The fundamental psychometric status of neuropsychological batteries

Elbert W. Russell a,*, Sally L.K. Russell a, B.D. Hill b

a Private Practice, 9350 South Dixie Highway PH-III, Miami, FL 33156, USA
b Louisiana State University, Department of Psychology, Baton Rouge, LA, USA

Accepted 5 May 2005

Abstract

A fundamental requirement for neuropsychological assessment is dependability. Neuropsychological knowledge is dependable only if it has been validated using psychometric methods. Since batteries are used for interpretations, the psychometric validation methods that are acceptable for individual tests must be applied to batteries to produce dependable information. While the standardized battery has been validated, the flexible battery has not. Due to the probability that some tests will be impaired by chance, a flexible battery cannot produce dependable interpretations by selecting or combining test results. Localization and diagnostic assessments are obtained by comparisons. Comparisons require that the tests in a battery are invariant or have equivalent norms along with a common metric. While standardized batteries do meet these criteria, flexible batteries do not. Consequently, clinical judgment applied to a flexible battery cannot provide dependable knowledge beyond that which could be provided by a single validated individual test.

© 2005 Published by Elsevier Ltd on behalf of National Academy of Neuropsychology.

Keywords: Neuropsychology; Assessment; Fixed battery; Standardization; Validation; Interpretation; Clinical judgment

The controversy between different approaches to the construction and utilization of assessment batteries is a major area of contention in clinical neuropsychology today. This debate often centers on a comparison between fixed or standardized batteries and flexible or heterogeneous approaches. Many have contended that both methods are acceptable and equally adequate for assessment (Bauer, 2000; Franzen, 2000, pp. 2–4; Goldstein, 1997; Incagnoli...
The main contention of these neuropsychologists is that one or the other method may have slight advantages under certain circumstances but it is generally agreed that the acceptance of the standardized or flexible method is primarily a matter of personal preference.

However, to adequately address this controversy requires an understanding of the most fundamental basis of assessment. The discussions mentioned above have seldom dealt with two fundamental concepts related to neuropsychological assessment, although they are often implied. The first of these is the concept of dependability. The second concerns the application of psychometric methods to interpretations derived from batteries of tests (AERA, 1999; Anastasi & Urbina, 1997). These two concepts will be the focus of this paper.

1. Dependability

The most fundamental concept related to the application of scientific methodology can be designated as dependability. Dependability means that information will be invariant from one situation to another. The justification for all standardized scientific procedures is the assurance of dependability. The basic purpose of the neuropsychological assessment is to provide dependable information concerning the functioning of the human brain for both forensic and medical purposes. It is generally accepted that validated psychometric procedures insure that the information derived from neuropsychological assessment can be dependably generalized to the person tested as a basis for interpretation. The concept of dependability is especially important in forensic situations. As neuropsychologists have become increasingly involved in forensic cases, the dependability of their information has been increasingly challenged.

The preeminent standard by which dependability is now assessed in forensic situations is Daubert v. Merrell (1993). This standard specifies that dependability, which is termed “reliability” by the court, is derived through the use of the scientific method. The Daubert standard states that information presented in legal settings must have been tested and validated. This eliminates information that fails to meet these criteria even if it is apparently correct. In Chapple v. Ganger (1994), the Daubert standard was applied to neuropsychological assessment. The judge’s ruling was that the results of a fixed battery as a whole could be accepted as evidence but the results of only some tests were acceptable from a flexible battery (Reed, 1996). This is due to the fact that the flexible battery has not as a whole been tested and validated. The Chapple application, as this ruling has since been called, does not specifically eliminate the use of flexible batteries. Rather the emphasis is on the dependability, or reliability, of the expert’s testimony.

The question that has been forced upon us is how do we insure the dependability of neuropsychological interpretations? In neuropsychological assessment, scientific methodology is, for all practical purposes, psychometric methodology (AERA, 1999; Anastasi & Urbina, 1997). That is, psychometrics, with all its limitations, is the only assessment procedure that at the present can produce dependable information for interpretation. Qualitative and inferential interpretation may, however, begin where psychometric analysis leaves off. The purpose of psychometric validation is to confirm that a procedure actually provides dependable...
knowledge as opposed to assuming it is accurate without evidence. A concept may be correct but if it has never been proven to be correct using psychometric validation methods it cannot be considered to be dependable. Therefore, neuropsychological knowledge is dependable if, and only if, it has been validated using psychometric methods. However, while the application of dependability has been well established for individual tests, it has not been applied to test batteries to any great extent.

2. Test batteries

From a psychometric point of view, there is a difference between a group of tests and a battery. A battery is defined as two or more tests that are related by an assessment method, such as by combination or comparison, which is utilized for a neuropsychological interpretation. In other words, if one uses the relationship between two or more tests to reach an interpretative conclusion then one is employing a battery. The advantages of using a testing battery for neuropsychological assessment are well known (Lezak et al., 2004; Mitrushina et al., 1999). A battery provides information about a person that cannot be obtained from individual tests used alone. This information comes from combinations and patterns that are derived from the relationships between the tests.

Many neuropsychologists contend that the validation of batteries is unnecessary since the individual tests used have already been validated. This contention is correct when tests are only used individually without relating them to each other. This is done when a hypothesis-testing format is employed and a single test is used to answer the hypothesis. However, heterogeneous or flexible battery users, as observed in forensic cases, almost never consistently use a hypothesis-testing format. They often gain much of their information from the entire battery in a manner that could not be derived from the individual tests. When used in this way, statistical constraints apply to the interpretation of the battery results.

Standardization of a battery generally applies the procedures that are used for standardizing individual tests to the whole battery (Anastasi & Urbina, 1997, pp. 6–7). There are three reasons that batteries should be validated as thoroughly as individual tests. First, interpretations derived from tests are validated, not the tests themselves (AERA, 1999, p. 9). When interpretations are derived from batteries the whole battery must be validated. Second, batteries produce information that is uniquely different and impossible to obtain from individual tests. The results of the battery are not simply the sum of the individual tests. Finally, the use of test batteries creates certain psychometric difficulties that reduce the effectiveness and accuracy of the individual tests that comprise the battery (Rosenfeld, Sands, & VanGorp, 2000; Russell & Russell, 2003). In other words, using individual tests in a battery places constraints on the validity of the results from the individual tests. This last caveat is particularly important and will be discussed in greater detail.

2.1. Capitalizing on chance

The major problem with using interpretations derived from a single test in a battery is that tests do not statistically operate the same in a battery as they do individually. When not used in
a battery, a test has a certain probability that it is correct. For instance, if one standard deviation is used as a cutting point for impairment, an individual test with a normal distribution will be impaired about 1/6 of the time. Consequently, it is relatively safe to state that impairment on the test indicates brain damage. The clinician will be correct about 5/6 of the time.

However, when a test is used in a battery, the probability of correctly identifying impairment is transformed into a percentage of the tests in the whole group that are generally impaired (Ingraham & Aiken, 1996). If all of the tests in the battery have a normal distribution, utilizing one standard deviation as the indication of impairment will mean that about 1/6 of the tests in the battery will probably fall into the impaired range. A random error of 1/6 of the tests means that in a battery of 12 test scores, two of the scores can be expected to fall in the impaired range simply by chance. In a battery of 24 test scores, four of the scores should be expected to fall within the impaired range when no impairment is actually present. In forensic practice, it is common for neuropsychologists to employ batteries containing many more than 24 test scores. This phenomenon is similar to what occurs in research when one examines conclusions from many different tests. There is a certain probability that a test result will be significant due to random variation. Researchers are taught to avoid capitalizing on chance by instituting statistical controls in their analyses, such as the Bonferroni correction method or the Scheffe’ test (Gravetter & Wallnau, 2000). Unfortunately, many clinical neuropsychologists fail to extend this same logic to their assessment procedures.

2.2. Battery interpretation

The implication of this psychometric battery characteristic is that in a battery the clinician cannot interpret the impairment of any single test score as an indication of brain damage. The flexible battery provides no psychometric justification for selecting any particular test results in the impaired range to support the hypothesis that a person has brain damage. Consequently, using a flexible battery, there is no way to provide dependable knowledge concerning whether a particular impaired test score indicates brain damage.

Another implication is that the proportion of impaired test scores that are required to indicate the existence of brain damage varies among batteries depending on the tests selected for the battery. In any battery, there is no way of knowing how many impaired test scores are required to indicate the existence of brain damage unless there is a validated index. Consequently, the user of a flexible battery has no way of knowing how many impaired results are necessary to reliably constitute a finding of brain damage since no indexes have been created. Additionally, these indexes probably cannot be obtained for the flexible battery since its construction varies among clinicians. Dependability is sacrificed.

Conversely, a standardized battery has validation studies and indexes to dependably determine the number of impaired test results that are necessary to indicate brain dysfunction. The psychometric reliability of indexes in standardized batteries appears to be sound (Franzen, 2000, pp. 121–123, 140–141; Goldstein & Watson, 1998; Russell, 1992). Additionally, an index has a significant advantage over individual tests in that, when properly constructed, the index is generally more accurate than any single test in a battery (Russell, 1995). In conclusion, a validated index from a standardized battery is more accurate and dependable in assessing the existence of brain dysfunction than any other psychometric method.
There are additional problems with using a flexible battery that has not been validated. Perhaps the most significant of these is that the validity of the battery as a whole is no better than the single most valid test in the battery. Flexible batteries that have not been subject to validation have no method for the combining of test scores other than clinical judgment. From a psychometric point of view, the examiner may as well only administer the single most valid test since the examiner has no data to show that the additional tests increase the validity of the flexible battery.

Another concern is that the user of the flexible battery has no demonstrated dependable means of determining which tests correctly assess brain damage when the test results contradict each other. The clinician who interprets the impaired results while ignoring those results that are in the normal range is capitalizing on chance and engaging in bad science. The only way to avoid this is to know both the relative accuracy of the various tests in the battery and the relative accuracy of any group of tests in the battery. The most accurate test or, usually, group of tests will provide the most accurate assessment. This information has been derived for the Halstead–Reitan Battery (HRB) and Halstead Russell Neuropsychological Evaluation System-Revised (HRNES-R; Russell & Starkey, 1993) but not for most other neuropsychological tests.

3. The logic of test batteries

Shortly after 1950, both Reitan and Teuber developed methods to localize lesions with neuropsychological tests. Essentially, they both applied a similar method, which was to compare functions by means of two or more tests. Previously, assessment had depended upon the score from single tests. Teuber (1955, 1975) developed the method of double dissociation, which involves a comparison of the effects of at least two tests as applied to the two hemispheres of the brain in order to determine the relationship between tests and the hemispheres. This method can be expanded into multiple dissociations for global brain assessment (Russell, 1999, pp. 461–464).

At about the same time, Reitan independently began using comparisons of tests to localize brain lesions. He found that a minimum of two tests, which have an invariant relationship with each other and that have differing sensitivities to the area damaged, were required to localize lesions accurately. The invariant relationship was needed in order to provide a fixed ratio of their scores. This method works because different tests have differing sensitivities to different locations and brain conditions. While focal brain damage may affect both tests to some degree, it will affect the test most sensitive to the damaged area more than tests that are less sensitive to the affected area. Thus, it is the ratio between test scores that is crucial for localization of injury, not the absolute score. This assessment relationship may be referred to as Reitan’s rule. This method is essentially a multiple dissociation of tests technique. The more tests that are known to be sensitive to a location that are used in conjunction with tests that are not sensitive to the same location, the more accurate is the assessment of localization. In this method of comparison tests, which are less sensitive to lesions in a specific area are just as important as tests, which are more sensitive to damage in the same area. This same technique can be applied to differential diagnosis as well (Russell & Polakoff, 1993).
3.1. The problem of stability

In order to compare scores from tests, the test scores must either be invariant or equivalent (Anastasi & Urbina, 1997; Ghiselli, Campbell, & Zedeck, 1981). Invariance creates stability or consistency across time. The relationship (ratio) between tests remains stable with a fixed battery of raw scores. Comparisons using raw scores can therefore occur in fixed batteries because the test scales are invariant and so retain the same relationship from testing to testing. In fact, much of the multivariate research in psychology has used raw scores that remain fixed during the research (Nunnally, 1978). This invariance in fixed batteries has the same stabilizing effect as equating the tests. The same cannot be said of flexible batteries.

3.2. The problem of equivalence

However, even with fixed batteries, equivalent scores are desirable. Fixed raw scores do not have equivalent scales so that the expected ratios between the raw score scales vary from one relationship to another. Due to the differences in scale units the relationship between tests varies from pair to pair. For instance a ratio of .5 may indicate impairment for one pair of tests but may be normal for another pair of tests. As such each of the ratios used in a battery must be learned, which may require years of experience.

Equivalency insures that the relationships between all test scales in the battery are the same. The same score indicates an equivalent amount of ability for all tests. A score, such as 100, indicating an average ability will be the same for every test. Equivalency is obtained by means of standardized scale scores. The following are necessary for standardization: the relationship between test scores must be invariant, the norms must be derived from the same normative sample (co-norming) or equated by some method, and there must be common metric units.

Unlike flexible batteries, tests in standardized batteries are co-normed. This procedure norms all of the tests in a battery simultaneously by employing the same sample and the same norming procedure. The “fixed” status of the battery is established by the norming process. Once the tests have been co-normed, they may be used individually or in groups without losing their “fixed” status since the norming process establishes the consistency of the relationships.

As norms are so critical to a battery the concept of standardization deserves to be examined in more detail. Different types of norms have been created for various fixed batteries. When choosing a set of norms to use, care should be taken to ensure that the norms are truly representative (Russell, 1997, pp. 15–65). There are three methods used for creating norms for a standardized battery. The first method is to use stratified scores in which a proportion of the total sample was represented in each of the various demographic variables included in the battery. The second method is to use linear regression to predict the mean scores for the demographic variables of each subgroup or cell. The third method, used by Reitan to set norms for the Neuropsychological Deficit Scale (NDS; Reitan, 1986), is to derive scale norms from experience rather than from statistical data. The NDS has since been thoroughly validated (Russell, 1995). The primary problem that neuropsychologists have when utilizing the various sets of norms occurs when the norms are not stratified or adjusted for age and education.
In flexible batteries, when different norms for each test are used, the assumption is that the norms for the different tests are equivalent regardless of their norming sample. Most methods of placing tests in flexible batteries assume that the various test norm samples all represent the entire United States. However, no neuropsychological sets of norms have come close to representing the whole average population of the United States, with the exception of the Wechsler tests (Russell, 1997). This assumption of normative equivalence is contrary to all of the rules of standardization in the literature (AERA, 1999, pp. 49–60; Anastasi & Urbina, 1997, pp. 66–70; Ghiselli et al., 1981, pp. 37–55). The assumption of normalcy is especially questionable for “voluntary” norms (Russell, 2005). So far, two of the criteria necessary for equivalency have been satisfied by fixed batteries but not flexible batteries. We will now look at the final criteria, the common metric.

Initially, standardization requires a common metric, such as $z$-scores or $t$-scores. Standardized batteries contain a common metric for the scale scores. In order to overcome this limitation in a flexible battery, Lezak et al. (2004) and Mitrushina et al. (1999) have advocated transforming scores to standard scores. This would provide the common metric or scale scores for the tests. One such method of using $z$-scores to combine test results is the Rohling Interpretative Method (RIM; Miller & Rohling, 2003). It is a definite advancement in establishing an equivalency between tests in neuropsychology, which could be used in a flexible manner. However, at this point the primary problem with RIM is that the scores from various tests do not have equivalent norms (Ghiselli et al., 1981, pp. 37–55). In fact, for equating tests, adequate common norms are more important than equivalency of metric units (Anastasi & Urbina, 1997, pp. 66–70). Dependable knowledge can only be produced by batteries if they are either fixed or standardized, which at present means that they are co-normed. Co-norming on its own overcomes most of the problems of equating tests. It would be difficult to find a flexible battery that is composed entirely of co-normed tests. By comparison, fixed batteries such as the Luria Nebraska Neuropsychological Battery (LNNB; Golden, Purisch, & Hammeke, 1991) and the group of standardized versions of the HRB all use either simultaneous norming or co-norms. Having failed all three criteria for equivalence, the information derived from the flexible battery cannot be considered dependable.

4. Clinical judgment

An alternative validation method to formal psychometric analysis is clinical judgment. Many neuropsychologists using flexible batteries claim that they are not concerned about the formal psychometric requirements for combining and comparing tests in a battery since clinical judgment is the basis for any forthcoming interpretations. While it is not a formal psychometric method, clinical judgment can be subject to validation in the sense that the accuracy of such judgments may be validated using accepted psychometric methods. Consequently, a complete examination of the psychometric status of neuropsychological batteries requires a discussion of clinical judgments. The most complete examinations of clinical judgments have been published in book form by Howard N. Garb (1998). He also published an article looking specifically at clinical judgments in neuropsychology (Garb & Schramke, 1996). Garb found good interrater reliability between clinical judgments when fixed batteries were used. However, he found no
evidence that clinical judgments using flexible neuropsychological batteries were reliable. He was additionally able to locate interrater reliability data for flexible batteries related to regular psychological tests. The results of these studies found very poor reliability between test raters (Garb, 1998, p. 13).

There have been many studies concerning the validity of clinical judgments in neuropsychology. A meta-analysis by Garb and Schramke (1996) found an overall hit rate of 84% for standardized batteries assessing the existence of brain damage and a hit rate of 89% for studies comparing right versus left hemisphere damage. Conversely, no validation studies of clinical judgment using a flexible battery have been published (Garb, 1998, pp. 157–162). The information derived from flexible batteries using clinical judgment must consequently be considered undependable.

Many neuropsychologists claim that it is their experience that allows them to interpret the results from flexible batteries. In regard to experience, the few studies in the area that have been published indicate that beyond the graduate school level, there is almost no relationship between the amount of experience of the clinician and the accuracy of their assessment (Garb, 1998, pp. 167–169; Garb & Schramke, 1996). This is true of all psychologists. Additionally, as far as expertise is concerned, ABPP certified neuropsychologists have not been found to be more accurate than neuropsychologists without this accreditation (Garb, 1998, pp. 169–244).

5. Conclusion

The conclusion of this review is that testing batteries must be examined and found valid in order to produce dependable information. In regard to psychometric validation, published studies have demonstrated that all major fixed and standardized batteries are valid but there are no studies in the literature that support the validity of flexible batteries. Flexible batteries often rely on the interpretation of individual tests. However, some test results in a battery will fall in the impaired range simply because of random variation rather than being a reflection of brain dysfunction. As there is no known method that can determine which specific tests in the flexible battery are truly indicative of brain damage, the clinician who interprets these results is capitalizing on chance and cannot dependably diagnosis brain injury.

Further, localization and diagnostic assessments utilize test comparisons. Dependable comparisons of tests requires a standardized battery in which tests are invariant, have equivalent norms, and a common metric. Studies have validated the ability to accurately compare tests in a standardized battery but not in flexible batteries. Additionally, only standardized batteries have been demonstrated to foster reliable and valid clinical judgments.

In summary, in forensic situations, the expert witness using a standardized battery is the only psychologist who can provide dependable testimony interpreting psychometric test data as a whole to the court. The expert witness who utilizes a flexible battery cannot provide dependable evidence derived from the battery as a whole, though they may present the results of individual tests. Overall, this review should be accepted as supporting the continued development of standardized batteries and be viewed as a challenge to those who use flexible batteries to validate their work with empirical methods.
References


