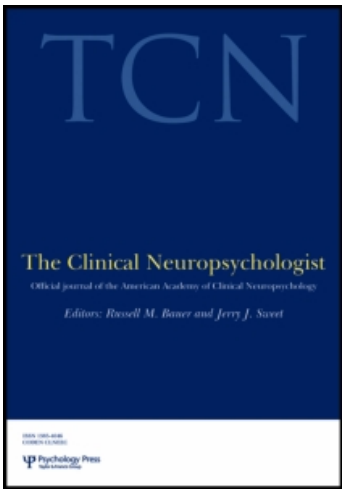


The Clinical Neuropsychologist
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713721659

# On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores

John R. Crawford [a]; Paul H. Garthwaite [b]; Daniel J. Slick [c]
[a] School of Psychology, University of Aberdeen, Milton Keynes, UK [b] Department of Mathematics and Statistics, The Open University, Milton Keynes, UK [c] Departments of Pediatrics and Clinical Neurosciences, University of Calgary and Alberta Children's Hospital, Canada

First Published on: 26 March 2009

**To cite this Article** Crawford, John R., Garthwaite, Paul H. and Slick, Daniel J.(2009)'On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores',The Clinical Neuropsychologist,23:7,1173 — 1195

**To link to this Article: DOI:** 10.1080/13854040902795018

**URL:** http://dx.doi.org/10.1080/13854040902795018

*The Clinical Neuropsychologist*, 23: 1173–1195, 2009
http://www.psypress.com/tcn
ISSN: 1385-4046 print/1744-4144 online
DOI: 10.1080/13854040902795018

Psychology Press
Taylor & Francis Group



# ON PERCENTILE NORMS IN NEUROPSYCHOLOGY: PROPOSED REPORTING STANDARDS AND METHODS FOR QUANTIFYING THE UNCERTAINTY OVER THE PERCENTILE RANKS OF TEST SCORES

**John R. Crawford[1], Paul H. Garthwaite[2], and Daniel J. Slick[3]**
[1]*School of Psychology, University of Aberdeen,* [2]*Department of Mathematics and Statistics, The Open University, Milton Keynes, UK, and* [3]*Departments of Pediatrics and Clinical Neurosciences, University of Calgary and Alberta Children's Hospital, Canada*

***Normative data for neuropsychological tests are often presented in the form of percentiles. One problem when using percentile norms stems from uncertainty over the definitional formula for a percentile. (There are three co-existing definitions and these can produce substantially different results.) A second uncertainty stems from the use of a normative sample to estimate the standing of a raw score in the normative population. This uncertainty is unavoidable but its extent can be captured using methods developed in the present paper. A set of reporting standards for the presentation of percentile norms in neuropsychology is proposed. An accompanying computer program (available to download) implements these standards and generates tables of point and interval estimates of percentile ranks for new or existing normative data.***

***Keywords***: Neuropsychological assessment; Interval estimates; Confidence intervals; Credible intervals; Test norms; Non-normal data; Percentile ranks; Bayesian methods; Reporting standards; Statistical reform; Computer scoring.

## INTRODUCTION

There are many examples of the use of percentile norms in neuropsychology (e.g., see Mitrushina, Boone, Razani, & D'Elia, 2005; Strauss, Sherman, & Spreen, 2006). They tend to be used when the distribution of raw scores departs markedly from a normal distribution, particularly when the range of scores obtained by the normative sample is limited, or when there is a limited number of test items in the first place. The norms may consist of full tables that present the percentile rank for each raw score, or it may be restricted to presenting the raw scores corresponding to various landmark percentiles (e.g., the raw scores corresponding to the 1st, 5th, 10th, 50th percentile etc.).

It could be argued that even greater use could be made of percentile norms, as they would often be more appropriate than the common practice of presenting

Address correspondence to: Professor John R. Crawford, School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 2UB, UK. E-mail: j.crawford@abdn.ac.uk

Accepted for publication: February 2, 2009. First published online: March 26, 2009.

normative data in the form of raw score means with accompanying standard deviations. When using this latter type of normative data, neuropsychologists are either explicitly or implicitly encouraged to estimate an individual's standing on the test by expressing her/his scores as a *z* score and referring it to a table of areas under the normal curve. Quite apart from the problems with this approach when the normative sample is modest in size (Crawford & Garthwaite, 2008; Crawford & Howell, 1998), it also assumes that the normative data are drawn from a normal distribution. However, normative data for neuropsychological tests will often reveal evidence of extreme skew and/or leptokurtosis (Crawford, Garthwaite, Azzalini, Howell, & Laws, 2006); often this is because the tests are comfortably within the competence of most healthy participants so that scores at or near ceiling predominate.



## ARE PERCENTILE RANKS THE POOR RELATIONS OF STANDARDIZED SCORES?

For test scores that are normally distributed (or can be transformed to normality) many commentators on neuropsychological assessment emphasize the use of standardized scores (e.g., IQs, Indexes, or *T* scores, etc.) rather than percentiles as the preferred means of representing an individual's performance (Crawford, 2004; Lezak, Howieson, Loring, Hannay, & Fischer, 2004; Strauss et al., 2006). Indeed some are distinctly hostile towards the use of percentiles (Bowman, 2002).

Percentiles, however, have one distinct advantage over any alternative metric: They tell us *directly* how common or uncommon a patient's test score is in the normative population. In contrast, a standardized score, such as (say) an Index score of 82, has no intrinsic meaning: It only becomes meaningful when we know that approximately 12% of the normative population are expected to obtain a lower score. Thus a standardized score can only be interpreted if a neuropsychologist has developed an internalized look-up table mapping standardized scores on to percentile ranks or has consulted an external conversion table (Crawford & Garthwaite, 2009).

The focus of the present paper is on how to represent performance when test scores are not normally distributed (and cannot be adequately transformed to normality) and thus, in such cases, the relative merits of standardized scores and percentiles is irrelevant: Standardized scores are not an option and so the normative data need to be presented in the form of the percentile ranks.

## SOME PROBLEMATIC PARTICULARS OF PERCENTILES

Most neuropsychologists, even those who would have a preference for standardized scores (were they an option) acknowledge that a further positive feature of percentiles is their simplicity: Their meaning is apparently unequivocal. For example, it is often noted that this simplicity makes them useful for communicating results to non-psychologists (Lezak et al., 2004). We are broadly in agreement with this position. However, the devil can be in the detail.

Although it may not be widely recognized as such, one significant problem with using percentile norms in neuropsychology is the co-existence of different definitions of a percentile. The most recent edition of the *Standards for Educational and Psychological Testing* offers the following definitions, ''Most commonly the percentage of scores in a specified distribution that fall below the point at which a given score lies. Sometimes the percentage is defined to include scores that fall at the point; sometimes the percentage is defined to include half of the scores at the point'' (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999 p. 179). Hereafter these three definitions will be referred to as definition A (the percentage of scores that fall below the score of interest), B (the percentage of scores that fall at or below the score of interest), and C (the percentage of scores that fall below the score of interest, where half of those obtaining the score of interest are included in the percentage).

In some circumstances the differences between the percentile ranks of a score obtained by applying these three different definitions will be marginal. However, the differences can be very considerable when there is only a modest number of test items, or when, despite a large number of items, the scores obtained by the normative sample are nevertheless concentrated across a narrow range (as will occur, for example, when many members of the sample score at or near ceiling). It will be appreciated that these are the very characteristics that lead neuropsychologists to opt to present normative data in the form of percentiles. Thus, in neuropsychology, we cannot afford to be casual in our attitude towards which particular method has been used to calculate percentiles. Furthermore, small normative sample sizes are still quite common in neuropsychology (Bridges & Holler, 2007; Crawford & Howell, 1998) and this will often further exacerbate differences between the percentiles provided by the three methods.

To illustrate the potential magnitude of differences between percentiles calculated using the three definitions, Table 1 records a hypothetical frequency distribution of raw scores obtained by a normative sample of 100 persons on a neuropsychological test consisting of 12 items, together with the percentile ranks of the raw scores calculated using the three methods. It can be seen that the differences are substantial: For example, the percentile rank for a raw score of 6 is 2 according to definition A, but is 6 according to definition B; for a raw score of 8, the corresponding percentile ranks are 10 and 24.

This problem is really two problems. First, it is unsatisfactory that three mutually exclusive methods of calculating percentiles are currently in use, particularly when the differences between them can be marked. This makes working with percentile norms more complicated than it need be. However, the second problem is more serious: When percentile norms are presented the method used to calculate them is often unspecified. That is, the method section of a normative study may state only that ''the raw scores were expressed as percentiles'' or something similar. Even where a definition is provided, there can still be room for ambiguity. Definition A in the *Standards for Educational and Psychological Testing* is ''the percentage of scores in a specified distribution that fall below the point at which a given score lies'' (p. 179). This definition is unequivocal in its meaning only because it is then explicitly contrasted with the alternative definitions. If such a definition

**Table 1** Applying three different definitions of a percentile rank to the raw scores

| | | Percentile ranks | | |
|---|---|---|---|---|
| Raw score | *n* obtaining | Definition A: $m/N$ | Definition C: $(m+0.5k)/N$ | Definition B: $(m+k)/N$ |
| 0 | 0 | <1 | <1 | <1 |
| 1 | 0 | <1 | <1 | <1 |
| 2 | 0 | <1 | <1 | <1 |
| 3 | 0 | <1 | <1 | <1 |
| 4 | 0 | <1 | <1 | <1 |
| 5 | 2 | <1 | 1 | 2 |
| 6 | 4 | 2 | 4 | 6 |
| 7 | 4 | 6 | 8 | 10 |
| 8 | 14 | 10 | 17 | 24 |
| 9 | 16 | 24 | 32 | 40 |
| 10 | 20 | 40 | 50 | 60 |
| 11 | 30 | 60 | 75 | 90 |
| 12 | 10 | 90 | 95 | >99 |

Illustration of the effects of applying three different definitions of a percentile rank to the raw scores from a normative sample of 100 ($m$ = number scoring below, $k$ = number obtaining given score, and $N$ = overall normative sample size).



appeared in isolation in a study presenting percentile norms it would not be clear whether the percentiles were calculated using definitions A or C (that is, in calculating the percentage of scores below the point at which a given score lies, half of those obtaining the integer valued score may have been included).

Therefore it is not just the case that neuropsychologists have to struggle to recall which definition has been used for a particular set of norms, but they often do not have access to this information in the first place. In this latter scenario there is the possibility that some neuropsychologists may define a percentile according to, say, definition A while being unaware that some of the percentile norms they use have been generated according to, say, definition B. This will often lead to erroneous conclusions concerning the standing of a case's scores as can be seen by referring back to the examples presented in Table 1.

## THE CASE FOR ADOPTING A SINGLE DEFINITION OF PERCENTILES IN NEUROPSYCHOLOGY

Having seen that the percentile ranks for raw scores can differ substantially depending on how percentiles are defined, and having discussed the practical problems that ensue, we suggest that the neuropsychological community should attempt to settle on a single, agreed, definition of a percentile for neuropsychological test scores. It could be argued that which of the three definitions should fulfill this role is of secondary importance compared to the desirability that one (any one) should prevail over the two others. We would have some sympathy with this position, nevertheless we strongly advocate the use of definition C. In doing so we are in august company (e.g., Guilford, 1954; Ley, 1972).

Although test scores are discrete (i.e., integer valued), the underlying cognitive abilities they index are generally taken to be continuous, real-valued quantities. Thus a raw score of, say, 7 is regarded as a point estimate of a real valued score which could lie anywhere in the interval 6.5 to 7.4999 (plus an infinite number of additional 9s after the fourth decimal place). Put another way, theoretically we could distinguish among individuals obtaining the same raw score were we to introduce tie-breaking items. In the absence of knowledge of the shape of the underlying distribution, our best estimate is that half of the real-valued scores will be below 7 and half above. This, then, is the rationale for definition C.

Moreover, if neuropsychology is to adopt a single definition of a percentile, and if this single definition is to encompass all scenarios in which percentiles are used, then definition C is the only candidate. That is, in the present paper the emphasis is on the use of percentile norms with test scores that depart markedly from normality. However, as previously noted, scores on standardized tests (such as IQs or Index scores from Wechsler tests) can also be expressed as percentiles. Definition C is implicit in the latter scenario: Approximately 2.66% of the population are expected to obtain an IQ of 100 so that, although the percentile rank for this score is 50 using definition C, the use of definitions A or B gives percentile ranks of 49 and 51 respectively. The discrepancies are even larger when Wechsler subtest scores are expressed as percentiles: The percentile rank for a subtest score of 10 is 50 using definition C but is 43 and 57 respectively using definitions A and B.

Before leaving this topic, and in keeping with our emphasis on providing an explicit definition of a percentile, the formula (Ley, 1972) for obtaining the percentile rank of a given score using definition C is presented below:

$$\text{Percentile rank} = \left(\frac{m + .5k}{N}\right)100, \tag{1}$$

where $m$ is the number of members of the normative sample scoring below a given score, $k$ is the number obtaining the given score, and $N$ is the overall size of the normative sample.

## UNAVOIDABLE UNCERTAINTY OVER PERCENTILE RANKS

In a previous section we highlighted the uncertainty that can arise over percentile norms because (a) there are multiple definitions of a percentile, and (b) it is often not clear what definition has been used to generate a particular set of percentile norms. This uncertainty could be avoided if explicit definitions of a percentile were provided by those presenting percentile norms or, even better, if the neuropsychological community were to adopt a single definition of a percentile. In contrast, there is a further uncertainty when using percentile norms that is unavoidable. We turn now to a consideration of this uncertainty.

When neuropsychologists refer a case's score to percentile norms, their interest is in the standing (percentile rank) of the case's score in the normative *population*, rather than its standing in the particular group of (say) 150 participants who happen to make up the normative *sample*. That is, the percentile rank of a score obtained from normative data is a point estimate of its standing in the normative population

and there is uncertainty over this quantity (in the foregoing example it is very unlikely that the percentile ranks of raw scores would be the same were they obtained from an alternative sample of 150 members of the normative population).

Fortunately point estimates can potentially be supplemented with *interval* estimates of a score's percentile rank. These interval estimates would serve the useful, general, purpose of reminding us that normative data are fallible and serve the specific, concrete, purpose of quantifying the degree of fallibility expected for a particular normative data set.

It might be thought that the utility of interval estimates will lie mainly in quantifying the uncertainty over the percentile ranks of scores when a normative sample is modest in size. It is certainly true that interval estimates would be particularly useful in this situation, as they would provide a stark reminder of the lack of precision of the point estimate. However, a neuropsychologist may have an unrealistic sense of the precision of point estimates of percentile ranks when using norms from more moderately sized samples and so an interval estimate would still be valuable in this latter situation. Moreover, when an interval estimate of a percentile rank *is* narrow (as will occur when the normative sample size is very large), it will still have served the useful purpose of providing reassurance that the standing of a test score has been quantified with a high degree of precision. Finally, in neuropsychological assessment much emphasis is placed on examining the profile of a patient's scores across a range of cognitive tests. When (as is often the case) these tests have been normed on different samples, there will be differing degrees of uncertainty over the standing of the various scores; the provision of accompanying interval estimates would help the neuropsychologist weigh the evidence from these different tests when arriving at a formulation.

The foregoing arguments suggest that the point estimates of a score's percentile rank should be supplemented with an interval estimate. However, although percentile norms are widely used in neuropsychology, we could find no example in any of the three major collections of published normative data for neuropsychological tests (Lezak et al., 2004; Mitrushina et al., 2005; Strauss et al., 2006) in which the percentile ranks for raw scores (i.e., the point estimates of the standing of raw scores) were accompanied by interval estimates (i.e., confidence intervals quantifying the uncertainty over the true standing of a raw scores in the normative population). The next sections are concerned with how such estimates can be obtained.

## METHODS FOR OBTAINING INTERVAL ESTIMATE OF A TEST SCORE'S PERCENTILE RANK

When test scores are assumed to be normally distributed (either because the raw scores of the normative sample tend to a normal distribution, or the scores can be transformed to normality, or the test in question is part of a standardized battery), then a parametric method developed by Crawford and Garthwaite (2002) can be used to provide an interval estimate of the percentile rank of a test score (see also Crawford & Garthwaite, 2008). It is possible to calculate this interval because, when test scores are normally distributed, their percentile ranks follow a non-central *t*-distribution. The methods employed are those of classical statistics but

it has recently been shown that a Bayesian approach to the same problem gives identical results (Crawford & Garthwaite, 2007).

The foregoing parametric methods have a wide range of potential applications in neuropsychology. However, they are not appropriate for the problem at hand because neuropsychologists usually opt to present normative data in the form of percentiles precisely when it is clear that the normality assumption cannot be met. Alternative methods are therefore required: Consideration will now be given to non-parametric approaches to obtaining an interval estimate of a percentile rank. There is no shortage of potential methods for this purpose because a percentile rank is a proportion multiplied by 100. Non-parametric methods of obtaining interval estimates for proportions can therefore be used to provide interval estimates of the percentile ranks of raw scores. In the next sections we consider existing classical and Bayesian methods that could be used to obtain interval estimates for percentile ranks and go on to develop variants upon them that are better suited to the particulars of the present problem.



## CLASSICAL (FREQUENTIST) INTERVAL ESTIMATE FOR A PERCENTILE RANK

There is a host of classical methods for obtaining interval estimates for a proportion; see Newcombe (1998) and Brown, Cai, and DasGupta (2001) for reviews and quantitative evaluations of these methods. Among the oldest and most widely used method is that developed by Clopper and Pearson (1934). It ensures a coverage probability of at least 95% and is often referred to as an "exact" method, because it is based on the inversion of a binomial test, which is an exact test of a binomial proportion (Brown et al., 2001).

This method, or the other alternative classical methods referred to above, could be used to provide interval estimates for the percentile rank of an individual's test score. However, there is a complication. As previously noted, normative data will almost always contain a sizeable number of tied scores; that is, a large number of people in the normative sample will often obtain the same raw test score. Indeed, if the normative sample is very large and the number of different possible scores is limited, then there could literally be hundreds of such ties for a given raw score. The present problem therefore differs from standard binomial sampling in which there can be no possibility of multiple ties. One solution would be to simply ignore the problem of ties and obtain an interval based on the number of "successes" as provided by definition C. However, regardless of which existing classical method was used to obtain this estimate, it would not incorporate all of the uncertainty involved. Fortunately we can do better. The nature of the problem and its solution is best illustrated with a concrete example.

Suppose that a case obtains a particular raw score on a neuropsychological test. Also suppose that, in a normative sample consisting of 80 people, $k = 4$ people obtained this score and a further $m = 10$ people obtained a lower score. We want to obtain both a point estimate of the case's percentile rank and a two-sided 95% interval estimate of the percentile rank. Applying definition C the point estimate of the percentile rank for the case's score is $100 \times (10 + 2)/80 = 15$; expressed as a proportion it is 0.15.

To obtain the interval estimate for the proportion we could define these data as constituting $x = m + 0.5k = 12$ "successes" out of 80 and treat them as an observation from a binomial distribution. A binomial distribution has two parameters: $N$ the number of trials, and $p$, the probability of success on any one trial. To find the lower limit of the interval, we use a search procedure to find the binomial distribution, with 80 as its first parameter (as there are 80 "trials", i.e., members of the normative sample), that has $x = 12$ as its 0.025 percentile point: The second parameter of this distribution ($p$) provides us with the upper limit and, in this example, the distribution turns out to be bin(80, 0.0800). For the upper limit we find the binomial distribution (again with 80 as its first parameter) that has $x = 12$ as its 0.975 percentile point and this turns out to be bin(80, 0.2474). Therefore, the 95% confidence interval on the proportion is from 0.0800 to 0.2474. The procedure just described is the Clopper-Pearson method referred to earlier. Multiplying the endpoints of this interval by 100 provides the two-sided confidence interval on the percentile rank for the given raw score, which is from 8.00 to 24.74.



## DEVELOPMENT OF A CLASSICAL METHOD THAT INCORPORATES THE UNCERTAINTY ARISING FROM MULTIPLE TIES

We have $k$ members of the normative sample with the same integer valued score as the case ($k = 4$ in the numerical example). On the underlying continuous score the number of these members ($x$) scoring below the individual could range from 0 to $k$ (with each of these $k + 1$ possibilities being equally probable), and the interval estimate should reflect this uncertainty. We assume that each of these possibilities is equally likely, based on the premise that the case is very similar to the controls with which it is tied, so that the results of any tie-breaking are almost random.[1] To obtain the lower endpoint of the confidence interval, we find the binomial distribution (again with $N$ as its first parameter) that has $A$ as its 0.025 quantile, where $A =$ the average of the $k + 1$ probabilities that $x$ is less than $m + 0$, $m + 1, \ldots, m + k$. In the current example there are five such probabilities (e.g., the probability that $x$ is less than 10, through to the probability that $x$ is less than 14). The binomial distribution having this characteristic is bin(80, 0.0739) and thus 0.0739 is our lower endpoint. A similar approach is used to find the upper endpoint and the associated binomial distribution is bin(80, 0.2548), so the upper endpoint is 0.2548. Thus the interval estimate of the percentile rank for the case's score is from 7.39 to 25.48. A fuller, more formal, description of the method is set out in the Appendix, including a worked example (Section A.1) and details of the search procedure required to find the appropriate binomial distribution (Section A.4).

It can be seen that, for the current example, the interval estimate of the percentile rank obtained using this method (7.39 to 25.48) is wider than that obtained using the standard classical (Clopper-Pearson) approach (8.00 to 24.74). This occurs because it incorporates the additional uncertainty arising from the presence of ties. The difference between these intervals is very modest in this

[1] Note that treating these possibilities as equally probable is consistent with using definition C, i.e., $(m + 0.5k)/N$, as the point estimate of the percentile rank.

example but that is because there are relatively few ties (the number of ties was chosen so that it was practical to provide a worked example in the Appendix). However, the differences can be substantial: A later section compares the interval estimates obtained when the number of ties is varied.

The current problem with tied scores has many similarities to that encountered when applying the Mann-Whitney test to compare the scores of two samples. However, because in the present case the problem is one of referring an *individual's* score to a single normative sample containing ties, obtaining a satisfactory solution for dealing with ties is much simpler than when comparing two groups.

Before leaving this issue it is worth stressing that the methods developed here are motivated by the assumption of an underlying continuous score, so that ties can be broken, and the assumption that all possible results from tie-breaking are equally likely. These assumptions also lead to definition C for a percentile rank and render a standard approach based on binomial sampling non-optimal. When ties cannot be broken because the underlying quantity of interest is discrete (the quantity is unlikely to be a test score), a lower-bound or upper-bound percentile may be appropriate and the methods developed here would be unnecessary.



## A MID-*p* VARIANT OF THE CLASSICAL METHOD

The classical method developed in the foregoing section can be construed as an extension to the standard Clopper-Pearson method to incorporate the additional uncertainty introduced by the presence of multiple ties. It is commonly thought that the standard Clopper-Pearson method is unnecessarily conservative (Agresti & Coull, 1998) and, instead, a mid-*p* variant of the standard method is often used to obtain interval estimates (Brown et al., 2001). The Clopper-Pearson method bases confidence intervals on the probability of the observed result plus the probability of more extreme results, while the mid-*p* variant bases them on *half* the probability of the observed result plus the probability of more extreme results; see Berry and Armitage (1995) for a general discussion of the rationale for using a mid-*p* approach to confidence intervals.

The approach giving a mid-*p* variant of the standard method can also be applied to obtain a mid-*p* variant of the classical method for dealing with multiple ties developed here. The technical details are set out in section A.2 of the Appendix and the option of applying the mid-*p* variant is incorporated into the computer program that accompanies this paper (see later for details of how to download). The 95% interval estimate calculated using the mid-*p* version of the classical interval is (7.79 to 24.86) and it can be seen that this interval is shorter (i.e., less conservative) than that obtained earlier (7.39 to 25.48).

## BAYESIAN CREDIBLE INTERVALS ON A PERCENTILE RANK

Bayesian statistics provide an alternative to the classical approach to inference. The essential difference between the classical and Bayesian approaches is that the classical approach treats parameters as *fixed* but unknown whereas, in the Bayesian approach, parameters are treated as random variables and hence have probability distributions. For example, suppose $\theta$ is the (unknown) proportion in a

specified normative population. Then a Bayesian might say "the probability that $\theta$ is less than 0.3 is 0.95" while classical statistics does not permit a probabilistic statement about $\theta$. With the classical approach, $\theta$ is either less than 0.3 or it is not, and there is no random uncertainty, so the probability that $\theta$ is less than 0.3 is either 1 or 0, and nothing in between.

In the Bayesian approach, a *prior* distribution is used to convey any information about model parameters that was available before the sample data were gathered. This is combined with the information supplied by the data, which is contained in the *likelihood*, to yield a *posterior* distribution. Formally,

$$\text{posterior} \propto \text{prior} \times \text{likelihood},$$

where $\propto$ means "is proportional to". Prior distributions can enable background knowledge to be incorporated into a statistical analysis. This is one of the central sources of dispute between the classical and Bayesian schools. For the Bayesian the ability to incorporate prior opinion or knowledge is a strength of the approach and is consistent with general principles of scientific reasoning (Howson & Urbach, 1993). For opponents, the incorporation of prior opinion allows results to be, in part, determined by the subjective biases of the investigator. Bayesians can marshal a variety of arguments against this latter position but, in the present context, this debate need not overly concern us: we assume no prior knowledge of the normative population and therefore use a non-informative prior distribution.

In the basic Bayesian approach to obtaining interval estimates of a proportion (that is, when the problem involves standard binomial sampling with no ties) the prior distribution for the proportion is represented by a single beta distribution: A beta distribution has two parameters, denoted $a$ and $b$ and the beta distribution is denoted beta$(a, b)$. Three beta distributions have commonly been used to represent a lack of prior knowledge or opinion: the Jeffrey's prior, beta$(0.5, 0.5)$; a uniform prior, beta$(1, 1)$; and a beta$(0, 0)$ distribution. We will use the Jeffrey's prior, although the choice among these priors makes little practical difference as their effect on the posterior distribution is swamped by the data unless the sample providing the data is very modest in size.

As in any Bayesian analysis, the prior is combined with the data to yield the posterior distribution. For the present problem, conveniently, the posterior is another beta distribution thereby avoiding the need for numerical integration (when prior and posterior distributions are of the same distributional form they are referred to as conjugate distributions). To obtain this posterior distribution denote the number in the normative sample scoring below a case's score as "successes" $(x)$ and the remainder as "failures" $(N-x)$, then the posterior distribution is a beta$(0.5+x,\ 0.5+N-x)$ distribution.

This standard Bayesian approach could be used to obtain an interval estimate of the percentile rank of the case's score. Take the example used to illustrate the classical method, in which 10 members of a normative sample of 80 scored below a case's score and 4 obtained the same score as the case. Thus, as in the classical example, we set $x$ to $m+0.5k=12$. Then the posterior distribution is beta$(0.5+12,$ $0.5+80-12)$. To obtain the 95% two-sided, equal-tailed interval estimate of the proportion scoring below the case simply requires finding the quantiles

corresponding to the 0.025 and 0.975 percentile points of this beta distribution. These quantiles can be obtained from tables (e.g., Phillips, 1973) or, more accurately (as the need for any interpolation is avoided), by a computer algorithm. In this example the interval estimate of the proportion is from 0.0847 to 0.2401 and hence the interval estimate of the case's percentile rank is from 8.47 to 24.01. Bayesians use the term "credible interval" rather than "confidence interval" for such interval estimates.

Just as with the standard classical method, for the problem at hand this standard Bayesian approach can be improved upon to incorporate the additional uncertainty introduced by multiple ties. This is achieved by using a mixture of beta distributions. As in the standard Bayesian approach, we use a Jeffrey's prior for the prior distribution. However, unlike the standard approach, the posterior distribution is a mixture of $k+1$ beta distributions in which each element of the mixture has probability $1/(k+1)$. These $k+1$ beta distributions are all beta$(0.5+x, 0.5+N-x)$ but with $x=m+i$, where $i$ takes values of 0, 1, ..., $k$. The interval estimates are obtained by finding the 0.025 and 0.975 quantiles of this posterior distribution. In the specific example used earlier the interval estimate for the proportion is 0.0739 to 0.2548 and hence the interval estimate for the case's percentile rank is from 7.39 to 25.48. It can be seen that this interval is wider than the interval obtained using the standard Bayesian method. Fuller details of the new method are provided in section A.3 of the Appendix along with a worked example.

## COMPARISON OF INTERVAL ESTIMATES

Table 2 allows the reader to compare the interval estimates of a percentile rank generated by the standard Bayesian and classical methods (in which the uncertainty arising from tied scores is ignored) with the methods developed here (that incorporate this source of additional uncertainty). Crucially, Table 2 was constructed so that the point estimate of the percentile rank for a hypothetical raw score (obtained using definition C) was always 16. However, this point estimate can be arrived at in a variety of ways. Take the specific example in Table 2 of a normative sample of 50 people in which the point estimate of 16 was arrived at by eight people scoring below the given raw score and none obtaining the score: The point estimate can also be arrived at if two people obtained a lower score and 12 obtained the given score.

Table 2 demonstrates a number of important points. First it can be seen that, at each sample size, the standard methods give the same interval estimates, regardless of whether there are tied scores (because they are blind as to how the point estimate was obtained). Second, when there are no tied scores (i.e., no members of the normative sample obtained the score of interest), the methods developed here necessarily give exactly the same intervals as the corresponding standard methods. For example, with a sample size of 50 and no tied scores, the present Bayesian method gives the interval (7.9, 27.9), as does the standard Bayesian method. Thus for this problem, the three standard methods are essentially subsumed under each of the three corresponding methods developed here.

Third it can be seen that, when there *are* tied scores (as will normally be the case), the intervals are substantially wider in order to incorporate the additional uncertainty. For example, in the example just referred to (i.e., a normative sample

**Table 2** Comparison of standard Bayesian and classical interval estimates of a percentile rank

| | Standard methods | | | Methods capturing uncertainty from presence of ties | | |
|---|---|---|---|---|---|---|
| | Bayesian | Classical | Classical mid-*p* | Bayesian | Classical | Classical mid-*p* |
| $N=50$, $m=8$, $k=0$ | (7.9, 27.9) | (7.2, 29.1) | (7.7, 28.1) | (7.9, 27.9) | (7.2, 29.1) | (7.7, 28.1) |
| $N=50$, $m=6$, $k=4$ | (7.9, 27.9) | (7.2, 29.1) | (7.7, 28.1) | (6.8, 29.4) | (6.1, 30.5) | (6.6, 29.5) |
| $N=50$, $m=2$, $k=12$ | (7.9, 27.9) | (7.2, 29.1) | (7.7, 28.1) | (2.5, 35.0) | (1.9, 36.1) | (2.4, 35.1) |
| $N=100$, $m=16$, $k=0$ | (9.8, 24.1) | (9.4, 24.7) | (9.8, 24.2) | (9.8, 24.1) | (9.4, 24.7) | (9.8, 24.2) |
| $N=100$, $m=12$, $k=8$ | (9.8, 24.1) | (9.4, 24.7) | (9.8, 24.2) | (8.5, 25.7) | (8.1, 26.3) | (8.4, 25.8) |
| $N=100$, $m=4$, $k=24$ | (9.8, 24.1) | (9.4, 24.7) | (9.8, 24.2) | (3.3, 31.9) | (2.9, 32.4) | (3.3, 31.9) |
| $N=200$, $m=32$, $k=0$ | (11.4, 21.6) | (11.2, 21.8) | (11.4, 21.6) | (11.4, 21.6) | (11.2, 21.8) | (11.4, 21.6) |
| $N=200$, $m=24$, $k=16$ | (11.4, 21.6) | (11.2, 21.8) | (11.4, 21.6) | (9.8, 23.4) | (9.6, 23.7) | (9.8, 23.4) |
| $N=200$, $m=8$, $k=48$ | (11.4, 21.6) | (11.2, 21.8) | (11.4, 21.6) | (3.9, 29.9) | (3.7, 30.2) | (3.9, 29.9) |
| $N=500$, $m=80$, $k=0$ | (13.0, 19.4) | (12.9, 19.5) | (13.0, 19.4) | (13.0, 19.4) | (12.9, 19.5) | (13.0, 19.4) |
| $N=500$, $m=60$, $k=40$ | (13.0, 19.4) | (12.9, 19.5) | (13.0, 19.4) | (10.9, 21.6) | (10.8, 21.7) | (10.9, 21.6) |
| $N=500$, $m=20$, $k=120$ | (13.0, 19.4) | (12.9, 19.5) | (13.0, 19.4) | (4.3, 28.5) | (4.2, 28.6) | (4.3, 28.5) |

Comparison of standard Bayesian and classical interval estimates of a percentile rank with the methods developed in the present paper to capture the additional uncertainty arising from tied scores: Importantly the point estimate of the percentile rank is 16 in all examples, the sample size ($N$) is varied as is the number of tied scores ($k$) and thus also the number scoring below a given score ($m$).

size of 50), for the Bayesian method, the interval estimate of the percentile rank when there are 12 tied scores is (2.5, 35.0) compared to (7.9, 27.9) when there are no ties. Similar differences resulting from variation in the number of ties can be observed for the classical methods.

Fourth, it can be seen that the three methods developed to incorporate the uncertainty arising from ties (Bayesian, Classical, and the Classical mid-*p* variant) yield fairly similar intervals; the degree of convergence increases with the size of the normative sample. Note also that the convergence between the Bayesian and classical methods is particularly close for the classical mid-*p* variant. It can be proved mathematically that convergence for large sample sizes is a property of these methods (Garthwaite & Crawford, 2009). (Table 2 shows that the *standard* Bayesian and classical methods also exhibit convergence, but this is of less interest for present purposes.)

Fifth it can be seen that, in general, there is substantial uncertainty over the percentile rank of a raw score when normative samples are moderate in size. The uncertainty is reduced with larger sample sizes but is still appreciable, particularly when there are a large number of ties. These observations underline that, when working with percentile norms, neuropsychologists should have access to interval estimates of the percentile rank for a case's score, not just a point estimate.

## BAYESIAN VERSUS CLASSICAL (FREQUENTIST) INTERPRETATIONS OF INTERVAL ESTIMATES OF A PERCENTILE RANK

As Antelman (1997) notes, the classical (frequentist) conception of a confidence interval is that "It is one interval generated by a procedure that will

give correct intervals 95% of the time. Whether or not the one (and only) interval you happened to get is correct or not is unknown'' (p. 375). Thus, in the present context, the frequentist interpretation is as follows: ''If we could compute confidence intervals for a large number of normative samples collected in the same way as the present normative sample, about 95% of these intervals would contain the true percentile rank of the score obtained by the case.''

In contrast, the Bayesian interpretation of the interval estimate is that ''there is a 95% probability that the true percentile rank of the score obtained by the case lies within the stated interval'' (as noted, for the Bayesian, probability statements can be attached to parameters). This statement is not only less convoluted but, we suggest, it also captures what a neuropsychologist would wish to infer from an interval estimate. Indeed, as Howell (2002) observes, most psychologists who use frequentist confidence limits probably construe these in what are essentially Bayesian terms.

For this particular problem the two sets of limits converge, and therefore the upshot is that neuropsychologists can place a Bayesian interpretation on the interval estimate of a percentile rank, regardless of whether it was obtained using Bayesian or frequentist methods.



## ONE-SIDED INTERVAL ESTIMATES OF PERCENTILE RANKS

In the foregoing treatment of interval estimates, attention has been limited to two-sided intervals. However, there will be occasions in which a one-sided interval for a percentile rank may be preferred (just as there are situations in which a one-tailed rather than a two-tailed test is appropriate). For example, a neuropsychologist may be interested in whether a patient's score is less extreme than is indicated by the point estimate but not particularly interested in whether the score is even more extreme (or vice-versa). Both the classical and Bayesian methods developed here can easily be adapted to provide a one-sided limit. However, without prior knowledge of which limit is of interest (the situation here, as the aim is to provide intervals for use by others) it is more convenient to generate $100(1 - a + [a/2])$ two-sided intervals which then provide $100(1 - \alpha)$ one-sided lower and upper limits. For example, if a 95% lower limit on the percentile rank is required, then a 90% two-sided interval can be generated: the user then simply disregards the upper limit of the two-sided interval and treats the lower limit as the desired one-sided 95% limit.

## PRECISION WHEN REPORTING PERCENTILE RANKS

The precision with which percentile ranks are reported is a relatively minor issue in comparison to others dealt with in earlier sections. However, the issue does warrant a brief discussion. Our view is that one size does not fit all. That is, for percentile ranks that are not at the extremes (i.e., percentile ranks greater than 5 and less than 95), it is sufficient to report these as integers. Greater precision than this adds nothing and is potentially a distraction. On the other hand, for percentile ranks that *are* extreme, reporting to one decimal place has clear advantages. For example, it is useful to make a distinction between a patient with a score that is estimated to be exceeded by only one in a thousand members of the general adult population and a patient whose score is estimated to be exceeded by one in a hundred. By adopting

the present proposal this form of distinction can be made while simultaneously preserving the convenience and simplicity of using integer values for the percentile ranks of less extreme scores.

To avoid any potential confusion, it is not the case that the normative sample must be 1000 or greater for this proposed convention to be useful. For example, suppose that the normative sample consists of 500 individuals and that only one member obtained a score lower than that obtained by a patient (suppose also, in the interests of simplicity, that none obtained the same score as the patient). Then the estimated percentile rank is 0.2; this is more informative (and more a cause for concern) than knowing only that the score is below the 1st percentile.

## TWO RADICALLY DIFFERENT CLASSES OF INTERVAL ESTIMATES FOR TEST SCORES

The interval estimates on the percentile ranks provided here, using either Bayesian or classical methods, should not be confused with interval estimates that attempt to capture the effects of measurement error on an individual's score. These latter intervals, which usually assume normally distributed scores, are based on classical *test* theory.

When these latter interval estimates are used, the neuropsychologist is posing the question "ignoring the error in estimating the population mean, standard deviation and reliability of the test, and assuming a normal distribution of scores, how much uncertainty is there over an individual's score as a function of *measurement* error in the instrument?" In contrast, when using the interval estimates presented in the present paper, the concern is solely with the score in hand and no assumptions concerning the distribution of raw scores need be invoked (other than the ubiquitous assumption that the underlying scores are continuous, albeit measured as integers). The more concrete question posed is "how much uncertainty is there over the standing (i.e., percentile rank) of the score the case obtained as a function of error in using a normative sample to estimate its standing rather than directly establishing its standing in the normative population?" That is, the limits set out here do not address the issue of what score a case might obtain on another occasion, or on a parallel version of the test, but simply provide an interval estimate of the percentage of the normative population that would score below the score obtained by the case.

## A COMPUTER PROGRAM FOR THE GENERATION OF POINT AND INTERVAL ESTIMATES OF PERCENTILE RANKS

A computer program (Percentile_Norms_Int_Est.exe) was written for PCs in the Delphi programming language to accompany this paper. It can be downloaded (either as an uncompressed executable or as a zip file) from the first author's web pages (at www.abdn.ac.uk/~psy086/dept/Percentiles_Int_Est.htm). The program implements all of the suggested reporting standards set out in the final section of this paper, including the provision of interval estimates based on the Bayesian and classical methods developed here. It can be used to generate tables of percentile

norms (point and interval estimates) from the frequency distribution of raw scores for a normative sample.

To use the program the user needs to first prepare a text file consisting of two columns (using Notepad or a similar application to enter these data would be the easiest option; if using a word processor the file must be saved as a raw text file). The first column should list the raw scores in ascending order, and the second the corresponding number of participants from the normative sample obtaining each of these raw scores. The columns must be separated by a blank space.

The user is prompted to enter the overall sample size of the normative sample ($N$) and the minimum and maximum raw scores. It is important to note that it is assumed that the raw scores are expressed as integers. (Note also that if, within the range of the minimum and maximum scores, some scores are unobtainable or are obtainable but were not obtained by any members of the normative sample, then these must still be entered and accompanied by a zero in the second column). The user then selects their preferred method of generating the interval estimates for the percentile ranks, Bayesian, classical, or the classical mid-$p$ variant. Our preference is for the Bayesian method but the classical methods are made available for those who are wedded to the classical approach to inference (as noted, typically the limits converge in any case). Finally, identifying information on the test and/or normative sample can be entered (in a field entitled ''User's Notes'') for future reference (these notes are reproduced in the program output).

The output from the program consists of a table of six columns: the first column records the raw scores, the second the percentile rank (i.e., point estimate) for these raw scores calculated using formula 1 (i.e., definition C). The third and fourth columns list the corresponding 95% lower and upper endpoints for the interval estimates of the percentile rank. The fifth and sixth columns present the corresponding 90% lower and upper endpoints (as noted, these can serve as the 95% one-sided endpoints if these are required). The output of the program can be viewed on screen, saved to a file, and/or printed.

Although this program can be used to provide point and interval estimates of percentile ranks for new (i.e., previously unpublished) normative data, we hope that it will also be used with existing (previously published) normative data sets. There are many such datasets in neuropsychology and, in many cases, the effort expended in their collection was clearly considerable. We considered writing a second program that would provide interval estimates using existing (i.e., published) tables of percentile ranks rather than raw frequency distributions as input. This would have allowed neuropsychologists to obtain interval estimates for percentile norms without requiring access to the raw frequency distribution (and allow them to apply definition C when calculating percentile ranks regardless of the definition used to generate the original norms). However, we decided that there were a number of problems with this.

First, and as noted, there are three definitions of percentiles in current usage and the definition applied to obtain percentile ranks is often unspecified; this would preclude application of the reporting standards, including the provision of the interval estimates. Second, noise would be introduced (particularly when calculating interval estimates for extreme scores) because of the common practice of rounding percentile ranks to integers; our proposed standards also recommend that

non-extreme percentile ranks should be rounded and reported as integers but the calculations performed to provide the interval estimates are based on the raw counts.

Finally, note that, on occasion, a neuropsychologist may only require a point and interval estimate of the percentile rank for a single raw score, rather than a full table of percentile norms (for example, a neuropsychologist may wish to compare the score of a single-case to a control or normative sample). A companion program, Single_PR_PIE.exe, was written to deal with this scenario: the user need only enter the overall $N$ for the normative sample together with the number of members of the sample obtaining the score of interest and the number scoring below this score.



## CONCLUSIONS AND PROPOSED REPORTING STANDARDS

There are two important but very different sources of uncertainty when using percentile norms in neuropsychology, one avoidable, the other not. The first uncertainty is over how percentiles have been defined, and hence calculated, by the originator of the norms. The second uncertainty stems from the use of a normative sample to represent the normative population. Although this uncertainty is unavoidable its effects can be quantified. The conceptual and numerical analysis of the issues surrounding percentile norms in the foregoing sections of the present paper leads to the following proposed reporting standards for the presentation of percentile norms for neuropsychological test scores:

(1) When presenting tables of percentile norms the definition of a percentile rank used to generate the norms should be made explicit.
(2) Definition C (Equation 1) should be used when calculating percentile norms. Not only is this the method most commonly employed by experts in measurement, it is also consistent with the ubiquitous assumption of an underlying continuous score, and it avoids practical problems. It is to be hoped that even those who may have a preference for either of the two alternative definitions will consider that the advantages of establishing a single definition outweigh these preferences. Furthermore, to our knowledge, this recommendation is not in conflict with existing advice to neuropsychologists, as we are unaware of any previous attempt to develop a set of reporting standards.
(3) When presenting the percentile ranks of raw scores, 95% interval estimates should also be provided. This not only serves the generally useful purpose of reminding users that the percentile ranks are point *estimates* (i.e., it highlights that norms are fallible and thereby counters any tendency to reify the percentile rank), but it also serves the eminently practical purpose of directly quantifying this uncertainty. We further suggest that the methods developed in the present paper should be used to generate these intervals, as they allow for the full range of uncertainties involved. Although they are more complicated to implement than potential alternative approaches this is at no cost either to those developing norms or to end users as the accompanying computer program (see earlier for details) automates the process of obtaining them.

(4) A brief statement concerning the interpretation that can be placed on interval estimates should be provided for the user. This would avoid any potential confusion with intervals that estimate the effects of measurement error on test scores. We assume that the Bayesian interpretation of the intervals featured here is the most relevant to a neuropsychologist's concerns (as noted, a Bayesian interpretation is defensible even if the classical methods are used to obtain the intervals given the convergence between the methods).
(5) Consideration should be given to supplementing 95% intervals with 90% intervals when presenting tables of percentile norms. This would allow users to obtain 95% one-sided lower or upper intervals should they be desired (see earlier section for the rationale and use of such one-sided intervals).
(6) When reporting the point and interval estimates of percentile ranks the number of significant digits recorded should be varied with the extremity of the percentile ranks. For percentile ranks in the range 5 to 95 it is sufficient and more convenient for the end user to round the percentile ranks to the nearest integer whereas, for more extreme scores, the percentile ranks should be reported to one decimal place.

The foregoing reporting standards are straightforward and easy to implement (particularly with the assistance of the computer program). Their adoption when presenting percentile norms would have the additional positive effect of raising awareness among neuropsychologists of the general issues surrounding the use of any set of existing percentile norms.

Finally, the present paper has been concerned with determining the standing of a case's score with regard to those obtained from a healthy normative population. One way of looking at percentile ranks is that they indicate how likely it is that a case's score belongs to this latter population. However, this does not address the question of whether the case's score belongs to another specific population. As noted by a reviewer, if the concern is with determining the probability that a case has a condition of interest (COI) then the percentile rank only provides some of the information required. We would also need to know the percentile rank of a case's score in the population of persons having the COI, the sample sizes from which percentiles were determined, and base rates that give the relative sizes of the normative population and the population with the COI. Bayes' theorem could then be applied to evaluate the probability that the case has the COI. For recent treatments of this issue see Chelune (in press) and Crawford, Garthwaite, and Betkowska (in press).

In closing, it was demonstrated earlier that the uncertainty over the percentile rank of a test score can be considerable, particularly when the normative sample is modest and there are many tied scores. Those providing percentile norms may be tempted to restrict themselves to presenting point estimates because of concerns that accompanying interval estimates might cast doubt on the value of their data. Similarly, neuropsychologists might choose to ignore interval estimates of percentile ranks because they find them unsettling. Jacob Cohen, who did so much to make the case for the use of interval estimates in psychology, was acutely aware of this issue. He noted (Cohen, 1994) that interval estimates are "... rarely to be found in the literature. I suspect that the main reason they are not reported is that they are so

embarrassingly large!'' (p. 1002). The crucial point, of course, is that ignoring the uncertainty does not make it go away: It is better to acknowledge its existence, both to ourselves and to others, and take solace in the fact that its effects can at least be quantified.

## APPENDIX

### A.1. Classical confidence interval

Suppose that $l$ controls have larger scores than the case, $k$ controls have the same score as the case, and $m$ controls have a lower score. Define a ''success'' to be a score for a control that is less than the score of the case were ties eliminated by a tie-breaking mechanism that breaks ties randomly. The number of successes is thus $m$, $m+1, \ldots,$ or $m+i$ and each of these possibilities is equally likely. Assume we require an equal-tailed 95% confidence interval.

We have that $n=k+l+m$. When $p$ is the probability of success, the probability of $x$ successes is

$$\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}, \tag{2}$$

as the number of successes has a binomial distribution. Hence, the probability of $m+i$ successes or fewer is

$$P_i = \sum_{x=0}^{m+i}\left(\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}\right). \tag{3}$$

For $i=0, 1, 2, \ldots, k$, each $P_i$ has probability $1/(k+1)$ of being the appropriate $P_i$. Hence, if $p_b$ is the upper limit of the equal-tailed 95% confidence interval, then

$$\sum_{i=0}^{k} P_i/(k+1) = 0.025.$$

when $p_b$ is substituted for $p$ in Equation (3). That is,

$$0.025 = \sum_{i=0}^{k}\sum_{x=0}^{m+i}\left(\frac{n!}{x!(n-x)!}p_b^x(1-p_b)^{n-x}\right)/(k+1). \tag{4}$$

Similarly, for the lower endpoint we will seek a value $p_a$ that satisfies

$$0.025 = \sum_{i=0}^{k}\sum_{x=m+i}^{n}\left(\frac{n!}{x!(n-x)!}p_a^x(1-p_a)^{n-x}\right)/(k+1). \tag{5}$$

When $m+k=n$, the right-hand side of (4) may exceed 0.025 even when $p_b \to 1$, in which case we set $p_b$ equal to 1. Similarly, when $m=0$, the right-hand side of (5) may exceed 0.025 even when $p_a \to 0$, in which case we set $p_a$ equal to 0.

To illustrate the method we use the example provided in the main body of the paper. In a normative sample of size $n = 80$, $k = 4$ members obtained the same raw score as the case, and $m = 10$ obtained a lower score. Thus $x$ ranges from 10 through to 14. The binomial distribution that provides the upper endpoint for the interval estimate is bin(100, 0.2548) because, for this binomial distribution,

$$\frac{1}{5}[P(X \leq 10) + P(X \leq 11) + P(X \leq 12) + P(X \leq 13) + P(X \leq 14)]$$
$$= \frac{1}{5}[.003544 + 0.008281 + 0.017593 + 0.034247 + 0.061497] = 0.025.$$

Similarly, the binomial distribution that provides the lower endpoint is bin(100, 0.0739) as for this distribution,

$$\frac{1}{5}[P(X \geq 10) + P(X \geq 11) + P(X \geq 12) + P(X \geq 13) + P(X \geq 14)]$$
$$= 1 - \frac{1}{5}[P(X \leq 9) + P(X \leq 10) + P(X \leq 11) + P(X \leq 12) + P(X \leq 13)]$$
$$= 1 - \frac{1}{5}[0.929862 + 0.966897 + .985699 + .994324 + .997923] = 0.025.$$

Hence, as recorded in the main body of the paper, the 95% interval estimate of the proportion is (0.0739, 0.2548) and for the percentile rank it is (7.39, 25.48).

### A.2. A mid-*p* variant of the classical interval when there are multiple ties

We first demonstrate the mid-$p$ variant of the classical method for the numerical example used in the previous sections and then go on to provide the general formula and the procedure to deal with extremes.

**Mid-*p* upper limit.** As previously noted for the classical limit with multiple ties we consider

$$\frac{1}{5}[P(X \leq 10) + P(X \leq 11) + P(X \leq 12) + P(X \leq 13) + P(X \leq 14)].$$

The corresponding probability for the mid-$p$ variant with multiple ties is

$$\frac{1}{5}\left[\begin{array}{l}\frac{1}{2}\{P(X \leq 9) + P(X \leq 10)\} + \frac{1}{2}\{P(X \leq 10) + P(X \leq 11)\} + \frac{1}{2}\{P(X \leq 11) \\ + P(X \leq 12)\} + \frac{1}{2}\{P(X \leq 12) + P(X \leq 13)\} + \frac{1}{2}\{P(X \leq 13) + P(X \leq 14)\}\end{array}\right].$$

This can be expressed more succinctly as

$$\frac{1}{5}\left[\frac{1}{2}P(X \leq 9) + P(X \leq 10) + P(X \leq 11) + P(X \leq 12) + P(X \leq 13) + \frac{1}{2}P(X \leq 14)\right]$$

The general formula for the mid-$p$ variant is as follows. Suppose we have $n$ trials with $m$ values below and $k$ values equal. Then, if $m$ is not equal to 0, and $m+k \neq n$, for the mid-$p$ method we want

$$\frac{1}{k+1}\left[\frac{1}{2}P(X \leq m-1) + \sum_{i=m}^{m+k-1} P(X \leq i) + \frac{1}{2}P(X \leq m+k)\right] \tag{6}$$

**Mid-*p* lower limit.** Using the same numeric example, for the classical lower limit we consider $\frac{1}{5}[P(X \geq 10) + P(X \geq 11) + P(X \geq 12) + P(X \geq 13) + P(X \geq 14)]$. The corresponding probability for the mid-$p$ variant is

$$\frac{1}{5}\left[\begin{array}{l}\frac{1}{2}\{P(X \geq 10) + P(X \geq 11)\} + \frac{1}{2}\{P(X \geq 11) + P(X \geq 12)\} + \frac{1}{2}\{P(X \geq 12) \\ + P(X \geq 13)\} + \frac{1}{2}\{P(X \geq 13) + P(X \geq 14)\} + \frac{1}{2}\{P(X \geq 14) + P(X \geq 15)\}\end{array}\right].$$

Again this can be expressed more succinctly as

$$= \frac{1}{5}\left[\frac{1}{2}P(X \geq 10) + P(X \geq 11) + P(X \geq 12) + P(X \geq 13) + P(X \geq 14) + \frac{1}{2}P(X \geq 15)\right]$$

For the general formula suppose we have $n$ trials with $r$ values below and $k$ values equal. Then if $m$ is not equal to 0 and $r+k \neq n$, for the mid-$p$ variant we want

$$\frac{1}{k+1}\left[\frac{1}{2}P(X \geq m) + \sum_{i=m+1}^{m+k} P(X \geq i) + \frac{1}{2}P(X \geq m+k+1)\right] \tag{7}$$

**Mid-*p* variant: Extremes.** When $i=0$, the standard mid-$p$ method sets $P_i$ equal to 1; this is not very satisfactory as it essentially mixes one- and two-tailed intervals. Similarly, when $i=n+1$, rather oddly, the standard mid-$p$ method sets $P_i$ equal to 1. Rather than adopt this approach for extremes, we set $P_0$ equal to $P_1$ and set $P_{n+1}$ equal to $P_n$ when these probabilities are required in Equations (6) and (7).

### A.3. Bayesian credible intervals

We suppose the prior distribution for $p$ is the Jeffrey's beta(0.5, 0.5) distribution. Having observed $x$ successes from $n$ controls, the posterior distribution for $p$ is a beta$(0.5+x, 0.5+n-x)$ distribution.

As $x=m+i$, where $i$ is equally likely to be $0, 1, \ldots,$ or $k$ with equal probability, the posterior distribution is a mixture of $k+1$ beta distributions: beta$(0.5+m+i, 0.5+n-m-i)$ for $i=0, 1, \ldots, k$. Each element of the mixture has probability $1/(k+1)$. Given a value $p^*$, we determine the quantile of the posterior distribution that this corresponds to, as follows.

For $i=0, 1, \ldots, k$ determine the quantile of the beta$(0.5+m+i, 0.5+n-m-i)$ distribution that $p^*$ corresponds to. Let $p_i^*$ denote that quantile. Then the quantile that $p^*$ corresponds to is

$$\left(p_0^* + p_1^* + \cdots + p_k^*\right)/(k+1). \tag{8}$$

Using a search procedure described below, we find values of $p^*$ that correspond to the 0.025 and 0.975 quantiles. If $p_a^*$ and $p_b^*$ denote these quantiles, then $(p_a^*, p_b^*)$ is our 95% equal-tailed credible interval. Note that, unlike the classical method, no modifications are required for extreme values.

To illustrate using the previous example, for this problem we have five (i.e., $k+1$) beta distributions: beta(10.5, 70.5), beta(11.5, 69.5), beta(12.5, 68.5), beta(13.5, 67.5), and beta(14.5, 66.5). For the upper endpoint of the interval, 0.2477 is the 0.996594, 0.991911, 0.982593, 0.965785, and 0.938122 quantile, respectively, of these beta distributions. Also $(0.996594 + 0.991911 + 0.982593 + 0.965785 + 0.938122)/5 = 0.975$ and so $p_b^* = 0.2477$ is the upper endpoint. For the lower endpoint, 0.0786 is the 0.069191, 0.033190, 0.014619, 0.005932, and 0.002225 quantile, respectively, of these five beta distributions. Also $(0.069191 + 0.033190 + 0.014619 + 0.005932 + 0.002225)/5 = 0.025$ and so the lower endpoint of the interval for the proportion is $p_a^* = 0.0786$. Thus, as reported in the text, the interval estimate for the percentile rank of the case's score is (7.86, 24.77).



### A.4. Search procedure to find the endpoints

We first consider how to search for the endpoints of the classical confidence interval; that is, values of $p_b$ and $p_a$ that satisfy Equations (4) and (5) respectively. There are a number of potential search procedures. For example, we could use a start value provided by the standard Clopper-Pearson method. That is, we could compute the Clopper-Pearson interval based on $m + 0.5k$ successes in $N$ trials (i.e., 12 successes in 100 trials in the foregoing worked example) and then successively increase or decrease $p_b$ until it satisfies Equation 4 for the upper endpoint and use the same approach for the lower endpoint. Alternatively we could use a brute force approach and simply specify initial lower and upper bounds for $p_b$ of 0 and 1 (and similarly for $p_a$) and use bisection to successively reset these bounds until the required value of $p_b$ is trapped between them with an acceptable level of precision (i.e., if the initial candidate value of 0.5 for $p_b$ is too large then this candidate value becomes the new upper bound; if it is too small then it becomes the new lower bound etc.). This was the approach adopted here and the details for finding the upper endpoint are set out below:

(1) Let $w$ denote the lower endpoint.
(2) Let $L$ be an upper bound for $w$ and let $U$ be a lower bound.
(3) Initialise the algorithm by putting $L = 0$ and $U = 1$.
(4) Let $p = (L + U)/2$.
(5) Using a routine that gives the cumulative distribution function of a binomial distribution, determine for bin$(N, y)$ the probabilities for each of the $k+1$ possible values of $x$. Let $A$ denote the average of these $k+1$ quantities.
(6) If $A$ is greater than 0.025 then we have a new upper bound and we put $U = p$. If $A$ is less than or equal to 0.025 we have a new lower bound and we put $L = p$.
(7) We repeat steps 4 to 6 a number of times (in the present case 15 times gives sufficient precision).
(8) Using the final values of $L$ and $U$, we set the lower endpoint equal to $(L + U)/2$.

The procedure for finding the lower endpoint is identical to that above but with 0.025 replaced with 0.975.

The same search procedure is used for the Classical Mid-*p* method, except that in step 5 of the algorithm $A$ is set equal to the quantity in expression (7) when searching for the lower limit and to the quantity in expression (6) in a search for the upper limit. To search for the endpoints of the Bayesian credible interval, in step 5 the quantity in expression (8) is equated to $A$.

## REFERENCES


Agresti, A., & Coull, B. A. (1998). Approximation is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*, 119–126.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

Berry, G., & Armitage, P. (1995). Mid-*p* confidence intervals: A brief review. *The Statistician*, *44*, 417–423.

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

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative studies in pediatric neuropsychology. *Child Neuropsychology*, *13*, 528–538.

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

Chelune, G. J. (in press). Evidence-based research and practice in clinical neuropsychology. *The Clinical Neuropsychologist*.

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

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. (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. (2009). Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *The Clinical Neuropsychologist, 23*, 193–204.

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., Garthwaite, P. H., & Betkowska, K. (in press). Bayes theorem and diagnostic tests in neuropsychology: Interval estimates for post-test probabilities. *The Clinical Neuropsychologist*.

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.

Garthwaite, P. H., & Crawford, J. R. (2009). *Inference for a binomial proportion in the presence of ties*. Manuscript in preparation.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

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

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2nd ed.). Chicago: Open Court.

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

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.

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.

Phillips, L. D. (1973). *Bayesian statistics for social scientists*. London: Nelson.

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.