The operating characteristics of the major HRNES-R measures

Elbert W. Russell*

VA Medical Center, Emeritus, 9350 S. Dixie Highway, Suite PH3, Miami, FL 33156-2944, USA

Accepted 17 December 2003

Abstract

The operating characteristics and base rate effects of tests and indexes in the Halstead Russell Neuropsychological Evaluation System—Revised (HRNES-R) were obtained to determine its accuracy in assessing brain damage. Since operating characteristics and base rate problems are not well understood they were discussed in some detail. The operating characteristics of Sensitivity, Specificity, Positive Predictive Power, Negative Predictive Power and Overall Predictive Power along with base rate effects were obtained for 2 HRNES-R indexes, 10 index tests and 3 other Halstead Reitan Battery (HRB) tests. The indexes were found to be as accurate as the most accurate indexes from other HRB studies. The accuracy of the various tests were high but varied according to their function and design.

© 2004 Published by Elsevier Ltd on behalf of National Academy of Neuropsychology.

Keywords: Operating characteristics; HRNES-R; HRB tests

The operating characteristics of an assessment measure are primary methods for demonstrating the scientific validity and accuracy of that measure (Gouvier, 1999; Retzlaff & Gibertini, 2000). In this study, the operating characteristics of the major tests and indexes in the Halstead Russell Neuropsychological Evaluation System—Revised (HRNES-R; Russell & Starkey, 2001a, 2001b) were obtained to both examine the validity of this battery and to illustrate the utility of operating characteristics and base rates.

*Tel.: +1-305-6673821; fax: +1-305-670-2285.
E-mail address: ewr@bellsouth.net (E.W. Russell).
1. Utility of operating characteristics

Operating characteristics more accurately evaluate the situation occurring in assessment than more traditional statistics such as T-tests or ANOVAs. These traditional methods are not specifically designed to help the clinician’s situation in which the N is 1 (Retzlaff & Gibertini, 2000, pp. 277–299). Since traditional statistics compare an entire group with another group, the patients at the extremes of the groups exert a effect on the statistic as well as those that are close to the cut point. In fact, generally the effect exerted by a score is greater the further from the cut point the score lies. Thus, the conventional statistics may give the impression that the scale is more accurate in separating two groups than it actually is.

When a neuropsychologist is dealing with a specific individual, the examiner wants to know whether that patient falls within one group or another, such as within a brain-damaged or a control group. The cut point is crucial. When using operating characteristics the severity of impairment has no effect on the statistic other than determining which side of a cut point the subject falls. All members of a group on each side of the cut point are treated the same statistically no matter how extreme is their impairment. Consequently, for clinical purposes, in which the question is determining the existence of a condition, statistics based on the operating characteristics of a cut point are more accurate.

In forensic cases, these characteristics are particularly important. The Daubert Standard (Daubert v. Merrell, 1993) established that scientific expert testimony must be derived by scientific methods to qualify as “scientific knowledge” (Reed, 1996). One of the major criteria of the Daubert standard for determining whether an expert’s testimony was based on scientifically reliable studies was whether a technique considered “… known or potential rate of error, and existence and maintenance of standards controlling the technique’s operation.” (Daubert v. Merrell, 1993, 28, 2789). In a neuropsychological or medical setting the primary method for demonstrating potential rate of error is obtaining operating characteristics.

2. Types of operating characteristics

There are a number of different operating characteristics which may be used to evaluate a criterion. Each characteristic provides different information concerning the accuracy of the measure. Several of these operating characteristics have now become fairly standard (Retzlaff & Gibertini, 2000). These characteristics consist of Prevalence, Sensitivity, Specificity, Positive Predictive Power, Negative and Overall Predictive Power. (The term “specificity” means the same as the word “selectivity”). These cover the major measures that can be used to determine the accuracy of a test in terms of correct and incorrect assessments of its criteria. The accuracy of a cut point for a test should be evaluated for each characteristic since each characteristic provides different information concerning the measure. For instance, if we want to know how accurate a test is in predicting a condition the Positive Predictive Power is the most important characteristic. However if we want to know how accurate a test is in predicting individuals who do not have the condition (normal subjects) then the Negative Predictive Power is required.

The operating characteristics of a test, using a particular cut point for the criterion, such as brain damage, are statistically described in a $2 \times 2$ matrix (see Table 1). The scores in
Table 1

HRNES-R Average Impairment Score (AIS) operational characteristics with obtained and percent scores

<table>
<thead>
<tr>
<th>1 Test</th>
<th>AIS original scores</th>
<th>AIS percent (prevalence = 50%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Brain damage</td>
<td>Brain damage</td>
</tr>
<tr>
<td></td>
<td>A Present (B)</td>
<td>Absent (C)</td>
</tr>
<tr>
<td>2</td>
<td>Positive</td>
<td>444</td>
</tr>
<tr>
<td>3</td>
<td>Negative</td>
<td>132</td>
</tr>
<tr>
<td>4</td>
<td>Totals</td>
<td>576</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Prevalence (base rate)</td>
<td>0.74</td>
</tr>
<tr>
<td>7</td>
<td>1-BR</td>
<td>0.26</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.77</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.94</td>
<td>Specificity</td>
</tr>
<tr>
<td>Positive Predictive Power</td>
<td>0.97</td>
<td>Positive Predictive Power</td>
</tr>
<tr>
<td>Negative Predictive Power</td>
<td>0.59</td>
<td>Negative Predictive Power</td>
</tr>
<tr>
<td>Overall Predictive Power</td>
<td>0.81</td>
<td>Overall Predictive Power</td>
</tr>
<tr>
<td>BRE</td>
<td>0.18</td>
<td>BRE</td>
</tr>
<tr>
<td>1-BRE</td>
<td>0.82</td>
<td>1-BRE</td>
</tr>
</tbody>
</table>

Note. The formulas for the various operating characteristics are as follows: Prevalence (BR): B4/D4; 1-BR reverse of BR 1-C6; Sensitivity B2/B4; Specificity: C3/C4; Positive Predictive Power: B2/D2; Negative Predictive Power: C3/D3; Overall Predictive Power: (B2 + C3)/D4; Base Rate Effectiveness (BRE): (B3 + C2)/D4; Base Rate Effectiveness limits: <50% = BRE; >50% = 1-BRE.
each cell vary independently. An operating characteristic is defined by the ratio between the number representing a characteristic and the total number of subjects (marginal totals) relevant to that characteristic. For instance, “Sensitivity” indicates how responsive a measure is to a condition when the condition is present, such as brain damage. Thus, Sensitivity is the ratio of subjects correctly classified as brain damaged compared to all of the brain-damaged subjects.

At this point each operating characteristic needs to be described since most neuropsychologists are only familiar with a few characteristics. Table 1 provides examples and the formulas for deriving each characteristic.

“Prevalence” is the proportion of subjects in the study who have a condition, such as brain-damage. It is the number of subjects that demonstrate the condition divided by the total number of subjects. Prevalence is equivalent to base rate.

The two most familiar measures, “Sensitivity” and “Specificity”, are essentially derived from the criteria (e.g. brain damage and no brain damage), so they refer to that condition. Sensitivity is concerned with how accurately a test assesses the existence of a condition. This measure is the number of true positives divided by the total number of positives. Specificity or selectivity provides an indication of how accurately a measure identifies subjects that do not have the condition or disorder. This is the number of true negatives divided by the total number of negatives.

It should be noted that Sensitivity and Specificity are not directly affected by base rate (prevalence). They measure the ratio of correct prediction for each category of a condition independently. The numbers of subjects having a condition (brain damage) vary independently from the number which do not have the condition. Each condition is treated as a separate sample. That is, the number of subjects with brain damage vary independently from those that are normal. If there are 60 subjects with brain damage and the test correctly designates 45 of those subjects as having brain damage Sensitivity is 75%. Since this measure is not concerned with the normal control subjects, there do not need to be any control subjects in the study to obtain this characteristic.

“Specificity” only concerns the subjects who do not have brain damage, that is the normals. It specifies the accuracy of the test in correctly determining the number of the normal subjects that do not have brain damage. If the normal group has 100 subjects and the test designates 88 of them as not having brain damage its Specificity is 88%.

The two tests of predictive power are less well known but they are of greatest concern to the examiner who wants to know the accuracy of a measure. They are more directly concerned with the accuracy of a test than Sensitivity and Specificity. Predictive power measures determine how accurately a test can predict whether a subject is a member of the brain-damaged or normal group. Positive Predictive Power indicates the probability that a person with a positive score on a test actually has a particular disorder. This measure is the number of true positives divided by the total number of test positives.

Negative Predictive Power indicates the probability that a person with a negative score on the test will not have the disorder. This is the number of true negatives divided by the total test negatives. This provides the most direct indication of the accuracy of a test in determining whether a subject is normal, that is they do not have the condition.
The Overall Predictive Power is the total ability of a test to indicate both the existence and nonexistence of a condition by combining the true positives and true negatives. The measure is the number of true positives plus the number of true negatives divided by the total \( N \) of the sample.

It should be noted that these measures of predictive power are directly affected by base rate, so base rate is critical for determining their accuracy. To put this another way, both Positive and Negative Predictive Power distinguish brain-damaged subjects from control subjects so the relative number in each group affects the measure. The relative number is prevalence or base rate.

3. Base rate

Since Meehl’s (Meehl & Rosen, 1955) introduction of the concept of base rates into psychology its importance has been known but often ignored. There is a paucity of literature concerned with this subject in neuropsychology even though base rates are an important consideration in determining the predictive power of the operating characteristics of neuropsychological measures (Gouvier, Hayes, & Smiroldo, 1998).

3.1. Cut points and base rates

Less well known but perhaps just as important is the understanding that base rates affect the initial setting of cut points. The cut point is that number which most accurately separates two groups, such as brain-damaged versus controls. However, the cut point will vary depending upon the base rate of the two groups. If one group is much larger, often the brain-damaged group, the cut point will be shifted in the direction of the smaller group or the controls. Consequently, the most accurate cut point will vary from one setting to another depending upon the base rate of the condition in that setting.

3.2. Base rate and operating characteristics

In regard to the accuracy of operating characteristics for an established measure, the predictive power of the proposed cut point will be affected by the base rate (Gouvier et al., 1998). Consequently, it is important to obtain a certain resolution of the base rate problem before either the cut point or the operational characteristics can be determined.

The base rate problem in neuropsychology has several aspects. The base rate utilized in most neuropsychological studies in the literature is simply the proportion of normals to the number of subjects having a particular pathology used in a particular study. However, a specific study seldom represents the prevalence of brain-damaged and normal subjects even in the particular clinic where the research was conducted.

Prevalence will vary from study to study due to factors that have little to do with the accuracy of a test or prevalence of the condition in neurological groups. What is obvious is that the base rate from any validating study will seldom be the same as the base rate of the institution in which a particular test or battery is used for assessment. In order to
counter this situation, often the advice from statisticians is for each laboratory to gather its own data to determine the institutional base rate. The “normal subjects” would be neurologically normal patients, that is patients with negative neurological examinations. However, even the base rate from a single laboratory may vary over time depending upon the situation and the referral system. The consequence of all of these factors is that the size of the control and brain-damaged groups and thus their base rates vary for many reasons. The particular base rates used for comparing brain-damaged with normal subjects depends upon so many unessential factors that the particular proportion of controls to brain-damaged subjects in a study is almost arbitrary.

This study proposes and uses two methods to deal with the base rate problem. First, it proposes the use of a neutral base rate of 50/50 to establish cut points. One of the major research methods of determining the accuracy of a test is the matched pair design in which brain-damaged and control subjects are matched. In this case the base rate is established by the research methodology, which is, of course, 50/50. This base rate is not determined by the prevalence of various kinds of cases in a particular situation but by the methodology utilized. In neuropsychology research, a matching technique is one of the most approved methods of comparing brain-damaged to controls.

One resolution to the base rate assessment problem can be derived from this matching methodology. The statistical characteristics of the two groups can be designed to be equal by simply utilizing the percent of subjects in each group rather than the \( N \) for each group. That is the \( N \) for each interval is divided by the total of each group and multiplied by 100 (Guilford, 1965, pp. 34–36). This percent transformation method provides an assessment base rate that is equivalent to comparing two groups that have 100 subjects each. In this way, the size of the groups will be equal for both normal controls and brain-damaged subjects rather than depending on variable base rates.

While generally the relevance of the unaltered group scores will vary from one location to another in an unknown manner, the test cut score accuracy based on percent will not vary due to base rate changes. Thus, such a percent transformation provides a measure that is the most representative and the most stable method of determining cut points when the conditions and base rates of a pathological population or condition are unknown.

3.3. Base rate effectiveness

The second method of dealing with the base rate problem is to calculate the maximum and minimum prevalence rates within which the test is more accurate than the base rate. If the base rate for either group being distinguished is greater or less than the accuracy of the measure then the base rate is more accurate in predicting the measure.

Gouvier et al. (1998) describe this situation and propose a method of determining the limits for base rate. They term this method “base rate effectiveness”, which they measure by a Base Rate Effectiveness (BRE) index. The BRE index is the number of false positives plus false negatives divided by the total \( N \). This index is exactly opposite of the Overall Predictive Power measure. This effectiveness index permits the comparison of BRE to other operating characteristics. If the BRE Index is larger than the base rate, then the number of brain-damaged subjects is so low that a better prediction can be made by utilizing the base rate.
When two groups of equal size are compared in a study the chance of a measure of a characteristic being correct is 50%. Gouvier et al. (1998, pp. 62–63) offer two formulas for determining the limits within which the score of a measure is more accurate than the base rate. The lower limit, which is below 50% is the BRE. The upper limit is 1-BRE. In regard to the AIS using the prevalence of 50% (Table 1) the BRE is 14% and the 1-BRE is 86%. That is the base rate for the AIS is a better predictor of brain damage than the measure, when the base rates go to 13% or less and when they go to 87% or more.

4. HRB operating characteristics

Almost all the studies presenting operating characteristics and statistical analysis of neuropsychological batteries have been performed on the Halstead Reitan Battery (HRB) (Franzen, 2000; Reitan & Wolfson, 1993; Russell, 1995). This is primarily because it is one of the only two widely used fixed adult batteries in neuropsychology. Some work (McKinzie, Podd, Krehbiel, Mensch, & Trombka, 1997) has been published concerning the Luria Nebraska Neuropsychological Battery (Golden, Hammeket, & Purisch, 1991) but there has been only a minimal amount published recently.

The consequence of the large amount of research on the HRB is that its traditional statistical characteristics have been better delineated than those of any other test battery. This extensive research of the various measures used with the HRB has been recently reviewed in several places (Franzen, 2000, pp. 116–137; Russell, 1995). However, other than studies of Sensitivity and Specificity (Goldstein & Shelly, 1982) there has been no analysis of the full operating characteristics of the HRB or any of its derivative batteries.

Another important consideration in determining the operating characteristics of neuropsychological measures is base rates. (Gouvier, 1999; Willis, 1984). Unfortunately little research has been performed in this area in neuropsychology so that the base rates of neuropsychological conditions are not known (Duncan & Snow, 1987). With few exceptions (Willis, 1984) examination of the effect of base rates on the HRB battery’s operating characteristics have not been determined. However, this is also true of almost all single tests used in neuropsychology.

In this study, both operating characteristics and base rate measures will be examined for the HRNES-R (Russell & Starkey, 2001b). Since the HRNES-R is derived from and utilizes the HRB, many of these operating characteristics also apply to the HRB tests. The HRNES-R provides scale scores for the HRB and 12 other tests commonly used in neuropsychology, including the WAIS-III and relevant portions of the WMS-III. The battery scale scores were obtained by co-norming all of the tests. The advantage of co-normed scale scores is that neuropsychologists can compare scores without needing years of experience learning the relationships between raw scores. Also, the Average Index Score (Russell & Starkey, 1993, pp. 19–20) averages the 10 index scores. This provides a direct measure of the amount of brain damage impairment, which the Halstead Index does not do.
5. Method

This present study is designed to examine the operating characteristics of the HRNES-R, a recent revision, of the HRNES (Russell & Starkey, 2001a, 2001b). A previous study designed to indicate the accuracy and validity of the HRNES was prepared for the HRNES manual (Russell & Starkey, 1993, p. 37). Nevertheless, the previous study did not provide the full set of operating characteristics for the HRNES. Consequently, this present study was undertaken to obtain the full set of operating characteristics for the HRNES-R. Since the HRNES utilizes the same raw score data base as the HRNES-R these operating characteristics apply to both versions of this battery. As such, these operating characteristics would support the validity of the HRNES-R and the HRNES, while providing new information that can be used in neuropsychological interpretations of these tests.

5.1. Subjects

The subjects for this study were derived from the norming sample used for the HRNES and HRNES-R (Russell & Starkey, 1993, pp. 27–32). The norms consist of 200 neurologically normal subjects and 576 brain-damaged subjects. The normal subjects were patients at the Miami VA Medical Center who were referred for a neuropsychological examination. They were found to be free of any neurological condition by an independent neurological examination. This control group sample had 176 males and 24 females. Of the 200 normals subjects, 12 were black. The mean age was 44.6 (S.D. 13.3) and the mean education level was 12.7 (S.D. 2.9) years. Its mean WAIS-R FSIQ was 102.

The brain-damaged sample consisted of 576 subjects with a mean age of 47.1 (S.D. 14.3) and a mean education level of 12 years (S.D. 3.2). The group’s mean WAIS-R FSIQ was 85. There were 54 female and 63 black subjects in the brain damage group. The major neurological diagnoses included, 52 tumors, 125 head trauma, 182 vascular conditions, 63 degenerative diseases, 75 alcoholics and 79 others. A more complete description of this sample may be found in the HRNES manual (Russell & Starkey, 1993, pp. 27–32).

5.2. Measures utilized

The major measures of the HRNES and HRNES-R are the primary indexes and the tests utilized by those indexes. These measures indicate the accuracy of the HRNES and the HRNES-R in assessing the existence of brain damage. The measures included 2 indexes of brain damage, 10 index tests and 3 other HRB tests.

The two indexes are the AIS and the Percent Impaired Index (PII). The AIS is the average of the scale scores for the 10 index tests. The PII is similar to the Halstead Index (HI) in that the most accurate cut point for differentiating brain damage from normals was determined for each index test. The PII is the percent of those index tests falling into the impaired range. These indexes provide the primary indicators of the existence of brain damage in the HRNES and HRNES-R.

In constructing the original HRNES, 10 tests were selected as index tests. The 2 indexes were derived from these 10 tests. The index tests include 5 of the tests that are part of the
HI as it is employed today (Reitan & Wolfson, 1993). These are the Category Test, Tactual Performance Test (TPT) Total Time, Index Finger Tapping Test, TPT Memory and Speech Perception. The TPT Location and the Rhythm Test were not included in the HRNES Index since unpublished research indicated that they were not as sensitive to brain damage as the other HI tests. However, the Finger Tapping Test was retained, although it was found to be less sensitive to brain damage. Since it was the only motor test in the HI or AIS it evidently added unique variance to the indexes. In addition to the HI tests, Trail Making B, the Aphasia Screening Test and the Perceptual Disorders Examination were added to the index tests. The latter two tests, while not being particularly sensitive to the existence of brain damage provide a measure of the severity of damage due to their extremely extended range (Russell, 1991).

In addition, two tests that were WAIS-R (Wechsler, 1981) subtests were utilized. They were Digit Symbol and Block Design. The two HI tests, Rhythm Test and TPT Location, that were not utilized in the HRNES-R index tests were also examined in this study, as was the Cross Drawing from the Aphasia Screening test (Russell & Starkey, 1993, 2001a, 2001b).

The percent transformation method was initially utilized to establish raw score cut points between brain-damaged and normal controls for the HRNES. These same cut points are used in the HRNES-R. They best discriminated the brain-damaged and control groups. The cut points were generally equivalent to a scale score of \( \leq 95 \) (Russell & Starkey, 1993, p. 37) which was used to indicate the existence of brain damage in the HRNES-R and HRNES.

In this study, raw scores were used in obtaining operating characteristics for all of the measures except the indexes, since the cut points for the indexes were derived from raw scores. The equivalent scale score cut points are provided in the HRNES manual (Russell & Starkey, 1993, p. 37). The indexes utilized index test scale scores since they were derived from scale scores rather than raw scores (Russell & Starkey, 1993, pp. 19–20).

Scoring systems had been constructed for the Aphasia Screening Test and Cross Drawing for the Neuropsychological Key (Russell, Neuringer, & Goldstein, 1970). The revised version of these scoring systems is described in the HRNES manual (Russell & Starkey, 1993). The rest of the HRB tests used the HRNES scoring method (Russell & Starkey, 1993), which is generally the same as that used in the HRB (Reitan & Wolfson, 1993). The WAIS-R tests used WAIS-R scale scores (Wechsler, 1981).

The linear regression method for deriving the HRNES and HRNES-R scale scores is described in the HRNES manual (Russell & Starkey, 1993, pp. 33–35) and elsewhere (Russell, 1991). Norm tables for the HRNES and HRNES-R are provided in Appendix F for the HRNES-R (Russell & Starkey, 2001a).

5.3. Methods of analysis

Two types of analysis were performed in this study. These were calculating the operating characteristics for each measure and examining the effect of base rate on the measures.

The operating characteristics utilized in this study were those described by Retzlaff and Gibertini (2000). The characteristics that were obtained were Sensitivity (True positives/total positives), Specificity (True negatives/total negatives), Positive Predictive Power (true
positives/total test positives), Negative Predictive Power (true negatives/total test negatives) and Overall Predictive Power (true positives + true negatives/the total N).

In addition the base rates, the BRE and the limits of BRE are presented. The limits of the BRE both above and below 50% were obtained separately. BRE provides the lower limit while 1-BRE provides the upper limit.

6. Results

The results of this study are provided in two parts. The first part is illustrative of the effect of the percent transformation method as well as varying base rates on the AIS. The method utilized to obtain operational characteristics and base rates for the original scores and for the percent transformation is provided in Table 1. The effect of extreme base rates is provided in Table 2.

The second part presents the operating characteristics for all of the various HRNES-R index measures and the base rate measures. These are provided in Table 3. This table includes two indexes of brain damage, the AIS and a Percent Impaired Index, as well as the 10 tests utilized in the HRNES-R index. The three tests that are commonly used with the Halstead battery and the HI are also be presented.

6.1. AIS operating characteristics for varying base rates

Due to the large number of measures employed for obtaining the various operating characteristics and BRE limits the methods utilized need to be illustrated. An initial table will help clarify the procedures used in this study. The methods for obtaining these characteristics are provided in Table 1 in the form of formulas. These formulas and the abbreviations for the various operating characteristics are provided in the table’s notes. These formulas utilize the marginal numbers and letters in Table 1.

This table for the Average Index Score (AIS; Russell & Starkey, 2001a, 2001b) illustrates how the various operating characteristic measures were obtained and how they are related to various base rates. Table 1 is composed of two sections that are related to percent transformation. In the left section, the operating characteristics and effectiveness measures are presented using the original norming sample data. Since the norming sample contained 576 brain-damaged subjects and 200 control subjects, these original scores have a base rate for the control subjects of 74% so this section represents the effect of base rates at about 75%.

The second section (right) presents the data using the percentage method described above. This is the AIS data after it has been transformed so that the prevalence is 50/50. Several results related to operating characteristics are evident. First it should be noted that prevalence or base rate has no effect on Sensitivity or Specificity. That is the ratio that constitutes Sensitivity and Specificity remains the same regardless of the proportion of subjects in the present and absence categories.

Second, as prevalence reaches 50%, the various operating characteristics become their most accurate. In this regard AIS scores related to the positive and Negative Predictive Power are quite strong being 0.93 and 0.85. This of course is quite acceptable.
Table 2
HRNES-R Average Impairment Score (AIS) operational characteristics with varying base rates

<table>
<thead>
<tr>
<th>Test</th>
<th>Prevalence = 5%</th>
<th>Prevalence = 80%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Brain damage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test A Present</td>
<td>Absent</td>
</tr>
<tr>
<td>2 Positive</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>3 Negative</td>
<td>2</td>
<td>179</td>
</tr>
<tr>
<td>4 Totals</td>
<td>10</td>
<td>190</td>
</tr>
<tr>
<td>5 Prevalence (base rate)</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>7 1-BR</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Positive Predictive Power</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Negative Predictive Power</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Overall Predictive Power</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>Base Rate Effectiveness (BRE)</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>1-BRE</td>
<td>0.93</td>
<td></td>
</tr>
</tbody>
</table>
Table 3
Operating characteristics with a prevalence of 0.5 for major HRNES-R measures of normal and brain-damaged subjects in percents

<table>
<thead>
<tr>
<th>Measure</th>
<th>Cutting point&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Raw score</th>
<th>HRNES-R</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Predictive Power</th>
<th>Base rate limits</th>
<th>Base rate limits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Positive</td>
<td>Negative</td>
<td>Overall</td>
</tr>
<tr>
<td>AIS</td>
<td>≤95 ≤95</td>
<td>77</td>
<td>94</td>
<td>93</td>
<td>80</td>
<td>85</td>
<td>14</td>
<td>86</td>
</tr>
<tr>
<td>Percent Impaired Index</td>
<td>≥50</td>
<td>81</td>
<td>85</td>
<td>84</td>
<td>82</td>
<td>83</td>
<td>17</td>
<td>83</td>
</tr>
<tr>
<td>Category Test</td>
<td>&gt;60 ≤95</td>
<td>79</td>
<td>71</td>
<td>73</td>
<td>77</td>
<td>75</td>
<td>25</td>
<td>75</td>
</tr>
<tr>
<td>Trails B</td>
<td>&gt;100 ≤95</td>
<td>82</td>
<td>75</td>
<td>77</td>
<td>81</td>
<td>78</td>
<td>21</td>
<td>78</td>
</tr>
<tr>
<td>Tapping Test&lt;sup&gt;b&lt;/sup&gt;</td>
<td>&lt;39 ≤95</td>
<td>64</td>
<td>69</td>
<td>67</td>
<td>66</td>
<td>66</td>
<td>33</td>
<td>66</td>
</tr>
<tr>
<td>Digit Symbol&lt;sup&gt;c&lt;/sup&gt;</td>
<td>≤5 ≤92</td>
<td>81</td>
<td>76</td>
<td>77</td>
<td>80</td>
<td>78</td>
<td>21</td>
<td>78</td>
</tr>
<tr>
<td>TPT Total Time</td>
<td>≥24 ≥95</td>
<td>80</td>
<td>85</td>
<td>84</td>
<td>81</td>
<td>82</td>
<td>17</td>
<td>82</td>
</tr>
<tr>
<td>TPT Memory</td>
<td>≥6 ≤95</td>
<td>72</td>
<td>74</td>
<td>73</td>
<td>73</td>
<td>73</td>
<td>27</td>
<td>73</td>
</tr>
<tr>
<td>Speech Perception</td>
<td>≥10 ≥95</td>
<td>80</td>
<td>74</td>
<td>75</td>
<td>79</td>
<td>77</td>
<td>23</td>
<td>77</td>
</tr>
<tr>
<td>Block Design&lt;sup&gt;d&lt;/sup&gt;</td>
<td>≤6 ≤90</td>
<td>62</td>
<td>87</td>
<td>83</td>
<td>70</td>
<td>74</td>
<td>25</td>
<td>74</td>
</tr>
<tr>
<td>Aphasia Test</td>
<td>≥10 ≤90</td>
<td>49</td>
<td>94</td>
<td>89</td>
<td>65</td>
<td>71</td>
<td>28</td>
<td>71</td>
</tr>
<tr>
<td>Perceptual Disorders</td>
<td>≥16 ≤95</td>
<td>56</td>
<td>95</td>
<td>92</td>
<td>68</td>
<td>75</td>
<td>24</td>
<td>75</td>
</tr>
<tr>
<td>Rhythm Test</td>
<td>≥6 ≤92</td>
<td>62</td>
<td>76</td>
<td>72</td>
<td>67</td>
<td>69</td>
<td>31</td>
<td>69</td>
</tr>
<tr>
<td>TPT Location</td>
<td>≤2 ≤90</td>
<td>74</td>
<td>65</td>
<td>68</td>
<td>71</td>
<td>69</td>
<td>30</td>
<td>69</td>
</tr>
<tr>
<td>Cross Drawing</td>
<td>≥4 ≤87</td>
<td>39</td>
<td>89</td>
<td>78</td>
<td>59</td>
<td>64</td>
<td>35</td>
<td>64</td>
</tr>
</tbody>
</table>

<sup>a</sup> Cutting point indicating brain damage.
<sup>b</sup> Index Finger Tapping mean male score.
<sup>c</sup> Digit Symbol, WAIS-R.
<sup>d</sup> Block Design, WAIS-R.
Since this study is comparing two conditions, the possibility that a subject might fall into either condition is 50%. Consequently, when any of the operating characteristics drops below a 0.5 score the particular characteristic is below the level of chance. Thus, any of the operating characteristics must be above 0.5 to demonstrate that this measure is providing scores that are better than chance.

6.2. The effect of base rates

Various situations may change the base rate so that the operating characteristics will change. Consequently, the base rate (prevalence) is examined under four conditions. Tables 1 and 2 present the AIS data transformed to illustrate differing BRE measures. The BRE index (Gouvier, 2001, pp. 62–63) is provided at the bottom of each section of the table. The data from the 50% prevalence section was transformed in Table 2 to be equivalent to more extreme base rates, which mark extremes of the effectiveness index.

The left section of Table 2 provides the characteristics when the lower limit of BRE is reached. This is a prevalence of 0.05 for the AIS, that is the number of brain-damaged cases is 5% of the total number of cases. The right section provides the AIS characteristics when the prevalence has almost reached the maximum effectiveness base rate for the brain-damaged subjects. In this table, 80% of the subjects have brain damage, which is equivalent to a prevalence of 0.8.

When the base rate is low at 5% (Table 2, left) the Positive Predictive Power drops to below 50% \((P = .42)\) and the difference between the Positive and Negative Predictive Powers becomes quite large. The Negative Predictive Power is 0.42, which is below chance. In addition the BRE is greater than the base rate indicating that the test adds no discriminating power beyond chance.

In contrast, as prevalence reaches 80% (Table 2, right), the difference between the Positive and Negative Predictive Powers remains quite large but is reversed. The Negative Predictive Power drops to 51 which is essentially chance. The Positive Predictive Power is 98%. Almost no brain-damaged cases are missed. However, the limits of BRE (1-BRE) are 0.81 while the prevalence is 0.80 thus the measure adds almost no predictive accuracy to the prediction of brain damage beyond base rate.

The results of this study indicate that according to predictive power and BRE, the AIS is effective when the brain damage base rate is between about 80 and 5%. This provides the range in which one can reliably interpret the AIS.

6.3. Individual measures

Table 3 presents the operating characteristics and BRE limits for all of HRNES-R indexes and primary tests in a condensed tabular form. This table utilizes the prevalence set at 50% to provide the most unbiased base rate for these measures. Since the prevalence is the same for all measures, it is not specified in the table.

6.3.1. Index measures

The results provided in this table indicate that the total accuracy in assessing the existence of brain damage by the HRNES-R was most adequately measured by the operating characteristics
of the AIS and secondarily by the PII. The Overall Predictive Power of the AIS, when the percent transformation was utilized, was 86%. The Positive Predictive Power of the AIS was 93% and the Negative Predictive Power was 80%. Thus, according to Positive Predictive Power the AIS will assess the existence of brain damage when it exists quite accurately. The AIS has some tendency to miss brain damage when it exists so that its Sensitivity, the ability to predict the existence of brain damage, was somewhat low at 77%. However, its Specificity was 94% which means that an AIS score of >95 will seldom label a person as brain damaged when they are not.

The PII was also quite accurate in that its Overall Predictive Power was 83%. In addition, its predictive power, Sensitivity and Specificity were approximately equal, running between 82 and 84%. Thus, it indicated the existence of brain damage and Specificity at about the same level of accuracy.

6.3.2. Individual tests

Although none of the various individual tests utilized by the HRNES-R index were as accurate as the indexes, they were generally relatively accurate as indicated by their Overall Predictive Power. The major exceptions were Finger Tapping, with an Overall Predictive Power of 66%, indicating that it was relatively insensitive to the existence of brain damage. The most sensitive test in this study was the TPT total time, with an Overall Predictive Power of 82%. This is approaching the predictive power of the indexes.

The two tests from the WAIS-R, Digit Symbol and Block Design, were about equally accurate with an Overall Predictive Power of 78 and 74%, respectively. Thus, their accuracy was equivalent to the predictive power of the HRB tests in this index.

Neither of the two tests, that were not selected to be part of the Index Tests but are utilized by the HRB, Rhythm and TPT Location, were highly accurate measures of the existence of brain damage. The Rhythm test, when employing the most accurate cut point of ≥6 errors, was relatively inaccurate in indicating the existence of brain damage, with an Overall Predictive Power of 69%. It was moderately specific in not indicating brain damage when the subject was normal. TPT Location was also relatively insensitive to brain damage with a raw score cut point set at ≤2 correct placements to indicate the existence of brain damage, while it was more accurate in indicating the absence of brain damage. The Cross Drawing with a cut point of ≥4 errors, was also highly insensitive to the existence of brain damage (39%) but generally indicated brain damage when the impairment reached or exceeded this cut point. Its Positive Predictive Power was 78%.

In addition, the limits for BRE are also provided in Table 2. The scores are the lowest or highest percent (rounded two digits) that is within the lower and upper effectiveness range. The base rate limits vary at the lower end from 14 to 35%. The 14% was the AIS, which had the widest range (14–86%) while the 35% was the Cross Drawing with the narrowest range (35–64%). In regard to the upper limits the highest score was 86% for AIS. The PII had almost the same upper limit of 83%. In general, the effectiveness range parallels the accuracy of the measures and the widest ranges were related to the measures with the highest levels of accuracy.
7. Discussion

The results of this study provided a method of judging the utility of various major measures in the HRNES-R as indicators of brain damage. As will be mentioned these statistically obtained results often support the neuropsychological lore related to these measures.

The operating characteristics generally indicate that the HRNES and thus the HRNES-R Indexes were highly accurate. The Overall Predictive Power of the AIS was 85% and the other general index of brain damage, the PII, was almost as accurate as the AIS, with an Overall Predictive Power of 83%. Other studies (Russell, 1995) have indicated that the AIS accuracy obtained in this study was approximately equivalent to the highest accuracy of the HI and the Neuropsychological Deficit Scale (Reitan & Wolfson, 1993, pp. 347–397). However, no other studies have compared various tests and indexes using measures of predictive power.

In regard to the specific operating characteristics, the Positive Predictive Power of the AIS of 93% means that it correctly indicated the existence of brain damage in 93% of the cases. Specificity, the indication that a particular subject does not have brain damage, when they do not have it, was quite accurate (94%). The predictive power for the AIS was excellent with an Overall Predictive Power of 85%. Although at the designated cut point the AIS has some tendency to miss brain damage when it exists (Sensitivity 77%) the probability of brain damage chart in the HRNES-R Appendix F (Russell & Starkey, 2001a, p. 26) indicates that an AIS score of 90 and below was 90% accurate in predicting brain damage. Scores of 85 and below were 95% correct and no normal subject had a score below 80.

The PII was almost as accurate as the AIS but it was more sensitive and a little less specific. Its Positive Predictive Power, the ability of the measure to predict the existence of brain damage was 84%, was about equivalent to its Overall Predictive Power of 83%. Thus, this study tended to support the slightly more accurate status for the method of averaging index scores than utilizing impairment cut points. However, the difference is probably not significant.

In regard to individual tests, this examination of the operating characteristics of the major HRNES-R index tests supported the decision not to include the Rhythm Test and TPT Location among the index tests even though they were in the HI. Both the Rhythm test and TPT Location were among the least sensitive of the HI tests. The Rhythm Test may be too short to be highly accurate.

Examination of the TPT Location operating characteristics was interesting in that a score of 3 or more in this study tended to be associated with the normality or a lack of brain damage. Certainly, a score above 4 or 5 would almost always indicate a lack of damage. However, a low score, below 3, was not strongly indicative of brain damage (Specificity of 65%). This finding supports the clinical observation that normal subjects sometimes perform quite poorly on this test.

The three HRB tests that had the lowest Overall Predictive Power among the index tests were the Finger Tapping Test, the Aphasia Screening Test and the Perceptual Disorders Examination. The relatively low accuracy of the Finger Tapping Test for indicating the existence of brain damage, was supported by a low Overall Predictive Power (66%). This relatively poor accuracy has also been found in other studies (Lezak, 1995, pp. 680–682). However, the advantage of the Tapping Test in the HRB is that it is the most sensitive test to lateralization (Russell & Starkey, 1993, p. 40). In addition, its accuracy is sufficient not to detract from the total sensitivity of
the index and as the only motor test in the index it probably adds some unique variance to the index.

The Aphasia Screening Test and the Perceptual Disorders Examination were retained in the index even though they were less accurate than other tests. The Negative Predictive Power of the Aphasia test was poor (65%). That is the Aphasia test tended to miss brain damage to some extent. This was possibly due to both a relative lack of sensitivity to the existence of brain damage for the cut point and its being a fairly focal left hemisphere test. Nevertheless, both the Aphasia Screening Test and the Perceptual Disorders Examination, have a very low basal, that is many standard deviations below the normal mean (Russell, 1991). Thus, they furnished a measure of the severity of impairment which many of the other tests could not provide.

The most accurate single test in the index was TPT Total Time with an Overall Predictive Power of 82%, a Positive Predictive Power of 84%, a Sensitivity of 80% and a Specificity of 85%. This accuracy approached that of the two index measures. Such accuracy supports studies (Goldstein & Shelly, 1972) and clinical lore that indicated its high level of sensitivity to brain damage.

The exact nature of the brain functions that the TPT is measuring is not well understood. The TPT is obviously related to stereognosis. However, it appears to be measuring a function that is considerably more complex than simply stereognosis. The test requires that the subject be aware of spatial relationships involved in the TPT Board as well as the shape of the blocks. It appears to be a measure of a person’s knowledge of their own relation to their spatial surroundings. It is known that body schema is intimately involved with external spatial awareness and orientation (Kolb & Whishaw, 1990, pp. 421–4320). Research has demonstrated that the TPT time is more related to the right than left hemisphere (Russell, 1974). Thus, the TPT may be a measure of what has been called body schema. Body schema includes spatial awareness and orientation within the body as well as external to it. As such, the TPT is the only known measure of body schema functions and they are functions that are obviously extremely important.

The Trail Making Test B was also an extremely accurate measure of brain damage according to the test’s Overall Predictive Power (79%). Its Sensitivity (82%) is as great as that of the two index measures; however, its Specificity (75%) and Positive Predictive Power (77%) was not as great. In this study, a Trails B raw score above 100 was a good indicator of the existence of brain damage but only a moderate indicator that brain damage does not exist. These results certainly justify Reitan’s selection of this test to be part of his battery (Reitan & Wolfson, 1993, p. 279).

The two tests derived from the WAIS-R (Wechsler, 1981), Digit Symbol and Block Design, appeared to make a contribution to this index due to their relatively high accuracy. Digit Symbol, with the Overall Predictive Power of 78%, tied for the second most accurate individual test along with Trails B. Apparently, in accordance with other research, slowed psychomotor speed is one of the most accurate indicators of brain damage.

Block Design was also a quite accurate indicator of brain damage, with an Overall Predictive Power of 74%. Its lower Sensitivity (62%) may be related to its lateralized location in that it is more sensitive to right hemisphere damage in right handed people. Consequently, it may tend to miss subjects with lateralized left hemisphere damage.

This study indicates that these WAIS-R tests of cognitive functions are also tests of brain damage. This finding supports the contention that neuropsychological tests are, in fact, tests
of cognitive functions (Ardila, 1999; Larrabee, 2000). The difference between the usual cognitive ability tests and neuropsychological tests lies in the design of the tests (Russell, 2001). Cognitive ability tests are designed to have a normal distribution for a normal population. Neuropsychological tests, on the other hand, are designed to measure impairment so that they often have a low ceiling and greatly extended basal (Russell, 2001).

In regard to the general implications of this study, it is clear that the operating characteristics of various neuropsychological tests vary according to several parameters. First, the base rate will affect many of these characteristics both in terms of the selection of a cut point and certain operating characteristics. Perhaps what is surprising is that base rate has less effect on operating characteristics than might be expected. When these characteristics were obtained using the base rate of the original sample, an Overall Predictive Power for the AIS was 81% whereas when the percent transformation was used the predictive power became 86%. Sensitivity and Specificity did not change. What this means is that the operating characteristics hold up well even with a fairly large variance in base rates. Consequently, neuropsychologists can feel secure in utilizing these operating characteristics even when the base rate of the population from which the subject is drawn may vary to a relatively unknown degree from that used in this study. Obviously if one is dealing with extreme base rates the accuracy of the measure will suffer.

In understanding these results, one must keep in mind that operating characteristics apply to a particular cut point for a test. The cut point was set at the position which most accurately separated the groups having a particular characteristic and a group that did not. In this study, the groups were those with brain damage and without damage.

In some circumstances, the cut point may be somewhat optional since the distributions of the control and brain-damaged groups may overlap for a number of scores. For instance, in regard to Perceptual Disorders an alternate cut point would be a raw score of \( \geq 16 \) points (equivalent to a scale score of 100). This cutting point provides approximately the same Overall Predictive Power of 0.75 as the cut point of \( \geq 16 \). However, with the \( \geq 16 \) cut point the Positive and Negative Predictive Power scores are very asymmetrical at 92 and 68%. In regard to interpretation the relatively low predictive power for Perceptual Disorders test with a cutting point of \( \geq 16 \) indicates that, while the score of 16 and above almost always indicates brain-damaged, a score less than 16 may fairly often be obtained by people with brain-damage.

By contrast, when the \( >7 \) cut point is used the Positive and Negative Predictive Power for operating characteristics for percentage scores are almost equivalent. The Positive Predictive Power is 0.74 and the Negative Predictive Power is 0.73. Thus, the Positive and Negative Predictive Powers are equivalent rather than being asymmetrical as they are in Table 3. Thus, for interpretation it is important to take into consideration not only the Overall Predictive Power but the specific Positive and Negative Predictive Powers of a test.

The use of a cut point for interpretation has several implications. First, as the patient’s score moves away from the cut point the accuracy of two operating characteristics becomes greater while that of the other two are reduced. For instance, as the AIS score decreases below 0.95, Positive Predictive Power and Sensitivity increase while Negative Predictive Power and Specificity decrease. This effect can be observed in the in the HRNES-R Appendix F, Table F-5 (Russell & Starkey, 2001a, p. 26) labeled “Percentage of brain-damaged subjects . . . for each
AIS level”. This provides the percent of, what are essentially operating characteristics, for each major AIS score.

Secondly, there will always be a certain proportion of subjects who although their scores fall in the normal range in fact have brain damage and a proportion of normal subjects will have scores that fall in the brain damage range. In regard to the AIS, with the cut point at $\leq 95$ (Table 3) 23% of the subjects that were designated as not having brain damage, in fact had damage and 6% of those designated as having brain damage did not have it. While this condition involves the usual understanding of statistics, in forensic situations a specific subject falling into false positive or false negative ranges, especially near the cutting point, may present considerable difficulty in interpretation. In such a case, the neuropsychologists should make use of the tests in the HRNES-R which are not index tests. Focal residual impairments of a brain insult should be taken into consideration even though they may not be extensive enough to lower the general index. It is beyond the scope of this paper to discuss these factors. Nevertheless because operating characteristics only provide the statistical likelihood that a patient has brain damage, the entire history of a particular patient must be taken into consideration in an interpretation before determining how these statistical operating characteristics apply to an individual case.

Acknowledgments

I would like to thank Paul Retzlaff for initially reading this writing and making suggestions that considerably improved the paper.

References