Journal of Clinical and Experimental Neuropsychology

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

Demographic Corrections with Comprehensive Norms: An Overzealous Attempt or a Good Start?

Robert K. Heaton, Charles G. Matthews [b], Igor Grant [a c] & Nanci Avitable [d]

[a] University of California at San Diego

[b] University of Wisconsin

[c] Veterans Administration Medical Center, La Jolla, CA

[d] Denver Public School District

Available online: 04 Jan 2008

To cite this article: Robert K. Heaton, Charles G. Matthews, Igor Grant & Nanci Avitable (1996): Demographic Corrections with Comprehensive Norms: An Overzealous Attempt or a Good Start?, Journal of Clinical and Experimental Neuropsychology, 18:3, 449-458

To link to this article: http://dx.doi.org/10.1080/01688639608409001

Journal of Clinical and Experimental Neuropsychology
1996, Vol. 18, No. 3, pp. 449-458

1380-3395/96/1803-449$12.00
© Swets & Zeitlinger

COMMENT

Demographic Corrections with Comprehensive Norms: An Overzealous Attempt or a Good Start?*

Robert K. Heaton[1], Charles G. Matthews,[2] Igor Grant,[1,3] and Nanci Avitable[4]
[1]University of California at San Diego, [2]University of Wisconsin, [3]Veterans Administration Medical Center, La Jolla, CA, and [4]Denver Public School District

Fastenau and Adams (1996) have provided a detailed and highly critical review of our *Comprehensive Norms* manual. In addition to criticizing the format of the manual and several of the test variables that were included or not included, they have challenged the conceptual and statistical underpinnings of our demographic correction process. They suggest that "many of the scientist-practitioners in clinical neuropsychology have embraced this book in an uncritical manner" (p. 447) unaware of the quirks and serious problems that abound in this work. They surmise that this has occurred because users have failed to read or understand the methods described in the manual, or somehow have been duped by the convincing marketing of the publisher and/or the reputations of the authors. After exposing the numerous presumed deficiencies of the *Comprehensive Norms*, Fastenau and Adams propose that responsible use of them would require such restrictive caveats as to render clinical interpretations meaningless. An example suggested is: "Compared to one other well-educated woman her age, this patient's BNT score was (in a word) lower." (p. 446).

The reviewers conclude by recommending that we redeem our essentially useless work by re-analyzing the normative data base in a more "psychometrically sound" way, and by presenting the results in a format that they consider preferable to the current one. They also suggest that we and our publisher offer to exchange all purchased copies of the current manual for copies of their prescribed revision, at a "substantially discounted" rate. Finally, Fastenau and Adams stress the importance of validating our norms with demographically diverse samples of brain-damaged as well as normal subjects "to demonstrate whether sensitivity is improved in comparison to non-corrected scores and to show the degree to which clinical decisions are affected" (p. 448).

In this article, we address the criticisms and questions raised by Fastenau and Adams. In fact, their review contains numerous misrepresentations of our work, and it is our opinion that their conclusions are unfounded. The following presentation will use the general organization of the Fastenau and Adams review. After briefly clarifying several points about the test variables and format used in our manual, the statistical challenges will be discussed in more detail, since these latter criticisms form the basis of the negative conclusions about the overall usefulness of the *Comprehensive Norms*. Ultimately, the merit of the reviewers' criticisms and conclusions should be determined empirically, rather than by theoretical speculation. To this end, we recount the relevant analyses that are included in our manual, but that are ignored or misrepresented

* The authors thank Carol Demong and Thomas Marcotte for their assistance in the preparation of this manuscript.

Address correspondence to: Robert K. Heaton, Ph.D., UCSD Department of Psychiatry, 3427 Fourth Avenue, San Diego, CA 92103, U.S.A.

Accepted for publication: January 9, 1996.

in the review by Fastenau and Adams. In addition we present some data to help answer the reviewers' questions about how the current norms function with neurologically abnormal as well as normal subjects.

TEST VARIABLES INCLUDED IN THE COMPREHENSIVE NORMS

Fastenau and Adams point out that the *Comprehensive Norms* manual includes norms and demographic corrections for the original (1955) version of the WAIS and an early version of the Boston Naming Test, which they describe as "outdated." As the reviewers surmised, these versions of the tests were included because the studies on which the norms were based were begun before the revised versions were published. However, Fastenau and Adams fail to mention that a WAIS-R supplement to the *Comprehensive Norms* has been published (Heaton, 1992); this provides demographic corrections based upon the data from the WAIS-R standardization sample (Wechsler, 1981). With regard to the 60-item versus 85-item versions of the Boston Naming Test, there is evidence that the scores can be made essentially equivalent by prorating (Thompson & Heaton, 1989); this suggests that the *Comprehensive Norms* can be used with both versions.

The *Comprehensive Norms* includes only one score from the Wisconsin Card Sorting Test (WCST). This is the perseverative response score, which appears to be the most sensitive WCST measure to brain disorders generally and to focal frontal lobe lesions. For neuropsychologists who wish to use other WCST measures in their work, the recently revised WCST Manual and associated software permit conversions from raw scores to demographically corrected *T* scores for all WCST measures that have appropriate distributional properties (Heaton, Chelune, Talley, Kay, & Curtiss, 1993).

The reviewers note that the PIAT "raw" scores used in *Comprehensive Norms* actually are centiles for the oldest age group in the PIAT standardization sample (18-year-olds). This occurred because we had used the centiles in our clinical work, and they were in our data base. It turned out that these centiles for our adult normal sample were readily converted to normally distributed scaled scores. We agree, however, that we could have entered the raw scores for the demographic correction project, and this would have saved a step for users of our norms.

We provided norms for the Thurstone Word Fluency Test, instead of the currently more popular oral fluency tests, because the former are what we had at the time. The *Comprehensive Norms* manual clearly states the nature of the tests that were used.

FORMAT ISSUES

Index Tabs

We agree with the reviewers that it would have been helpful to include in our manual index tabs for frequently used sections.

Subdivision without Justification

The reviewers point out that there is a look-up table in the *Comprehensive Norms* manual for each of 120 Age × Education × Sex combinations. Each of these tables includes scaled score to *T*-score conversions for all 54 test variables. Fastenau and Adams consider this to be inappropriate because performances on many tests are not significantly related to all three demographic characteristics.

The issue here is entirely one of format, not "appropriateness." We could have presented a separate table for each test variable, with breakdowns only for the demographic characteristics that are relevant for that variable. However, the user then would have to consult as many as 54 tables, instead of one in order to score every subject's test protocol. This, we felt, would be much more time-consuming and involve a higher risk of error in the look-up process. Again, there are different ways to organize a table, and the choices are not a matter of right and wrong.

STATISTICAL CONCERNS

Conversion from Raw to Scaled Scores

Fastenau and Adams suggest that the *Comprehensive Norms* manual provides no compelling reason for the conversion of raw scores to scaled scores. The reviewers assert that these conversions restrict the natural range of test scores, reduce the fidelity of test measures, and are time-consuming and cumbersome for clinicians.

Our primary reason for this transformation was to normalize the distribution of the raw test scores, prior to beginning the process of demographic correction. This substantially improved the distribution of the final, demographically corrected *T*-scores. It is true that other transformations might have been as effective or even more effective, but the current transformations resulted in *T*-score distributions that did not deviate significantly from normal for 51 of 54 test variables. Also, the interpretation of scaled scores is more straightforward than that of other possible transformations (e.g., what does the square root of a Category Test error score mean?). Because of their experience with the Wechsler Intelligence Scales, psychologists are used to dealing with scaled scores, and to making raw score to scaled score conversions.

It is true that conversion to scaled scores reduces the possible range of scores to 20, which could reduce the fidelity or sensitivity of some test measures to some degree. However, we doubt that such reductions are large or clinically meaningful. Empirical evidence to the contrary was not provided in the review by Fastenau and Adams.

Finally, the time involved in making these conversions is about 5 to 7 minutes (if the entire, 8-hr test battery is used – less if only components are used). Setting up the tables with raw scores instead of scaled scores would have been possible, but the tables would have been longer and "busier" (more numbers!). We agree that reasonable people could have chosen to do this the other way.

Possible Violations of Multiple Regression Assumption of Homoscedasticity

Fastenau and Adams have described this basic assumption well, but may have exaggerated its likely impact on the demographic correction process. Also, they fail to mention our efforts (described in the manual) to minimize this potential problem and to check for major deviations from homoscedasticity.

Violation of the homoscedasticity assumption in multiple regression does not invalidate the analysis, but rather weakens it to some degree: "The linear relationship between the variables still is captured by the analysis but there is even more predictability if the heteroscedasticity is accounted for, as well" (Tabachnick & Fidell (1989), p. 83; see also p.133).

One of the methods for improving heteroscedasticity, if it exists, is to transform variables to better achieve multivariate normality. Our transformation of the (frequently non-normal) test raw scores to normalized scaled scores should have helped in this regard, and it definitely improved the distribution characteristics of the final *T* scores.

For the clinician, the major issue here is whether the *T*-score conversions work similarly at all levels of age and education (which determine the "predicted" scores). Although Fastenau and Adams did not mention this or any other of our efforts to test the validity of our *T* score conversions (pp. 11-19 in the *Comprehensive Norms*), we did visually and statistically examine means and standard deviations of *T* scores at three age levels (<40, 40-59, and 60+ years) and three levels of education (<12, 12-15, 16+ years). For each test variable, a series of 3 x 3 ANOVAS was performed on initial *T* scores for the Base sample, the Validation sample, and the Total sample of subjects.[1] Although age and ed-

[1] The Base sample is the subject sample used to develop the demographically corrected *T* scores. The Validation sample is a randomly selected subgroup, which was excluded from analyses used in developing *T* scores in order to permit independent checks on several features of the *T*-score conversions bearing on validity. The Total sample is the combined Base and Validation samples.

ucation main effects were negligible in all cases, there were some significant interaction effects. Consequently, we made further adjustments of the prediction algorithms for the Base sample, which resulted in elimination of significant interaction effects for all but two variables. Separate comparisons between the *T* scores of males and females also revealed no significant differences.

For all three samples (Base, Validation, and Total), the final *T*-score distributions were normal for 51 of the 54 test variables. The predicted (from perfect normality) and actual numbers of subjects in seven *T* score ranges are reported in the *Comprehensive Norms*, so that the user may examine the closeness of fit (Table 7). Also reported for all variables are the amounts of test score variance accounted for by demographics (linear predictions) before and after demographic corrections with the *T*-score system (Table 5). What that table shows is elimination of large amounts of demographic influence (or bias) on this test battery.

Questionable Accuracy and Value of Demographic Prediction

Fastenau and Adams suggest that the accuracy of the demographically corrected *T* scores depends upon the accuracy of the predictions made using the multiple regressions. They claim that the appropriate measure of such accuracy is the standard error of the estimate (SEe), and point out that these standard errors were not reported in the manual. In addition, Fastenau and Adams predict that the SEe measures are likely to be "much larger than a clinician would want to see" because "predicted values rarely yield improvement over the total sample mean when the multiple correlation coefficient is less than .70" (p. 446).

The issue being addressed here is the desirability of including demographic corrections within the norming process. Fastenau and Adams imply that, unless the multiple correlation between demographics and NP test scores is very high (.70 or higher), one should ignore demographics and simply use the mean and standard deviation of the total subject group as the normative guide.

In fact, there are a few NP variables in our test battery ($n = 5$) that have little or no demonstrable demographic influence; these have minimal or no demographic "correction" in the *Comprehensive Norms.* Most variables have moderate associations with demographics (Rs from .32 to .69; $n = 43$), whereas only a few ($n = 6$) have multiple correlations higher than .70. Available evidence suggests that even NP variables that have modest relationships with demographic characteristics show clinically significant demographic biases when used to classify subjects as neurologically normal or abnormal (Heaton, Grant, & Matthews, 1986; Heaton, Ryan, Grant, & Matthews, in press). Thus, our decision was to attempt to minimize or eliminate *any* demographic biases (small, medium, or large) in interpreting the test variables.

Parenthetically, the use of SEe as the sole criterion for determining the value of psychological test prediction is not universally endorsed. As Anastasti (1982) points out: "When examined in the light of error of the estimate, most tests do not appear to be very efficient ... A test may appreciably improve predictive efficacy if it shows *any* significant correlation with the criterion, however low" (p. 160).

Demographic variables do not need to predict test scores of normals precisely in order to be important in interpreting test scores. The question remains: How helpful are the *T*-score conversions in removing the demographic influences? As noted above, Table 5 of the *Comprehensive Norms* manual shows minimal linear relationships between the demographic variables and *T* scores, even in the independent validation sample. This provides some assurance that Table 7 and 8 reflect a demographically unbiased assessment of where a person's test score falls within the distribution of normals' scores. The same tables can estimate the likelihood that a person's test score does *not* fall within the normal range (i.e., is abnormal). Presumably, the accuracy of these latter judgments, critical in neurodiagnostic work, also will have greatly reduced demographic biases. However, Fastenau and Adams correctly suggest that this assumption should be tested. Demographic biases involved in raw scores versus *T* scores should be compared with both normal and brain-damaged subject groups. To help address this need, we present the following results of 949 normal and brain-damaged subjects on the Average Impairment Rating (AIR) from the Halstead-Reitan Battery (Russell, Neuringer, & Goldstein, 1970).

Our total subject sample includes the 486 normal adults whose data formed the basis of the *Comprehensive Norms*, and 463 adults who had structural brain lesions verified by appropriate methods (CT or MR brain scans in most cases, and/or neurosurgical reports). The etiologies in the brain-damaged group included: extrinsic tumors ($n = 29$), intrinsic tumors ($n = 48$), cerebrovascular accidents ($n = 68$), vascular malformations ($n = 6$), trauma ($n = 82$), infection ($n = 11$), epilepsy ($n = 5$), toxic/metabolic disorders ($n = 11$), hydrocephalus ($n = 25$), multiple sclerosis ($n = 108$), Alzheimer's disease ($n = 44$), and other neurodegenerative diseases ($n = 26$). There were 319 males and 167 females in the normal group, whereas the brain-damaged group included 263 males and 200 females. The normal group's mean age of 42.0 years (SD = 16.81) was comparable to the mean of 43.27 (SD = 15.09) of the brain-damaged group (t (947) = 1.23, $p = .22$). The education levels of the two groups also were similar: mean for normals = 13.55 years (SD = 3.51) versus for brain-damaged = 13.10 years (SD = 2.99), (t (947) = –1.09, $p = .04$).[2]

Table 1 shows the correlations between two demographic variables (age and education) and the AIR raw scores versus T scores for the normal and brain-damaged groups. Higher AIR raw scores, indicating worse neuropsychological test performances, were associated with older age and lower education levels for both subject groups. Age and education together accounted for 64% of the AIR raw score variance in normals, as compared to 33.6% in the brain-damaged subjects. Clearly, a 64.% demographic bias is likely to have a major impact on accuracy of diagnostic classification of normals (i.e., *specificity*) at different levels of age and education. The fact that less AIR raw score variance is accounted for by demographics in the brain-damaged group is to be expected due to the additional impact of their neurologic disorders, the severity of which may be unrelated to age and education. Nevertheless, as will be seen presently, even a 33.6% demographic bias can have substantial effects on the diagnostic accuracy (i.e., *sensitivity*) of the AIR in detecting brain damage in subjects who differ in age and education. Moreover, as can be determined from the Table 1 data, the T-score corrections reduce the AIR variance accounted for by demographics to less than 1% for both subject groups.

To supplement the correlational approach, we conducted a series of 4, two-factor ANOVAs, with one factor being diagnostic group (brain-damaged or normal) and the second being level of age or education (low, middle, or high). The respective levels for age were <40, 40-59, and 60+ years; for education, <12, 12-15, and 16+

Table 1. Bivariate (Pearson) and Multiple Correlations Between Demographic Variables and Average Impairment Rating Raw Scores Versus Demographically Corrected T Scores for Normal ($n = 486$) and Brain-damaged ($n = 463$) Groups.

	Normal Group		Brain-Damaged Group	
	Raw scores	T scores	Raw scores	T scores
Age	.71***	.00	.50***	.07
Education	.60***	.02	.29***	.02
Both	.80***	.02	.58***	.08

*** $p < .0001$

[2] In conducting statistical comparisons of these large groups, we use an alpha of .01 in order to avoid concluding that minor differences are "significant."

years. These 2 × 3 ANOVAs were performed using first raw scores and then *T* scores as dependent variables, permitting for each the assessment of diagnosis and demographic level main effects as well as possible diagnosis by demographic level interaction effects. Such interactions could occur, for example, if the neurobehavioral effects of brain damage were more severe for people with older age or lower education levels.

Both ANOVAs using AIR raw scores revealed highly significant ($p < .0001$) main effects for demographic level ($F(2,943) = 216.78$ for age; and $F(2,943) = 108.10$ for education) as well as for diagnosis (respective *F*s(1,943) of 579.28 and 481.17); the demographic level by diagnosis interaction effects were nonsignificant, using an alpha of .01 (Age Level × Diagnosis $F(2,943) = 3.26, p = .04$; Education level × Diagnosis $F(2,943) = 1.55, p = .21$).

In the analyses of AIR *T* scores, there were again highly significant diagnosis main effects (*F*s(1,943) = 624.90 and 623.61), but the main effects for demographic level were negligible (*F*s(2,943) = 1.15, $p = .32$ for age, and 0.23, $p = .80$ for education); the interaction effects were nonsignificant as well (*F*s(2,943) = 2.48, $p = .08$ for Age × Diagnosis, and 0.24, $p = .79$ for Education × Diagnosis).

Figure 1 displays the AIR raw score and *T*-score means of normal and brain-damaged subgroups at increasing levels of age and education (see Table 2 for subgroup *n*s). The large effect of brain damage can be seen at all demographic levels, regardless of whether raw scores or *T* scores are used. The upper graphs also show substantial demographic effects on the AIR raw scores, both for normal and brain-damaged subjects. However, the lower graphs show minimal demographic influence on mean *T* scores, which approximate 50 for all normal subgroups and range from approximately 32 to 35 for the brain-damaged subgroups. Finally, consistent with the ANOVA results, no substantial diagnosis by demographic level interaction effects are apparent.

To the clinician, the most important aspect of any normative system is its diagnostic accuracy: how well and consistently it allows for correct classification of neurologically normal and abnormal persons, regardless of their demographic characteristics. Table 2 shows how these classification rates differ for our age-level and education-level subgroups, depending upon whether the AIR criterion used is the standard raw score cutoff of 1.55 or a demographically corrected *T* score cutoff of 40.

Starting at the most global level, for the 949 normal and brain-damaged subjects, the total correct classification rate of 80.9% for the *T* scores is only marginally better than the rate of 74.5% for the raw score cutoff. Overall specificity was about the same for raw scores and *T* scores (82.5% vs. 85.8%, respectively) whereas sensitivity was somewhat better for the *T* scores (75.8% vs. 66.1%).

However, the largest diagnostic performance differences of AIR raw scores and *T* scores relate to consistency across the age-level and education-level subgroups. This can be illustrated by considering, for the six age and education subgroups, the deviations in classification rates from the overall specificity or sensitivity rates of the respective AIR score-type (*T* versus raw). Ideally, of course, the classification rate at each level of age and education would deviate little, if at all, from the overall rate for normal or brain-damaged subjects. For *T* scores, overall specificity (correct classification of normals) was 85.8%; across the six demographic subgroups, the mean absolute deviation from this rate was 4.2% (range = 0.8 to 8.2%). Specificity did not fall below 82.5% for any group. By contrast, for uncorrected AIR raw scores, the mean deviation from their overall specificity rate of 82.5% was 19.5% (range = 3.1 to 43.5%). Here, the worst subgroup classification accuracy rates were unacceptable: 39.% for the oldest subgroup, and 47.1% for the subgroup with less than a high school education.

Interpretation of subgroup variation in *sensitivity* (correct classification of brain-damaged subjects) is complicated by the fact that no effort was made to match brain-damaged subgroups with respect to the nature of brain disorders that were included. Nevertheless, there are large subgroup differences for AIR raw scores that are entirely consistent with a demographic bias ex-

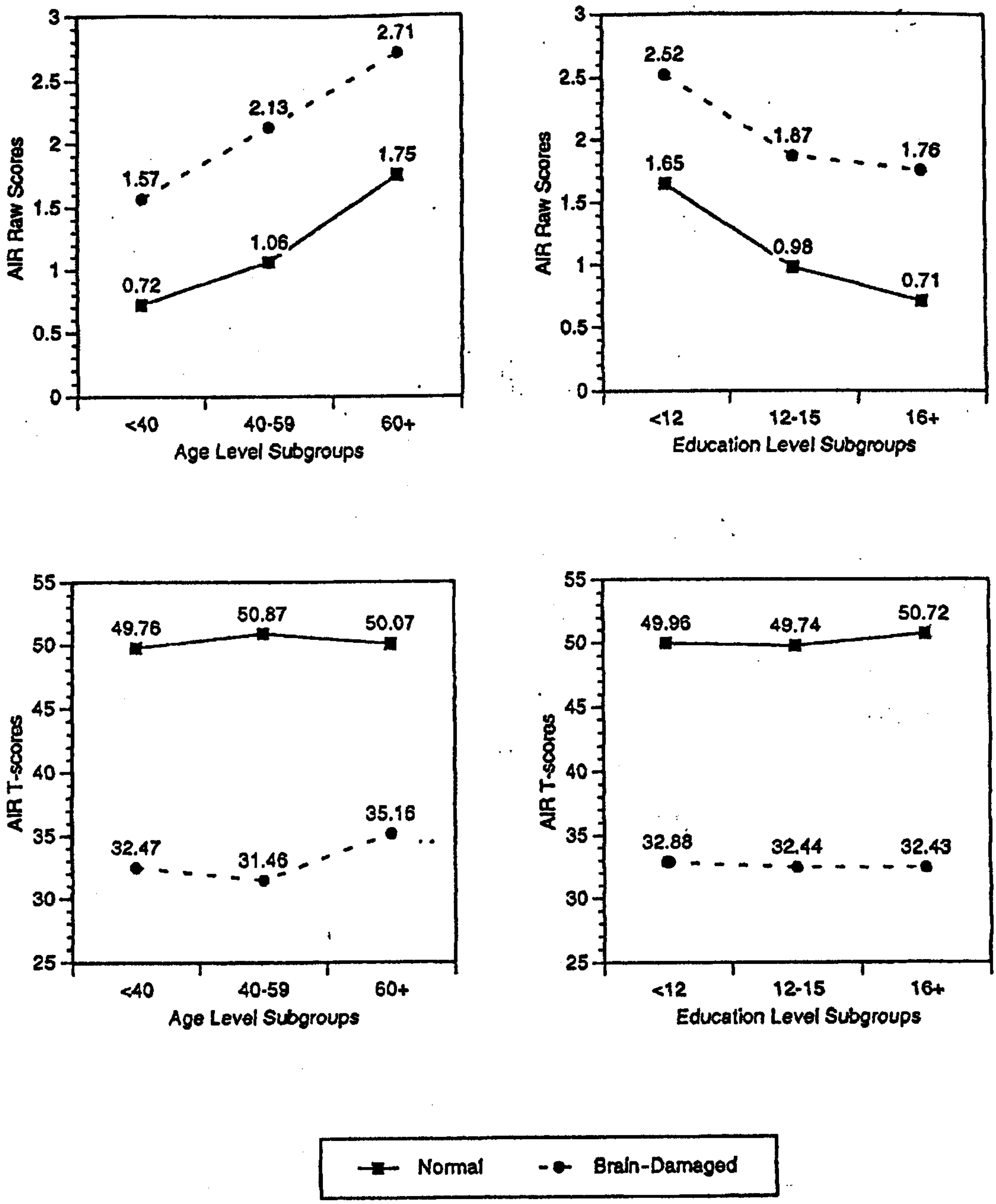

Fig. 1. Means of Average Impairment Rating (AIR) raw scores and demographically corrected *T* scores for normal and brain-damaged subgroups at three levels of age and education.

planation: that is, progressively older and progressively less educated subjects show the most impairment and are most likely to be diagnosed as brain-damaged (see Figure 1 and Table 2).

Table 2. Percentages of Subjects at Three Age Levels and Three Education Levels Who Were Correctly Classified as Normal or Brain-Damaged by the Uncorrected Raw Score Cutoff Versus Demographically Corrected *T*-Score Cutoff on the Average Impairment Rating.

	Age Subgroups for Normal Sample			Age Subgroups for Brain-Damaged Sample		
	< 40 (*n* = 254)	40-59 (*n* = 132)	60+ (*n* = 100)	< 40 (*n* = 210)	40-59 (*n* = 177)	60+ (*n* = 76)
Raw Scores	98.0	85.6	39.0	46.2	77.4	94.7
T scores	81.5	87.9	94.0	72.4	81.9	71.1

	Education Subgroups for Normal Sample			Education Subgroups for Brain-Damaged Sample		
	< 12 (*n* = 102)	12-15 (*n* = 216)	16+ (*n* = 168)	< 12 (*n* = 94)	12-15 (*n* = 250)	16+ (*n* = 119)
Raw Scores	47.1	90.3	94.0	88.3	63.2	54.6
T scores	91.2	86.6	81.5	79.8	75.6	73.1

Note. Age and education ranges are in years. The cutoffs that define the abnormal range on the Average Impairment Rating are ≥ 1.55 for raw scores and < 40 for *T* scores.

There is no apparent reason why subgroup differences in nature of brain disorders would produce such an orderly association with levels of age and education (especially the latter).

In any event, using the standard AIR raw score cutoff, the overall sensitivity with this group of brain-damaged subjects was 66.1%; on average, the absolute deviation from this figure for the demographic subgroups was 16.1% (range 2.9 to 28.6%). In the case of *T* scores, however, the mean subgroup deviation from the overall sensitivity rate of 75.8% is only 3.5% (range 0.2 to 6.1%). Clearly, these subgroup differences are much smaller and do not suggest any significant demographic bias.

Possible Inadequate Cell Sizes

Fastenau and Adams assert that the cell sizes of a series of presumed 10 (age groups) by 6 (education groups) by 2 (sex groups) ANOVAs were seriously inadequate. Here the reviewers apparently have misread the description of the ANOVAs on page 11 of the *Comprehensive Norms*, and have misunderstood their purpose. As mentioned above, the ANOVAs in question were performed on the various normal subject groups (Base, Validation, and Total samples) to ensure that the *T*-score conversions were operating appropriately at different levels of age and education. These were *two*-factor (age and education) ANOVAs, using *three* levels of each; thus, the total number of cells was 9, not 120, and the cell sizes were much larger than those inferred by Fastenau and Adams. Separate two- group comparisons were performed to assess possible gender differences. The results of these ANOVAs and *t* tests were generally quite reassuring regarding the consistency of mean *T* scores across demographic levels.

Fastenau and Adams also note that the *Comprehensive Norms* manual contains 120 separate tables, for use with males or females at 10 levels of age and 6 levels of education. With a presumed maximum sample size of 486, the cell sizes associated with these individual tables were said to be too small to form the basis for any clinical interpretation. Here the reviewers appear to have lost sight of the fact that the *T* scores in these tables were based upon regression analyses using the Base sample that, for most measures, exceeded 350 subjects. As was described in the manual, the most accurate *T*-score conversions would use the regression weights with age and education as continuous variables. However, as a practical matter, this requires the use of the available computer software: Tables for males and females at *every* year of age and education would be too numerous to publish, so we presented the data within age and education ranges that were small enough to provide reasonable concordance with the software. As described on page 21 of the manual, the *T* scores derived from the software and the manual rarely differed by more than two. In sum, the number of tables in the manual was driven entirely by the desire to provide a way to use the norms in a reasonably accurate manner without computer assistance.

CONCLUSIONS

Clearly, the *Comprehensive Norms* manual is less than ideal in several respects. It would be desirable to have data to present on several additional tests and test measures – especially on those tests that have been revised since our normative database was collected. We agree that it would be helpful to use PIAT raw scores instead of centiles, but at present this would require a fairly major chart search and review. We regret that our current database does not permit us to correct for racial/ethnicity performance differences, although efforts are under way to collect the needed data for African Americans. It is not clear to us that the format changes recommended by Fastenau and Adams would represent an improvement; however, we agree that it would be desirable to consult users about this issue before completing any revision of the manual.

Although some of the statistical issues raised by Fastenau and Adams could be problematic in the abstract, the proof of such problems would be evidence that the *T* scores: (a) do not eliminate substantial demographic biases inherent in the use of neuropsychological raw scores; (b) show substantially different error rates at extremes of the predicted scores that are based upon demographic regression weights; (c) behave very differently in clinical versus normative groups; and (d) ultimately, do not improve diagnostic accuracy. The relevant data presented in the *Comprehensive Norms* manual and above are, in general, reassuring with regard to these four possibilities. All available data strongly suggest that substantial demographic biases are eliminated or greatly reduced by using the demographically corrected norms. Clearly, further research is needed to examine the performance of the *T* scores with demographically diverse samples of subjects who have documented brain disorders; nevertheless, our results to date are encouraging regarding the overall *sensitivity* of *T* scores relative to raw scores.

Table 2 above indicates that the diagnostic performances of AIR *T* scores at different levels of age and education are not perfectly consistent, even for neurologically normal subgroups. On the other hand, the *T* scores' inconsistencies across demographic subgroups are not large, and represent a major improvement over the standard raw score cutoffs.

One rather striking omission in the Fastenau and Adams review is any consideration of the relative merits of using the *Comprehensive Norms* versus the currently available alternative technologies for interpreting performances on these neuropsychological tests. While we agree that this initial attempt at providing demographic corrections for several commonly used tests could have been more statistically sophisticated, and possibly could have been more user friendly, the evidence seems to indicate that the norms do have significant advantages for neuropsychological clinical work and research.

REFERENCES

Anastasi, A. (1982). *Psychological testing* (5th ed.) New York: Macmillan.

Fastenau, P. S., & Adams, K. M. (1996). Book Review. Heaton, Grant, and Matthews (1991) Comprehensive Norms: An overzealous attempt. *Journal of Clinical and Experimental Neuropsychology, 18*, 444-448.

Heaton, R. K. (1992). *Comprehensive norms for an expanded Halstead-Reitan Battery: A supplement for the WAIS-R.* Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting*

Test Manual, Revised and Expanded. Odessa, FL: Psychological Assessment Resources.

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.

Heaton, R. K., Ryan, L., Grant, I., & Matthews, C. G. (in press). Demographic influences on neuropsychological test performance. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (2nd ed.). New York: Oxford University Press.

Russell, E. W., Neuringer, C., & Goldstein, G. (1970). *Assessment of brain damage.* New York: John Wiley & Sons.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics,* (2nd ed.). New York: Harper Collins Publishers.

Thompson, L. L., & Heaton, R. K. (1989). A comparison of different versions of the Boston Naming Test. *The Clinical Neuropsychologist, 3,* 184-192.

Wechsler, D. (1981). *WAIS-R Manual.* New York: The Psychological Corporation.