A REVISED INSTRUCTION SET FOR THE
BOOKLET CATEGORY TEST

DISSERTATION

Presented to the Graduate Council of the
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Daniel M. Rockers, B.S., M.A.
Denton, Texas
August, 1996
A REVISED INSTRUCTION SET FOR THE
BOOKLET CATEGORY TEST

DISSERTATION

Presented to the Graduate Council of the
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Daniel M. Rockers, B.S., M.A.
Denton, Texas
August, 1996
Eighty-eight (N = 88) non-brain-injured adults were
tested with one of two versions of the Booklet Category Test
(BCT). Forty-four (N = 44) individuals were tested with the
standard version of the BCT, and forty-four (N = 44) were
tested with a revised BCT in which between-subtest cueing
was removed, called the Noncued Category Test (NCT).

The results of this study indicate that removal of
cueing instructions changes the Category test significantly.
Subjects administered the NCT scored significantly more
errors than those who were administered the standard
Category test.

While BCT scores correlated significantly with
nonverbal intelligence scores, NCT scores did not. However,
the difference in these correlations was not significant,
indicating that the intelligence aspect measured in the two
versions is not different. Neither the BCT nor the NCT
correlated significantly with the Wisconsin Card Sort, Word
Fluency, Stroop, or Trail Making Test. It is recommended
that the NCT be administered to circumscribed clinical
populations in order to best utilize present findings.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Original Version</td>
<td></td>
</tr>
<tr>
<td>1943 Version</td>
<td></td>
</tr>
<tr>
<td>1947 Version</td>
<td></td>
</tr>
<tr>
<td>Current Versions</td>
<td></td>
</tr>
<tr>
<td>Functions Measured</td>
<td></td>
</tr>
<tr>
<td>Neuroanatomic Measures</td>
<td></td>
</tr>
<tr>
<td>Psychometric Data</td>
<td></td>
</tr>
<tr>
<td>Alterations of the CAT</td>
<td></td>
</tr>
<tr>
<td>Category Test Theory</td>
<td></td>
</tr>
<tr>
<td>Present Study</td>
<td></td>
</tr>
<tr>
<td>Hypotheses</td>
<td></td>
</tr>
<tr>
<td>METHOD</td>
<td>23</td>
</tr>
<tr>
<td>Subjects</td>
<td></td>
</tr>
<tr>
<td>Instrumentation</td>
<td></td>
</tr>
<tr>
<td>Procedure</td>
<td></td>
</tr>
<tr>
<td>RESULTS</td>
<td>33</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>36</td>
</tr>
<tr>
<td>Methodological Issues</td>
<td></td>
</tr>
<tr>
<td>BCT-NCT Comparisons</td>
<td></td>
</tr>
<tr>
<td>BCT-KBIT Comparisons</td>
<td></td>
</tr>
<tr>
<td>NCT-KBIT Comparisons</td>
<td></td>
</tr>
<tr>
<td>Other Test Comparisons</td>
<td></td>
</tr>
<tr>
<td>Future Research Directions</td>
<td></td>
</tr>
<tr>
<td>Conclusions</td>
<td></td>
</tr>
<tr>
<td>APPENDICES</td>
<td>47</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>64</td>
</tr>
</tbody>
</table>

iii
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Age, Sex, and Language Demographics</td>
<td>51</td>
</tr>
<tr>
<td>2. Race and Handedness Demographics</td>
<td>52</td>
</tr>
<tr>
<td>3. Education and Marital Status Demographics</td>
<td>53</td>
</tr>
<tr>
<td>4. Illness and Alcohol Consumption Demographics</td>
<td>54</td>
</tr>
<tr>
<td>5. Category Test Descriptives</td>
<td>56</td>
</tr>
<tr>
<td>6. Other Test Descriptives</td>
<td>57</td>
</tr>
<tr>
<td>7. Category Test ANOVA Results</td>
<td>58</td>
</tr>
<tr>
<td>8. KBIT-Category Test Correlations</td>
<td>59</td>
</tr>
<tr>
<td>9. BCT/NCT-Other Test Correlations</td>
<td>60</td>
</tr>
<tr>
<td>10. WCST Correlations</td>
<td>61</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

Original Version

"Is not the basis of good judgement the ability to appreciate likenesses and differences? Is not the ability to group things one of the organism’s economizing processes?" (Halstead, 1940, p. 1264).

This quote was made in response to the original Halstead Category Test, which purported to measure grouping behavior. In this 1940 study, Halstead hoped to better correlate aspects of psychological behavior with specific cerebral anatomy. He presented a set of 62 actual objects to individuals with definite, localized brain lesions (most had excised brain sections). The task for the subjects was to: 1) group the objects according to a principle or rule, 2) respond to questions whether this or that object could be removed or added without violating the principle, 3) respond to questions about groupings made by the examiner (using the same objects), and 4) freely recall which objects were used in testing.

Although he used subjects with variously located lesions, Halstead found that those with frontal lesions, left or right, showed characteristically different grouping behavior than did his normal controls or other brain-damaged
subjects. Specifically, frontal lobe damaged subjects showed fewer objects used in spontaneous groupings, fewer objects in recall, and fewer numbers of groupings, even though they spent longer doing it. No feedback regarding correctness of response was given; in this study there were no correct or incorrect answers. In this original study by Halstead, we can see the beginnings of principles still used today in the Category test, namely: 1) the idea of establishing a rule for a group and then testing it, 2) testing the ability to use the same objects in a different group, organized according to a different rule, 3) a free recall section, and 4) the idea that this test is sensitive to frontal lobe damage.

Investigating the neurophysiological correlates of grouping and organizational ability seemed worthwhile, but it appears that the original test was too subjective--neither scoring nor presentation was standardized to the point of being useful for diagnosis. What was needed was something that was more quantifiable and more structured.

1943 Version

Three years later, Halstead and Settlage (1943) reported the creation of a new version of the test, which retained the principles mentioned earlier, and addressed the difficulty of subjectivity. How was the new version different? Halstead presented geometric figures to the subject by means of a special apparatus--a revolving drum.
Attached to the drum was a piece of white cloth, on which the figures were drawn. A subject could only see one figure set (four drawings) at a time. Subjects were told to flip a switch corresponding to a number from 1 to 4--whatever the picture might suggest.

Correct and incorrect feedback for the Category Test were introduced: if the subject picked correctly, the drum rotated to the new figure set. If an incorrect response was made, nothing happened. Subjects were coaxed, or sometimes even told the principle if they could not figure it out for themselves.

Nine subtests comprised the test. Each subtest was created according to a principle, either number, oddity, part-whole, or recognition. These same principles remain in use today.

Halstead's subjects were six patients from his 1940 study, three of which had frontal lobe extirpation. Interestingly, these three did not even finish the testing. Each refused to continue after subtest 6.

In changing the test, Halstead now cleared the way for a less subjective and more quantifiable method of testing. Item groups were preset, giving more structure to the situation and reducing required spontaneity from the subjects. This probably reduced the testing time somewhat--Halstead reported as many as 7 thirty-minutes sessions for
some patients--but more importantly, made it much easier to administer and record test results.

1947 Version

In 1947, Halstead produced a book regarding the human brain and intelligence, detailing research on 27 neuropsychological tests. One of these tests was the Category Test.

In this study, Halstead postulated an "impairment index," which was composed of scores from 10 of the tests most sensitive to brain injury. According to Halstead, "the functions reflected by our impairment index are maximally represented in the frontal lobes of the brain" (p. 133). The Category was included in this set of 10. When comparing the impairment index for the frontal lobectomies and non-frontal lobectomies, Halstead found a significant difference ($p < .001$). As well, differences between control group and frontal lobectomies were significant ($p < .001$). To Halstead, this was an indication that the Category was a sensitive indicator of frontal lobe involvement in certain cerebral functions.

What psychological functions did the Category Test measure? Through factor analysis, Halstead found significant loading on what he called factors of central integration (C) and abstraction ability (A). C was considered the psychological function which compares new information with familiar ground information, and either
accepts (incorporates) or rejects it. Abstraction ability (A) was considered to be the function of grouping according to a criterion.

Halstead felt that the problems of brain-injured patients on the Category Test were due to problems with both factors A and C. According to Halstead: "total performance of the brain-injured on our category tests tends to be determined almost as much by the influence of a pathological central integrative process (C) as by a basic ability for abstraction of essential similarities and differences--or grouping to criterion" (p. 46). These processes were considered by Halstead to be a fundamental growth principle of the ego. Normal ego development, or growth, then, is blocked by frontal lobe damage, measured by the Category Test.

Regarding test makeup, in this report we see the number of test questions reduced from 360 to 336, with the first 16 items being called warmup items. These first 16 items were Roman numerals, and their purpose was to fix the verbal instructions in the subject. The test categories remained the same as in the previous version.

The feedback principle was slightly altered for this version: subjects heard either a chime (correct) or a buzzer (incorrect response). Although it is not explicitly stated, it is assumed that this version adopted the current technique of allowing only one response per item. The test
unit itself was made more portable by putting it on a set of casters, which could be rolled from room to room.

In summary, the 1947 version of the Category Test was shorter than the previous version, it was more portable (though still large), and correct/incorrect feedback mechanisms were in place. More empirical evidence for frontal lobe involvement is presented, and an explanation of the postulated factors is presented.

**Current Versions**

Today the Category test exists in several different versions, including the standard projection version (CAT) (Reitan and Wolfson, 1985), the booklet version (BCT) (DeFilipis & McCampbell, 1979), and a recently developed computerized version (CVCT-A) (Miller, 1989). Common to all these versions is a 208 item stimulus set, correct/incorrect feedback for each item, and single response for each stimulus item. While the CAT and BCT contain colored stimulus items, the CVCT-A does not.

The 208 items are divided into 7 subtests, each with a common theme or principle running throughout. The first group consists of 8 items and is usually easily performed, even by those with serious brain lesions (Reitan & Wolfson, 1985). It requires matching stimulus-item Roman numerals with Arabic numerals.

Subtest 2 consists of 20 items and is a basic quantity test: if there are two squares shown, the answer is 2, if
there are 4 circles, the answer is 4, and so on. Subtests 3, 4, 5, and 6 each contain 40 items, and are based on a particular principle, including odd stimulus, ordinal position, color (Spreen & Strauss, 1991), quadrant, or proportion (Reitan & Wolfson, 1985). Subtest 7 is not based on any given principle as it contains items from the previous 6 subtests. Subjects must remember the correct response from previous viewings and give that same response.

Scoring for the Category test is based on number of errors. Halstead (1947) suggested a cutoff of 50: more than 50 indicated brain-damage, while a score of 50 or less could be considered normal. This appeared to be the optimal limit for distinguishing between brain-damaged and normals in his study. Fromm-Auch and Yeudall (1983) agreed with the 50-error cutoff, but only for those under 40 years of age. In a sample of normal adults aged 41-64 (N = 10), median score was 53. Bigler, Steinman, and Newton (1982) also indicated that the CAT was sensitive to age, noting that the mean error score rose dramatically for those over 50.

Yeudall, Reddon, Gill, and Stefanyk (1987) reported age-stratified scores in 225 normal adults between 15 and 40 years of age. Group means were found to range from 27 to 35 errors on the CAT. Reitan and Wolfson (1985) presented severity ranges for CAT scores: 0-25, Perfectly Normal; 26-45, Normal; 46-65, Mildly Impaired; 65+, Seriously Impaired. Thus, several scoring strategies for the CAT exist. Clinical
needs may differ from experimental needs, with appropriate scoring methods chosen, whether it is a scoring range or a cutoff.

Functions Measured

There is considerable debate over what is actually measured by the CAT (Bertram, Abeles, & Snyder, 1990; Boyle, 1988; DeFilipis & McCampbell, 1991; Finlayson, Sullivan, & Alfano, 1986; Halstead, 1940; Halstead, 1947; Reitan & Wolfson, 1985; Rothke, 1986; Sherrill, 1985; Taylor, Hunt, & Glaser, 1990). Halstead (1940) originally felt that the CAT measures "grouping" behavior--brain damaged subjects showed fewer and smaller groupings than did normals. In 1947, he stated that it measured integration and abstraction ability. Russell and Levy (1987) state that "abstraction" is a broad, encompassing term and should be broken into smaller elements. One element is extracting an element from a figure, whereas another is the ability to shift sets or principles. Vanderploeg and Logan (1989) state that in addition to the already mentioned set-shifting (between subtests) ability, there is another type of shifting required in the CAT: perceptual. This involves shifting within a subtest, for example, when the principle remains the same, but the element type changes.

Rothke (1986) compared the CAT with the Wisconsin Card Sorting Test and found that the CAT is not so much a measure of set-shifting skill, but more a measure of complex concept
acquisition, visual memory and visuospatial reasoning—in short, a general intellectual measure. Likewise, Boyle (1988) reported the CAT as a measure of general intellectual capacity.

Bertram, Abeles, and Snyder (1990) reported that the CAT could be viewed as a test of learning ability. Nussbaum and Bigler (1989) agreed that the CAT measures learning, but added the skills of nonverbal reasoning and concept formation. According to Reitan and Davison (1974) the CAT is not only a measure of concept formation, but also the ability to apply organizing principles in a performance mode.

Glyshaw (1990) compared performance of the Picture Arrangement WAIS-R subtest with the CAT to determine if it is a test of planning ability. Results indicated that it was a measure of abstract planning, as well as the ability to plan ahead and exercise judgement. Planning ahead suggests employing a strategy; Laatsch (1983) determined that the CAT involved the development and application of strategies and rules. Developing strategies indicates hypothesis formation and validation: Golden, Osmon, Moses, and Berg (1981) reported the CAT as a hypothesis forming and validating test.

**Neuroanatomic Measures**

Although Halstead originally reported the CAT as a measure of frontal lobe damage (Halstead, 1940; Halstead,
1947; Halstead & Settlage, 1943), subsequent researchers have failed to validate this (DeFilipis & McCampbell, 1979). Russell and Levy (1987) suggested that the abstraction ability measured by the CAT was a parietal cortical function, whereas the set-shifting ability was a frontal lobe function. Miller (1989) agreed, adding that planning and set-shifting involved the reticular activating system as well. However, CAT scores have not been found to correlate specific location or laterality of brain damage (Klove, 1974; Landsdell & Donnelly, 1977; Pendleton & Heaton, 1982; Taylor, Goldman, Leavitt, & Kleinman, 1984). Instead, researchers now consider it a measure of general cortical integrity (Finlayson, Sullivan, & Alfano, 1976; Sherrill, 1985; Taylor, Hunt, & Glaser, 1990; Bornstein, 1986). This approach makes sense in light of the many psychological functions tapped by the CAT.

In summary, the CAT likely does not measure specific cortical damage, but more diffuse or general damage, and is considered one of the more sensitive tests which discriminates between brain-damaged and non-brain-damaged subjects (Golden, 1978; Mercer, 1993). Regarding functions measured, it appears that the CAT is a measure of general intellectual capacity, and requires skill in concept formation and hypothesis testing. Performance on the CAT requires several operations: (a) detection of similarities and differences in stimuli, (b) hypothesis generation based
on (a), (c) testing of hypotheses through single response and feedback per item, (d) based on feedback, revision of hypotheses as needed, (e) abstraction to determine hypothesis or principle at a general level, (f) application of general hypothesis or principle through generation of appropriate responses for changing stimuli.

**Psychometric Data**

Pure psychometric data studies on the CAT are sparse. Many compute psychometrics on altered forms of the CAT.

Matarazzo, Wiens, Matarazzo, and Goldstein (1974) computed test-retest reliabilities by administering the CAT twice (20 weeks apart) to 29 normal males. Correlations between administrations were high ($r = .60, p < .001$) indicating high reliability. As a comparison group, 16 60-year old individuals with diffuse cerebrovascular disease showed test-retest correlations of .96 ($p < .001$), again showing very high reliability.

In an extension study (Matarazzo, Matarazzo, Wiens, Gallo, & Klonoff, 1976) 15 carotid endarterectomy patients (CEP) and 35 chronic schizophrenic patients (CSP) were added. Again, high reliabilities were found (CEP, $r = .82, p < .001$; CSP, $r = .72, p < .001$) though in the schizophrenic patients, this psychometrically high reliability is less meaningful in a clinical sense. This is because on original test, 5 of 35 schizophrenic patients were misclassified as normal, and on retest 14 were misclassified as normal.
MacInnes, et al. (1981) in a test-retest cross-validation of CAT and BCT, found low test-retest reliability for schizophrenic patients ($r = .521, p < .02$) and concluded that the Category test is unreliable for schizophrenics. In other populations, test-retest correlations were high: organic brain syndrome ($N = 16, r = .785, p < .001$), personality disorders ($N = 9, r = .947, p < .01$). Another test-retest study (DeFilipis & McCampbell, 1979) found high correlations for 30 normals tested 3 weeks apart ($r = .913, p < .001$) as well as for 30 alcoholics ($r = .804, p < .001$) tested 3 weeks apart. Kilpatrick (1970) computed split-half reliability by examining the protocols of 41 heterogeneous patients and found correlations between odd-item-errors and even-item-errors to be .97 ($p < .01$).

Charter, Adkins, Alekoumbides, and Secat (1987) also computed split-half reliability in a large unspecified sample ($N = 311$), corrected for age and education, and found high odd-even split-half reliability ($r = .95$). Moses (1985) evaluated internal consistency reliability in a sample of 285 diverse patients and found that the CAT as a single score appears to have little measurement error (Cronbach’s coefficient alpha = .96).

Thus, in terms of psychometrics, the CAT as a 208-item version shows high test-retest reliability for normals and organic brain-damage patients, but not for schizophrenics. Split-half and internal consistency computations also
indicate the CAT is both reliable and consistent in terms of individual items.

Alterations of the CAT

Criticisms. Although the CAT has been shown to be sensitive to brain damage, many complaints have been lodged against it. Some of the complaints include excessive length (Boyle, 1986; Calsyn, O'Leary, & Chaney, 1980; Kilpatrick, 1970; Russell & Levy, 1987) and excessive time (Boyle, 1986; Calsyn, et al., 1980; Gregory, Paul & Morrison, 1979; Kilpatrick, 1970; Sherrill, 1987; Summers & Boll, 1987; Taylor, Goldman, Leavitt, & Kleinman, 1984). It has been argued, however, that the time necessary for the CAT is not excessive. In a study of 43 severely impaired and 66 heterogeneously disabled subjects, Finlayson, Sullivan, and Alfano (1986) found that 93% completed the test in less than an hour, and 50% completed in less than forty minutes. Finlayson, et al. argue that most of the time-consuming reports are anecdotal, and that forty minutes to an hour is not excessive. Other complaints include the test being too expensive, not portable, or not good for patients with TV-set hallucinations (Defilipis & McCampbell, 1979; MacInnes, Forch, & Golden, 1981; Slay, 1984; Summers & Boll, 1987).

Abbreviated versions. These criticisms have also suggested solutions, including alternate presentation methods, as well as abbreviated forms of the test. Recall that the test originally presented subjects with actual
objects (Halstead, 1940) then pictures on a revolving drum (Halstead & Settlage, 1943) and finally in a somewhat portable projection unit (Halstead, 1947).

Summers and Boll (1987) reported both an alternate presentation method and item reduction. While they selected the first 20 items from subtests II through VI, eliminating tests I and II, other approaches have been used. Kilpatrick (1970) pulled files of 33 male and 8 female patients who had taken the CAT. He then computed odd-even split-half reliability estimates of the test, finding correlations of odd-item-errors to total-errors to be .90, and even-item correlation of .99. According to Kilpatrick (1970), using either odd or even items as a short form of the CAT is feasible.

Boyle (1975, 1986) warns that using only odd or even items will result in immediate repetition of some stimulus items. In studies of 16 normals and 35 brain-damaged subjects, Boyle tested an 84-item short-form, with half the items selected from sets I, II, III, IV, and VII. Sets V and VI were treated as one set, since they have the same solution principle. Boyle found that all subtests except V/VI differentiated significantly between brain-damaged and non-brain-damaged ($p < .001$). Thus, sets V and VI can be dropped from the CAT with no loss of discrimination power. Problems with the half-item short forms include unbalanced Ns and a failure in one case to specify which items were
selected for use. As well, a limitation of Kilpatrick's study is that the test was not independently administered to a sample, thereby removing a major source of error variance.

Another approach to shortening the test is by using only the first 4 subtests. Based on Boyle's (1975, 1986) finding that subtests V and VI have limited discriminative power, and that test VII is a memory test, Calsyn, et al. (1980) computed statistics for this short form using existing data from 150 patients undergoing treatment for alcoholism. Validation and cross-validation correlations of .89 and .88 were found, respectively. Calsyn et al. notes that the first four subtests accounted for 77-79% of the variance, and provide a reasonable estimate of the total score.

Dunn, Margolis, and Taylor (1984) investigated the utility of the Calsyn et al. short form, in a study of 60 geriatric patients. Files were pulled for a mixed-diagnosis group of 7 men and 23 women, and an Alzheimer group of 6 men and 24 women. Dunn, et al. found similar correlations as Calsyn, et al., and that the first four subtests account for 75% to 80% of the test variance. Because the CAT is sensitive to age as well as cerebral damage, Dunn suggests that the short form is but a gross estimate of the patients condition, and that ranges of impairment should be defined,
as opposed to the traditional impaired-unimpaired classification.

Another cross-validation of the first-four-subtest short form (Taylor, Goldman, Leavitt, & Kleinman, 1984) found a correlation of .91 for the short form vs. the original, although 19 of 30 normals were misclassified with the short form. As well, subjects with right focal lesions could not be differentiated significantly from normals.

Taylor, et al. warn that shortening the CAT may limit its utility with certain populations. More recently, Taylor, Hunt and Glaser (1990) found that the Calsyn et al. short form correlated at .91 with the CAT in a cross-validation study of 294 male veterans.

In all four of the previous studies (Calsyn, et al., 1980; Dunn, et al., 1984; Taylor, et al., 1990; Taylor, et al., 1984) investigators computed utility of a short-form based on existing data, removing an important source of error variance. The practical significance of this neglect is unclear, since comparison studies have not been done.

Truncating sets and dropping sets as a means of shortening the CAT has been mentioned previously. Gregory, et al. (1979) rescored protocols of 70 diverse Ss, and found a high correlation ($r = .95$) between a 120 item short form (CAT-120) and the standard CAT (208 item) protocol. Tests VI and VII were dropped, and the tests II-V truncated.
Independent validation studies were suggested by the authors.

Sherrill (1987) in an attempt to compose utility and validity of several short forms, looked at 35 CAT protocols and 64 CAT-120 protocols. He found that elimination of scoring of subtests I and II improves CAT predicted scores from the CAT-120. Also found was that the CAT-120 correlated highly \((r = .98)\) with the CAT. According to Sherrill (1987) the CAT-120 can sensitively detect disruptions in complex thinking, thereby serving as an effective substitute for the CAT.

Alternate methods of presentation. In 1975, Beaumont published results of a study using a computerized version of the CAT. In his study of 10 psychiatric controls and 10 brain-damaged subjects, he found that the computerized version of the CAT failed to discriminate between the groups, resulting in 50% misclassification. In this study, the CAT was administered via on-line computer, with the examiner reading the instructions. Problems with this study include small N, and no correlation or validation psychometrics reported for a computerized version. He simply states that it is feasible and can be considered equivalent to the standard CAT. More recently, Miller (1989) and Mercer (1993) reported results of using a computerized Macintosh version of the CAT. Little psychometric data is available on
this version, including test-retest reliability or validity studies.

A booklet version of the CAT (BCT) was made available in 1979 (DeFilipis & McCampbell, 1979). In this study, 30 normals and 30 alcoholics were administered the BCT and the CAT in counterbalanced order. DeFilipis and McCampbell found that the BCT had similar ability to discriminate normals and alcoholics as the standard version (false positives for normals at 18%, false negatives for alcoholics at 7%).

MacInnes, et al. (1981), in a test-retest cross-validation study of the BCT, collected data from 38 male inpatients (16 brain-damaged, 13 schizophrenic, 9 personality disorders) and found that the BCT correlated with the CAT ($r = .828$) as well as the CAT does with itself ($r = .934$). Thus, the BCT may be used in lieu of the standard version with no loss of reliability or validity.

Summers and Boll (1987), in a study of 14 male and 36 female undergraduates, found their own abbreviated booklet version of 100 items to be an acceptable alternative to the standard version. Their version was shortened by using only the first 20 items for subtests II through IV.

Slay (1984) presented plans for a slide projection system for the CAT. His complaint is that the CAT is too expensive and not portable. Psychometric validation has apparently not been established.
Still another version was suggested (Russell & Levy, 1987) in which tests I and II were shortened to 5 and 10 items, test VII was dropped, and V and VI were reorganized—each to its own principle. The CAT was given to 120 neurology patients but rescored for the proposed short form. Total CAT score correlated with the short form at .97. Criticisms of the study included the fact that alternative-form reliability and rescoring removed error variance (Vanderploeg & Logan, 1989). As well it was noted that Russell and Levy's (1987) version substantially reduces the perceptual set shifts (different format items within a subtest), which they stated was an important factor measured. Russell and Barron (1989) replied that it was acceptable to rescore precisely because it removed error variance, including test-retest and the learning effect. In addition, they conducted an additional study which supported their original contention that conceptual shifts (between subtests) were important, but that perceptual shifts were not nearly so.

Category Test Theory

Perceptual set-shifting. As noted previously (Russell & Levy, 1987; Russell & Barron, 1989), the BCT measures the ability to shift modes of thinking according to conceptual as well as perceptual demands. In the BCT, perceptual set shifting ability is measured within the subtests. For example, in any one subtest, the principle is the same
throughout the entire subtest, while the items used may differ. An illustration for this might be subtest III, where the principle is the concept of uniqueness. While an early item might consist of one triangle slightly larger than three others, a later stimulus item might be four figures differing in size and color, but one is in outline form, and the others are solid. In both stimulus items, the principle is the same, but implemented differently. Thus, the changing stimuli require perceptual set-shifting abilities. Items within a subtest typically progress in difficulty, and require the subject to utilize his or her perceptual set-shifting abilities to continue to apply one principle.

Conceptual set-shifting. Conceptual set-shifting refers to the ability to think flexibly. In the BCT, different subtests are organized according to different principles. In moving from one subtest to the next, the subject is required to shift his or her mode of thinking and determine the new rule. However, subjects are always notified when the subtests (and corresponding principle) change. Thus, subjects are "tipped off."

The Wisconsin Card Sorting Test (WCST) (Berg, 1948; Grant & Berg, 1948) is considered a shifting concepts test (Lezak, 1983) but one in which the subject is not warned of upcoming shifts. The subject must sort cards according to a particular category, which changes after the subject scores
ten consecutive correct answers. The subject is informed only that his or her response is correct or incorrect.

As such a test, the WCST is considered one of the best frontal lobe tests available (Kolb & Whishaw, 1985). Drewe (1974), Milner (1964), and Taylor (1979) all report that poor performance on the WCST is a good indicator of frontal lobe lesioning.

Although the BCT is considered to be very sensitive to diffuse cerebral damage, it does not appear singularly sensitive to frontal lobe compromise, as originally thought by Halstead. Perhaps this is due to the subjects always being warned of upcoming conceptual set shifts. If the instructions of the BCT were altered to be similar to the WCST, the subject's perceptual as well as conceptual set-shifting ability would be tested.

**Present Study**

The present study, therefore, proposes to alter the administration of the BCT by changing the directions given to the subject at the beginning of testing. Subjects will be informed that the stimulus item will remind them of a number from one to four, and the examiner will only indicate whether the response is "correct" or "incorrect." In this way, the BCT will be altered in such a manner as to measure conceptual set shifts as well as perceptual set shifts. It is hypothesized that this version of the BCT (NCT, or New
Category Test) will be a sensitive indicator of frontal lobe damage, depending on error type.

Before clinical populations can be tested, however, the NCT norms for non-compromised individuals must be established. Therefore, the proposed study will test non-brain-injured individuals using the NCT, and the BCT as a control group. The purpose of the study is to determine if changing the instruction set makes the Category test more difficult for a nonclinical population. In addition, the study will determine correlations with other known neuropsychological frontal lobe indicators, such as the Stroop test, Word Fluency test, WCST perseveration errors, Trails A and B, and the Kaufman Brief Intelligence Test (KBIT).

Hypotheses

1. The number of errors for normals on the NCT will be greater than for the BCT.

2. BCT-KBIT correlation coefficients will be significant.

3. NCT-KBIT correlation coefficients will be significant.

4. NCT-Stroop correlations will be significant.

5. NCT-WCST (perseveration errors) correlations will be significant.

6. NCT-WF correlations will be significant.

7. NCT-Trails B correlations will be significant.
CHAPTER II

METHOD

Subjects

Ninety college-aged students were tested for inclusion in this study. Participants volunteered from University of North Texas psychology classes and received extra credit points for participating. Exclusion criteria was Beck Depression Inventory scores exceeding 19 and/or prior cerebral trauma involving loss of consciousness or hospitalization as indicated on the Mercer screen (Mercer, 1993). On the BDI, raw scores greater than 19 indicate at least a moderated depression (See Instrumentation/Depression Inventory Section, below) which can affect neuropsychological functioning. As well, prior cerebral trauma also affects neuropsychological functioning. Screening out depressed individuals helps maintain a homogenous group of normally functioning college-age students. No subjects were excluded based on BDI scores or prior cerebral trauma. One protocol was rendered unusable due to improper test administration and another subject could not fully participate due to an inability to adequately communicate in English, leaving an N of 88 subjects.
Subjects' ages ranged from 18 to 26, with an average age of 21. Exactly half the subjects were male (N = 44) and half were female (N = 44). 86 subjects spoke English as their native language, while two were bilingual (Spanish-English, Persian-English). 73 subjects were white (83%), 9 were of Spanish descent (10%), 4 were Arabic (4%), and 2 were Asian (2%). Of the students tested, most were right-handed (81%), while a few were left-handed (6), and one was ambidextrous, all by self-report. See Appendix B, Tables 1 and 2 for summaries.

Subjects educational status ranged from 12 to 17 years, with 33% reporting 15 years. The average was 14.3 years of education. A total of 23 different majors were reported, with the most frequent being psychology (34%). 88 subjects (96.6%) were single, while 3 were married.

One-third (33%) of the participants reported being hospitalized at one time or another. In the only instance relating to neurological problems or mental illness, one individual reported prior hospitalization due to depression. As this individual's current level of functioning appeared normal according to the BDI screen, the protocol was included. No other individuals endorsed items relating to mental illness or hospitalization due to neurological problems. One subject reported being knocked down by a baseball when much younger, thus constituting the only individual endorsing loss-of-consciousness. As this
individual was not hospitalized and showed no ill effects, the protocol was included. A visual inspection of the protocol revealed no extreme scores.

Approximately half the subjects (43) reported that they did not drink alcohol, while 41 reported that they did so in moderated fashion. Four individuals refused to respond to this inquiry. The University of North Texas Institutional Review Board stipulated that all participants have the right to refuse to respond to any question.

Beck Depression Inventory scores ranged from a low of 0 to a high of 19. The average score was 3, with most respondents (N = 50) scoring below 3. While no subjects were excluded on the basis of BDI scoring, the extreme protocol (BDI = 19) was visually inspected to insure validity. See Appendix B, Tables 3 and 4 for a summary of this information.

Instrumentation

**Depression Inventory.** The 21 item Beck Depression Inventory (BDI) (Beck & Steer, 1987) was used to screen for depression. The BDI is a self-report inventory containing questions assessing symptoms and attitudes in areas of mood, pessimism, sense of failure, self-dissatisfaction, guilt, punishment, self-dislike, self-accusation, suicidal ideas, crying, irritability, social withdrawal, indecisiveness, body image change, work difficulty, insomnia, fatigability,
loss of appetite, weight loss, somatic preoccupation, and
loss of libido (Beck & Steer, 1987).

Scores on the BDI range from 0 to 63, with scores of 0-9 for asymptomatic individuals, 10-18 indicating mild to moderate depression, 19-20 indicating moderate to severe depression, and 30-63 indicating extremely severe depression. A cutoff score of 19 was used to exclude individuals; anyone with a score of 19 or greater was considered to not be functioning at potential.

**Neurological Insult Screen.** The Mercer screening survey (Mercer, 1993) is a demographic data survey for gathering subject background information. It was utilized by Mercer to assess medical history variables such as hospitalization, head injury, loss of consciousness, disease, physical symptoms, or cognitive symptoms. The survey is brief—it typically takes 5 to 10 minutes to orally administer. In this study, head injury and/or loss of consciousness was considered to be exclusionary criteria.

**Category Test, Booklet Version (BCT).** The Category Test is considered a test of concept formation (Reitan & Wolfson, 1985). Subjects are shown a stimulus item, and based on some organizing principle, must indicate whether the stimulus represents 1, 2, 3, or 4. Subjects must respond to each of 208 stimulus items by indicating a number 1, 2, 3, or 4. Subjects are allowed only one response per stimulus item.
before continuing to the subsequent item. Each of the seven subtests is created around a particular principle.

The Booklet Version Category Test (DeFilipis & McCampbell, 1979) stimulus items are contained and presented in two 3-ring notebook volumes. Subjects indicate their response either by pointing to a number grid containing the numbers 1, 2, 3, or 4, or by stating their answer aloud. Stimulus items are both color and black and white. Administration time is generally from 20 minutes to 1 hour. Scoring on the BCT is reported as a single number: total errors.

**Stroop Test.** The Stroop Test (Stroop, 1935) is widely used as a measure of frontal lobe functions (Lezak, 1983). It measures the ease with which a subject can shift perceptual sets. While Dodrill (1978) reported a modification of the test, the original format was utilized. The Stroop consists of three parts: one sheet containing 100 (5 columns, 20 rows) color word names printed in black, one sheet containing 100 XXXXs printed in color, and a third sheet containing 100 color word names randomly printed in these colors. The color of ink may not match the printed word; for example, the word "red" may be printed in blue ink. In the first part of the test, the subject reads the printed word. The second part of the test requires that the subject report the color in which each XXXX is printed. The third part required that the subject report the color in
which the color-word is printed. Total administration time is usually 5-10 minutes. Scoring for the Stroop test is in three parts: Part I total time, Part II total time, and Part III total time.

**Word Fluency Test.** A written word fluency test first appeared in Thurstone’s Primary Mental Abilities tests (1938, 1962). The subject must write words beginning with the letter S, as many as possible in five minutes. The second part involves writing four-letter words beginning with C, as many as possible in four minutes. According to Lezak (1983), the average 18-year old can write 65 words within the nine-minute total writing time.

The present study utilized the F-A-S word fluency test, in which the subject has 45 seconds to report as many words as possible for the letter F, 45 seconds for reporting words beginning with the letter A, and 45 seconds for reporting words beginning with the letter S. A single score is produced for the FAS test, the total number of words.

**Wisconsin Card Sorting Test (WCST).** The Wisconsin Card Sorting Test (Berg, 1948; Grant & Berg, 1948) is considered one of the best tests for the frontal lobes (Kolb & Whishaw, 1985). It measures conceptual set-shifting ability.

The WCST consists of a pack of 64 cards, each containing one to four symbols: triangles, stars, crosses, or circles. These symbols are colored either red, green,
yellow, or blue; neither colors nor symbols are mixed on the same card. No two cards are identical.

The subjects task is to place the cards one at a time under one of four stimulus cards: one red triangle, two green stars, three yellow crosses, or four blue circles, according to a deduced principle. For each card, the examiner indicates that the placement is either correct or incorrect. It is up to the subject to determine the principle, whether it be color, form, or number. For example, if the principle is number, the correct placement of a card with two objects on it is under two green stars, regardless of the color. After 10 consecutive correct responses, the principle changes, but the subject is not told. He or she must deduce this and respond by shifting sets. Scoring for the WCST is nine parts: number correct, number of errors, number of perseverative responses, number of perseverative errors, number of categories achieved, number of trials to complete the first category, failure to maintain set, and number of responses in series of 3 or more consecutive correct.

**Intelligence Test.** The Kaufman Brief Intelligence Test (KBIT) (Kaufman & Kaufman, 1990) is a brief, individually administered measure of both verbal and nonverbal intelligence. It is designed to be used with individuals aged 4 to 90, and takes about 15 to 30 minutes to administer. The KBIT Verbal subtest is comprised of two
sections: Expressive and Definitions. Matrices, the nonverbal section, is not subdivided. All answers for the KBIT are scored dichotomously as either 0 or 1, and the resulting raw scores are converted into standard scores. The KBIT ultimately yields age-based standard scores, designed to be compatible with similar tests, such as the Kaufman Assessment Battery for Children (K-ABC) (Kaufman & Kaufman, 1983), the Wechsler Intelligence Scale for Children Revised (WISC-R) (Wechsler, 1974) and the Wide Range Achievement Test Revised (WRAT-R) (Jastak & Wilkinson, 1984), among others.

Test norms were established on a sample of 2,022 children and adults aged 4 through 90. Split-half reliability coefficients of .92 and .87 were found for the Vocabulary and Matrices sections, respectively. Test-retest correlation coefficient for the KBIT was .94. The KBIT was validated against Wechsler's (1974, 1981) Verbal and Performance Scales (Parker, 1993).

*Trails A and B.* In its simplest description, the Trail Making Test (Reitan & Wolfson) is a connect-the-dots test. Part A consists of 25 circles printed on a sheet of white paper. Each circle contains a number from 1 to 25. The subject is required to connect the circles sequentially, beginning with 1, as rapidly as possible. Part B consists of 25 circles, containing a mix of numbers (1-13) and letters (A-L). In this section, the subject is required to
alternately connect numbers and letters, starting with 1, then A, then 2, B, and so on, until 13 is reached.

This test is considered one of the best measures of general brain functions (Reitan, 1955; Reitan, 1958), as it requires activity in both hemispheres (Reitan & Wolfson, 1985): usage of short-term memory as comparison reference, integration of numerical and alphabetic series, and performance under time pressure. Scores are reported in number of seconds to completion; separate scores are reported for Part A as well as Part B.

Procedure

Subjects participated in a single test session lasting approximately 120 minutes. The experimenter provided a brief explanation of the study and informed the subject of their right to participate or not participate. The experimenter then distributed a packet containing an informed consent form to be signed, as well as an informational sheet (See Appendix A). After the subject read and signed the informed consent form, random assignment to one of two groups followed.

In Group 1, individuals were administered these tests in the following order: CAT (BCT or NCT), Stroop, WFL, KBIT, WCST, Trails. Group 2 individuals were administered tests in the following order: WCST, Stroop, WFL, KBIT, CAT (BCT or NCT), Trails. Forty-four subjects were included in Group 1 (N = 22 males, N = 22 females) and forty-four subjects were
included in Group 2 (N = 22 males, N = 22 females). Within each group, half were given the BCT version of the CAT, and half were given the NCT version. This was also split evenly between males and females. Thus, administration orders were counterbalanced for CAT version (BCT, NCT), and equally weighted with males and females.

After test administration, subjects were debriefed and any remaining questions were answered. Subjects were thanked, given extra-credit points, and dismissed.
CHAPTER III

RESULTS

The overall mean score for Category test errors was 42.1, with a standard deviation of 22.1. Scores ranged from a low of 5 to a high of 95. By test version, error scores were as follows: BCT error scores mean was 35.4, with a standard deviation of 20.2. There were 11 scores greater than 50. NCT error scores mean was 48.8, with a standard deviation of 22.3. There were 22 scores greater than 50 in the NCT group. See Table 5 for a summary of Category test descriptives. Dividing subjects into groups by test administration order: Order 1 (CAT, Stroop, WFL, KBIT, WCST, Trails) yielded a mean of 42.5 errors, with a standard deviation of 23.2. Order 2 (WCST, Stroop, WFL, KBIT, CAT, Trails) yielded a mean of 41.7 errors, with a deviation of 21.4. These data are presented in tabular form, along with all other descriptive statistics from the study, in Appendix C, Tables 5 and 6.

A 2x2 analysis of variance was used to analyze effects of test version (BCT vs NCT) and administration order on Category test total error scores. The results of the analysis show that there was a significant difference in total error scores due to version administered ($F = 8.579$, $df = 1.84$, $p < .01$). Order of administration produced no
significant effect and there were no significant interaction effects. See Appendix C, Table 7 for a tabular summary.

Examining all subjects as one group (N = 88), Category test total errors correlated with KBIT Matrices scores significantly in a negative direction (r = −.37, p < .01), while KBIT Verbal scores did not show significant correlation. KBIT Composite scores correlated negatively (r = −.32) at a significant level (p < .01).

For the BCT (N = 44), KBIT Matrices scores showed significant correlation (r = .44, p < .01) while KBIT Verbal scores were not significantly correlated. KBIT Composite scores correlated with BCT error scores at a significant level (r = −.3391, p < .05).

KBIT Matrices scores were not significantly related to NCT error scores (r = −.29), nor were KBIT Verbal scores (r = −.11). As well, KBIT Composite scores were not significantly correlated to NCT error scores (r = −.27). See Appendix C, Table 8 for a summary of KBIT-Category test correlations.

Pearson-Product-moment correlations were computed for BCT-Stroop scores, BCT-WF scores, BCT-WCST scores, BCT-Trails A, and BCT-Trails B scores. Correlations between the BCT and Stroop were all low and non-significant: BCT-Stroop Part I (r = −.08), BCT-Stroop Part II (r = −.10), BCT-Stroop Part III (r = −.03). Correlation between the BCT and WF was low and not significant (r = −.04). Correlations between
the BCT and Trails Part A were nonsignificant ($r = .14$). Also, correlations between the BCT and Trails Part B were nonsignificant ($r = .04$). Nine scores for the WCST were computed and correlated with BCT error scores. All correlations were low and nonsignificant. See Appendix C, Tables 9 and 10 for summaries.

Pearson-Product-moment correlations were computed for NCT-Stroop scores, NCT-WF scores, NCT-WCST scores, NCT-Trails A, and NCT-Trails B scores. Correlations between NCT and Stroop were low and nonsignificant: NCT-Stroop Part I ($r = -.04$), NCT-Stroop Part II ($r = -.21$), NCT-Stroop Part III ($r = -.06$). Correlation between NCT and WF was low and nonsignificant ($r = -.07$). Correlations between NCT error scores and Trails Part A were not significant ($r = .24$). Also, correlations between NCT and Trails Part B ($r = .05$) were not significant. Nine scores for the WCST were computed and correlated with NCT error scores. All correlations were low and nonsignificant. See Appendix C, Tables 9 and 10 for summaries.
CHAPTER IV

DISCUSSION

Methodological Issues

Testing conditions. This study utilized college students who were tested in a Neuropsychological test room. Unfortunately, the test room was situated near an exercise room from which TV noise, treadmill noise, grunts, groans, and other body noises were emitted. Occurrences such as these were disruptive for the test situation and were cause for occasional interrupted testing. This could constitute a possible confounding factor. Replication of the study in a different testing situation would confirm or deny this. As it happened without regularity, however, the test results are likely robust with respect to these interruptions.

Samples and generalizeability. When conducting any study, there is always a tradeoff between homogeneity of sample and generalizeability of results. This study employed a non-clinical sample of male and female college students aged 18 to 26. This was done in order to limit extraneous factors from impinging on the results. For example, individuals who were depressed were not used. While this homogeneous sample limits generalizeability, it aids in purity of attributional factors. Nevertheless, the limits of generalizeability mean that results of this study may only
be applied to other non-clinical populations with great caution. For example, consider geriatric populations, who typically score more errors on the Category test simply because of age. In this respect, the study needs to be replicated with other non-clinical populations.

**BCT-NCT Comparisons**

This study investigated the effects of revising the Category test instruction set by removing set-shift cues. The primary goal of this study was to determine if normal subjects performed differently on the Category test when they were not notified of impending principle changes. By altering the instructions in this manner, it was expected that subjects would be required to perform conceptual set-shifts independently, thus increasing the difficulty and concomitant error scores.

The results from this study support the principle research hypothesis. Subjects administered the NCT performed significantly differently ($p = .004$) than did those administered the BCT. NCT test-takers typically scored 13.4 more errors than did those taking the BCT.

Rothke (1986), in a similar study utilizing a heterogeneous clinical population, found that removal of set-shifting cues made no significant in Category test scores. Rothke studied 52 male inpatients with a variety of diagnoses--cerebrovascular accident, schizophrenia, psychotic depression, bipolar disorder, multiple substance
abuse, multiple sclerosis. Mean age was 44.4 years (range 22-70) and average education was 12.1 years (range 8-19). Rothke's no-cue group mean was 90.5 errors, while the cued group averaged 84.8 errors. Rothke predicted that if a nonclinical population were sampled and tested, significant differences would not be found.

The current study's results contrasts directly with Rothke's prediction. Problems with Rothke's study include use of a widely varied clinical sample and use of an aged population: the Category test is sensitive to age effects (see Fromm-Auch, 1983; Bigler, Steinman, & Newton, 1982; Yeudall, Reddon, Gill, & Stefanyk, 1987) and is much less sensitive to certain clinical disorders, such as schizophrenia (see Matarazzo, et.al., 1976).

That there was no order effect for test presentations was rather surprising. One might suspect that after taking the WCST, in which there is no set-shift cueing, subjects would apply the same approach to the NCT. As the analysis showed no order effects (no practice effects) it seems likely that the tests measure separate skills. This is confirmed in the non-significance of correlations between the Category test and all measures of the WCST.

In changing the instruction set, how was the BCT changed specifically? In the BCT there exist a total of 61 perceptual shifts required, and a total of 5 conceptual shifts. Subjects taking the test must make perceptual set
shifts as well as conceptual shifts, whether cued or not. However, when not cued, subjects must generate these shifts independently—a self-cueing response. It seems likely that self-cueing would be a frontal lobe response, especially since the same task is required on the WCST, which is a widely used frontal lobe test. Recall though, that Category test scores were not altered by presentation order with the WCST, thus indicating measurement of a separate function.

However, the two tests differ significantly in another dimension not yet addressed. In the WCST, item stimuli differ on three dimensions: color, form, and number. These three dimensions also form the basis of concept set-shifts. In this sense, item stimuli remain constant throughout the test administration. In contrast, on the Category test, many stimuli are shown only once, and new stimuli appear throughout the test—even in subtest 7, where subjects are specifically told that all stimuli have been seen before! On the Category test, differing stimuli appearing throughout the test adds to the complexity. Part of what may be required on the Category test is the capacity to screen out unnecessary or unimportant details. Debriefings with several subjects who had especially difficult times revealed a seeming preoccupation with insignificant details. One subject reported counting the dashes in test items, while another admitted to trying to see the outline of a number
somewhere in the stimuli. These subjects were unable to screen out insignificant details.

With this in mind, one might expect a high correlation with Stroop scores, where subjects must screen unimportant information in order to complete the task. On the Stroop, the subject must read the colors of word letters while ignoring the printed word. Nonsignificance of Stroop-Category test correlations might initially suggest measurement of different skills. However, in the Category test, the subject must supply the higher order function of determining what to screen out. Determining what to screen out is a different ability than performing the actual screening. Subjects need not make this determination in the Stroop, and the categories in the WCST are elementary, usually apparent, and finite. Thus, the Stroop and WCST are likely not testing this higher-order function.

With the removal of cueing in the NCT, subjects must now generate both perceptual as well as conceptual set-shift criteria, as well as implementing both. In addition, subjects must simultaneously screen insignificant details. This combination could provide enough change to significantly increase the difficulty in normal subjects.

**BCT-KBIT Comparisons**

The results from this study supported the second hypothesis, that BCT-KBIT scores would correlate significantly. As an entire group, Matrices and Composite
scores correlated significantly negatively with Category test scores. The negative correlation exists because Category test scores are error scores, whereas KBIT scores are scaled "correct" scores. BCT scores correlated significantly with KBIT Matrices and Composite scores as well. While these correlations were significant, none were greater than .44, indicating moderate correlation strength at best. Matrices, a nonverbal test of intelligence, shares about 19% of variance with the BCT, while the Verbal portion shares about 1%.

Other studies have correlated Category test scores with intelligence test (WAIS) scores with mixed results. Some studies found that Verbal IQ (VIQ) correlated less strongly than Performance IQ (PIQ) did with the Category test. For example, Cullum, Steinman and Bigler (1984) found a lower correlation for VIQ (-.31) than for Performance IQ (-.52) in a study of head-injured patients. Thomas and Trexler (1982) found a similar trend in a study of rehabilitation patients.

In contrast, Reitan (1956) found higher correlations between Category test and PIQ than for VIQ (-.64 and -.58, respectively) for a group of brain-damaged patients. Correlations for controls were somewhat higher (VIQ -.65, PIQ -.67). Beaumont (1975), in a study of psychiatric patients, reported no significant correlation between IQ measures and Category test scores.
In sum, previous studies support correlations between Category test scores and VIQ, as well as PIQ. The extent of this relationship with VIQ in studies examined ranged from -.31 to -.65. The current study found CAT-VIQ correlations from .11 to .14. Previous studies found correlation between Category test and nonverbal IQ scores ranging from -.52 to -.67. The current study found Category test-nonverbal IQ correlations from -.29 to -.44. Reasons for the current study's lower correlations may be due to the difference in IQ tests used, as well as a difference in populations, differing in size, homogeneity, and clinical factors.

**NCT-KBIT Comparisons**

Hypothesis 3 proposed that NCT scores would also be highly correlated with KBIT scores. Results did not support this hypothesis. No scores correlated significantly. By removing the Category test cueing, the test was either made more difficult, or it began to measure a particular function more sensitively.

Of interest then, is the difference in correlations between Matrices-BCT scores and Matrices-NCT scores. Are the two correlations significantly different? A post-hoc analysis of correlations (Fisher Z coefficient, Guilford & Fruchter, 1978) shows this difference to be not significant. As well, visual inspection of the correlational values shows both to be in the low range. The BCT version shares about
19% of variance with KBIT Matrices, while the NCT version shares 8%.

Why were NCT or BCT correlations lower than previous study's findings? Perhaps variations in a sample of normals is not sufficient for strong correlations; however, inspection of the range of scores shows this not to be the case. On the BCT, scores ranged from 10 to 84; on the NCT, scores ranged from 5 to 95. Regarding IQ test results, scores ranged from 68 to 129.

That these tests would not share a significant amount of variance is rather surprising, considering that most are considered good frontal lobe measures. Recall, that the Category test, which was originally considered a frontal lobe test, is really a good diffuse damage indicator. Perhaps it does not measure any one function in enough depth to correlate with other tests measuring specific functions. However, if such be the case, one would expect fairly high correlations with a general test of intelligence such as the KBIT. Again, the question arises: does a population of normals vary enough to permit meaningful correlations? Visual inspection of the ranges shows a wide range of variations in all measures used.

An interesting finding was the significance of BCT-KBIT correlations and nonsignificance of NCT-KBIT correlations. Perhaps the BCT is a broad enough test to correlate with general nonverbal IQ measures, but when changed to the NCT
it measures a different function. Perhaps it measures frontal lobe functioning more sensitively, as it was originally designed. Or maybe it measures diffuse damage better than the BCT. Either way, the data from this study do not support the notion that the NCT is better at localizing frontal lobe functioning. On the other hand, the notion is not refuted. Further research with specific clinical populations is called for, in order to utilize the present results.

Other Test Comparisons

No correlations were significant between BCT and Stroop, Word Fluency, WCST perseverative errors, or Trails. As with the BCT, no significant correlations were found between the NCT and any of the other tests administered. Thus, hypotheses 4 through 7 were not supported. The Category test with cueing removed shares little variance with Stroop, Word Fluency, WCST, or Trails.

Future Research Directions

Other populations. Results of this study demonstrate that the Category test is significantly different with the removal of conceptual set-shift cueing. The next logical question deals with localization of function regarding the NCT. In order to best localize, homogeneous clinical populations need to be tested. Suggestions include those with frontal lobe, parietal lobe, or subcortical damage. In addition, severity of damage needs to be held consistent.
Other Alterations to the Category Test. The Category test, even in altered form, is still a complex test measuring many higher-order functions. Some functions that it measures constitute unnecessary sections of the Category test. Several studies have pointed out suggested changes (see Alterations of the CAT, above). These studies have indicated that there are unnecessary repetitions, as in subtests I and II which contribute little to the overall score, or subtests V and VI, which have the same concept, or subtest VII, which is purely a memory test (Sherrill, 1985; Sherrill, 1987;). Laatsch and Choca (1991) used item analysis to remove 45 items which served as poor discriminators, being either too easy or too difficult.

Psychometric techniques, such as item analysis or construct validation could be used to refine the Category test. At present, the Category test is a sensitive diffuse cerebral injury indicator. Using this as a base, and applying psychometric techniques to refine and delineate functions and localization, a truly new type of Category test could be generated. Redevelopment of the Category test should include improvements such as increasing the number of conceptual set-shifts and decreasing the perceptual set-shifts, while simultaneously applying psychometric techniques such as item-analysis and cross-validation.
Conclusions

The results of this study indicate that removal of cueing instructions changes the Category test significantly. In a non-clinical college-age sample, subjects who were not cued for conceptual set-shift changes (NCT) scored significantly more errors than those who were administered the standard Category test (BCT).

While BCT scores correlated significantly with nonverbal intelligence scores, NCT scores did not. However, the difference in these correlations was not significant, indicating that the intelligence aspect measured in the two versions is not different. Neither the BCT nor the NCT correlated significantly with the WCST, Word Fluency, Stroop, or Trail Making Test. It is recommended that the NCT be administered to circumscribed clinical populations in order to best utilize present findings.
Informed Consent Information

Investigators: E. H. Harrell, Ph.D., Daniel Rockers, M.A.

Purpose of the Study:
The purpose of this study is to examine the performance of normal individuals on two forms of a neuropsychological test.

Procedures: The participant will attend one session of neuropsychological testing. This session will take approximately 2 hours.

Safeguards: The participant’s answers will be kept completely confidential. Only first names and code numbers will be used. I am not interested in comparing individual responses, only group responses will be examined.

Participation in this study is entirely voluntary, and you may end your participation at any time you desire. To the best of our knowledge, participation will not cause any physical or psychological harm.

Benefits: Not only does participation in this study provide a new experience for the individual, it may provide an opportunity for new growth and development. This study provides an opportunity to increase our understanding of the performance of the frontal cerebral lobes in humans. Potentially this may benefit society in general as we learn more about these functions.

Additional questions:
Additional questions regarding your participation, or this study, may be directed to Daniel Rockers (817) 383-7450, or E. H. Harrell (817) 565-2671.
INFORMED CONSENT FORM

This is to certify that I

______________________________ (your name)

am participating as a volunteer in a scientific study as an authorized part of the educational and research program of the University of North Texas.

The investigation and my part in the investigation have been defined and fully explained to me by Daniel Rockers, and I understand his explanation. the procedures of this investigation and their risks and discomforts have been described in a separate statement, and are discussed in detail.

I have been given an opportunity to ask whatever questions I may have had, and all such questions and inquiries have been answered to my satisfaction.

I understand that I may refuse to answer any question on the questionnaires.

I understand that any data or answers to questions will remain confidential, and only group data will be reported.

I further understand that I am free to withdraw my consent and terminate my participation at any time, without penalty, prejudice, or effect on my class standing.

______________________________ (date)______________________________ (signature)
APPENDIX B

GROUP DEMOGRAPHIC INFORMATION
### Table 1

*Age, Sex, and Language Demographics*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frequency</th>
<th>Percent</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td>21.08</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>44</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>44</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>86</td>
<td>97.7</td>
<td></td>
</tr>
<tr>
<td>Spanish-Eng.</td>
<td>01</td>
<td>01.1</td>
<td></td>
</tr>
<tr>
<td>Persian-Eng.</td>
<td>01</td>
<td>01.1</td>
<td></td>
</tr>
</tbody>
</table>
Table 2

Race and Handedness Demographics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frequency</th>
<th>Percent</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>73</td>
<td>83.0</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>09</td>
<td>10.2</td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>04</td>
<td>04.5</td>
<td></td>
</tr>
<tr>
<td>Arabic</td>
<td>02</td>
<td>02.3</td>
<td></td>
</tr>
<tr>
<td>Handedness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>81</td>
<td>92.0</td>
<td></td>
</tr>
<tr>
<td>Left</td>
<td>06</td>
<td>06.8</td>
<td></td>
</tr>
<tr>
<td>Ambi</td>
<td>01</td>
<td>01.1</td>
<td></td>
</tr>
</tbody>
</table>
### Table 3

**Education and Marital Status Demographics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frequency</th>
<th>Percent</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td>14.3</td>
</tr>
<tr>
<td><strong>Major</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychology</td>
<td>30</td>
<td>34.0</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>58</td>
<td>66.0</td>
<td></td>
</tr>
<tr>
<td><strong>Marital Status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>85</td>
<td>96.6</td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>03</td>
<td>03.4</td>
<td></td>
</tr>
<tr>
<td><strong>Hospitalization</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>29</td>
<td>33.0</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>59</td>
<td>67.0</td>
<td></td>
</tr>
</tbody>
</table>
Table 4

**Illness and Alcohol Consumption Demographics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frequency</th>
<th>Percent</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurological problems/ Mental Illness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>00</td>
<td>00</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>87</td>
<td>98.9</td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>01</td>
<td>01.1</td>
<td></td>
</tr>
<tr>
<td>Alcohol</td>
<td></td>
<td></td>
<td>3.21</td>
</tr>
<tr>
<td>Drink</td>
<td>41</td>
<td>46.6</td>
<td></td>
</tr>
<tr>
<td>Not drink</td>
<td>43</td>
<td>48.9</td>
<td></td>
</tr>
<tr>
<td>No response</td>
<td>04</td>
<td>04.5</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX C

RELATED STATISTICS
Table 5

**Category Test Descriptives**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category Test</td>
<td>42.1</td>
<td>22.1</td>
<td>88</td>
</tr>
<tr>
<td>Category Test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCT version</td>
<td>35.4</td>
<td>20.2</td>
<td>44</td>
</tr>
<tr>
<td>NCT version</td>
<td>48.8</td>
<td>22.3</td>
<td>44</td>
</tr>
<tr>
<td>Category Test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order 1</td>
<td>42.5</td>
<td>23.2</td>
<td>44</td>
</tr>
<tr>
<td>Order 2</td>
<td>41.7</td>
<td>21.4</td>
<td>44</td>
</tr>
</tbody>
</table>

Note: Order 1: CAT, Stroop, WFL, KBIT, WCST, Trails  
Order 2: WCST, Stroop, WFL, KBIT, CAT, Trails
Table 6

Other Test Descriptives

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Min/Max</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>KBIT Matrices</td>
<td>104.2</td>
<td>8.8</td>
<td>68/129</td>
<td>88</td>
</tr>
<tr>
<td>KBIT Verbal</td>
<td>104.4</td>
<td>8.6</td>
<td>80/129</td>
<td>88</td>
</tr>
<tr>
<td>KBIT Composite</td>
<td>104.7</td>
<td>7.7</td>
<td>81/124</td>
<td>88</td>
</tr>
<tr>
<td>WCST Perseverative Errors</td>
<td>9.5</td>
<td>6.5</td>
<td>3/38</td>
<td>88</td>
</tr>
<tr>
<td>Stroop I</td>
<td>106.8</td>
<td>16.4</td>
<td>60/149</td>
<td>88</td>
</tr>
<tr>
<td>Stroop II</td>
<td>78.3</td>
<td>10.4</td>
<td>54/103</td>
<td>88</td>
</tr>
<tr>
<td>Stroop III</td>
<td>49.1</td>
<td>9.7</td>
<td>24/71</td>
<td>88</td>
</tr>
<tr>
<td>Trails A</td>
<td>22.9</td>
<td>7.0</td>
<td>13/47</td>
<td>88</td>
</tr>
<tr>
<td>Trails B</td>
<td>56.2</td>
<td>29.3</td>
<td>24/234</td>
<td>88</td>
</tr>
<tr>
<td>Word Fluency</td>
<td>39.8</td>
<td>9.6</td>
<td>18/69</td>
<td>88</td>
</tr>
</tbody>
</table>
Table 7

Category test ANOVA Results

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAT version</td>
<td>1</td>
<td>3,928.91</td>
<td>3,928.91</td>
<td>8.58*</td>
</tr>
<tr>
<td>Order</td>
<td>1</td>
<td>14.73</td>
<td>14.73</td>
<td>0.03</td>
</tr>
<tr>
<td>Interaction</td>
<td>1</td>
<td>463.68</td>
<td>463.68</td>
<td>1.01</td>
</tr>
<tr>
<td>Within (error)</td>
<td>84</td>
<td>38,469.55</td>
<td>457.97</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>87</td>
<td>42,875.86</td>
<td>492.84</td>
<td></td>
</tr>
</tbody>
</table>

* p = .004
Table 8

**KBIT-Category Test Correlations**

<table>
<thead>
<tr>
<th></th>
<th>KBIT-M</th>
<th>KBIT-V</th>
<th>KBIT-C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Combined groups</strong></td>
<td>-.37**</td>
<td>-.14</td>
<td>-.32**</td>
</tr>
<tr>
<td>BCT (N = 44)</td>
<td>-.44**</td>
<td>-.13</td>
<td>-.34*</td>
</tr>
<tr>
<td>NCT (N = 44)</td>
<td>-.29</td>
<td>-.11</td>
<td>-.27</td>
</tr>
</tbody>
</table>

Note: * = $p < .05$

** = $p < .01$
Table 9

**BCT/NCT-other test correlations**

<table>
<thead>
<tr>
<th></th>
<th>Stroop I</th>
<th>Stroop II</th>
<th>Stroop III</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCT</td>
<td>.08</td>
<td>-.19</td>
<td>-.03</td>
</tr>
<tr>
<td>NCT</td>
<td>.04</td>
<td>-.21</td>
<td>-.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Trails A</th>
<th>Trails B</th>
<th>Word Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCT</td>
<td>.04</td>
<td>.04</td>
<td>-.04</td>
</tr>
<tr>
<td>NCT</td>
<td>.05</td>
<td>.07</td>
<td>.07</td>
</tr>
</tbody>
</table>
Table 10

**WCST Correlations**

<table>
<thead>
<tr>
<th></th>
<th>WCST Cats Achd</th>
<th>WCST 3 Consec.</th>
<th>WCST Total errs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCT</td>
<td>-.08</td>
<td>-.12</td>
<td>-.21</td>
</tr>
<tr>
<td>NCT</td>
<td>-.12</td>
<td>-.07</td>
<td>.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>WCST Total corr.</th>
<th>WCST Fail mtn set</th>
<th>WCST Non pers. errs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCT</td>
<td>.01</td>
<td>.02</td>
<td>.16</td>
</tr>
<tr>
<td>NCT</td>
<td>.08</td>
<td>.09</td>
<td>.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>WCST Pers. errs</th>
<th>WCST Pers. resp.</th>
<th>WCST Trials to first</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCT</td>
<td>.23</td>
<td>.21</td>
<td>.05</td>
</tr>
<tr>
<td>NCT</td>
<td>.17</td>
<td>.17</td>
<td>.09</td>
</tr>
</tbody>
</table>
REFERENCES


Miller, D. C. (1989). *Relationships among three levels of measuring planning: Electrophysiological--the Category Test, and psychological--PASS scales in adolescent males*. Unpublished doctoral dissertation, Ohio State University, Columbus, OH.


