (2002) study reinforces our view that neuropsychological test scores should routinely be expressed in terms of their percentile rank as well as on a standard metric. The support provided by the percentile ranks when arriving at a formulation does away with the need to "guesstimate" how common or uncommon a given score is in the population and also deals with the risk that some neuropsychologists may exhibit the systematic biases observed in psychology students.

In pursuing her case against percentiles Bowman presents a further argument: that reporting scores in terms of their percentile ranks may lead to non-psychologists over-pathologizing scores. She gives the example of a score that is at the 25th percentile, and suggests there is a danger that non-psychologists may take this as evidence of "a severe deficit" (2002, p. 299) despite the fact that such a score is "within one standard deviation below the mean" (p. 299). This is a strange inverted logic: the only reason a score that is less than one standard deviation below the mean should not be regarded as pathological is precisely because it is not very unusual. If there is a tendency for non-psychologists (e.g., judges, lawyers, medics) to regard percentiles in this range as pathological, then neuropsychologists' educational efforts should be firmly aimed at correcting this view, rather than making an appeal to the fact that such percentiles are less than one standard deviation below the mean. The non-psychologist's attitude to such an appeal is liable to be "so what?"—and we can only agree.

DISADVANTAGES OF PERCENTILES REVISITED

In this section we return briefly to the most commonly cited disadvantages of expressing scores as percentile ranks. In considering the pros and cons of standard

scores versus percentile ranks, a clear distinction can be made between the question of which metric is best suited for the pre-interpretative stage of the assessment process, in which various computational procedures may be performed on scores, and the question of which is best suited to the interpretative stage, the point at which the neuropsychologist uses the test data in their final form to arrive at a formulation. It is undeniable that standard scores are the appropriate metric for the pre-interpretative stage. However, it does not in any way follow from this that standard scores are therefore the appropriate metric for the interpretative stage. For example, as noted earlier, standard scores should be used if a neuropsychologist wishes to obtain an index of an individual's averaged performance across related tasks. However, having obtained the average expressed as a standard score, the percentile rank of this averaged score is a more direct indication of the individual's cognitive standing.

It was also noted earlier that the simple differences between the percentile ranks of an individual's test scores are not helpful in arriving at a formulation because their clinical implications will depend on the absolute values of the percentile ranks. Given the emphasis in neuropsychological assessment on the analysis of an individual's profile of strengths and weaknesses (Crawford, 2004; Lezak et al., 2004; Strauss et al., 2006) this characteristic of percentiles may appear to place them at a serious disadvantage relative to standard scores. We argue below that this is not the case.

Take the most basic scenario in which a neuropsychologist is concerned with the potential clinical significance of a discrepancy between an individual's performance on two tasks. It is true that the difference in the percentile ranks of the two scores does not provide a sound basis for such judgments. Importantly, however, the difference between standard scores is also insufficient.

Suppose an individual's z scores on two tasks are 0.20 and -0.80 (or equivalently, that the individual's scores are 103 and 88 expressed on an IQ metric). Inferences should not be based on examination of this simple difference because its potential clinical significance will crucially depend on the strength of correlation between the two measures in the general population. A simple difference between z scores of 1.00 (as in the present example) could be very unusual if the two tasks are highly correlated, but will be quite common if the tasks are uncorrelated.

In order to estimate the level of rarity of a difference, the simple difference is divided by the standard deviation of the difference between z scores to obtain a z score for the difference and this in turn is referred to a table of areas under the normal curve in order to estimate the percentage of the population that will exhibit a larger difference (Crawford, Garthwaite, & Gray, 2003; Payne & Jones, 1957).[1] In the present example, if the correlation between the two tasks was 0.90 then it is estimated that only 1.3% of the population would exhibit a larger difference in the same direction as that observed, whereas if the correlation was 0.10 it is estimated

[1] We assume here that the statistics used to calculate this quantity come from very large normative samples so that the sample statistics provide good estimates of the population parameters. When this is not the case things become much more complicated: see Crawford and Garthwaite (2005) and Crawford and Garthwaite (2007) for a fuller discussion of this issue and for methods that are suitable for use with statistics obtained from modest normative samples.

that 22.8% would exhibit a larger difference. This (admittedly extreme) example underlines that examination of simple differences between standard scores is not sufficient. In this scenario the main advantage of standard scores is one of computational convenience: if the standard scores are already available, the percentile ranks of the two scores do not have to be converted to z in order to estimate the rarity of the difference. (Rounding errors may also have some effect if standard scores are converted to percentiles and then back to standard scores.)

The alternative to the foregoing statistical approach is to consult empirical tables of the base rates of differences obtained from a sample of the general population (usually the tests' standardization sample). Again, it is more convenient if the neuropsychologist has already expressed an individual's scores as standard scores, as these are required to use such tables. Note, however, that regardless of whether the statistical or empirical approach is used, it is the *percentile rank* of the difference (i.e., the estimated level of rarity of the difference) that provides the most pertinent information when assessing clinical significance.

In summary, a clear distinction should be drawn between the metric required for the pre-processing of scores and that best suited for the final stage at which the neuropsychologist directly interprets an individual's results. The requirement to use standard scores at the former stage should not blind us to the advantages of percentile ranks for the latter stage.

REPORTING CONFIDENCE LIMITS THAT CAPTURE THE EFFECTS OF MEASUREMENT ERROR ON TEST SCORES AS PERCENTILE RANKS

Up until this point we have dealt only with the pros and cons of *point* estimates of a score's standing expressed on either a standard score metric or as a percentile rank. We turn now to a consideration of confidence limits on a score that capture the uncertainty arising from measurement error. Confidence limits are useful because they serve the general purpose of reminding users that test scores are fallible (they counter any tendencies to reify the score obtained) and serve the very specific purpose of quantifying this fallibility (Crawford, 2004).

There are a variety of ways of setting confidence limits, but in the present paper we use what Charter and Feldt (2001) refer to as the traditional approach. In this approach confidence limits are centered on the observed score and are obtained by multiplying the standard error of measurement (SEM) by a value of z (a standard normal deviate) corresponding to the required degree of confidence. Thus, for a two-sided 95% confidence interval the SEM is multiplied by 1.96. The quantity obtained is then subtracted from and added to the observed score to form the lower and upper confidence limits respectively. The formula for the standard error of measurement is:

$$\text{SEM} = s\sqrt{1 - r_{XX}} \tag{1}$$

where r_{XX} is the reliability of the test and s is the test's standard deviation. To take a specific example, the reliability of the Perceptual Organization Index (PO) on the Wechsler Adult Intelligence Scale Third Edition (WAIS-III; Wechsler, 1997) is 0.93

and hence the standard error of measurement is 3.97. The 95% confidence interval on a PO score of, say, 84 is therefore $84 \pm 1.96(3.97)$; i.e., 76 to 92.

All authorities on psychological measurement agree that confidence intervals of this form, or variants upon them, should accompany test scores. However, it remains the case that some neuropsychologists do not routinely record confidence limits. There is also the danger that others will dutifully record the confidence limits but that, thereafter, these limits play no further part in test interpretation. Thus it could be argued that anything that serves to increase the perceived relevance of confidence limits should be encouraged. We suggest that expressing confidence limits as percentile ranks will help to achieve this aim.

Expressing confidence limits on a score as percentile ranks is very easily achieved: the standard score limits need only be converted to z and the probability of z (obtained from a table of areas under the normal curve or algorithmic equivalent) multiplied by 100. In the previous example (where the score was 84 and therefore at the 14th percentile) the lower and upper limits expressed on a standard score (Index score) metric (76 and 92) correspond to zs of -1.59 and -0.53. Thus the 95% confidence interval, with the endpoints expressed as percentile ranks, is from the 6th percentile to the 29th percentile. As a further illustration of percentile rank confidence limits, Table 1 presents confidence limits for a range of WAIS-III Verbal Comprehension and Processing Speed Index scores: the limits are expressed on an Index score metric and as percentile ranks.

To our knowledge no test publisher expresses confidence limits in the form of percentile ranks for any existing neuropsychological instrument, nor are we aware of previous recommendations that such a practice be adopted. However, we suggest that such limits are more directly meaningful than standard score limits and offer what is, perhaps, a more stark reminder of the uncertainties involved in attempting to quantify an individual's level of neuropsychological functioning. At the risk of

Table 1 Confidence limits

		Verbal Comprehension (VC)				Processing Speed (PS)			
		Lower		Upper		Lower		Upper	
Score	PR	SM	PR	SM	PR	SM	PR	SM	PR
122	93rd	116	86th	128	97th	112	78th	132	98th
114	82nd	108	71st	120	91st	104	60th	124	95th
103	58th	97	42nd	109	72nd	93	32nd	113	81st
91	27th	85	16th	97	42nd	81	10th	101	53rd
84	14th	78	7th	90	25th	74	4th	94	35th
68	2nd	62	0.6th	74	4th	58	0.3rd	78	7th

Confidence limits expressed on a standard IQ/Index score metric (SM) and as percentile ranks (PR) for a range of obtained scores on the Verbal Comprehension and Processing Speed Indices of the WAIS-III: the limits are obtained using the traditional approach.The values chosen for the Index scores may appear a little arbitrary but were constrained by the fact that they had to be obtainable from the sum of scaled scores for both VC and PS (these two Indices were chosen as they differ appreciably in their reliability).

laboring the point, the lower limit on the percentile rank in the previous example (the lower limit is at the 6th percentile) is clearly more tangible than the Index score equivalent (76) since, as noted, this latter quantity only becomes meaningful when we know 6% of the normative population are expected to obtain a lower score.

If we take it as a given that neuropsychologists should use measures that are as reliable as possible, then regular use of confidence limits expressed as percentile ranks may also have a secondary beneficial effect. Anything that makes the uncertainty stemming from unreliability more tangible is liable to lead to neuropsychologists giving more weight to the reliability of tests when selecting among competing measures of a given construct.

As the WAIS-III has featured in these examples it should be noted that the traditional method used to set confidence limits on scores is not that used in the manual for the WAIS-III, nor in the manual for the latest version of the Wechsler Intelligence Scale for Children (WISC-IV; Wechsler, 2003). The latter are formed using a method proposed by Glutting, McDermott, and Stanley (1987). We opted for the traditional approach because it is widely endorsed (e.g., Anastasi, 1988; Charter & Feldt, 2001; Feldt & Brennan, 1983; Hopkins, Stanley, & Hopkins, 1990), and the meaning of these limits are more transparent than the Glutting et al. limits, whether expressed as standard scores or percentile ranks. For a critique of the Glutting et al. method see Charter and Feldt (2001). These authors also point out that one of the authors of the Glutting et al. paper (J. C. Stanley) appears to have reverted to the traditional approach in subsequent treatments of this topic (Hopkins et al., 1990).

A further method for setting confidence limits is that proposed by Lord and Novick (1968). Like the Glutting et al. method, these confidence limits are centered around estimated true scores but they use a different standard error, the standard error of estimation. Estimated true scores are obtained using the formula:

$$\text{True score} = r_{XX}(X-\overline{X})+\overline{X} \tag{2}$$

where X is the observed score and $\overline{X}$ is the observed score mean. The formula for the standard error of estimation (see Lord & Novick, 1968) is:

$$\text{SEE} = s_X\sqrt{r_{XX}}\sqrt{1-r_{XX}}. \tag{3}$$

To obtain a 95% confidence interval the standard error of estimation is multiplied by a z of 1.96 and this quantity is added and subtracted to the estimated true score to form the upper and lower limits.

Before going on to consider the relationship between these limits and limits expressed as percentiles, it is worth discussing the relationship between the use of estimated true scores as *point* estimates of the standing of a score and the corresponding use of a percentile rank as a point estimate. The fundamental result is that, because the estimated true score is simply a linear transformation of the observed score, the percentile rank of an estimated true score is necessarily identical to the percentile rank of the observed score.

To illustrate, suppose that a test has reliability of 0.8 and that the mean and standard deviation for observed scores are 100 and 15 respectively. Suppose also that the interest is in an observed score of 85; this converts to a z score of -1.0 and hence the observed score is at the 15.9th percentile. The estimated true score in this example, using formula (3) is 88. The mean of true scores equals the mean of observed scores and the standard deviation of estimated true scores (and of the theoretical actual true scores) is obtained by multiplying the standard deviation of obtained scores by the reliability of the test, i.e., $r_{XX}s_X$, thus the standard deviation of true scores in this example is 12.0. Substituting the estimated true score, and the mean and standard deviation of true scores in the formula for z in place of the observed score and observed score mean and standard deviation yields exactly the same value (-1.0) as obtained for observed scores and thus the percentile rank for the estimated true score is identical to the percentile rank of the observed score.

The correspondence between the percentile ranks of an observed score and its corresponding estimated true score does not hold when the endpoints of the traditional confidence intervals and the Lord and Novick (1968) intervals are compared. The degree of divergence between the percentile confidence limits obtained using the two methods is modest when, as in the case of the Wechsler Indices, the tests have high reliability. However, when tests have more modest reliability the divergence can be marked.

To illustrate, using the previous example of a score of 84 on the Perceptual Organization Index of the WAIS-III, the percentile interval calculated using the traditional method was from the 6th percentile to the 30th percentile. Using Lord and Novick's method the confidence interval on this score is from 78 to 93. Converting these endpoints to percentiles (that is, expressing the endpoints as deviation scores, dividing the deviation scores by the standard deviation of true scores to obtain z and thereby the percentiles) yields an interval from the 6th to the 30th percentile. These percentile limits are only minimally different from the percentile limits obtained using the traditional method (indeed, as can be seen, following rounding to integers the limits are identical). However, suppose that all the quantities in the previous example remain the same (i.e., the score obtained was 84 and the standard deviation of obtained scores was 15), except that the test concerned had a reliability of only 0.6. Then the interval calculated using the traditional method is from the 1st to the 57th percentile, whereas the interval using Lord and Novick's method is from the 0.4th percentile to the 70th; the difference between the limits are substantial in this latter example.

As previously noted, our preference is to use the traditional method of forming limits because the Glutting et al. (1987) and Lord and Novick (1968) alternatives are more abstract (i.e., more removed from the score actually obtained by the client) and are therefore less intuitive for the end user. It could be argued therefore that their use is not in keeping with the generally recognized need to increase the perceived relevance of confidence intervals in neuropsychological practice.

Furthermore, it is questionable whether, in clinical populations (in which cognitive impairment will be common), it is reasonable to center limits around an estimated true score formed by regressing an observed score toward the mean of the normative (unimpaired) population. For example, in the population of patients who

have suffered a severe head injury, the mean score on the Processing Speed Index of the WAIS-III will not be 100 and yet if we center the confidence limits for a severely head-injured client on the estimated true score, we proceed as though it were (that is, in all likelihood we will have regressed the score towards the wrong mean).

OTHER FORMS OF CONFIDENCE LIMITS ON PERCENTILES

In the foregoing section we explicitly identified the proposed confidence limits (and their alternatives) as capturing the uncertainty arising from measurement error. This was to clearly differentiate them from confidence limits that capture the uncertainty arising from using normative sample statistics in place of population parameters when establishing the standing of a score (Crawford & Garthwaite, 2002). The former limits treat the normative sample statistics as population parameters. Crawford and Garthwaite's (2002) limits do not but, on the other hand, are concerned solely with the score as observed (Crawford & Garthwaite, 2008).

The present limits should also be distinguished from confidence limits on percentiles formed using the binomial distribution (Newcombe, 1998); these limits are more commonly used to set confidence limits on a proportion but are easily adapted for use with percentiles. They are akin to Crawford and Garthwaite's (2002) limits in that they capture the uncertainty arising from using a normative sample in place of a normative population and are not concerned with capturing the effects of measurement error (they yield wider limits than Crawford and Garthwaite's limits but are a useful fall-back when it is unreasonable to assume that the normative data are normally distributed).

COMPUTER PROGRAMS FOR SIMULTANEOUSLY EXPRESSING TEST SCORES AND ACCOMPANYING CONFIDENCE LIMITS ON THEIR STANDARD METRIC AND AS PERCENTILE RANKS

The arguments presented earlier make the case for expressing confidence limits on test scores on both a standard metric and as percentile ranks. It is clearly a simple task to convert limits expressed on a standard metric to percentile ranks. (As noted, the limits need only be expressed as z and the corresponding left-hand area under the normal curve multiplied by 100.) However, it is tedious and error prone. Research shows that clinicians make many more simple clerical errors than we like to imagine (e.g., see Faust, 1998; Sherrets, Gard, & Langner, 1979; Sullivan, 2000). It therefore makes sense to leave this task to a computer, particularly as the limits can then be obtained instantaneously.

Two computer programs for PCs accompany this paper. WAIS3_PRCLME.exe (**P**ercentile **R**ank **C**onfidence **L**imits reflecting **M**easurement **E**rror) is for use with the WAIS-III; a companion program WISC4_PRCLME.exe is for use with the WISC-IV (both can be downloaded from the following website address: http://www.abdn.ac.uk/~psy086/dept/PRCLME.htm). The programs generate confidence limits on Index scores (and FSIQ) for the WAIS-III or WISC-IV using the traditional method described earlier. The only inputs required are the Index scores and FSIQ for either of the two tests. The output consists of the original scores, the 95% confidence limits expressed on

an Index/IQ metric, the percentile rank of these scores, and the confidence limits expressed as percentile ranks.

It would be straightforward to develop similar dedicated programs for other tests in the neuropsychologist's armamentarium. Alternatively, it would also be simple to set up the required procedures in a generic spreadsheet application.

REFERENCES

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Bowman, M. L. (2002). The perfidy of percentiles. *Archives of Clinical Neuropsychology*, *17*, 295–303.

Charter, R. A., & Feldt, L. S. (2001). Confidence intervals for true scores: Is there a correct approach? *Journal of Psychoeducational Assessment*, *19*, 350–364.

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, UK: 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. (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. (2008). 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., & Gray, C. D. (2003). Wanted: Fully operational definitions of dissociations in single-case studies. *Cortex*, *39*, 357–370.

Faust, D. (1998). *Forensic assessment, Comprehensive clinical psychology volume: Assessment* (pp. 563–599). Amsterdam: Elsevier.

Feldt, L. S., & Brennan, R. L. (1983). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

Glutting, J. J., McDermott, P. A., & Stanley, J. C. (1987). Resolving differences among methods of establishing confidence limits for test scores. *Educational and Psychological Measurement*, *47*, 607–614.

Golden, C. J., Espe-Pfeifer, P., & Wachsler-Felder, J. (2002). *Neuropsychological interpretation of objective psychological tests*. New York: Kluwer.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation*. Englewood Cliffs, NJ: Prentice Hall.

Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, H. J., & Fischer, J. S. (2004). *Neuropsychological assessment* 4th ed.). New York: Oxford University Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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.

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

Payne, R. W., & Jones, G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology*, *13*, 115–121.

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

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.

Sullivan, K. (2000). Examiners' errors on the Wechsler Memory Scale–Revised. *Psychological Reports*, *87*, 234–240.

Wechsler, D. (1997). *Manual for the Wechsler Adult Intelligence Scale – Third edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: The Psychological Corporation.