**Author Queries**

*Journal title:*  **ASMNT**

*Article Number:*  **10.1177/1073191116669784**

Dear Author/Editor,

Greetings, and thank you for publishing with Sage. Your article has been copyedited, and we have a few queries for you. Please respond to these queries when you submit your changes to the Production Editor.

Thank you for your time and effort.

NOTE: Please answer the queries by editing within the article: follow the AQ hyperlink below and edit the text directly. Add a "New Comment" (in the "Review" tab), if the correction cannot be done directly or for further queries.

Below, please tick the boxes to the right of the query, to indicate the query has been dealt with.

Please assist us by clarifying the following queries:

|  |  |  |
| --- | --- | --- |
|  | Please confirm that all author information, including names, affiliations, sequence, and contact details, is correct.  Corrections have been made to the affiliation of the second author. Some additional corrections have been made to improve grammar and correct typographic errors. |  |
|  | Please review the entire document for typographical errors, mathematical errors, and any other necessary corrections; check headings, tables, and figures. |  |
|  | Please confirm you have reviewed this proof to your satisfaction and understand this is your final opportunity for review prior to publication. |  |
|  | Please confirm that the Funding and Conflict of Interest statements are accurate. |  |
| [AQ 1](#aq1) | Please provide complete reference details for Lezak, 1976 or allow us to delete the citation. |  |
| [AQ 2](#aq2) | Please provide complete reference details for Rey, 1964 or allow us to delete the citation. |  |
| [AQ 3](#aq3) | Please provide complete reference details for Benton et al., 1983. |  |
| [AQ 4](#aq4) | Please provide complete reference details for Hooper, 1952. |  |
| [AQ 5](#aq5) | Please provide complete reference details for Wechsler, 2001. |  |
| [AQ 6](#aq6) | Please provide complete reference details for Bulmer, 1979 |  |
| [AQ 7](#aq7) | per text citation, figures 3a AND 3b HAVE BEEN RENAMED TO FIGURE 4, PLEASE VERIFY. ALSO NOTE THAT THE FIGURE CAPTION IS CHANGED. |  |
| [AQ 8](#aq8) | Delis et al., 1987 is not cited in the text. Please indicate where a citation should appear or allow us to delete the reference. |  |
| [AQ 9](#aq9) | Spreen & Strauss, 1991 is not cited in the text. Please indicate where a citation should appear or allow us to delete the reference. |  |

How Small Is Big: Sample Size   
and Skewness

Assessment

1–7

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/1073191116669784

asm.sagepub.com



1University of Southern Queensland, Ipswich, Queensland, Australia

2USQ Centre for Psychological Assessment, Springfield Central, Queensland, Australia

**Corresponding Author:**

Adina Piovesana, University of Southern Queensland, 11 Salisbury Road, Ipswich, Queensland 4305, Australia.

Email: adina.piovesana2@usq.edu.au

Adina Piovesana1 and Graeme Senior2

Abstract

Sample sizes of 50 have been cited as sufficient to obtain stable means and standard deviations in normative test data. The influence of skewness on this minimum number, however, has not been evaluated. Normative test data with varying levels of skewness were compiled for 12 measures from 7 tests collected as part of ongoing normative studies in Brisbane, Australia. Means and standard deviations were computed from sample sizes of 10 to 100 drawn with replacement from larger samples of 272 to 973 cases. The minimum sample size was determined by the number at which both mean and standard deviation estimates remained within the 90% confidence intervals surrounding the population estimates. Sample sizes of greater than 85 were found to generate stable means and standard deviations regardless of the level of skewness, with smaller samples required in skewed distributions. A formula was derived to compute recommended sample size at differing levels of skewness.

Keywords

minimum sample sizes, normative data, skewness, psychometrics, psychological assessment

In psychological assessment, interpretations are made by comparing the performances of an individual being tested with a normative sample that is representative of an appropriate population and accommodates relevant demographic characteristics. When clinicians use adequate and representative normative data, they can be confident that the norms are an accurate reflection of the true population parameters. One psychometric issue important to ensuring adequate normative data is the size of the sample (*n*) with the general consensus being “the bigger the better.” However, a great deal of time and resources are required to obtain large normative samples (Crawford & Garthwaite, 2002).

Disciplines such as neuropsychology that utilize psychological tests to detect cognitive impairment frequently construct their measures to have greater item content in the lower range of item difficulty to better evaluate and accommodate varying levels of impairment. Additionally, many neuropsychological tests are subject to floor or ceiling effects resulting in distributions that deviate from a normal distribution (Brooks, Stauss, Sherman, Iverson, & Slick, 2009; Capitani & Laiacona, 2000; Crawford & Garthwaite, 2005; Crawford & Howell, 1998). These skewed distributions are of particular interest because of the predictive and discriminative power they yield (Brooks et al., 2009). Where completion times or numbers of errors are the variables of interest, positively skewed distributions are commonplace and negatively skewed distributions, such as in recognition measures or effort testing, allow for the highest discriminative power at the lower ability level (Mitrushina, Boone, Razani, & D’Elia, 2005).

Cognitive tests have a long history of development by practitioners. Their findings are presented in the research literature and often reflect the development of normative data for populations most relevant to the clinician’s area of interest. Additionally, much valuable information exists as “gray literature” referring to the many and varied types of information in print and electronic form that are not controlled by commercial publishers (Auger, 1998). Rather than serving as a barrier to their adoption, many such measures have become staples in individual psychological test batteries. For example, for many English-speaking practitioners, their first introduction to the Rey Auditory Verbal Learning Test (RAVLT) was in *Neuropsychological Assessment* (Lezak, 1976**[[AQ1]](#raq1)**) which provided André Rey’s (1964**[[AQ2]](#raq2)**) French normative data for manual laborers (*N* = 25), Professionals (*N* = 30), Students (*N* = 47), Elderly Laborers (*N* = 15), and Elderly Professionals (*N* = 15). The limitations and lack of representativeness of these normative data were clear and a wealth of normative data now exists for this test provided almost exclusively as publications in scientific journals (Mitrushina et al., 2005). However, until publications began appearing in the research literature providing more extensive and appropriate normative data, users of the RAVLT were restricted to these groups with very small sample sizes.

While some studies report what would appear to be large sample sizes, it is readily apparent that once the effect of stratification by demographic variables which influence test performance are accommodated, the sample sizes for each cell may become less impressive. For example, the children’s normative data published by Forrester and Geffen (1991) for the RAVLT had a total sample size of 80 but once these norms were stratified by four age groups and two genders, the resulting *n* for each cell was reduced to only 10.

So, what constitutes an adequate sample size? This issue seemed to be addressed in the standardization of the Wechsler Memory Scale–Revised (Wechsler, 1987): Page 43 of the manual indicates “Fifty cases are considered sufficient to provide stable estimates of the population mean (Guilford, 1965; Hays, 1963).” Review of these references, however, reveals no evidence to support this assertion. Hays (1963) does not indicate a specific minimum sample size but presents a method for determining the sample size required to provide a specific level of accuracy in population standard deviation units at a designated level of confidence utilizing the relationship between the standard error of the mean and sample size. Guilford (1965) utilized a worked example, the inkblot test data, with a sample size of 50, but makes no reference to 50 as the sample size necessary to provide stable population parameters. Indeed, when discussing the standard error of a standard deviation, he refers to sample sizes of less than 100 as “small.”

Mitrushina et al. (2005) echo this assertion on Page 19 of their seminal compilation of normative data:

Although there is a controversy in the field as to what constitutes a sufficient sample size to ensure precision of estimates of test psychometric properties . . . , a sample size of 50 is typically viewed as adequate (Crawford & Howell, 1998).

However, Crawford and Howell (1998) were not advocating a sample size of 50, but instead were recommending that their Crawford–Howell modified *t* test be used on small sample sizes instead of the conventional *z* scores noting that: “ . . . the modified *t*-test be used with an *n* of less than 50; with larger sample sizes either method is more rapid” (p. 485).

Bridges and Holler (2007) attempted to determine optimal sample sizes for normative studies. Their examination of pediatric norms first consisted of calculating confidence intervals and their equivalent *z* scores for four neuropsychological tests: the Boston Naming Test, RAVLT, Hooper Visual Organization Test (HVOT), and the Rey Complex Figure. These authors found that the confidence intervals around the normative sample means varied widely, especially when the normative data had a small *n*. They then recalculated the confidence intervals for the same pediatric normative studies, but with different sample sizes (*n* = 5, 10, 25, 50, 100, 200, 300, and 500) concluding that

Fewer than 50 subjects results in confidence intervals that are deemed too large to be of clinical utility to neuropsychologists. Alternatively, normative studies having more than 75 subjects per group may not significantly decrease the width of the confidence interval. (Bridges & Holler, p. 537)

These authors noted that their study focused primarily on the effect sample size has on normative sample means, and highlighted that sample standard deviations are also affected by sample size. They also acknowledged that the calculated confidence intervals around their normative means were based on a normal distribution (Bridges & Holler, 2007) raising the concern that their optimal sample size recommendations may not apply to tests with highly skewed distributions.

In the light of these issues, the minimum recommended sample size to generate stable means and standard deviations for normative purposes remains uncertain particularly where tests with highly skewed distributions are involved. It was for this purpose that the current study was conducted.

Method

Participants

Data for this study were sourced from a database of normative studies conducted in the Department of Psychology at the University of Southern Queensland (USQ) between 1996 and 2012. All studies received ethics approval through the USQ Human Research Ethics Committee in their respective years with all participants providing informed consent. Participants were generally from regional Southeast Queensland or metropolitan Brisbane areas and had volunteered to participate in studies designed to establish Australian norms for a number of psychological tests. Participants were recruited through social and occupational networks and were excluded if they reported a history of neurologic or psychiatric conditions. Those participants who had corrected hearing or vision wore their glasses or hearing aids during testing. The sample sizes and participant demographic characteristics are provided in Table 1. The samples did not differ in their mean age or education with the exception of the Wechsler Test of Adult Reading (WTAR) sample which was on average 3.8 years older than the other samples, and the Trail Making Test (TMT) sample which had an average of 0.6 years more education.

Measures

Tests were selected for inclusion in the study if 250 or more normative cases were available. This resulted in the retention of data for seven tests contributing 12 measures: the Controlled Oral Word Association Test (Benton, Hamsher, & Sivan, 1983**[[AQ3]](#raq3)**); HVOT (Hooper, 1958**[[AQ4]](#raq4)**); Form A of Speed of Comprehension (SOC) and Spot the Word from the Speed and Capacity of Language-Processing Test (Baddeley, Emslie, & Nimmo Smith, 1992); oral and written trials of the Symbol Digit

**Table 1.** Descriptive Statistics for the Twelve Measures.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test | *N* | *M* | *SD* | *Mdn* | Age | Education | Gender | |
| *M* (*SD*) | *M* (*SD*) | M | F |
| COWAT | 935 | 42.11 | 11.7 | 41 | 36.04 (14.17) | 12.78 (2.3) | 383 | 552 |
| HVOT | 272 | 26.43 | 2.55 | 27 | 35.49 (16.05) | 12.5 (2.4) | 150 | 229 |
| SCOLP |  |  |  |  |  |  |  |  |
| SOC | 787 | 67.09 | 18.01 | 67 | 36.89 (15.18) | 12.64 (2.32) | 328 | 459 |
| STW | 783 | 49.16 | 5.26 | 50 | 36.76 (15.07) | 12.66 (2.31) | 326 | 457 |
| SDMT |  |  |  |  |  |  |  |  |
| Oral | 628 | 64.33 | 12.35 | 64 | 35.61 (14.86) | 12.55 (2.34) | 256 | 372 |
| Written | 628 | 55.5 | 10.68 | 56 | 35.61 (14.86) | 12.55 (2.34) | 256 | 372 |
| STROOP |  |  |  |  |  |  |  |  |
| Word | 728 | 98.22 | 16.36 | 100 | 36.76 (15.07) | 12.66 (2.31) | 326 | 457 |
| Color | 728 | 74.3 | 12.02 | 74 | 36.76 (15.07) | 12.66 (2.31) | 326 | 457 |
| Color–Word | 728 | 45.33 | 10.73 | 50 | 36.76 (15.07) | 12.66 (2.31) | 326 | 457 |
| TMT |  |  |  |  |  |  |  |  |
| Part A | 507 | 25.95 | 9.2 | 24 | 35.66 (15.17) | 13.10 (2.37) | 212 | 295 |
| Part B | 507 | 59.61 | 22.69 | 55 | 35.66 (15.17) | 13.10 (2.37) | 212 | 295 |
| WTAR | 389 | 37.32 | 8.05 | 39 | 40.19 (14.12) | 12.82 (2.5) | 168 | 221 |

*Note.* COWAT = Controlled Oral Word Association Test; HVOT = Hooper Visual Organization Test; TMT = Trail Making Test; Part A = TMT Subtest Part A; Part B = TMT Subtest Part B; SCOLP = Speed and Capacity of Language-Processing Test; STROOP = Stroop Color and Word Test; STW = Spot the Word; SOC = Speed of Comprehension; SDMT = Symbol Digits Modalities Test; WTAR = Wechsler Test of Adult Reading.

Modalities Test (Smith, 1973); the Color, Word, and Color-Word trials of the Stroop Color and Word Test (STROOP; Golden, 1978); Parts A and B of the TMT (Reitan, 1979); and the WTAR (Wechsler, 2001**[[AQ5]](#raq5)**). All tests were administered according to their standardized instructions from their respective manuals.

Analyses

The Statistical Package for the Social Sciences (SPSS) version 22 was used for all analyses. All test distributions were analyzed for skewness and are presented in Table 2 in descending order. Skewness coefficients were computed using the adjusted Fisher–Pearson coefficient of skewness and classified as normal (−0.5 to +0.5), moderately skewed −1.0 to −0.5, +0.5 to +1.0, or highly skewed (<−1.0 or >+1.0) based on classifications developed by Bulmer (1979**[[AQ6]](#raq6)**). Nonnormality of the distributions was determined using the Kolmogorov–Smirnov statistic with all distributions reflecting significant differences from normality.

The 12 measures chosen for this study ranged in degree of skewness. Trail Making Test Subtest A (TMTA) and Trail Making Test Subtest B were both highly positively skewed, whereas the HVOT was highly negatively skewed. Two moderately positively skewed tests (STROOP Color Word; Controlled Oral Word Association Test) and two moderately negatively skewed tests (Spot the Word; WTAR) were also included. Despite the classification of the Symbol Digits Modalities Test (Oral and Written), STROOP Word and Color, and SOC as “Normal,” the significant Kolmogorov–Smirnov statistic which has high power in large samples (Pallant, 2010), indicates that the distributions of these five measures differ significantly from a normal distribution. In order to interpolate the sample

**Table 2.** Tests of Normality for 12 Measures.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | Skewness | *SE* | Classificationa | K-S statisticb |
| TMTB | 1.70 | 0.11 | High | 0.12 |
| TMTA | 1.40 | 0.11 | High | 0.12 |
| STROOP CW | 0.70 | 0.09 | Moderate | 0.07 |
| COWAT | 0.60 | 0.08 | Moderate | 0.05 |
| SDMT Oral | 0.20 | 0.10 | Normal | 0.04 |
| SDMT Written | −0.07 | 0.10 | Normal | 0.05 |
| STROOP Word | −0.09 | 0.09 | Normal | 0.06 |
| STROOP Color | −0.11 | 0.09 | Normal | 0.04 |
| SOC | −0.11 | 0.09 | Normal | 0.03 |
| STW | −0.70 | 0.09 | Moderate | 0.07 |
| WTAR | −0.81 | 0.12 | Moderate | 0.09 |
| HVOT | −1.64 | 0.13 | High | 0.12 |

*Note. SE* = standard error; TMTB = Trail Making Test Subtest B; TMTA = Trail Making Test Subtest A; STROOP CW = Stroop Color Word; COWAT = Controlled Oral Word Association Test; SDMT = Symbol Digits Modalities Test; SOC = Speed of Comprehension; STW = Spot the Word; WTAR = Wechsler Test of Adult Reading; HVOT = Hooper Visual Organization Test.

aClassification refers to Bulmer’s (1979) classifications. bAll Kolmogorov–Smirnov (K-S) statistics were significant at *p* < .05.

sizes necessary to generate stable population parameters for any given level of skewness, the TRENDLINES function of Microsoft Excel was employed to examine which nonlinear polynomial had the “best fit” in terms of accounting for the most variance (*R*2).

Results

Cases were randomly drawn with replacement from each of the 12 measures producing multiple sample sizes of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. For each sample size, the process was completed five times and the average mean and standard deviation calculated. Average standard

**Table 3.** Ninety Percent Confidence Intervals for Population Means and Standard Deviations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test | *SEM* | 90% Confidence intervals | | | |
| *M* | | *SD* | |
| Lower | Upper | Lower | Upper |
| TMTB | 1.01 | 55.95 | 61.27 | 20.61 | 24.83 |
| TMTA | 0.41 | 25.28 | 26.63 | 8.45 | 10.00 |
| STROOP CW | 0.40 | 44.67 | 45.99 | 10.07 | 11.34 |
| COWAT | 0.38 | 41.49 | 42.74 | 11.07 | 12.29 |
| SDMT Oral | 0.49 | 63.52 | 65.14 | 11.67 | 12.97 |
| SDMT Written | 0.43 | 54.79 | 56.21 | 10.05 | 11.31 |
| STROOP Word | 0.61 | 97.22 | 99.22 | 15.57 | 17.24 |
| STROOP Color | 0.45 | 73.56 | 75.04 | 11.43 | 12.59 |
| SOC | 0.64 | 66.04 | 68.14 | 17.37 | 18.65 |
| STW | 0.19 | 48.85 | 49.47 | 4.97 | 5.56 |
| WTAR | 0.41 | 36.65 | 38.00 | 7.48 | 8.57 |
| HVOT | 0.13 | 26.25 | 26.67 | 2.28 | 2.80 |

*Note.* *SEM* = standard error of mean; TMTB = Trail Making Test Subtest B; TMTA = Trail Making Test Subtest A; STROOP CW = Stroop Color Word; COWAT = Controlled Oral Word Association Test; SDMT = Symbol Digits Modalities Test; SOC = Speed of Comprehension; STW = Spot the Word; WTAR = Wechsler Test of Adult Reading; HVOT = Hooper Visual Organization Test.

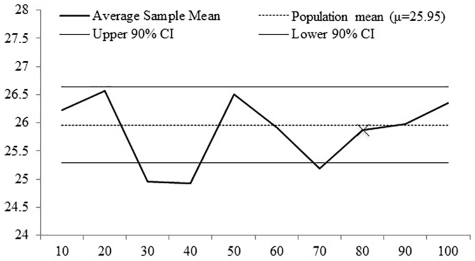
**Table 4.** Descriptive Statistics for TMTA and SCOLP SOC as a Function of Sample Size.

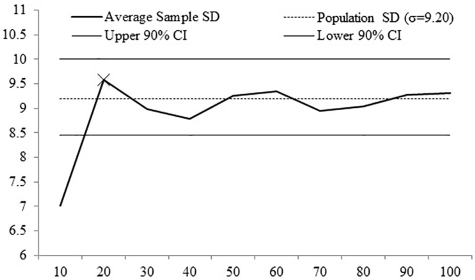
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *n* | | | | | | | | | |
| 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
| *TMTA* (μ = 25.95; σ = 9.20) | | | | | | | | | | |
| *M* | 26.22 | 26.57 | 24.96 | 24.93 | 26.5 | 25.92 | 25.18 | 25.87 | 25.98 | 26.35 |
| *SD* | 7.01 | 9.59 | 8.98 | 9.30 | 9.26 | 9.34 | 8.94 | 9.05 | 9.36 | 9.32 |
| *SOC* (μ = 67.09; σ = 18.01) | | | | | | | | | | |
| *M* | 67.96 | 66.6 | 68.10 | 66.21 | 67.08 | 69.28 | 66.29 | 67.60 | 67.54 | 67.99 |
| *SD* | 18.22 | 15.51 | 18.92 | 17.93 | 18.29 | 17.47 | 18.35 | 18.49 | 18.49 | 17.72 |

*Note. N* = sample size; TMTA = Trail Making Test Subtest A; SOC = Speed of Comprehension; SCOLP = Speed and Capacity of Language-Processing Test.

deviations were computed by averaging the variances and taking the square root of the result. For the purposes of this study, the mean and standard deviation of the entire sample for each measure represent the population estimates of μ and σ respectively, to which the sample statistics for each of the drawn samples were compared. To account for sampling error, 90% confidence intervals were calculated for the population mean for each measure using the standard error of the mean. Ninety percent confidence intervals for the population standard deviation were calculated for each measure using SPSS Bootstrap (IBM Corp, 2013) generating 1,000 random subsamples. The resulting 90% confidence intervals for the population statistics for each measure are presented in Table 3.

The minimum sample size necessary to generate stable means and standard deviations was determined by the sample size at which all subsequent sample means and standard deviations fell within the 90% confidence interval for the entire sample. Table 4 presents these data for two measures, TMTA and SOC, and are depicted in Figures 1 and 2, respectively.





**Figure 1.** Trail Making Test Subtest A (a) sample means with 90% confidence intervals (CIs) and (b) sample standard deviations with 90% CIs.



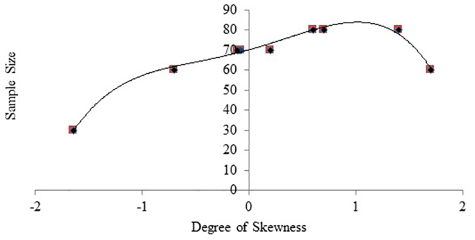


**Figure 2.** Speed of Comprehension (a) sample means with 90% confidence intervals (CIs) and (b) sample standard deviations with 90% CIs.

**Table 5.** Minimum Sample Sizes for 12 Measures.

|  |  |  |  |
| --- | --- | --- | --- |
| Test | Sample size required | | |
| *M* | *SD* | Minimum |
| TMTB | 60 | 50 | 60 |
| TMTA | 80 | 20 | 80 |
| STROOP CW | 80 | 70 | 80 |
| COWAT | 80 | 80 | 80 |
| SDMT Oral | 70 | 50 | 70 |
| SDMT Written | 70 | 70 | 70 |
| STROOP Word | 70 | 70 | 70 |
| STROOP Color | 70 | 70 | 70 |
| SOC | 70 | 40 | 70 |
| STW | 50 | 60 | 60 |
| HVOT | 30 | 30 | 30 |

*Note.* TMTB = Trail Making Test Subtest B; TMTA = Trail Making Test Subtest A; STROOP CW = Stroop Color Word; COWAT = Controlled Oral Word Association Test; SDMT = Symbol Digits Modalities Test; SOC = Speed of Comprehension; STW = Spot the Word; HVOT = Hooper Visual Organization Test.



**Figure 3.** Minimum sample sizes necessary to generate stable means and standard deviations as a function of skewness.

For TMTA, which has a highly positively skewed distribution (γ = 1.40), estimates of the population mean (Figure 1a) do not remain within the 90% confidence interval until a sample size of 80. A stable standard deviation, however, is obtained with a sample size as small as 20. Consideration of both of these parameters indicates a minimum sample size for this measure of 80.

For SOC, with a relatively normal distribution (γ = −0.11), stable means (Figure 2a) and standard deviations (Figure 2b) are obtained with sample sizes of 70 and 40, respectively. To ensure stable estimates of both parameters, a minimum sample size of 70 is recommended.

In this way, the minimum sample size needed for each measure was determined by considering the sample sizes necessary to obtain both a stable mean and standard deviation and these are presented in Table 5 for 11 of the 12 measures. The relationship between skewness and minimum sample size is depicted in Figure 3. Due to the clearly nonlinear nature of the relationship, higher order polynomials were examined to determine the equation that accounted for the greatest amount of variance. The data for the WTAR were excluded from this analysis to provide a measure to examine the accuracy of the derived equation.

A fourth-order polynomial was selected as the best fit for the data and accounted for 99% of the variance (*R*2 = .99), in comparison with linear (*R*2 = .44), quadratic (*R*2 = .92), and cubic (*R*2 = .94) functions. The formula to calculate optimum sample size (*y*) needed for tests with varying degrees of skewness (*x*) was

 (1)

The accuracy of the formula was evaluated with the WTAR, a test with a moderately negatively skewed distribution (γ = −0.81). Table 6 presents the sample means and standard deviations for different sample sizes for the WTAR which were compared with the 90% confidence interval for the population mean (Figure 4a) and standard deviation (Figure 4b). These comparisons indicate sample sizes of 60 and 40 to generate stable means and standard deviations, respectively, with a recommended minimum sample size of 60. Using Formula (1), a sample size of 60 (γ = 60.4) is indicated, yielding the same recommended sample size indicated in the confidence interval analysis, suggesting that the formula can be used to calculate the minimum sample size needed for normative data for psychometric tests at different levels of skewness.

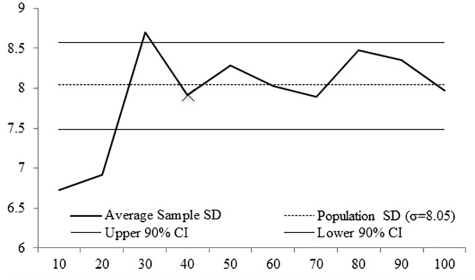
In practice, the application of a minimum recommended sample size for differing levels of skewness, should be then applied to the stratification of normative data based on the influence of relevant demographic variables. For example, in the current study, there was a significant correlation between Part A of the TMT (*N* = 507) and age (*r* = .30, *p* < .001) with Tukey post hoc analysis revealing three homogeneous subsets: 18-49, 49-59, and 60 and over with corresponding sample sizes of 398, 75, and 34 and test mean response times of 24, 30, and 36 seconds, respectively. The skewness coefficient for the 18- to 49-year-old group was 1.14, with 0.61 for 40- to 49-year-olds, and 0.65 for those 60 and over, indicating a greater level of skewness in the youngest group. Application of Formula (1) would indicate a recommended *n* of 84 for the youngest age group, suggesting that the sample of 398 cases is more than sufficient to obtain stable estimates of the population mean and standard deviation. For the 50- to 59-year-olds, the recommended sample size is 80 and indicates that a further 5 cases would be necessary to meet that minimum. The sample of 34 cases aged 60 years or older would be considered inadequate falling well-short of the recommended 81 cases. If the acquisition of this normative sample had been guided by these computations, resources would have been better expended in recruiting more cases for the older groups than the oversampling of 300 cases in the younger group. The recommended sample size for Part A of the TMT for the total unstratified sample was 80 based on a skewness coefficient of 1.4. Despite the differences in skewness between the highly skewed younger and moderately skewed older groups, the differences in recommended sample size were modest and indicated only 1 to 4 more cases over that of the total sample.

**Table 6.** Descriptive Statistics for WTAR as a Function of Sample Size (μ = 37.32; σ = 8.05).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *n* | | | | | | | | | |
| 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
| Sample *M* | 37.32 | 37.01 | 36.98 | 38.70 | 38.16 | 37.54 | 37.41 | 37.06 | 37.05 | 37.47 |
| Sample *SD* | 6.73 | 6.92 | 8.70 | 7.92 | 8.29 | 8.03 | 7.84 | 8.47 | 8.36 | 7.98 |

*Note. N* = sample size; WTAR = Wechsler Test of Adult Reading.





**Figure 4.** WTAR (a) sample means with 90% confidence intervals (CIs) and (b) sample standard deviations with 90% CIs. **[[AQ7]](#raq7)**

Discussion

An important issue when evaluating the appropriateness of normative samples is sample size. While the conventional wisdom is that sample sizes of 50 cases per cell are adequate for producing stable normative data, this seems to have been based on a misperception from a small number of statistical studies. Bridges and Holler (2007) found that sample sizes of 75 or more did not substantially alter mean or standard deviation estimates in normally distributed samples and recognized that this number may differ in skewed distributions. This proposition was evaluated in the current study by examining the minimum sample sizes necessary to estimate population parameters in measures of differing skewness. The sample sizes required for stable measures of central tendency and variance were influenced by skewness and, contrary to the recommended sample size of 50, varied from a minimum of 30 to a maximum of 80 depending on the extent of skewness. For normally distributed data, the sample size required for stable measures is *n* = 70. For negatively skewed distributions, the sample size ranged from *n* = 30 to *n* = 70 and for positively skewed distributions, the sample size varied from *n* = 60 to *n* = 80. An implication of these findings is that sample sizes of greater than 85 reflecting the maximum sample size score generated by the formula should be sufficient to generate stable means and standard deviations regardless of the level of skewness. It is important to note, however, that this minimum size relates to each relevant stratified cell and not the total sample size. For example, in practice, this means that if normative data for a particular test are stratified by four age groups and three education levels and skewness is unknown, each of the 12 cells must have the minimum computed sample size of *n* = 85, totaling 1,020 participants. However, where skewness coefficients are generated for individual stratified cells, fewer participants may be required to generate stable parameters.

This has interesting implications for the status of much normative data. For example, with the normative data provided for the TMT in the *Handbook of Normative Data for Neuropsychological Assessment* (Mitrushina et al., 2005), only 19% of cells met the default minimum recommended sample size of 85. With the recognition of a minimum recommended sample size, researchers may more effectively conduct normative studies by expending their resources and energies in increasing the sample size in inadequate cells rather than continuing to increase numbers in cells with already adequate numbers. This also has implications for census-matched approaches to normative sampling, and researchers would be well advised to oversample ensuring that samples sizes do not drop below the recommended number despite the low frequency of the particular cell in the population.

The findings of this study provide practitioners with an additional level of consideration in determining whether a set of published normative data is appropriate for clinical use. Where the level of skewness is unknown for a particular test, the practitioner can be confident that normative cells with *n*s of 85 or more will result in means and standard deviations that are representative of their respective population parameters. This may mean, however, that only some cells of normative data from published studies may reflect stable population estimates. Where skewness is known and the formula indicates stable estimates with smaller *n*s, then normative data with cells that meet the minimum recommended sample size could also be used with confidence. Use of data that do not meet the minimum recommended sample size introduces an additional level of error variance that may undermine the accuracy of test score interpretations.

Interestingly, despite the influence of skewness on the sample sizes necessary to obtain stable parameters, it is quite rare for skewness to be reported as one of the descriptive statistics in normative studies or test manuals. Provision of this information is recommended as it would permit clinicians to make informed decisions about their selection of normative data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Auger, C. P. (1998). *Information sources in grey literature.* London, England: Bowker-Saur.

Baddeley, A., Emslie, H., & Nimmo Smith, I. (1992). *The Speed and Capacity of Language-Processing Test*. Suffolk, England: Thames Valley Test.

Benton, A. L., Hamsher, K., & Sivan, A. B. (1983). Multilingual aphasia examination (3rd ed.). Iowa City, IA: AJA Associates.

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. doi:10.1080/09297040701233875

Brooks, B. L., Stauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology, 50*, 196-209.doi:10.1037/a0016066

Bulmer, M. G. (1979). Principles of Statistics. New York: Dover.

Capitani, E., & Laiacona, M. (2000). Classification and modelling in neuropsychology: From groups to single case. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 1, 2nd ed., pp. 53-76). Amsterdam, Netherlands: Elsevier.

Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test scores 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. doi:10.1037/0894-4105.19.3.318

Crawford, J. R., & Howell, D. C. (1998). Comparing an individual’s test score against norms derived from small sample. *The Clinical Neuropsychologist, 12*, 482-486. doi:10.1076/clin.12.4.482.7241

Forrester, G., & Geffen, G. (1991). Performance measures of 7- to 15-year-old children on the auditory verbal learning test. *The Clinical Neuropsychologist, 5*, 345-359.

Golden, C. J. (1978). *Stroop Color and Word Test: A manual for clinical and experimental uses*. Chicago, IL: Stoelting.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education*. New York, NY: McGraw-Hill.

Hays, W. L. (1963). *Statistics for psychologists*. New York, NY: Rinehart & Winston.

Hooper, H. E. (1958). The Hooper Visual Organization Test manual. Los Angeles: Western Psychological Services.

IBM Corp. (2013). IBM SPSS Statistics for Windows (Version 22.0). Armonk, NY: Author.

Lezak, M. D. (1976). Neuropsychological Assessment. 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.). Oxford, England: Oxford University Press.

Pallant, J. (2010). *SPSS: Survivor manual* (4th ed.). Berkshire, England:Open University Press.

Reitan, R. M. (1979). *Manual for administration of neuropsychological test batteries for adults and children.* Tucson, AZ: Neuropsychology Press.

Rey, A. (1964). L'examen clinique en psychologie. Paris: Presses Universitaires de France.

Smith, A. (1973). *Symbol Digits Modalities Test: Manual.* Los Angeles, CA: Western Psychological Services.

Wechsler, D. (1987). *Wechsler Memory Scale—Revised manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2001). Wechsler Test of Adult Reading. San Antonio, TX: Psychological Corporation.