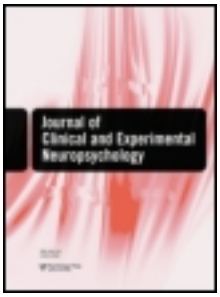

Journal of Clinical and Experimental Neuropsychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ncen20

Heaton, Grant, and Matthews' Comprehensive Norms: An Overzealous Attempt

Robert K. Heaton [a], Igor Grant [a], Charles G. Matthews [a], Philip S. Fastenau [a] & Kenneth M. Adams [a b]

[a] University of Michigan Medical Center

[b] Veterans Administration Medical Center, Ann Arbor, MI

Available online: 04 Jan 2008

To cite this article: Robert K. Heaton, Igor Grant, Charles G. Matthews, Philip S. Fastenau & Kenneth M. Adams (1996): Heaton, Grant, and Matthews' Comprehensive Norms: An Overzealous Attempt, Journal of Clinical and Experimental Neuropsychology, 18:3, 444-448

To link to this article: http://dx.doi.org/10.1080/01688639608409000

Journal of Clinical and Experimental Neuropsychology
1996, Vol. 18, No. 3, pp. 444-448

1380-3395/96/1803-444$12.00

BOOK REVIEW

Heaton, Grant, and Matthews' Comprehensive Norms: An Overzealous Attempt*

Robert K. Heaton, Igor Grant, & Charles G. Matthews (1991). *Comprehensive Norms for an Expanded Halstead-Reitan Battery: Demographic Corrections, Research Findings, and Clinical Applications.* Odessa, Florida: Psychological Assessment Resources, Inc., 344 pp. (No ISBN printed in the volume or in the catalog) $69.00 (U.S.)

Reviewed by Philip S. Fastenau [1] and Kenneth M. Adams [1,2]
[1] University of Michigan Medical Center, [2] Veterans Administration Medical Center, Ann Arbor, MI

Clinical neuropsychologists are always starving for good normative data for established neuropsychological measures. Unfortunately, too many studies (and even manuals) contain too few subjects and/or their samples are not representative of the target population on important demographic variables, especially age and education. This was the problem that Heaton, Grant, and Matthews attempted to address with *Comprehensive Norms for an Expanded Halstead-Reitan Battery* (1991). This practical product arose as a direct result of the authors' 1986 chapter in an edited book (Heaton, Grant, & Matthews, 1986). The project represents a substantial effort on the part of the authors, and it has many commendable qualities. However, the merits are accompanied by significant shortcomings.

Battery and Variable Selection

The authors present data on many measures that are widely used (e.g., Lezak, 1995), such as the Halstead-Reitan Battery (HRB), the Lafayette Grooved Pegboard Test, Digit Vigilance Test, Wisconsin Card Sorting Test (WCST), Peabody Individual Achievement Test (PIAT), and Boston Naming Test (BNT). Their battery also included the Wechsler Adult Intelligence Scale (WAIS), Thurstone Word Fluency Test, Seashore Tonal Memory Test, story and design memory tests, and the Boston Diagnostic Aphasia Exam (BDAE) Complex Ideational Material.

The clinician referencing the tables should carefully note the specific measures being used. First, the IQs and intellectual subtests are from the *original* WAIS, not the 1981 revision. Second, the BNT data are for the *experimental* version, which is based on 85 items, and not the 60-item version that was published in 1983. The authors imply that scores on the published version are equivalent to the experimental version (in terms of percent correct). However, application of these norms to the published edition would require a tedious conversion, and that transformation may not be uniform across age, sex, and education subgroups. It is understandable that norms can take time to collect and that some measures will inevitably be revised in the process; nonetheless, it is a problem that these two measures (WAIS and BNT) are so outdated.

A third concern is raised because the authors provide data for the WCST on perseverative responses only. Although this index may be the one most sensitive to cerebral disorders in some

* Address correspondence to: Philip S. Fastenau, Division of Neuropsychology, Department of Psychiatry, University of Michigan Medical Center, 1500 East Medical Center Drive, 480 MIB, Box 0840, Ann Arbor, MI, 48109-0840, USA.
Accepted for publication: October 26, 1995.

cases, other patients' (dis)abilities may be better characterized by the number of concepts achieved or the total number of errors. It is unfortunate that the authors excluded these other useful indices. The *Comprehensive Norms* volume and scoring software (available from this same publisher) could have profitably included an array of WCST performance aspects in this normative enterprise.

Fourth, the PIAT "raw" scores used in the conversion tables are *centile* scores based on 18-year-olds in the PIAT standardization sample. Thus, the raw score must first be converted to a PIAT standard score and its associated centile rank using those norms, which in turn must be carried through two more conversions required in the *Comprehensive Norms.*

Finally, "Word Fluency," as it appears in the table headings, is Thurstone's *written* fluency for the letters S and C and not the oral, 1-minute "F-A-S" and "C-F-L" trials that are in popular use with adults (e.g., Lezak, 1995).

Samples

The most commendable quality of this book is the size of the samples. Data are presented from almost 500 healthy adults on the WAIS, Grooved Pegboard Test, and HRB (an exception being the Aphasia Screening Test with a sample size of 352). Sample sizes for other measures in this battery were less impressive but useful. Digit Vigilance data were collected on 280 adults. The WCST, Thurstone Word Fluency Test, Seashore Tonal Memory Test, story and design memory, and BDAE Complex Ideational Material were administered to approximately 200 adults. PIAT and BNT samples were closer to 100.

Format

The *Comprehensive Norms* manual looks very attractive. The cover and spiral binding are durable to withstand frequent use. However, index tabs for frequently used sections would have been a helpful aid. Most impressive are the neatly tabulated columns of numbers, with every pair of opposing pages representing a separate age-education-sex group. The first half of the book is devoted to males; the latter half, to females. Within each half, there are six sections representing six education levels, and each of these is subdivided into 10 age levels.

The companion computer software for these norms does its basic job. One can also applaud the authors and publisher for their restraint in not levying the "per use" fees based upon the fiction that such funds go towards test development. This product also requires "Level C" purchaser qualifications; the publisher can be commended for limiting user access (and thus maintaining test security) in these expedient days of test publishing.

Statistical Foundations

In this volume, it is very striking to see the many tables of numbers. Where did all of these numbers come from? One might assume that each table summarizes a sufficiently large subgroup of subjects. However, obscured in the complicated methods lie statistical applications that sharply constrain the utility of these norms.

Excessive conversions

To generate the tables, the raw data underwent many conversions. First, all raw scores were converted to scaled scores (in comparison to the entire sample) with a mean of 10 and a standard deviation of 3. Demographically corrected *T* scores were predicted from these, which were then converted to residual scores (subtracting each subject's actual score from the predicted score). Residual scores were divided by the standard error of estimate (of the residuals) and then converted to *T* scores.

This method of deriving the *T* scores restricts the natural range of variability on every test to a maximum of 20 points (SS = 0 – 19) before the demographic corrections. It also creates an additional step for the clinician that is both unnecessary and cumbersome. That is, the clinician must use Appendix C to look up the scaled score for the total sample and then look up that scaled score in Appendix D to find the corrected *T* score. It is difficult to appreciate the benefit of this intermediate conversion, and the authors offer no convincing argument for doing so. As long as the authors were producing a separate table for each group anyway, it would have been

more parsimonious to keep the conversions limited to that same table, rather than have the user look up an intermediate value.

Too many cells, too few subjects

The authors predicted the demographically corrected *T* scores by use of multiple regression (MR). MR is robust to assumption violations when used for analytic or theoretical investigations. However, when used for prediction, as it is used here, the assumptions require closer attention (Cohen & Cohen, 1983, chap. 3). For multiple regression, it is assumed that, for every raw score, the predicted demographically corrected *T* scores distribute normally, with their mean on the regression line. It is also assumed that the variance of the predicted scores will be identical for every raw score (homoscedasticity). With greater departure from this state of affairs (heteroscedasticity), either a transformation needs to be conducted or additional variables should be added to the equation to improve predictive accuracy. Without these extra steps, predicted scores will be more inaccurate and confidence intervals on a given score will be misleading.

To illustrate these principles graphically, one could plot the predicted scores on the horizontal axis and the residual scores (the predicted score minus the actual raw score) on the vertical axis. The points on the plot should form a horizontal band of uniform width, centered on a residual score of 0. Violations of MR assumptions can be detected through diagnostic patterns on such plots (e.g., Cohen & Cohen, 1983, chap. 3). The authors provide no information regarding the satisfaction of these critical assumptions.

The accuracy of predictions that are based on multiple regression are reflected in the standard error of estimate (SE_e), which is itself sensitive to those same violations of assumptions. The authors failed to provide the SE_es. Without these, we cannot determine the reliability of an individual estimate, nor can we generate and examine the confidence interval for a patient's test score. It is highly probable that the SE_es were much larger than a clinician would want to see. Cohen and Cohen (1983, Chap. 2) showed by case illustration that predicted values rarely yield improvement over the total sample mean when the multiple correlation coefficient is less than .70, even for moderately large samples. Most of the multiple correlations in the *Comprehensive Norms* were at or below this level (Table 5, p. 13). Only six of 54 multiple correlations were sufficiently robust (up to .80), and half of those were only marginally stronger (up to .74).

Even without the SE_es, we can look at this issue in another way, one that might be easier to understand. The authors divided age and education into discrete levels, and then crossed all those levels with each other and with the two levels of sex. Thus, they conducted a three-way analysis of variance (ANOVA), a 10 (Age Group) × 6 (Education Group) × 2 (Sex) ANOVA. The assumptions of MR become identical to ANOVA in this application: homogeneity of variances across cells (which is synonymous with homoscedasticity in MR) and normal distribution of scores for each subgroup. ANOVA is robust to violations when cell sizes are equal in size and sufficiently large. A minimum of 15 subjects per cell is a rule of thumb for theoretical applications; 30 is a good minimum for norms in clinical decisions.

What is the sample size for each subgroup in the *Comprehensive Norms*, from which each table has been derived? The best estimate is the total sample size (107 to 486, depending on the test) divided by the number of subgroups ($10 \times 6 \times 2 = 120$ subgroups). Simple mental arithmetic yields a *maximum* of four people per subgroup, with as few as 1 person or even none at all representing at least some cells for at least some measures!

Thus, interpretations using these norms should begin with statements like, "On the HRB Category Test, this patient performed in the Superior range, having made fewer errors than three other 60-year-old men with 10th-grade education." Or, "Compared to one other well-educated woman her age, this patient's BNT score was (in a word) lower." It is our responsibility to report the adequacy of our comparison groups. Can neuropsychologists feel adept with statements like these in their reports? Would not consumer professionals be shocked by such qualifiers? And how many payors would be

pleased with a report that was riddled with these clauses? This leads to the question: Were all of these subdivisions necessary?

Subdivisions without justification

Data are subdivided by age, sex, and education for all tests under the broad defense that differences on *some* of these demographic variables exist on *some* of the tests. There is *no* support for *all* of these divisions on *all* of these tests. In fact, effects of all three variables (either uniquely or in interaction) were evident on only 14 of 54 measures (Table 5, p. 13). Most (33 of 54) were affected by only two demographic variables; six showed a main effect for only one demographic characteristic; and one showed no age, education, or sex influences at all. Based on this information alone, it is apparent that 120 subdivisions were not appropriate for the vast majority of these measures.

CONCLUSION

Clearly, Heaton, Grant, and Matthews (1991) recognize and address a critical issue in our field: Neuropsychologists need good normative data that are adjusted for demographic influences in order to draw more reliable clinical inferences about our patients' neuropsychometric performances. The authors assembled a relatively strong battery, they administered it to a very good sample, and they presented their data in attractive tables.

Our own informal inquiries have indicated that many of the scientist-practitioners in clinical neuropsychology have embraced this book in an uncritical manner. The strong and unreflective nature of such acceptance of these norms tells us how good an idea this kind of normative project is, in the abstract. Unfortunately, this particular product does not contribute as much as our expectations might lead us to anticipate. The format and marketing are so convincing that few would comb the introductory pages to analyze the test selection quirks and statistical/design problems that abound. Furthermore, Heaton, Grant, and Matthews have such outstanding reputations in the field that few would question the integrity of their work. In fact, at the writing of this review (4 years after the publication of the *Comprehensive Norms*) no one has printed a single critique of the substance of these norms. The only other review in print is an evaluation of the technical features of the software companion (Fuerst, 1993), which is also less than flattering.

There is a way to redeem this work if the authors would consider a few modifications. First, they could combine the data into fewer subgroups, each supported by a minimum of 30 to 50 people. This should be done on a test-by-test basis, reviewing the evidence for age, sex, and education (or IQ) differences for each test.

Second, Appendix C could be cut, and each table in Appendix D could contain actual raw scores; this would eliminate the intermediate conversion that distorts the data and detracts from scoring efficiency. Third, the margins of Appendix D could be scaled directly to *T* scores rather than using scaled scores; this would maintain more precision (potentially a 60-plus point range) instead of artificially reducing the data to a 20-point range.

As a fourth modification, the data could be presented using overlapping cells, as proposed by Pauker (1988). This method uses demographic midpoints so that each table is maximally descriptive of the patient under consideration. This could be especially helpful for those patients who would fall on the border of two age and/or education bands.

In addition, the authors should provide standard errors of estimate for each test and for each demographic subgroup. These should be provided for the *raw* scores and/or for the *T* scores, not for the residual scores. The SE_e could be presented under the test name in each conversion table, together with the mean and standard deviation for that same test for that demographic group. On a practical note, placing tabs on the edge of the pages to delineate the beginnings of major demographic sections would make the tables easier to locate.

Finally, a not-so-insignificant question has arisen since the debut of the *Comprehensive Norms* in 1991. Can data generated on non-patient volunteers provide a useful benchmark to

make statements about brain damage or dysfunction? Reitan and Wolfson (1995) have raised this and some derivative objections to this use of these norms. While their conclusions may be overstated and warrant considerable qualification (see Shuttleworth-Jordan [1995] for a critical review of that study), the importance of cross-validating these norms on a mixed sample that includes both neurologically impaired and neurologically normal subjects cannot be overstated. Simple deviance from neurologically normal performances as a brainless (pun intended) definition of impairment has simply gotten out of hand, along with neuropsychology's love affair with face validity. Consequently, at a minimum, the authors should include the application of the re-formulated norms to demographically diverse patient samples so as to demonstrate whether sensitivity is improved in comparison to non-corrected scores and to show the degree to which clinical decisions are affected.

With a little work, these same data could be reanalyzed and reformatted into a more psychometrically sound and very useful volume. Hopefully, the authors will take this challenge. In the best of possible worlds, the publisher would take its share of the responsibility and exchange purchased copies of the existing book for a substantially discounted copy of the revision. In the meantime, consumers should recognize the very significant limitations of the current volume and, if they use these norms, they should qualify their interpretations with professionally responsible statements such as those described in this review.

REFERENCES

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Fuerst, D. R. (1993). A review of the Halstead-Reitan Neuropsychological Battery norms program. *The Clinical Neuropsychologist, 7*, 96-103.

Heaton, R. K., Grant, I., & Matthews, C. G. (1986). Differences in neuropsychological test performance associated with age, education and sex. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment in neuropsychiatric disorders: Clinical methods and empirical findings* (pp. 100-120). New York: Oxford University Press.

Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an expanded Halstead-Reitan Battery: Demographic corrections, research findings, and clinical applications.* Odessa, FL: Psychological Assessment Resources.

Pauker, J. D. (1988). Constructing overlapping cell tables to maximize the clinical usefulness of normative test data: Rationale and an example from neuropsychology. *Journal of Clinical Psychology, 44*, 930-933.

Reitan, R. M., & Wolfson, D. (1995). Influence of age and education on neuropsychological test results. *The Clinical Neuropsychologist, 9*, 151-158.

Shuttleworth-Jordan, A. B. (1995). *Age and education effects on brain-damaged subjects: "Negative" findings revisited.* Manuscript submitted for publication.